VISUALISATION OF QUALITY INFORMATION FOR GEOSPATIAL AND REMOTE SENSING DATA

PROVIDING THE GIS COMMUNITY WITH THE DECISION SUPPORT TOOLS FOR GEOSPATIAL DATASET QUALITY EVALUATION

VICTORIA LUSH

Doctor of Philosophy (by Research)

ASTON UNIVERSITY

April 2015

Aston University

Visualisation of Quality Information for Geospatial and Remote Sensing Data

Victoria Lush

Doctor of Philosophy (by Research), 2015

Thesis Summary

The evaluation of geospatial data quality and trustworthiness presents a major challenge to geospatial data users when making a dataset selection decision. The research presented here therefore focused on defining and developing a GEO label – a decision support mechanism to assist data users in efficient and effective geospatial dataset selection on the basis of quality, trustworthiness and fitness for use. This thesis thus presents six phases of research and development conducted to: (a) identify the informational aspects upon which users rely when assessing geospatial dataset quality and trustworthiness; (2) elicit initial user views on the GEO label role in supporting dataset comparison and selection; (3) evaluate prototype label visualisations; (4) develop a Web service to support GEO label generation; (5) develop a prototype GEO label-based dataset discovery and intercomparison decision support tool; and (6) evaluate the prototype tool in a controlled human-subject study.

The results of the studies revealed, and subsequently confirmed, eight geospatial data informational aspects that were considered important by users when evaluating geospatial dataset quality and trustworthiness, namely: producer information, producer comments, lineage information, compliance with standards, quantitative quality information, user feedback, expert reviews, and citations information. Following an iterative user-centred design (UCD) approach, it was established that the GEO label should visually summarise availability and allow interrogation of these key informational aspects. A Web service was developed to support generation of dynamic GEO label representations and integrated into a number of real-world GIS applications. The service was also utilised in the development of the GEO LINC tool – a GEO label-based dataset discovery and intercomparison decision support tool. The results of the final evaluation study indicated that (a) the GEO label effectively communicates the availability of dataset quality and trustworthiness information and (b) GEO LINC successfully facilitates 'at a glance' dataset intercomparison and fitness for purpose-based dataset selection.

Keywords: geospatial data quality and trust indicators, fitness for use, user-centred design, quality labelling, decision support systems.

Publications

Scientific Papers

[1] Lush, V., Bastin, L. and Lumsden, J., 2012. Geospatial Data Quality Indicators. In: *Proceedings of the 10th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*, Florianópolis, SC, Brazil, 10–13 July 2012, pp. 121-126.

> This paper presented the results of the initial investigation and discussed high-level user requirements in geospatial data selection and data quality and trustworthiness evaluation. It reported on the geospatial data quality indicators which were identified as user priorities, and which can potentially be standardised to enable intercomparison of datasets against user requirements.

[2] Yang, X., Blower, J.D., Bastin, L., Lush, V., Zabala A., Masó, J., Cornford, D., Díaz, P., Lumsden, J., 2012. An Integrated View of Data Quality in Earth Observation. *Philosophical Transactions A*, 371 (1983).

> This paper was produced in collaboration with the GeoViQua project partners and focused on the broader GeoViQua research. The paper presented user needs on data quality, reviewed existing standards and specifications on data quality, and proposed an integrated model for data quality in the field of Earth observation. The contribution to this paper consisted of a section presenting a wide-ranging picture of user needs on geospatial data quality.

[3] Lush, V., Bastin, L., Lumsden, J., 2013. Developing a GEO Label: Providing the GIS Community with Quality Metadata Visualisation Tools. In: *Proceedings of the 21st GIS Research UK (GISRUK 3013)*, Liverpool, UK, 3–5 April 2013.

> This paper presented the results of the Phase I and Phase II questionnaire-based studies.

[4] Zabala, A., Riverola, A., Serral, I., Díaz, P., Lush, V., Masó, J., Pons, X. and Habermann, T., 2013. Rubric-Q: Adding Quality-related Elements to the GEOSS Clearinghouse Datasets. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 6(3), pp. 1676–1687.

> This paper was produced in collaboration with the GeoViQua project partners. The paper presented the Rubric-Q tool designed for visualisation of the metadata elements in a structured tabular format and provided a review of the completeness of metadata records provided in the GEOSS Clearinghouse. The contribution to this paper consisted of (a) a literature review of the challenges in geospatial dataset selection and quality evaluation, and (b) a discussion about the GEO label as an alternative approach to visualisation of geospatial data quality information.

[5] Bastin, L., Lush, V., Nüst, D., Sevillano, E., Zabala, A., Maso, J., van den Broek, M., Lush, C., Thum, S., 2014. Models and Tools to Support Dynamic, Effective Communication of Geospatial Data Quality. *Environmental Modelling & Software*. [**under review**]

This paper was produced in collaboration with the GeoViQua project partners. The paper presented Producer Quality Model (PQM), User Quality Model (UQM) and models' integration into widely-used GIS tools. The contribution to this paper consisted of an overview of how the PQM and UQM models are utilised by the GEO label service to generate dynamic GEO label representations.

Poster presentations

[6] Lush, V., Lumsden, J., Bastin, L., Cornford, D., 2011. User Perspectives on Geospatial Data Quality. In: *QA4EO Workshop on 'Providing Harmonised Quality Information in Earth Observation Data by 2015'*, Oxford, UK, 18–20 October 2011.

This poster presented user stories and themes elicited from the results of the initial investigation interviews with geospatial data experts.

[7] Lush, V., Lumsden, J., Masó, J., Díaz, P., McCallum, I., 2012. GEO Label: User and Producer Perspectives on a Label for Geospatial Data. In: *EGU General Assembly 2012*, Vienna, Austria, 22–27 April 2012.

This poster introduced a concept of a GEO label, discussed the results of Phase I questionnaire-based study and presented potential GEO label visualisations based on the survey results.

[8] Lush, V., Lumsden, J., Bastin, L., 2013. GEO Label - Quality Information Interrogation Tool for Geospatial Datasets: Towards Effective Visualization of Quality Metadata. In: *EGU General Assembly 2013*, Vienna, Austria, 7 - 12 April 2013.

This poster presented the concept of a dynamic GEO label based on the results of the Phase I and Phase II studies and discussed the GEO label integration into GEOSS.

[9] Lush, V., Lumsden, J., Bastin, L., 2013. GEO Label Web Services for Dynamic and Effective Communication of Geospatial Metadata Quality. In: *EGU General Assembly 2014*, Vienna, Austria, 27 April–02 May 2014.

This poster presented the GEO label service and provided examples of the GEO label integration into a prototype extension of the GEOSS portal, George Mason University (GMU) portal, and the GeoNetwork catalogue.

Dedicated to my mum, Nelli Sutupova.

Acknowledgements

These four years of my Ph.D. research were an amazing journey full of incredible new experiences. Like any research, this work was not without its challenges, but looking back, I am extremely happy and grateful that I was given this unique research opportunity. While a Ph.D. is a work of an individual, I cannot say that I was alone in this research journey. There are a lot of people without whom this Ph.D. experience would not be the same or would not even be possible. I would like to take this opportunity to thank them all.

Foremost, I extend my gratitude to Dr. Jo Lumsden who was a dedicated, encouraging and simply amazing supervisor. I remember starting this Ph.D. research being a scared inexperienced student who had a vague idea of what real research actually means. Jo guided me through this journey and helped me grow, not only as a professional, but as a person. I am very grateful for all her support and for being the best supervisor any Ph.D. student can wish for.

I also want to extend my sincere appreciation to my associate supervisor, Dr. Lucy Bastin. Lucy was my GIS guru and supported me in the technical side of this Ph.D. work. I want to thank Lucy for all the help in preparing and delivering conference papers and also for guiding my GeoViQua work. Lucy is an amazing researcher and academic and I am very thankful for having an opportunity to work with her. My appreciation also goes to Dr. Dan Cornford who was my associate supervisor during the first year of the Ph.D. It was a real pleasure working with Dan; the GeoViQua trips would not be as fun without his presence.

Working as part of the GeoViQua project was a remarkable and unique experience. I would like to thank all the GeoViQua project partners for the incredible memories and for their friendship. In particular, I want to express my gratitude to Joan Masó – our outstanding project leader – for all the encouragement and for promoting the GEO label work among the GEO community. I want to thank Ivette Serral, Paula Diaz, Simon Thum and Eva Sevillano for becoming my good friends and for all their moral support during and after GeoViQua. I also want to thank Daniel Nüst and Chris Lush for all the technical support in developing the GEO label service; and Maud van den Broek for assisting in the integration of the feedback server. Moreover, I would like to thank all the interviewees and study participants who dedicated their time to participate in this research.

Finally, my sincere gratitude goes to my family for all the moral support and for surviving this Ph.D. with me. I want to thank my mum and my boyfriend Chris for all the cooking and cleaning that I did not have to do, all their understanding and care, and for putting up with my Ph.D. life schedule.

Thank you all for making this Ph.D. a truly remarkable time of my life.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AIP-6	GEOSS Architecture Implementation Pilot Phase 6
Ajax	Asynchronous JavaScript and XML
API	Application Programming Interface
AV	Audio-Video
BA	Business Analytics
BADC	British Atmospheric Data Centre
BI	Business Intelligence
B2C e-Commerce	Business-to-Customer e-Commerce
CEN	European Committee for Standardization
CF	NetCDF Climate and Forecast
CLASS	Comprehensive Large Array-data Stewardship System
CMECS	Coastal and Marine Ecological Classification Standard
CPM	Corporate Performance Management Systems
DAAC	NASA Distributed Active Archive Centers
DAB-Q	Discovery and Access Broker
DAGIS	Discovering Annotated Geospatial Information Services
DCMI	Dublin Core Metadata Initiative
DG	Directorate-General
DOM	Document Object Model
DSS	Decision Support Systems
DW	Data Warehousing
EEA	European Environmental Agency
EFTA	European Free Trade Association
EISs	Executive Information Systems
ENs	European Standards
EO	Earth Observation
EOLISA	Earthnet Online ESA
ERASs	Enterprise Reporting and Analysis Systems
FAO	Food and Agriculture Organisation
FGDC	Federal Geographic Data Committee
FICCDC	Federal Interagency Coordinating Committee on Digital Cartography
GCI	GEOSS Common Infrastructure
GEO	Group on Earth Observation
GEO LINC	GEO Label INterComparison Tool
GEOSS	Global Earth Observation System of Systems
GI	Geographic Information
GML	OpenGIS Geography Markup Language
GMU	George Mason University
GSS	Group Support Systems
GSWS	Geospatial Semantic Web Services
HCI	Human-Computer Interaction
HTTP	Hypetext Transfer Protocol
IDSS	Intelligent Decision Support Systems
IS	Information Systems
ISO	International Organisation for Standardisation
JRC	Joint Research Centre

KMDSS	Knowledge Management-Based DSS
KML	Keyhole Markup Language
LTER	Long-Term Ecological Research
MODIS	Moderate Resolution Imaging Spectroradiometer
MUM	Multidimensional User Manual
NASA	National Aeronautics and Space Administration
NEODAAS	Earth Observation Data Acquisition and Analysis Service
NERC	Natural Environment Research Council
NGOs	Non-Governmental Organisations
NIEeS	National Institute for Environmental eScience
NSDI	National Spatial Data Infrastructure
NSDIs	National Spatial Data Infrastructures
NSIDC	National Snow and Ice Data Centre
NSS	Negotiation Support Systems
NWS	National Weather Service
OGC	Open Geospatial Consortium
OLAP	Online Analytical Processing Systems
OMB	US Office of Management and Budget
OO	Object-Oriented
OWS	OGC Web Services
PDSS	Personal Decision Support Systems
PQM	Producer Quality Model
PSI	Public Sector Information
QC	Quality Control
QoS	Quality of Service
REST	Representational State Transfer
SBA	Societal Benefit Areas
SDI	Spatial Data Infrastructure
SDIs	Spatial Data Infrastructures
SDSS	Spatial Decision Support Systems
SDTS	Spatial Data Transfer Standard
STC	GEO Science and Technology Committee
SVG	Scalable Vector Graphic
TCs	ISO technical committees
UCD	User-Centred Design
UI	User Interface
UQM	User Quality Model
URI	Uniform Resource Identifier
URR	GEO User Requirement Registry
USGCRP	US Global Climate Research Program
VGI	Volunteered Geographic Information
VIS	Visual Information Seeking
WIYBY	What's In Your Back Yard
XLink	XML Linking Language
XML	Extensible Markup Language
XSL	Extensible Stylesheet Language
XSLT	XSL Transformations

Chapter 1 Introduction

1.1 Motivation

A Geographic Information System (GIS) can be defined as a system of hardware, software, procedures, people, organizations, and institutional arrangements to support the capture, management, manipulation, analysis, modelling, and display of geospatially-referenced data for solving complex planning and management problems (Chrisman, 2001; Heikkila, 2007). While numerous definitions of GIS exist in the literature, they all encapsulate the concept of geospatial-referencing – the association with locations in the real world and their features (Goodchild *et al.*, 2007). Within the GIS community, geospatial data quality and quality visualisation has always been an area of active research. Subjected to processes of generalisation, abstraction, and aggregation, geospatial data can only provide an approximation of the real world, and therefore almost always suffers from imperfect quality (Goodchild, 1995; Devillers and Jeansoulin, 2006). Objective quality measures of geospatial data (*internal* quality) relate to the "*difference between the data and the reality that they represent*" (Goodchild, 2006, p. 13). Subjective measures of quality (*external* quality) relate to a dataset's '*fitness for use*', meaning that, in order to assess the quality of data, we need to have information about the data to be used as well as the actual user needs (Chrisman, 1991).

To address issues of geospatial data quality, international organisations, initiatives, and working groups such as the International Organisation for Standardization (ISO), the Open GIS Consortium (OGC), INSPIRE, and many more, are actively working to establish, improve and extend geospatial data and metadata standards. Despite detailed recommendations of standardisation bodies, and despite the existence of formal metadata standards such as ISO

19115:2003 and FGDC, data quality information is, however, often not communicated to users in a consistent and standardised way (Boin and Hunter, 2006). While standardisation efforts have significantly improved metadata interoperability, the increasing choice of metadata standards poses a number of unresolved questions: Which standards are best to follow? How much metadata to provide? How to make metadata 'useful' and not just 'usable'? (Gahegan, 2005; Longhorn, 2005; Comber *et al.*, 2007b; Goodchild, 2009; Brown *et al.*, 2013). Since metadata standards are mostly focused on data production rather than potential data use and application, a typical metadata document is not sufficient to effectively communicate this fitness for purpose to users from a variety of domains and expertise levels (Boin and Hunter, 2007; Comber *et al.*, 2007b; Goodchild, 2009; Brown *et al.*, 2013).

With increased use of geospatial datasets across heterogeneous user groups and domains, the need to assess fitness for use becomes ever more essential yet complicated (Triglav *et al.*, 2011). Users are essentially presented with an increasing choice of data available from various data portals, repositories, and clearinghouses. This means that the intercomparison of dataset quality and the evaluation of a dataset's fitness for use can present a major challenge for geospatial data users. Over the past decade, many researchers and scholars have therefore made active attempts to address the challenge of communicating geospatial data fitness for purpose information, proposing a more "user-centric approach" to geospatial metadata (Goodchild, 2009). Researchers argued the case for enriching metadata records with: references to relevant literature (citations information); less formal opinions from the data producers; expert opinions of data quality; and user feedback regarding previous data use (Comber *et al.*, 2007a; 2007b). Recent reviews, however, suggest that these recommendations have not yet been put into practice, with no practical means for collating and searching user-focused metadata, and many of the metadata records that are available being incomplete (Goodchild, 2012; Ellul *et al.*, 2013).

Trust significantly influences our decision making. In the field of e-Commerce trust is considered to be a crucial enabler for online transaction decisions (Kim *et al.*, 2008a). Compared to traditional commerce, e-Commerce transactions are more impersonal, anonymous, and automated (Hassanein and Head, 2004); trustworthiness cannot be assessed by means of body language and traditional environmental cues (Gefen, 2002), making trust especially significant. Impersonality of geospatial data selection decision closely mirrors that of the e-Commerce transaction experience. Transactional risk is a vital precondition of e-Commerce trust (Chopra and Wallace, 2003), and, as such, the similar risks involved in dataset selection and use (i.e., the importance of selecting the right dataset for a given purpose) establish a need for dataset users to trust in dataset providers to deliver a reliable, quality dataset to meet their needs. In e-Commerce, trust indicators are used to engender consumer trust in websites; it can, therefore, be argued that it should be possible to establish and deploy similar trust indicators in the GIS domain to convey information about the trustworthiness of geospatial datasets and dataset providers. In essence, drawing on the parallels with e-Commerce, it can be argued that representation and visualisation of key trust indicators associated with geospatial datasets and their producers has the potential to support more effective, informed, and trust-based selection of quality datasets. Surprisingly, research into mechanisms of representing trust in the GIS domain has not yet received the same level of attention as it has in the e-Commerce domain (Skarlatidou *et al.*, 2011a; Skarlatidou *et al.*, 2013).

To tackle the challenge of data quality assessment and dataset selection decision making, in 2009, the Group on Earth Observation (GEO) Science and Technology Committee (STC) proposed to establish a GEO label – a label that could potentially improve user recognition of the quality of geospatial datasets and promote trust in datasets that carry the established GEO label (ST-09-02, 2010). The STC believed that a GEO label could assist in dataset searching and selection activities by providing users with visual cues of dataset quality and possibly relevance; a GEO label could effectively operate as a decision support mechanism for dataset selection. Initial STC attempts to define a concept of a GEO label focused primarily on establishing a certification body and, unfortunately, did not produce any tangible results.

1.2 Research Questions and Objectives

The issues outlined in the previous section lead to the following research questions:

1. What are the key informational aspects of geospatial datasets upon which users rely when evaluating quality and trustworthiness of geospatial datasets and making a dataset selection decision?

2. What role should a GEO label serve to effectively support evaluation of geospatial dataset quality and trustworthiness?

3. How should a GEO label summarise and represent dataset quality and trustworthiness information in a way which permits a user to easily assess the relevance of a dataset for their needs, and interrogate the specific aspects which are key to their application?

4. How can a GEO label be applied to facilitate an innovative approach to decision support in geospatial dataset intercomparison and selection?

1.3 Approach

Recognising the subjective nature of fitness for purpose evaluation, this research adopted an iterative user-centred design (UCD) approach in order to develop a GEO label that is likely to garner user acceptance once deployed. UCD is a multi-stage philosophical approach to technology design that places the user at the centre of the design process. The main characteristic of UCD in that it attempts to optimise a product around user needs, abilities and desires, rather than forcing the users to change their behaviour to suit the product. UCD recognises that end user involvement in the design of tools is critical to the success of the developed solutions.

Rather than focusing on obtaining quantitative statistical results, this research embraced a qualitative approach to collecting and analysing rich subjective research data from a large number of geospatial data users and producers with diverse backgrounds, to ensure a broad and inclusive picture of user needs as they relate to quality assessment of geospatial datasets. This research utilised various tried-and-tested UCD methods of collecting and analysing research data, including: semi-structured interviews; complex questionnaire-based studies; community voting; semi-formal feedback and recommendations from experts and peers; prototype development; deployment of developed solutions into real-world GIS applications; and prototype evaluation using a controlled lab-based human-subject study. Qualitative data analysis was adopted to identify, analyse, and report patterns (themes) within collected data. Since the research questions directly aim at understanding geospatial data users and developing solutions tailored to their needs, this qualitative research approach is beneficial in that it provides better appreciation of user needs, experiences, and perceptions while also offering validated practical solutions that should facilitate improved geospatial data quality and trustworthiness evaluation.

The work presented in this thesis followed six main phases of exploration, development, evaluation and validation, with each phase building upon the knowledge gathered in the previous phases. These phases are discussed in more detail in Chapter 1.

1.4 Thesis structure

The next chapter presents the important concepts underpinning this research problem via a review of relevant scientific literature. The chapter provides a background in trust, geospatial data quality, decision making, and GEO label initiatives. It reviews the literature on the concept of trust, various models of trust, trust indicators, and trust within the geosciences domain and explores how established trust concepts from business-to-consumer (B2C) e-Commerce can be drawn on to develop effective trust representations for geospatial data. The chapter then outlines current progress and challenges in the representation and evaluation of geospatial data quality via a review of the literature on geospatial data quality, quality standards and indicators, and fitness for use. The focus then shifts to decision making, decision support systems (DSS) and visual information seeking using starfield displays, with the aim to provide an overview of processes and tools that can be adopted in the GIS domain to better support users in evaluation of geospatial data quality. The final section of Chapter 2 describes the initial proposals that started the initiatives to conceptualise and define the notion of a GEO label.

Chapter 3 outlines the qualitative method adopted for the development and evaluation of a GEO label. The chapter presents investigation techniques employed for collecting and analysing data and describes 6 main phases of research conducted to design, develop and evaluate the GEO label and a GEO label-based prototype decision support tool.

Chapter 4 outlines the process of the preparatory phase of this research – an initial investigation that was conducted to (a) elicit high level user requirements regarding geospatial data quality and trustworthiness assessment and evaluation, and subsequent dataset selection and (b) identify the informational aspects of geospatial datasets upon which users rely when assessing dataset quality and trustworthiness. The chapter then presents 11 geospatial data informational aspects that were identified as important for evaluation of geospatial datasets' quality and trustworthiness, and discusses the parallels between study findings and research on trust conducted in the field of B2C e-Commerce.

Chapter 5 presents the first main phase of this research study which was conducted to solicit geospatial data producers' and users' views on the concept of a GEO label and the role it should serve. The chapter outlines the mechanism by which this phase of the research was conducted, discusses the study results, and presents three prototypic graphical GEO label representations which were derived and informed by the results of this and the initial study as potential means to convey availability of geospatial dataset quality information.

Chapter 6 presents the studies that were conducted as part of the second main phase of this GEO label research. The chapter first outlines the process and the results of a questionnaire-based study conducted to evaluate the effectiveness of, or potential issues with, the proposed GEO label designs. It then describes the evolution of the GEO label visualisations based on experts' feedback and recommendations. The chapter concludes with the results of the geospatial community voting for the final GEO label design and a definitive specification for the GEO label visualisation.

Chapter 7 introduces a GEO label Web service that was developed to support the generation of GEO label representations for geospatial datasets by combining producer metadata with structured user feedback. The chapter describes the method adopted for the evaluation of information availability in supplied producer and feedback documents; presents the service Application Programming Interface (API) and the resources offered by the service; and gives an overview of the technical architecture of the service.

Chapter 8 presents a GEO label-based dataset discovery tool – GEO LINC (GEO Label INterComparison tool) – that was designed and developed as a prototype online system to support geospatial dataset intercomparison and selection. The chapter overviews the interface design and functionality of the GEO LINC tool and describes the server-side and client-side technologies used in the development of the prototype system.

Chapter 9 presents a human-subject evaluation study that was conducted as part of the third and final phase of this research. The chapter describes the techniques applied to generate a set of test metadata XML records for use in the GEO LINC evaluation study and outlines the mechanism by which the evaluation study was conducted. It then reviews the study results and describes eight themes that emerged from the study data analysis. The chapter concludes with a user-dictated specification for modifications and improvements to the GEO LINC tool.

Finally, Chapter 10 provides thesis conclusions. This chapter answers the research questions, outlines practical and scientific implications and discusses contributions to knowledge and future research directions.

Chapter 2 Background Literature Review

> "*No one is willing to make use of data they do not trust, or whose accuracy they do not understand*" (Goodchild, 1995, p. 421).

This chapter outlines the important concepts underpinning this research problem via review of relevant scientific literature. The chapter provides a background in trust, geospatial data quality, and decision making. Section 2.1 reviews the literature on the concept of trust, various models of trust, and trust indicators. Having reviewed research on trust as reported across a range of domains, the focus then shifts to research into trust within the geosciences domain; in particular, Section 2.2 outlines trust-related research that has been conducted in the GIS domain. Section 2.3 explores how established trust concepts from business-to-consumer (B2C) e-Commerce can be drawn on to develop effective trust representations for geospatial data. Section 2.4 reviews the literature on geospatial data quality, quality standards and indicators, and fitness for use. This section outlines current progress and challenges in representation and evaluation of geospatial data quality. Section 2.5 provides a review of the literature on decision making, decision support systems (DSS) and visual information seeking using starfield displays. The main aim of this section is to provide an overview of processes and tools that can be adopted in the GIS domain to better support users in evaluation of geospatial data quality. Finally, Section 2.6 introduces the concept of a GEO label and describes the initial proposals that started the initiatives to conceptualise and define the notion of a GEO label.

2.1 Trust

The way in which trust is defined and conceptualised across different domains (e.g., sociology, psychology, etc.) and the importance of trust and how that importance differs according to context are discussed in Section 2.1.1. Various models of trust are then presented in Section 2.1.2, demonstrating the multidimensionality and complexity of the trust concept. Finally, Section 2.1.3 provides an overview of research conducted in the e-Commerce domain to identify trust indicators that can be embedded within websites to engender user willingness to engage in online transactions.

2.1.1 Concept of Trust

Trust is a fundamental part of our everyday life (Hosmer, 1995; Barbalet, 2009; Cvetkovich, 2013). Without the presence of trust, society would experience a loss of effectiveness, task performance and dynamism leading to its inevitable destruction (Marsh, 1994b). A crucial component of human life (Hosmer, 1995; Uslaner, 1999; Kim *et al.*, 2005), trust enables relationships and outcomes that would not be possible without its presence (Barbalet, 2009). Trust significantly influences our decision making (Luhmann, 1979; Yamamoto, 1990; Gefen, 2000), risk perception (Mayer *et al.*, 1995; Gefen *et al.*, 2003; Verhagen *et al.*, 2006; Zhu *et al.*, 2011), and willingness to rely on others (Moorman *et al.*, 1993; Chopra and Wallace, 2003; Colquitt *et al.*, 2007). Trust allows our lives to be more sociable, helping us to connect to people and cooperate with others. While trust is not the only element that enables cooperation, it makes cooperation with strangers easier and more lasting (Mayer *et al.*, 1995; Ratnasingham, 1998; Uslaner, 2002). Trust makes our life better – "*it brings us all sorts of good things*" (Uslaner, 1999, p. 1).

There exist many types of trust (Marsh, 1994b; Mayer *et al.*, 1995; Mcknight and Chervany, 1996; Schoorman *et al.*, 2007) – with trust being viewed as a multi-dimensional concept (Costigan *et al.*, 1998; Gefen, 2002; Poortinga and Pidgeon, 2003; Tan and Sutherland, 2004; Kim *et al.*, 2008a; Jelenc *et al.*, 2013) – and there are many disciplines and research fields (e.g., economics, social psychology, sociology, management, marketing, information systems, commerce, and e-Commerce) that study this phenomenon (Hassanein and Head, 2004). The definition of trust largely depends on the nature of relationships (Head and Hassanein, 2002) and contexts (Hosmer, 1995; Kim and Prabhakar, 2000; Dahwa *et al.*, 2013) to which it applies. For these reasons it is difficult to establish and agree on a single definition of trust. As articulated by McKnight and Chervany (2001, p. 29), "*trust is conceptually, like the elephant, massive in terms of the meanings it conveys*". Their analysis of Webster's, Random House, and Oxford unabridged dictionaries showed that, on average, trust had 17.0 definitions. While there is no universal agreement on definition and components of trust, any trusting relationship always involves two specific parties: "*a trusting* *party (trustor) and a party to be trusted (trustee)*" (Wang and Emurian, 2005, p. 111). The *trustor* is a person, while the *trustee* may be another person, a commercial or governmental organisation, or a piece of technology. (Skarlatidou *et al.*, 2013).

In sociology, trust is described as a mechanism for coping with the freedom of others (Luhmann, 1979; Dunn, 2000; Gambetta, 2000). Famous sociologist Niklas Luhmann (1979) defined it as a mechanism for reducing the complexity of society, referring to trust as a "*basic fact of human life*" (Luhmann, 1979, p. 4). In society, which is inherently dynamic and uncertain, the presence of risk is unavoidable. Indeed, Luhmann (1979) states that the presence of trust itself presupposes a situation of risk. He points out that trust is used not only to reduce the complexity of the world but also to help society to handle risk. Risk, where one cannot control and predict a future outcome, is an essential precondition of trust (Ratnasingham, 1998; Chopra and Wallace, 2003), yet trust does not necessarily imply taking risk, but rather it is a *willingness* to take risk (Mayer *et al.*, 1995).

In psychology, trust is viewed as a personality characteristic (interpersonal trust) (Rotter, 1980) or a "*psychological state*" (Rousseau *et al.*, 1998, p. 398) which carries the notion of "*humanness and commitment to beneficent action*" (Yamamoto, 1990, p. 453). Trust reflects an optimistic world view (Marsh, 1994a; Hosmer, 1995; Gefen, 2002), "*a belief in the goodness of others*" (Rotter, 1980, p. 1) and an anticipation that others share one's fundamental values (Jones, 1996). Trust is mostly learned during childhood: the extent of one's trust as a child largely determines the extent of one's trust as an adult (Rotter, 1967; 1980). It stems from an upbeat world view that is transmitted early in life from one's family (Rotter, 1967). Social psychologist Deutsch (1958) argues that human trust is affected both by past preoccupations and current concerns. He suggests that a trustee will suffer more if trust is not fulfilled than if the trustee does not trust in the first instance. He defines trust as follows: "*an individual may be said to have trust in the occurrence of an event if he expects its occurrence and his expectation leads to behaviour which he perceives to have greater negative motivational consequences if the expectation is not confirmed than positive motivational consequences if it is confirmed*" (ibid, p. 266). Since trust presupposes a situation of risk, a person is willing to take a risk if his perceived potential gains are much higher than his potential losses. Hence, according to Deutsch, trust and risk are two sides of the same coin.

In e-Commerce, trust is accepted as a crucial element that enables personal and market transactions (Hoffman *et al.*, 1999; Jarvenpaa *et al.*, 1999; Manchala, 2000; Pavlou, 2003; Teo and Liu, 2007; Kim *et al.*, 2008a; Guo *et al.*, 2011; Hong and Cha, 2013). Many researchers in this field adopt a common definition of trust as being a belief or positive expectation that a vendor will fulfil promised obligations and that the vendor will not take any actions that will negatively affect the trustee (Geyskens *et al.*, 1996; Ratnasingham, 1998;

Chen and Dhillon, 2003; Corbitt *et al.*, 2003; Hong-ling and Guang-xing, 2011). Mayer *et al.* (1995, p. 712) formally define trust as "*the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party*". Being vulnerable in this context implies taking a risk where there is scope to lose something of importance. Consequently, this widely accepted definition assumes the presence of risk, as in Luhmann's concept of social trust (Luhmann, 1979), and anticipation of the trustee's good intentions, as discussed in Yamamoto's work (Yamamoto, 1990).

The breadth of domains across which trust is researched has resulted in there being no absolute agreement on exactly what trust is. Head and Hassanein (2002) state that the notion of trust depends both on the context in which it is used and the type of relationship to which it applies. Kini and Choobineh (1998, p. 51) describe trust as a "*belief*", while Daignault *et al.* (2002, p. 3) and Jarvenpaa *et al.* (2000, p. 49) identify trust as "*expectation*", with Grandison and Sloman (2000, p. 2) asserting that trust can be "*specified in terms of a relationship between a trustor, the subject that trusts a target entity, and a trustee (i.e., the entity that is trusted)*". As will be discussed in section 2.2, in comparison with other domains there is still lack of concrete exploration of trust as it applies within the GIS domain. For this reason, this thesis adopts the general definition of trust proposed by Grandison and Sloman (2000) as previously quoted.

2.1.2 Models of Trust

A series of models of trust have been proposed (e.g., Ganesan and Hess, 1997; McKnight *et al.*, 2002b; Mccord and Ratnasingam, 2004; Tan and Sutherland, 2004; Lee and Yu, 2009; Moyano *et al.*, 2012; Balakrishnan and Majd, 2013). Ganesan and Hess (1997) present two dimensions of trust – *credibility* and *benevolence*. The *credibility* dimension is based on a trustee's intention and ability to keep promises and a trustor's belief in the trustee's competencies, reliability in terms of good service, and predictability in job-related behaviour. The *benevolence* dimension, according to Ganesan and Hess (1997), is based on qualities, intentions and characteristics of the trustee and the trustee's concern and care for the trustor that is not purely based on profit. Business studies of trust have identified *credibility* (the belief that the vendor has the necessary capacity to complete a task effectively and reliably) and *benevolence* (the belief that the vendor has good intentions and will behave in a favourable manner even in the absence of existing commitment) as critical factors of trust (Doney and Cannon, 1997).

McKnight *et al.* (1998; 2000; 2002; 2006) focus on *initial* trust and define a model of initial trust formation. They argue that "*almost every relationship begins with an initial phase*" (McKnight and Chervany, 2006, p. 29). *Initial* trust can be defined as "*trust in an unfamiliar* *trustee*" (Hassanein and Head, 2004, p. 16) and willingness to rely on a trustee after a first-time interaction (Koufaris and Hampton-Sosa, 2004) . *Initial* trust between two parties will not be based on any kind of previous experience but will be based on the trustor's disposition to trust (see more on disposition to trust later). "*Initial trust is fragile*" (McKnight and Chervany, 2006, p. 33) and is particularly critical in e-Commerce because e-vendors need to engender sufficient trust to convince the customers to engage in a transaction when they visit and explore websites for the first time. *Initial* trust is, indeed, considered to be the most significant factor in the first stage of vendor-customer relationships (McKnight *et al.*, 1998; Kim and Prabhakar, 2000; Koufaris and Hampton-Sosa, 2004).

McKnight *et al.* (2002a) subsequently proposed a multidisciplinary, multidimensional trust model (Figure 2.1) that has become one of the most sited trust models in e-Commerce literature. Their model is composed of four high-level constructs – *disposition to trust*, *institution-based trust*, *trusting beliefs*, and *trusting intentions* which, when combined together, lead to *trust-related behaviours*. *Disposition to trust* is defined by McKnight *et al.* (2002a) as a general propensity to trust others. *Institution-based trust* refers to a trustor's perceptions of the institutional environment and belief that effective third-party mechanisms are in place to facilitate a successful transaction. *Trusting beliefs* are defined as a trustor's perceptions of the trustee's "*competence (ability of the trustee to do what the truster needs), benevolence (trustee caring and motivation to act in the truster's interests), and integrity (trustee honesty and promise keeping)*" (McKnight *et al.*, 2002b, p. 297). *Trusting intentions* refer to a trustor's willingness to depend on, or to be vulnerable to, the trustee. Finally, *trust-related behaviours* are trustor's actions that express dependencies on, and vulnerability to, a trustee. In e-Commerce, trust-related behaviours may include disclosing personal information, engaging in a transaction, or acting on information provided by the trustee.

Aston University

Illustration removed for copyright restrictions

Figure 2.1: Web Trust Model – Overview (McKnight *et al.*, 2002a, p. 337).

Tan and Sutherland (2004) define a multidimensional model that includes three major types of trust: *institutional*; *interpersonal*; and *dispositional* trust (see Figure 2.2).

Figure 2.2: Multi-Dimensional Trust Model (Tan and Sutherland, 2004, p. 47).

Institutional trust comes from sociology and refers to trust in institutions, such as laws and regulation in society (Pennanen *et al.*, 2008) and the presence of essential structural conditions (McKnight *et al.*, 2002a). In e-Commerce, *institutional* trust denotes "*trust in the Internet as a whole*" (Tan and Sutherland, 2004, p. 49) and particularly trust in the technology that it offers (Pennanen *et al.*, 2008). *Interpersonal* trust is an individual's trust in another specific party (Tan and Sutherland, 2004); in e-Commerce, this type of trust can represent a customer's trust in an e-vendor, trust in third-party assurances of e-vendor trustworthiness and integrity, or a friend's recommendation of an online vendor (Pennanen *et al.*, 2008). *Dispositional* trust was defined in the area of psychology, and refers to an individual's ability and willingness to trust in general. Mayer *et al.* (1995, p. 715) refer to this type of trust as "*propensity to trust*". *Disposition* or *propensity to trust* is determined by an individual's disposition to form trust in general and is developed throughout an individual's lifetime (Tan and Sutherland, 2004). *Dispositional* trust is particularly important in the initial stages of a relationship and in novel situations where familiarity is absent (Gefen *et al.*, 2003; Pennanen *et al.*, 2008).

Further trust classifications include *initial* and *experiential* trust (Marsh and Meech, 2000), *vertical* and *horizontal* trust (Lee and Yu, 2009), and *technological* and *relational* trust (Mccord and Ratnasingam, 2004). *Initial* or '*grabbing'* trust refers to a first trusting judgement at the commencement of a novel situation or relationship and is highly influenced by an individual's disposition to trust. *Experiential* trust, as the name implies, comes with experience and familiarity and is considered to be much more complex than *initial* trust. In e-Commerce, initial trust plays an important role in attracting new customers: if an e-vendor fails to engender customers' trust at the initial trust-building stage, it will be very hard (if not impossible) to move the customer to a higher, more established level of trust (Marsh and Meech, 2000). Lee and Yu (2009) describe the notion of *vertical* and *horizontal* types of trust: "*vertical trust captures the trust relationships that exist between individuals and institutions,* *while horizontal trust represents the trust that can be inferred from the observations and opinions of others*" (Lee and Yu, 2009, p. 9). They speculate that these two types of trust are involved in everyday decision making and are interplayed and combined by humans. For instance, when making a decision about where to make a purchase, a customer may consider the third-party certifications that a potential vendor holds (*vertical* trust) as well as consider the recommendations and experience of his/her friends (*horizontal* trust). Mccord and Ratnasingam (2004) discuss *technological* and *relational* types of trust. They define technological trust as "*the subjective probability by which an individual believes that the underlying technology infrastructure and control mechanisms are capable of facilitating inter-organizational transactions according to its confident expectations*" (Mccord and Ratnasingam, 2004, p. 921); they refer to relational trust as "*a consumer's willingness to accept vulnerability in an online transaction based upon positive expectations of future e-retailer behaviours*" (ibid, p. 921).

2.1.3 e-Commerce Trust Indicators

Meziane and Nefti (2007) agree with other researchers that trust is a vital aspect of B2C e-Commerce and suggest that customers' security and privacy concerns directly relate to trust. Customers often distrust e-vendors due to numerous perceived potential risks; the possibility of fraud, impersonality of the service, and the need to provide private information to potentially unsafe parties push customers away from engaging in online transactions (McKnight and Chervany, 2006; Kim *et al.*, 2008a; Chang *et al.*, 2013). As previously discussed, the situation is further aggravated by the fact that e-Commerce essentially lacks the traditional environmental and interpersonal cues relied upon in the physical world for trustworthiness assessment. The prevailing acknowledgement that trust is considered to be vital to the establishment of B2C relationships has brought increased recognition of the difficulty yet importance of developing mechanisms for supporting trust formation in e-Commerce (Grandison and Sloman, 2000). In their 'call to arms', Marsh and Meech (2000) went so far as to challenge website designers to start thinking about how trust can be facilitated in the initial ('grabbing') stages of online engagement, claiming that websites can be designed in such a way that trust is an integral part of the design rather than an afterthought.

Extensive research has subsequently been conducted in the e-Commerce domain to identify trust indicators that can be embedded within e-Commerce websites to engender user willingness to engage in online transactions (Jarvenpaa *et al.*, 2000). Many researchers (e. g., Jarvenpaa *et al.*, 2000; Egger, 2001; Riegelsberger and Sasse, 2001; Yang *et al.*, 2005; Chen *et al.*, 2010; Chang *et al.*, 2013; Liu *et al.*, 2013) either directly or indirectly illustrate that these trust indicators (known as *trust triggers*) can be effective in engendering consumer trust in e-Commerce and, hence, in promoting online transactions. In essence, an online trust trigger is an element of a website that acts as an indicator of the trustworthiness of the website (Lumsden, 2009). Chang and Chen (2008, p. 818) imply that trust triggers are "*environmental cues*" that influence customers' intentions to engage in an online transaction. Lumsden and MacKay (2006) identify nine of the most generally agreed-upon trust triggers that are commonly used in e-Commerce, namely: *customer testimonials and feedback*; *professional website design*; *branding*; *third party security seals*; *up-to-date technology and security measures*; *alternative channels of communication between consumers and the vendor*; *clearly stated policies and vendor information*; *consistent (professional) graphic design*; and *ease of navigation*. They classify these trust triggers into two subsets – *immediate* and *interaction-based* trust triggers – where the latter requires extended interaction with the website in order to be exposed to the trust trigger whereas the former has influence on first sight of the website. Out of the listed trust triggers, *ease of navigation* and *consistent professional graphic design* can be classified as interaction-based; the remainder of the triggers belong to the immediate subset. Each trigger will be discussed in turn below.

Customer Testimonials and Feedback

Customer testimonials and feedback are becoming widely used by online vendors to engender consumer trust in the products and/or services that a vendor provides and consequently increase consumer purchase intention (Zhang *et al.*, 2010; Ye *et al.*, 2011; Utz *et al.*, 2012). Feedback facilities and customer reviews are used by many online retail businesses – perhaps best known is Amazon.com, which was one of the first companies to publish user reviews on books (Lee and Yu, 2009; Mudambi and Schuff, 2010; Mudambi *et al.*, 2014). Many e-Commerce consumers seek product reviews provided by other customers mainly because peer reviews are perceived as more trustworthy than the information supplied by e-vendors and experts (Smith *et al.*, 2005). As some studies indicate, even negative consumer reviews can in some cases have a positive effect and increase purchase likelihood (Berger *et al.*, 2010). Specifically, negative testimonials increase consumer awareness of products and services that were previously unknown; consumers are more likely to consider reviewed products than products that carry no reviews (Utz *et al.*, 2012).

Online reviews typically consist of two components – star ratings and review text – where star ratings act as simple aggregate indicators of consumers' overall satisfaction (Mudambi *et al.*, 2014). Researchers recognise the importance of rating systems for online businesses, and studies show that retailers who get higher customer ratings can expect greater sales (Livingston, 2005; Resnick *et al.*, 2006; Ye *et al.*, 2011; Spillinger and Parush, 2012). That said, some studies indicate considerable inconsistency in customers' interpretation of the rating scales, namely, "*one person's view of a "3" may be considerably different from another's*" (Mudambi *et al.*, 2014, p. 1). Consequently, many online consumers consider review comments that support numerical ratings as more informative and valuable when making a purchase decision (Utz *et al.*, 2012).

Professional Website Design

Professional website design, as discussed in the literature, relates to the appearance of a website, including presentation of information and general layout (Wang and Emurian, 2005). Yang *et al.* (2005) suggest that the quality or professionalism of design of a website can significantly impact consumers' first impression of a website and can increase their satisfaction and re-purchase intention. A professionally developed interface suggests that the vendor invested considerable amounts of money in producing the website. This is important because, as illustrated by Jarvenpaa *et al.* (2000) and Egger (2001), consumers are more willing to trust companies that invest more money back into their businesses. Satisfied online users are more likely to spend additional time browsing the website, revisit the website for further purchases and recommend the website to other customers (Zhang and von Dran, 2000). Consequently, high-quality websites that utilise trust-oriented interface design can expect greater levels of trust and higher revenues (Skarlatidou *et al.*, 2013).

Branding

Studies suggest that branding plays an influential role in terms of customers' trust towards vendors, both on- and offline (Jarvenpaa *et al.*, 2000; Ha, 2004; Broutsou and Fitsilis, 2012). Consumers believe that if a company is mature and large in size, then it is trustworthy because the trust and business of previous customers will have served as a foundation for the company's growth (Jarvenpaa *et al.*, 2000; Chakraborty *et al.*, 2010; Fulmer and Gelfand, 2011; Broutsou and Fitsilis, 2012). Lohse and Spiller (1998) also argue that the size of the company in the offline world affects customers' perceptions of the company's online business. This suggests that larger, more recognised and established brands will instil greater levels of trust among customers when online.

Perceived vendor reputation exerts particularly high influence on initial trust formation (Koufaris and Hampton-Sosa, 2004). New customers who have not previously visited a website will be more willing to engage in an online transaction with a reputable e-vendor (Chakraborty *et al.*, 2010). As proposed by Jarvenpaa *et al.* (2000), less-known online vendors might be able to establish and promote their reputation by describing their history, stating their policies, and including customer testimonials and feedback "*regarding the quality, value, and efficiency of their service*" (ibid, p. 65).

Third Party Security Seals

Despite the fact that B2C e-Commerce is rapidly growing, customers' privacy, security and business integrity concerns remain the main bottlenecks affecting ongoing development of e-Commerce (Belanger *et al.*, 2002; Moores, 2005; Rifon *et al.*, 2005; Kaihong and Mingxia, 2007; Mascha *et al.*, 2011; Hartono *et al.*, 2014). To resolve such customer concerns, third-party seals of approval certification programs were introduced by the e-Commerce industry (Moores, 2005; Kim and Kim, 2011). Third-party assurance seals, also referred to as trust seals, are provided to online vendors by independent third parties, such as accountants, banks or consumer unions (Noteberg *et al.*, 2003). The providers of the seals assure that online vendors adopt and comply with fair information practices, privacy policies, security, business practices integrity, etc. depending on the seal function. In order to acquire a seal of approval, vendors typically need to write a privacy policy which is then submitted to the organisation providing the seal for review and approval. The cost of acquiring a trust seal will commonly depend on the applicants' annual income, but different seal providers apply different rates and often include additional charges for inspection of applicants' websites (e.g., CPA WebTrust). The trust seals themselves are represented with graphics that have a click-to-verify function: this ensures that when the seal graphic is clicked the customer is navigated to the validation page that provides the description of the seal program and outlines the compliance procedure (Moores and Dhillon, 2003). The most commonly used third-party assurance seals in e-Commerce include TRUSTe, VeriSign, BBBOnline, and WebTrust (example graphics are shown in Figure 2.3 below).

Aston University

Illustration removed for copyright restrictions

Figure 2.3: Example graphics of commonly used e-Commerce third-party seals of assurance.

Although third-party seals of approval are adopted by online vendors to engender customers' trust, research on trust seals provides conflicting results regarding whether such seals actually engender trust and influence customers' purchase intentions (Kim *et al.*, 2008b; Lowry *et al.*, 2008). A large number of studies propose that trust seals have a strong impact on customer trust in online vendors (e.g., Noteberg *et al.*, 2003; Rifon *et al.*, 2005; Aiken and Boush, 2006; Kaihong and Mingxia, 2007). In contrast, other studies suggest that the level of consumer recognition of seals is relatively low (Head and Hassanein, 2002; Moores, 2005) and seals of approval do not influence customers' purchase intentions (Belanger *et al.*, 2002; Mcknight *et al.*, 2004; Utz *et al.*, 2012). Despite the fact that the seals of approval are largely used by e-Commerce vendors (Jiang *et al.*, 2008), studies have shown that customers often fail to understand and appreciate their role in privacy, security or business integrity assurance and any professional-looking graphic can be confused with an official third-party seal of assurance (Moores, 2005; Utz *et al.*, 2012). To investigate the level of customer recognition, Moores (2005) compared users' reactions to three official privacy seals with a fictitious seal graphic. The results showed that the customers identified a fictitious seal to be an official seal of approval. Moores suggests that online vendors fail to display privacy seals prominently enough to aid user recognition of the official graphics and increase customers' education about the role of seals in engendering trust. Other studies indicate that customers do not actually spend the time necessary to check third-party seal certificates because they have no trust in the assurance seals themselves (Head and Hassanein, 2002). In addition, the presence of a privacy assurance seal can have a negative effect on consumer trust; customers sometimes view their presence as vendors' attempt to create an illusion of privacy protection (Mcknight *et al.*, 2004). When making a purchase decision, customers often value aesthetic aspects of a website (e.g., website design, convenience, ease of use, cosmetics) (Belanger *et al.*, 2002) and customer testimonials (Utz *et al.*, 2012) more than the presence of seals of approval. To enhance the understanding of how trust seals affect customer trust, some studies have investigated whether the stage at which a seal is encountered has any effect on the extent to which trust seals influence customers (Head and Hassanein, 2002; Mcknight *et al.*, 2004). Head and Hassanein (2002) define four stages of the trust building lifecycle, namely *chaos*, *establish*, *enhance* and *maintain*. During the *chaos* phase, a customer is unaware of a vendor or is untrusting; during the *establish* stage, the customer becomes aware of the vendor by browsing the vendor's website and seeking information; at the *enhance* stage, the customer engages in and completes a transaction; and at the *maintain* stage the customer performs regular successful transactions with a given vendor. Using this trust-building model, Head and Hassanein (2002) showed that trust seals can be more influential during the *establish* and *enhance* trust-building phases, when the customers seek information about the vendor and start engaging in the first transactions. In contrast, during the *maintain* stage, when customers have become familiar with the vendor, the trust seals become unimportant to the customers. Contradicting this discovery, Mcknight *et al.* (2004) showed that privacy seals of approval have no influence on customers' trust even at an early stage of trust building.

Due to conflicting evidence regarding whether third-party seals of approval have a positive effect on trust, researchers are trying to identify the root cause of the inconsistencies in findings and what is actually influencing the effectiveness of seals. Zhang (2005) examined whether there is a connection between types of seals and their trust-engendering effect on different product types. He used four different seals to address (1) information privacy, (2) information security, (3) reliability issues, and (4) money back guarantee and studied two product types (1) commodity (e.g., oil, paper clips) and (2) look-and-feel (e.g., suits, homes). Commodity products' quality can be inferred from limited information whereas look-and-feel products' quality is more subjective and can be hard to assess without direct access to the item (Lumsden, 2009). The results of the study demonstrated that, depending on product type, some seals were more effective than others; reliability-assurance seals were the most effective across all product types. To take this further, Hu *et al.* (2010) explored whether combining several assurance functions (privacy, security and transaction-integrity) in one seal would have a greater effect on engendering customers' trust online. The results of their study demonstrated that seals with multiple functions did not necessarily have greater effect than seals that carried a single function. Interestingly, the study suggested that the privacy function had a weakening effect when it was combined with any other assurance functions. This suggests that combining several roles in one label can have a detrimental effect and the functions to be combined have to be selected and evaluated very carefully. Studies also suggest that educating customers about third-party assurance seals can increase customers' awareness of the importance of the seals (Kim *et al.*, 2008b) and that customers who have some knowledge are more willing to buy when seals are present (Hu *et al.*, 2001).

Up-to Date Technology

The up-to date technology trust trigger typically manifests through the use of secure internet protocols for private/secure transactions, data encryption and authentication techniques (Lumsden, 2009; Kim *et al.*, 2010). Customers can identify this trigger when making a purchase by the presence of a padlock in a corner of the browser and "https" in front of the current page address. Although various security mechanisms and measures have been developed to minimise the risks of online transactions, recent studies report that 95% of surveyed customers are to some extent concerned about privacy of their information and six in ten survey respondents fear theft of their credit card details (Kim *et al.*, 2010).

The importance of up-to date technology as a trigger for trust directly relates to customers' concerns regarding the privacy of transactions (Büyüközkan *et al.*, 2010). Although it might be assumed that consumers rely on the presence of the padlock symbol to assure them of the safety of their payment details, studies have shown that a large number of customers are not actively searching for this trust trigger when making a purchase online (Lumsden, 2009). Whilst this may suggest a general lack of consumer awareness of the technical aspects of online transaction and electronic payment security, studies indicate that perceived technical protection is strongly associated with perceived security and consumers' perceived trust (Kim *et al.*, 2010). Thus, to engender and maintain customer trust, it is essential that e-vendors employ effective up-to date mechanisms for secure transactions and ensure privacy of personal customer data (Hasan and Abuelrub, 2011).

Egger (2001, p. 322) provides a list of recommendations that can be adopted by e-vendors to address customers' security concerns:

- *"List the measures taken to ensure that data is transferred, processed and stored securely;*

- *Provide prominent links to the security policy;*
- *Mention what hardware and software solutions are used: provide external links to providers;*
- *Complement browser feedback with text to inform users that they are on a secure page;* [and]
- *Provide several payment options."*

Alternative Channels of Communication

Alternative channels of communication on a website typically refer to the provision of a company's address details and/or a phone number such that consumers can communicate with the vendor via more than just email (Egger, 2001; Fogg *et al.*, 2001; Araujo and Araujo, 2003; Chen and Dhillon, 2003; Parasuraman *et al.*, 2005). Yang *et al.* (2005) state the importance of not only online but also offline channels of communication in engendering trust. Availability of an address or a phone number suggests that the vendor exists in the offline world and can be reached through more traditional means of communication. Via their qualitative credibility evaluation of two live websites conducted with 2,684 participants, Fogg *et al.* (2002, p. 41) discovered that "*a Web site wins credibility points by giving information about the organization behind the Web site: who they are, what they do, and how to contact them*". Researchers further concluded that availability of contact details and photographs of people behind the company provide website customers with a real-world feeling. More recent studies also confirm that traditional offline channels of communication are still widely used by online consumers (van Dijk *et al.*, 2007), meaning that it is important for e-vendors to consider offering both online and offline communication channels to their customers (Sivaji *et al.*, 2011).

Consistent Graphic Design and Ease of Navigation

These two trust triggers require user interaction with the website in order to influence trust or distrust. In the case of e-Commerce, since the website is sometimes the only means of communication between an e-vendor and its customers (Chen and Dhillon, 2003), the website's appearance and ease of interaction are critical factors that contribute to vendor's credibility and trustworthiness (Fogg *et al.*, 2002). Efficient navigation schemes, organized layouts and consistent graphic design aid in developing trust towards e-vendors (Koufaris and Hampton-Sosa, 2004; Yang *et al.*, 2005), increase consumer sense of control (Chen and Dhillon, 2003), positively affect perceived usefulness of the website (Kim, 2012), improve the consumer shopping experience (Sivaji *et al.*, 2011), and significantly impact consumer purchase intentions (Zhu *et al.*, 2011).

2.2 Trust in GIS

Having reviewed research on trust as reported across a range of domains, this section focuses on trust-related research that has been conducted in the GIS domain. The section thus presents literature on trust in geospatial data, volunteered geographic information (VGI), and geospatial systems and services.

There are a number of parallels between consumers' decisions to transact with a given e-Commerce vendor and dataset users' decisions to adopt one from *n* datasets for their given needs. Despite the significant growth in availability of geospatial data and the fact that geospatial datasets can, in many respects, be considered commercial products that are available for purchase online, trust has to date received relatively little attention in the GIS domain (Skarlatidou *et al.*, 2011a; Skarlatidou *et al.*, 2013). Given that millions of datasets from different producers are currently available through numerous Spatial Data Infrastructure (SDI) catalogues and clearinghouses (Crompvoets *et al.*, 2004; Maguire and Longley, 2005; Steiniger and Hunter, 2012; Sui, 2014; Tumba and Ahmad, 2014), it is surprising that research into mechanisms of representing trust in the GIS domain has not yet received the same level of attention as it has in the e-Commerce domain (Skarlatidou *et al.*, 2011a; Skarlatidou *et al.*, 2013). Some GIS researchers and scholars have made attempts to highlight the importance and relevance of trust to geospatial data and systems. Harvey (2003), for instance, evaluated effects of trust on development of the National Spatial Data Infrastructures (NSDIs) in the United States. He assessed the influence of trust on willingness of local governmental bodies to share geospatial data with the NSDIs. The results of telephone and face-to-face interviews, surveys and workshops with local government agencies' staff indicated that trust directly impacts willingness to share data and therefore "*is as integral to the development of SDIs*" (Harvey, 2003, p. 35). The surveys revealed that local government officials "*commonly share data on an informal basis with people and institutions they trust*" (ibid, p. 33). In this particular situation, government data producers actually inherit the role of trustors and the users (individuals, institutions and NSDIs) with which government producers share their data become the trustees. This is an example in which trust has to exist in both parties (trustor and trustee) to enable user-producer interaction and data sharing.

Bertino *et al.* (2008) discuss the role of trust in terms of managing, accessing and sharing of geospatial data that is used for safety-critical applications. The researchers argue that "*since geospatial data is used for critical applications such as emergency response, it is important that users can trust the retrieved geospatial data*" (Bertino *et al.*, 2008, p. 15). They propose that, to engender user trust in geospatial data, geospatial data repositories should: maintain complete logs of data provenance including data source and the submission date; utilise mechanisms for dynamic verification of the data source; and introduce privacy policies for protecting sensitive information from privacy violations. Bertino *et al.* (2008) also highlight the importance of past feedback about the data source (e.g., feedback on accuracy of datasets supplied in the past), the creation time (data freshness), and the data content (data correctness, i.e., has data been modified/tampered with) in perceived trustworthiness of geospatial data.

With recent growth in production and availability of *volunteered* geographic information (VGI), trustworthiness of VGI is increasingly attracting attention of the GIS community. Unlike traditional geospatial information, VGI is commonly generated by non-experts who have no formal training or expertise in collecting and describing geospatial data; such diversity in data origin directly affects the quality of produced geospatial information, creating a need for new quality evaluation techniques (Christian, 2010). Flanagin and Metzger (2008), in their discussion about credibility of VGI, suggest that credibility (or believability) of VGI can be evaluated in terms of the trustworthiness and expertise of the data source. In the case of VGI, data is often contributed by anonymous sources, reused and repurposed (data mashups) making evaluation of source credibility nearly impossible. Consequently, the authors argue a need for research into effective ways of tracking the *provenance* of VGI to ensure that VGI sources are more explicit. With a focus on filtering more reliable socially-generated geospatial content, Bishr and Kuhn (2007), Bishr and Mantelas (2008), and Bishr and Janowicz (2010) go as far as to propose using trust as a proxy measure of VGI quality. Collectively, they argue that quality is a subjective measure, but if trust-rated geospatial information is useful and relevant to many users then it is of satisfactory quality. To compute an aggregated trust rating for a geospatial observation, their proposed trust model takes into account (a) the location of the observer while making an observation (i.e., the closeness of the observer to the phenomenon), (b) reliability ratings provided by information users, and (c) the number of times the observation has been reported by other observers. Keßler and de Groot (2013) also support the idea of using trust as a proxy measure of VGI quality and identify five provenance-based trust parameters of VGI observations, namely: *versions*; *users*; *confirmations*; *tag corrections*; and *rollbacks*. A high number of *versions* indicates observation trustworthiness because the observation went through a certain number of iterations to improve its quality. A high number of *users* involved in the creation of the observation increases its trust measure. Updates to the immediate neighbouring features (within a 50m buffer of an observation) can be considered as indirect *confirmations* of the observation's accuracy which increase observation trustworthiness; the rationale is that when contributors edit an observation they also look at the features in the vicinity. *Tag corrections* indicate uncertainties in the observation classification and thus decrease trustworthiness. *Rollbacks* which revert an observation to its previous state decrease observation trustworthiness. To test whether VGI quality can be assessed using a provenance-based trust model, the authors carried out statistical analysis of correlation between trust assessments calculated for 74 OpenStreetMap (http://www.openstreetmap.org) observations and quality measurements of these observations derived from a field survey. The results of the analysis indicated a significant correlation between model-based trust evaluations and the ground truth quality measures of the OpenStreetMap observations. As can be observed, trust research in VGI largely focuses on provenance information as a measure of data quality and trustworthiness. Indeed, availability of complete provenance information is essential in evaluation of VGI credibility.

Relation of trust to geospatial systems and services has also received the attention of researchers and scholars in the GIS community. Alam *et al.* (2007) and Umuhoza *et al.* (2008) discuss trustworthiness of Geospatial Semantic Web Services (GSWS) – services that allow querying of geospatial information by applying analytical processing to geospatial data. Alam *et al.* (2007) suggest that Quality of Service (QoS) of GSWS directly relates to service trustworthiness. The authors present DAGIS (Discovering Annotated Geospatial Information Services) – "*a semantic Web services based framework for* [the] *geospatial domain that has* [a] *graphical interface to query and discover services*" (ibid, p. 268). The framework is designed to support users in selecting the best and most trustworthy service based on QoS evaluation. To address the issues of reliability, Umuhoza *et al.* (2008) propose using trust based on the quality of data that services provide as a means of filtering GSWS. The authors also discuss concepts of a trust ontology for GSWS which includes: data characteristics (accuracy, completeness, theme, etc.); service descriptions as supplied by the service provider; information about the service provider; direct evidence of past service performance (past experiences of service user); and indirect evidence of past service performance (past experiences provided by third parties).

Despite the fact that GIS researchers have been discussing the importance of trust in different GIS contexts (e.g., trustworthiness of geospatial data, credibility of VGI, trust in geospatial services, etc.), thus far no consideration has been given to user-centred design (UCD) in determining measures of trust in the GIS domain, with the only exception being work carried out by Skarlatidou *et al.* (2010b; 2011b; 2013). Skarlatidou *et al.* (2010a; 2010b; 2011a; 2011b; 2013) focus their research on trust-oriented interface design for Web GIS applications and tools. Using an HCI-based approach, they attempted to identify the trustee attributes that influence non-experts' trust perceptions and to define a set of guidelines to improve trust in Web GIS. Via a literature review of trust in electronic online environments and heuristic evaluations and cognitive walkthrough of the UK Environment Agency 'What's In Your Back Yard' (WIYBY) website (http://apps.environment-agency.gov.uk/wiyby/default.aspx), Skarlatidou *et al.* (2010a) identified a set of trust guidelines organised into five design dimensions, namely: *graphic*; *structure*; *content*; *functionality;* and *trust cue* design. The *graphic* design dimension refers to the quality of the graphic design of the user interface (UI) and GIS visualisations; this dimension includes such aspects as high-quality professional graphics, appropriate colour combinations, design consistency, and map size (should be larger than 400×600 pixels). The *structure* design dimension refers to the structure of the provided information and aims at simplifying navigation and promoting the system's perceived transparency. In the Web GIS context, to adhere to structure design guidelines, a legend should always be provided with a map, even if a map is considered as simple. Additionally, hiding map features demotes the trustworthiness of a system. Although accuracy and reliability of information are important in user perception of a system's trustworthiness, non-expert users do not always have the knowledge to assess these elements. The *content* dimension therefore aims at improving such attributes as vocabulary used, map scales provided (a map should provide at least 4 zoom levels), and availability of information about data and potential issues. It is also recommended that users are offered instructions and tutorials about the system or system components. The *functionality* design dimension focuses on improving functionality-related attributes of the Web GIS systems. Due, however, to the fact that the functionality of such systems is highly dependent on the context of their use, the only recommendation is to ensure that the system does not have 'broken' links and 'not found' pages. Finally, the *trust cue* design dimension refers to trust-inducing features that can be incorporated in the Web GIS interface design. Such trust-inducing features include: logos and branding; copyright and data issues information; privacy policy information; external links to additional information (these have to be regularly checked for validity); contact details; blogs which connect people and create user networks; information about data accuracy and data provenance; and map tutorials, ideally provided below the map. As can be seen, these trust cues largely mirror the e-Commerce trust indicators described in Section 2.1.3. To further examine how interface design and adherence to trust guidelines affect perceived trustworthiness of the Web GIS applications, Skarlatidou *et al.* (2010b; 2011b; 2013) conducted a number of evaluation studies with non-expert users. Usability studies of the Web GIS applications indicated that specific design attributes, such as map size, colours and presence of a legend, increase perceived trustworthiness (Skarlatidou *et al.*, 2010b; 2011b). The studies also showed that, in the absence of the trust design attributes, non-expert users base their trust perceptions on the reputation of the application provider – despite low usability of the WIYBY website the majority of study participants trusted the application because it was provided by a governmental organisation. Interestingly, metadata was completely ignored by non-expert users. A controlled user study further indicated that the Web GIS interfaces which incorporated all the trust design guidelines were perceived as the most trustworthy by non-expert users (Skarlatidou *et al.*, 2013).

Despite recent research efforts extended to highlight the importance of trust in the GIS domain, there is still a lack of formal identification/definitions of GIS-specific trust indicators and cues. There is also lack of transition of trust knowledge from other domains such as psychology, sociology and e-Commerce where notions of trust and trust cues have been established and empirically confirmed. In particular, very little attention has been given to the impact of trust on geospatial data selection and use. Only in recent years, since the availability of VGI has significantly increased, have researchers begun to discuss trustworthiness of volunteered geospatial data. This still leaves the trustworthiness of standard geospatial data as yet overlooked. Consequently, the next sections of this chapter attempt to bridge this gap and apply trust knowledge from other domains to the geospatial data, its selection and use.

2.3 Applying Trust Knowledge to the GIS Domain

The previous sections have reviewed the concept of trust, trust models, and trust indicators from various research domains. The following section moves on to discuss how this trust-related research applies to the GIS domain and to geospatial data selection and use.

2.3.1 Risk as a Precondition of Trust

As was discussed in Section 2.1.1, risk is a vital precondition of trust. Since risk can arguably be considered high in GIS dataset use – e.g., risks associated with selecting inappropriate data, risks of data misuse and risks of data misinterpretation – it follows that there is an inherent need for trust at some level in order to facilitate effective dataset selection and use. Misuse of geospatial data and use of datasets that are not fit for an intended purpose can potentially lead to high financial costs, lead to legal actions being taken, have ecological or social impact, and even result in a loss of human life. In recent years, when production and public availability of geospatial data has significantly increased, risks associated with selecting inappropriate data have also increased, meaning that dataset users' trust in dataset producers, dataset providers and the datasets themselves plays a crucial role in establishing effective producer-consumer relationships, which in turn facilitate effective use of geospatial data. As with e-Commerce, this means that it is important that effective means of communicating the trustworthiness of datasets and their producers and providers are identified within the GIS domain.

2.3.2 Initial and Experiential Trust

In the last decade the production, availability, and sharing of geospatial data has significantly increased (Wang and Huang, 2007; Brown *et al.*, 2013), with a corresponding increase in the availability of Spatial Data Infrastructures (SDIs), web-based catalogues, portals, standards and services. For instance, since 1993 when the concept of a Spatial Data Infrastructure (SDI) was formally initiated, more than 100 local, regional, national, and global SDIs have been established across the world (Maguire and Longley, 2005). Geospatial data can now be accessed via catalogues and portals which document data from different providers. In parallel, there are increasing numbers of professional and non-professional geospatial data *consumers* searching for data to fit their specific needs. Geospatial data producers and providers, especially the ones that are new and have no established reputation, face a challenge in terms of engendering sufficient trust to convince data consumers to acquire and use their datasets. Geospatial data providers have to ensure that first-time data consumers can and will establish sufficient trust (in this case, fragile *initial* trust) to establish a new consumer-provider relationship. As with e-Commerce, it is therefore essential to identify informational aspects (i.e., geospatial-specific trust triggers) that can help to promote user trust in geospatial data providers and the datasets that they offer. Conversely, when searching for datasets to meet their needs data consumers face a challenge in terms of assessing the trustworthiness of data providers and the quality of the datasets that they offer. If data providers do not engender enough trust at the initial trust-building stage, first-time consumers will not engage in a provider-customer relationship. While *initial* trust is critical in attracting new data consumers, maintaining trusting relationships can be even more demanding. Where there is a choice, geospatial data users will be highly unlikely to continue any relationship after having a negative experience with a data provider, either directly or via the quality of their datasets. *Experiential* trust is, therefore, an essential part of more established and long term GIS consumer-provider relationships.

2.3.3 Vertical and Horizontal Trust

International organisations and initiatives such as the OpenGIS Consortium Inc. (OGC) (OGC, 2014b), International Organisation for Standardisation Technical Committee (ISO/TC 211) (ISO/TC211, 2014a), INSPIRE (INSPIRE, 2014d), A Quality Assurance Framework for Earth Observation (QA4EO) (QA4EO, 2014), Dublin Core (DCMI, 2014a), and many more, are actively working on establishing and supporting geospatial data and metadata standards. In the GIS community, international standards on geospatial data quality largely concentrate on providing guidelines for, and quality control of, metadata records. These standardisation activities directly relate to *vertical* trust: when making a dataset selection decision, consumers (or users) of geospatial data consider the international standards supported by datasets, drawing on and establishing *vertical* trust (recall, this is trust between an individual and an organisation) in so doing. Adherence to international standards may indicate to data users that datasets are either of good quality or are at least supported with adequate documentation to enable effective fitness-for-use evaluation. Although standardised supporting documentation is important in the dataset assessment phase, data users usually seek advice and recommendations from their colleagues and peers in the data discovery phase – the formulation of and reliance on *horizontal* trust.

2.3.4 Technological and Relational Trust

Technological trust in the GIS context relates to data consumers' trust in technologies that allow access to, acquisition and use of geospatial data. These technologies can include:

- national and regional SDI systems that provide access to the catalogues of metadata;
- data portals that provide an interface for data discovery and acquisition;
- producers' websites that allow purchase or free download of datasets;
- technologies and algorithms used to collect and compute the data in the datasets; and
- Web GIS applications.

Relational trust in the GIS context refers to a geospatial data user's belief that a dataset provider will demonstrate favourable behaviour in the future – in essence, it is a measure of a users' willingness to accept vulnerability (perhaps in terms of the perceived suitability of a dataset for a given purpose) based upon positive expectations concerning the dataset provider's future behaviour. This may be assessed by means of the availability of valid contact information and the ability to contact the data provider in the future, or the availability of a customer support service to respond to user queries regarding datasets and services that a provider offers.

2.3.5 Credibility and Benevolence

In the GIS domain *credibility* relates to a data consumer's belief that a given data provider is capable of reliably supplying data of good quality and providing services of a high standard. GIS dataset providers can include commercial companies that sell data, governmental organisations, SDIs providing catalogues of available geospatial data, spatial data clearinghouses, data portals, or providers of VGI; the reputation of such providers directly relates to the *credibility* dimension of trust. The corresponding *benevolence* dimension of trust in the GIS domain relates to a data provider's good intentions towards data consumers, even where there is no direct financial gain. It is hypothesised that dataset users might be more likely to acquire data from a provider if they have established a belief that the provider will assist in any further queries regarding the data that the provider has supplied.

2.3.6 Web Trust Model

As discussed in section 2.1.2, the *Web Trust Model* (McKnight *et al.*, 2002a) is composed of four high-level constructs – *disposition to trust*, *institution-based trust*, *trusting beliefs*, and *trusting intentions* – which, when combined together, lead to *trust-related behaviours*. In the GIS context *disposition to trust* will especially affect trusting beliefs of novice users that are not very familiar with the domain. It is anticipated that, when selecting a suitable dataset, users with generally low propensity to trust others will be particularly cautious and considerate when assessing datasets and their quality. Presence of well-defined and established geospatial data trust indicators (trust triggers) could be particularly important to this type of users because they would probably require as much information as possible available to them to make a data selection decision. *Disposition to trust* will also be highly influenced by a data consumer's previous experiences. *Institution-based trust* in the GIS context concerns producers' compliance with international standards or any applicable certification programmes; dataset repositories, data portals and clearinghouses may be recognised by geospatial data users as third-party institutions that can be trusted to facilitate successful data acquisition. *Trusting beliefs* in the GIS context can be defined as geospatial data consumers' confidence that data providers will supply data of good quality (provided with well-documented metadata records) and will act in a favourable manner – for instance, will reply to further queries, send notifications about any dataset updates or discovered issues, warn about potential errors, suggest dataset application areas, etc. *Trusting intentions* refer to a geospatial data consumer's belief that a dataset is fit for its intended purpose and willingness to acquire geospatial data from a data provider. Finally, *trust-related behaviours* in the GIS context will be witnessed when data consumers acquire and use datasets from data providers.

2.3.7 Multi-Dimensional Trust Model

The *Multi-Dimensional Trust Model* (Tan and Sutherland, 2004) (see section 2.1.2) is very similar to the *Web Trust Model* already discussed. This model also includes *institutional* and *dispositional* trust, *trusting intentions* and *trust related behaviours* (online purchase behaviour), but additionally presents *interpersonal* trust. *Interpersonal* trust in this model's context, and when applied to the GIS domain, would likely manifest in geospatial data consumers' trust in given data providers, in peer advice and recommendations, and in journal papers or technical reports that provide dataset quality checks.

Trust clearly has the potential to have a major impact on users' geospatial dataset selection and quality evaluation processes. When searching for a suitable geospatial dataset, users may come across new data repositories or unknown data providers, in which case they have to decide whether to engage into a trusting relationship (*initial trust*) with that provider. Users may reflect on their previous experiences with geospatial data producers and providers to decide whether or not to return to a dataset provider to acquire future data sets (*experiential trust*). Furthermore, in any decision to trust a dataset provider, a user is essentially making an assessment of the provider's *credibility*, the *technological* trustworthiness of the provider or producer, and the observance of standards set by higher orders (*vertical* trust). Users may also contact their peers, work colleagues or friends to get advice and recommendations on what data could be suitable for given tasks (*horizontal* and *interpersonal* trust), they may seek information on projects or companies who have previously used a given dataset (*horizontal* trust), or they may look to journal papers, expert reviews and technical reports where dataset quality checks have been reported when making a selection decision (*horizontal* trust). When selecting from several dataset options, some users may be more keen on datasets that adhere to international standards and are supported with standardised metadata documentation (*institutional* and *vertical* trust). In contrast, in situations where consequences of data misuse are very severe, users may choose not to select datasets themselves but to use a third-party organisation to select datasets for them (*institutional* trust, *risk precondition*, third-party *credibility*).

The mapping between trust concepts and GIS dataset selection and use suggests that trust plays a vital role in geospatial data selection and use. Every time geospatial data users select a dataset to use, they are likely to have to make a trusting decision, often without even realising they are doing so. Consequently, the research reported in this thesis aims at identifying geospatial data trust and quality indicators upon which users rely when selecting a dataset that fits their needs.

2.4 Geospatial Data Quality

This section reviews the literature on geospatial data quality, quality standards and indicators, and fitness for use. The aim of this section is to outline current progress and challenges in representation and evaluation of geospatial data quality. Section 2.4.1 thus discusses the notion of geospatial data quality and discusses some widely accepted definitions. Section 2.4.2 discusses current international attempts to standardise and communicate geospatial data quality. This section also presents commonly accepted geospatial data quality indicators. Section 2.4.3 discusses the concept of fitness for use evaluation and presents research that has attempted to improve geospatial data quality evaluation in terms of user requirements.

2.4.1 The Notion of Geospatial Data Quality

The quality of geospatial data can be defined as *"a measure of the difference between the data and the real world that they represent"* (Goodchild, 2006, p. 13). The greater this difference, the poorer the quality of data and the smaller its true value. Having to go through processes of generalisation, abstraction and aggregation, geospatial data can only provide an approximation of the real world and therefore almost always suffers from imperfect quality (Goodchild, 1995; Li *et al.*, 2012). Perfect representation of the real world with all its unlimited complexity and level of detail cannot be achieved (Devillers and Jeansoulin, 2006; Goodchild, 2006). Consequently, almost all geospatial data has limited accuracy (Goodchild, 1995) and is inevitably uncertain (Couclelis, 2003). Goodchild (1995) identifies six major sources of error in spatial data: *measurement*; *definition*; *lack of documentation*; *distortion in physical media*; *processing*; and *interpretation*. The error of *measurement* takes place when raw geographic data is produced and stays with the "*product through the entire process of dataset creation and use*" (Goodchild, 1995, p. 414). The data acquisition process, for example the technique or instruments used, generally determines the quality of produced data. *Definition* errors arise from variations between different observers in terms of the definitions of variables being measured. Such definition variations cause misclassifications and data inaccuracy. For instance, what one observer records in the dataset as 'hill' can be defined as 'mound' by other observer. *Lack of adequate documentation* supporting spatial data can also contribute towards error propagation; having insufficient metadata means that discrepancies and other errors become harder to identify. *Distortion in physical media* occurs as a result of the digitalisation of paper maps and is probably less relevant at present when most spatial data is collected with satellite and digital technologies. *Processing* geospatial data to create different data products introduces more errors and thereby increases data uncertainty. When data lacks adequate documentation the responsibility for *data interpretation* lies more with the data users, which can lead to interpretation errors. Having analysed the major sources of spatial data errors, Goodchild concludes that imperfection in spatial data makes measuring and documenting its quality essential.

Worboys (1998) identifies five factors that can lead to spatial data quality deficiencies: *inaccuracy* and *error*; *vagueness*; *incompleteness*; *inconsistency*; and *imprecision*. He describes the *inaccuracy and error* factor as "*deviation from true values*" (Worboys, 1998, p. 258) – i.e., a difference between produced data and the real world. These quality deficiency factors can arise from imprecise *measurements*, *distortion in physical media* or *processing* of data as identified by Goodchild (1995). The *vagueness* factor – "*imprecision in concepts used to describe the information*" (Worboys, 1998, p. 258) – occurs when different producers and data analysts use different terms to describe the same concepts, objects and object properties. This factor directly relates to the Goodchild's error in *definition* discussed above.
The *data inconsistency* factor arises when information conflicts exist in the data produced. For example, inconsistency can emerge from a dataset which encloses data from multiple sources or where data producers are not complying with common standards and best practices (Mohammadi *et al.*, 2009). The *inconsistency* factor affects many spatial datasets as they are aggregated from multiple sets of data which may each have different levels of quality or even come from different sources. *Lack of documentation* and *interpretation* errors, as described above, can contribute to the *inconsistency* of the dataset. The *imprecision* factor implies low resolution and granularity of spatial data; when data does not provide a sufficient level of detail and/or precision it leads to data uncertainty. This factor typically arises from low resolution *measurement* instruments and *distortion in physical media*.

Collins and Smith (1994) classify the errors depending on the phases of data collection and use: *data collection* (e.g., errors of measurement, errors produced by data collection equipment, etc.); *data input* (e.g., data digitalisation errors); *data storage* (e.g., errors caused by numerical imprecision, rounding); *data manipulation* (e.g., error propagation, map mash-ups); *data output* (e.g., errors produced from scaling by output device); *data usage* (e.g., data misuse or misinterpretation). Their classification schema closely mirrors classifications proposed by Goodchild (1995) and Worboys (1998) but introduces new concepts such as *data input*; *storage*; *manipulation*; and *output* errors.

Beard (1989) proposes to classify map errors into three main categories: errors produced through data acquisition (*source* errors); errors introduced by data processing (*process* errors); and errors caused by data misuse (*use* errors). It can be argued that this classification scheme is a high level categorisation of the geospatial data errors and factors proposed by Goodchild (1995), Worboys (1998) and Collins and Smith (1994). The *source* errors category encompasses *data collection* errors; errors of *measurement*; *lack of documentation*; *interpretation* (or *vagueness*) errors; *incompleteness*; *inconsistency*; and *imprecision*. The *process* errors can include errors of *processing*; *input*; *storage*; *manipulation*; *output*; *inaccuracy*; *inconsistency*; and *imprecision*. Finally, *use* errors comprise errors of *interpretation* and *misuse*.

While a number of classifications have been identified to categorise the factors that affect geospatial data quality, the proposed categories are generally consistent and interrelated. Spatial data quality deficiency can arise from any combination of the numerous factors described above; hence, as already mentioned, almost all the spatial data that is being produced is uncertain and imprecise to some degree, and the aim must be to control and document the errors such that they do not adversely affect the use to which such data is put.

Devillers *et al.* (2005) and Devillers and Jeansoulin (2006) discuss two categories of data quality: *internal quality* and *external quality*. *Internal quality* refers to the level of similarity between the data produced and the "perfect" data that should have been produced. In the GIS domain, *internal quality* is often described in terms of the 'famous five' elements of geospatial data quality (see more on these elements in section 2.4.2). As argued by Devillers and Jeansoulin (2006), the internal quality of data can be improved during the course of data creation. *External quality* refers to how well a product meets user's needs or expectations, in a given context. *External quality* is not absolute and is subjective; it largely depends on user requirements and therefore the same product can be of different quality to different users. Due to its subjective nature, *external data* quality is often defined as 'fitness for use' or 'fitness for purpose' (see more on 'fitness for use' in section 2.4.3). Despite the diversity in notions of *internal* (objective) and *external* (subjective) data quality, these two categories are closely linked together because, in order to evaluate *external* data quality, users will often require objective data quality descriptions. While there exist methods for evaluation of *internal quality* of geospatial data, evaluation of *external quality* still remains an open issue in the GIS domain (Ivánová *et al.*, 2013).

2.4.2 Geospatial Data Standards and Quality Indicators (Internal Quality)

Spatial data is being increasingly produced, shared, evaluated and used by GIS professionals and non-expert users who are always interested in data of high quality (Wang and Huang, 2007; Brown *et al.*, 2013). Unsurprisingly, spatial data quality and uncertainty has been an area of active research within the Geographic Information (GI) community during the past three decades (Devillers *et al.*, 2010). The international geospatial communities are actively working on imposing quality standards, improving quality of spatial data and providing tools for data quality assessments (Triglav *et al.*, 2011). International organisations, initiatives, and working groups, as well as data producers, largely focus their attention on quantitative (*internal*) quality of geospatial data (Gervais, 2006). Thus, this section will review research on quantitative geospatial data quality indicators (or attributes) and standardisation attempts to improve quality of geospatial data and its descriptions (metadata records).

2.4.2.1 Geospatial Data Quality Indicators and Attributes

The importance of spatial data quality indicators is widely recognised in scientific literature (e.g., Caprioli *et al.*, 2003; Devillers *et al.*, 2007; Wang and Huang, 2007). Devillers *et al.* (2002, p. 50) argue that quality indicators are "*a way of seeing the big picture by looking at a small piece of it*". They suggest that quality indicators can inform users of a global measure of quality without them having to examine the data in much detail. Indicators significantly simplify quality evaluation, decision-making and justification processes by providing a number of quality cues that are easy to manage and avoiding information overflow (Devillers *et al.*, 2007). Many researchers and scholars refer to the 'famous five' as the common criteria for evaluating spatial data quality (Duckham, 2000; Pundt, 2002), namely: *lineage*; *completeness*; *consistency*; *positional accuracy*; and *attribute accuracy*. Devillers *et al.* (2007), refine the 'famous five' to be: *positional accuracy*; *attribute accuracy*; *temporal accuracy*; *logical consistency*; and *completeness* as common spatial data quality elements. Caprioli *et al.* (2003) identify four major elements of spatial data quality: *accuracy*; *resolution*; *consistency*; and *completeness*. They further refine accuracy into *spatial*, *temporal* and *thematic accuracy*. Each of these commonly accepted spatial data quality attributes are discussed in more detail below.

Accuracy

Accuracy is a measure of difference between the produced spatial data and the real world that the data represents (Chrisman, 1991). The level of geospatial data accuracy varies significantly because highly accurate data can be costly and complex to produce. The concept of geospatial data accuracy can be refined to *horizontal*, *vertical*, *attribute*, *conceptual*, and *logical* accuracy. Accuracy is a relative measure and always depends on some defined specification of a true value.

Attribute/Thematic Accuracy

Attribute or thematic accuracy denotes the correctness of object classifications and the level of precision of attribute descriptions in the produced data (Cockcroft, 1997). The data produced can have high positional accuracy but objects can be misclassified or a low level of detail is provided. For instance, a line in a dataset that denotes a river can be misclassified as a road. On the other hand, the classification of the object can be correct but the description of it can be insufficient; for instance, a farm object can have the farmer or crops descriptions missing from it.

Positional/Spatial Accuracy

Positional or spatial accuracy is the level of accuracy of the spatial objects in a dataset (Stein and van Oort, 2006). It is defined as *"the difference between the recorded location of a feature in a spatial database or in a map and its actual location on the ground, or its location on a source of known higher accuracy*" (Tucci and Giordano, 2011, p. 453). Positional accuracy can be refined to *horizontal* and *vertical* accuracy as it applies to horizontal and vertical positions of captured data.

Temporal Accuracy

Temporal accuracy is the difference between encoded dataset values and the true temporal values of the measured entities (Veregin, 1999; Devillers and Jeansoulin, 2006). It only applies when the dataset has a temporal (time) dimension in the form of [x, y, z, t]. Temporal accuracy indicates the time stamp applied to the entities in the dataset. It is often mistaken with data currency – up-to-dateness of the data – even though these two concepts are quite distinct since currency refers to the degree to which a database is up to date (Veregin, 1999).

Lineage

The lineage of geospatial data is the historical information about the data which refers to how the data has been collected and processed to arrive at the final data product (Stein and van Oort, 2006). Geospatial data lineage includes information on data source, data producer, data content, capturing effort, the methodology applied to collect the data, processing steps that were applied to derive the data product, algorithms applied, geographic coverage of the data, and other historic information.

Completeness

Geospatial data completeness measures the omission error in the data and its compliance with data specification. From a data supplier's point of view, it can be defined as "*a measure of the degree to which data content corresponds to the real world in accordance with the data capture specification, dataset coverage, and at the level of currency required by the update policy*" (Harding, 2006, p. 150). Highly generalised data can be accepted as complete if it complies with its specification of coverage, classification and verification.

Consistency

Geospatial data consistency can be defined as the absence of conflicts or contradictions in a dataset (Caprioli *et al.*, 2003). Geospatial data consistency includes logical consistency, topological consistency, temporal consistency, and thematic consistency. Logical consistency relates to structures and attributes of geospatial data and defines compatibility between dataset objects – e.g., variables used adhere to the appropriate limits or types (Servigne *et al.*, 2006). Topological consistency is the dataset compliance with topological rules – e.g., no objects can have x-coordinate values below 0, polygons cannot intersect (Caprioli *et al.*, 2003). Temporal consistency is conformance to temporal topology rules – e.g., the dataset rules can specify that only one event can occur in one place a given time. Temporal consistency relates to dates of data acquisition, types of updates, and validity of periods (Servigne *et al.*, 2006). Thematic consistency measures conflicts in thematic attributes – e.g., a population density value must be correct given population and area (Caprioli *et al.*, 2003).

Resolution

Resolution is the amount of detail that geospatial data contains and is also known as precision or granularity (Caprioli *et al.*, 2003). Resolution is always finite because no measurement system can be infinitely precise. High resolution does not necessarily mean better fitness for use: in some cases low resolution may be required to formulate more general models.

The broad scientific acceptance of the common spatial quality elements does not imply their applicability to all the cases of quality or fitness for use evaluation (Pundt, 2002) since user requirements can go far beyond the widely accepted 'famous five'. While no tangible user-defined quality indicators to specifically assist fitness for use evaluation have been identified, there are many existing forms of metadata (such as documentation describing subjective quality measures outlined in this section) which can potentially be used to this end if they are consistently supplied, and can be easily viewed by a user through the prism of their own priorities.

2.4.2.2 International Attempts to Standardise Geospatial Data

A standard, as defined by the International Organization for Standardization, is "*a document that provides requirements, specifications, guidelines or characteristics that can be used consistently to ensure that materials, products, processes and services are fit for their purpose*" (ISO, 2014c). With ever growing production and availability of geospatial data, the importance of quality standards increases, requiring effective quality control to standardise geospatial data quality. Yet, as a result of the fact that data comes from different sources, varying from private researchers to national agencies, adherence with quality standards is hard to guarantee. International standards on spatial data quality largely concentrate on providing guidelines for, and quality control of, metadata records – supporting documents that supply information about the data they represent. The main purpose of metadata records is to assist end users in understanding the quality of data and assessing a dataset's fitness for use. Data producers do not, however, always follow metadata standards, resulting in metadata records being too general and incomplete to enable effective assessment of dataset quality (Devillers *et al.*, 2007; Brown *et al.*, 2013).

There exist a number of organisations and initiatives that are working to create and support geospatial data standards: International Organisation for Standardization (ISO/TC 211); INSPIRE; Open Geospatial Consortium (OGC); Dublin Core Initiative; Federal Geographic Data Committee (FGDC); CEN/TC 287; and many more. These standardisation bodies and initiatives are described in more detail below.

International Organisation for Standardization ISO/TC 211

International Organization for Standardization (ISO) is "*the world's largest developer of voluntary International Standards*" (ISO, 2014a). ISO carries out the work of establishing International Standards through its technical committees – a panel of experts from relevant industry, consumer associations, academia, non-governmental organisations (NGOs), and government. ISO does not make decisions on establishing new standards, it instead responds to the requests from industry or other stakeholders. In 1994, ISO established Technical Committee 211 – Geographic Information/Geomatics – with the aim to "*establish a structured set of standards for information concerning objects or phenomena that are directly or indirectly associated with a location relative to the Earth*" (ISO/TC211, 2014b). The establishment of geographic information and geomatics standards has been required to: support understanding and usage of geospatial data; to increase the availability, access and sharing of geospatial data; and to enable geospatial data interoperability. ISO 19115-1:2014 Geographic information – Metadata – Part 1: Fundamentals (ISO, 2014b) (a revised version of the ISO 19115:2003 Geographic Information – Metadata) is a geospatial metadata standard which defines the schema required for describing geographic information and services. The ISO 19115-1:2014 standard can be used to fully describe geospatial datasets and services, catalogue geospatial data, perform clearinghouse activities (i.e., provide access to digital spatial data and related services), and is applicable to geographic services, geographic datasets, dataset series, and individual geographic features and feature properties.

The ISO 19115-1:2014 standard defines:

- mandatory and conditional metadata sections, metadata entities, and metadata elements;
- the minimum set of metadata required to serve most metadata applications (data discovery, determining data fitness for use, data access, data transfer, and use of digital data and services);
- optional metadata elements to allow for a more extensive standard description of resources, if required;
- a method for extending metadata to fit specialized needs (ISO, 2014b).

ISO 19115 standard is presented in UML packages where each package contains one or more entities (UML classes); entities contain elements (UML class attributes) which identify the discrete units of metadata. Figures 2.4 – 2.8 provide some examples of ISO 19115:2003 metadata UML schema for describing geospatial dataset information, such as: identification information; dataset's responsible party (i.e., information about data producer, data distributer, data owner, etc.); supplemental information about a dataset (i.e., semi-formal producer comments); compliance with standards; dataset quality information; and lineage information. As will become apparent from the following chapters, these examples directly relate to the work presented in this thesis.

`MD_Metadata` (Figure 2.4) is a mandatory ISO 19115:2003 root entity which defines metadata about a resource or resources. This entity consists of elements such as `fileIdentifier`, `contact`, `parentIdentifier`, `metadataStandardName`, `metadataStandardVersion`, etc.

Aston University

Illustration removed for copyright restrictions

Figure 2.4: ISO 19115:2003 schema definition of the MD_Metadata entity (ISO/TC211, 2003, p. 19).

Responsible party information package (Figure 2.5) provides a standardised method for describing authoritative reference information, including responsible party and contact information for a resource. The `CI_ResponsibleParty` datatype[1] contains information about person(s), and/or position, and/or organisation(s) associated with the resource.

Aston University

Illustration removed for copyright restrictions

Figure 2.5: ISO 19115:2003 schema definition for describing responsible party (ISO/TC211, 2003, p. 34).

`MD_DataIdentification` (Figure 2.6) is a metadata entity that provides a standardised method for describing information required to uniquely identify a dataset. This entity includes `supplementalInformation` element (highlighted in Figure 2.6) which allows recording of any additional descriptive information about the dataset.

[1] A data type is a descriptor of a set of values that lack identity.

Aston University

Illustration removed for copyright restrictions

Figure 2.6: ISO 19115:2003 schema definition for describing supplemental information (ISO/TC211, 2003, p. 20).

`DQ_DataQuality` (Figure 2.7) is a package that contains a general assessment of the quality of the dataset. The `DQ_DataQuality` entity is optional and is an aggregate of `LI_Lineage` (discussed below) and `DQ_Element` entities. `DQ_Element` can be specified as five entities that represent elements of data quality, namely: `DQ_Completeness`, `DQ_LogicalConsistency`, `DQ_PositionalAccuracy`, `DQ_ThematicAccuracy` and `DQ_TemporalAccuracy`. As can be noted, these five metadata entities represent the geospatial data quality indicators discussed in Section 2.4.2.1.

Aston University

Illustration removed for copyright restrictions

Figure 2.7: ISO 19115:2003 schema definition for describing dataset quality information (ISO/TC211, 2003, p. 22).

The `LI_Lineage` entity (Figure 2.8) is optional and defines metadata required to describe the sources and production processes used in producing a dataset. `LI_Lineage` is an aggregate of `LI_Source` and `LI_ProcessStep`. `LI_Source` represents information about the source data used in creating the dataset and `LI_ProcessStep` represents information about an event or transformation in the life of the dataset.

Aston University

Illustration removed for copyright restrictions

Figure 2.8: ISO 19115:2003 schema definition for describing lineage information (ISO/TC211, 2003, p. 23).

Figure 2.9 provides an example of an ISO-complaint metadata record showing a practical implementation of the ISO 19115:2003 schema definition.

Figure 2.9: An example of an ISO 19115:2003 compliant metadata record (FAO, 2015).

INSPIRE

The INSPIRE (INSPIRE, 2014d) initiative was launched by the European Commission in September 2001 with the aim to establish a European spatial data infrastructure that would provide users with integrated spatial information linked by common standards and protocols.

The initiative is working towards facilitating public access to and sharing of relevant, harmonised and quality geographic information across Europe and is based on a number of common principles:

- "*data should be collected only once and kept where it can be maintained most effectively;*
- *it should be possible to combine seamless spatial information from different sources across Europe and share it with many users and applications;*
- *it should be possible for information collected at one level/scale to be shared with all levels/scales; detailed for thorough investigations, general for strategic purposes;*
- *geographic information needed for good governance at all levels should be readily and transparently available; and*
- [it should be] *easy to find what geographic information is available, how it can be used to meet a particular need, and under which conditions it can be acquired and used"* (INSPIRE, 2014a).

To ensure interoperability and harmonisation, the INSPIRE Directive requires common implementation rules in the following areas: metadata; data specifications; network services; data and service sharing; and monitoring and reporting. INSPIRE Metadata Implementing Rules (INSPIRE, 2008) are technical guidelines based on ISO 19115:2003 and ISO 19119:2005 (ISO, 2005). The aim of the guidelines document is to define how the INSPIRE Regulation can be implemented using ISO 19115:2003 and ISO 19119:2005 metadata standards. The INSPIRE Technical Guidelines Annex I, II and III (INSPIRE, 2014c) are a set of documents that describe the INSPIRE data specifications for 34 spatial data themes, including: coordinate reference systems; geographical grid systems; geographical names; elevation; land cover; geology; statistical units; buildings; soil use; etc. These guidelines highlight mandatory and the recommended elements related to the implementation of INSPIRE and often provide examples for the technical provisions and the underlying concepts. The INSPIRE network services area (INSPIRE, 2014f) includes technical reports and technical guidelines regarding the INSPIRE services such as: Download Services; View Services; Discovery Services; Schema Transformation Network Service; and Coordinate Transformation Services. The INSPIRE data and services sharing area (INSPIRE, 2014b) specifies implementing rules to regulate the provision of access to spatial datasets and services. The main points of the INSPIRE data and services sharing Regulation are as follows:

- "*Metadata must include the conditions applying to access and use for Community institutions and bodies; this will facilitate their evaluation of the available specific conditions already at the discovery stage.*

- *Member States are requested to provide access to spatial data sets and services without delay and at the latest within 20 days after receipt of a written request; mutual agreements may allow an extension of this standard deadline.*
- *If data or services can be accessed under payment, Community institutions and bodies have the possibility to request Member States to provide information on how charges have been calculated.*
- *While fully safe-guarding the right of Member States to limit sharing when this would compromise the course of justice, public security, national defence or international relations Member States are encouraged to find the means to still give access to sensitive data under restricted conditions, (e.g. providing generalized datasets) Upon request, Member States should give reasons for these limitations to sharing.*" (INSPIRE, 2014b).

To ensure progress and future evolution of INSPIRE, continuous monitoring of the implementation of the INSPIRE Directive and regular reporting are necessary. Consequently, the INSPIRE monitoring and reporting area (INSPIRE, 2014e) provides documents and guidelines on monitoring and reporting procedures. This area covers the 4 main fields of the INSPIRE Directive: metadata; spatial datasets and services; network services; and data sharing. Monitoring takes place annually and follows a quantitative approach, while reporting takes place every 3 years and involves more qualitative aspects.

The INSPIRE team consists of staff of the European Commission from the Directorate-General (DG) Environment, Eurostat and Joint Research Centre (JRC), with the European Environmental Agency (EEA) becoming increasingly involved in the INSPIRE activities. DG Environment carries the role of a legislative and policy co-ordinator. The JRC acts as a technical co-ordinator ensuring the evolution of the technical infrastructure and also coordinating INSPIRE work with other relevant international initiatives. Between 2007 and 2013 Eurostat acted as an implementation co-ordinator. Jointly with JRC, Eurostat prepared the programme of work for the INSPIRE initiative.

Open Geospatial Consortium (OGC)

The Open Geospatial Consortium (OGC) is "*an international industry consortium of 477 companies, government agencies and universities participating in a consensus process to develop publicly available interface standards*" (OGC, 2014a). The Open GIS Consortium Inc. (now the Open Geospatial Consortium) was established in 1994 with a vision of developing diverse geoprocessing systems. Since 1994, the membership has grown from 20 to 477 government, academic, and private sector organizations. OGC is actively working on establishing and implementing geospatial standards – the technical documents that detail interfaces or encodings designed to address specific interoperability challenges.

The OGC currently provides over 70 standards, specification documents and standards implementations, these include: OGC netCDF encodings; OGC Web Services Common Standard; ISO 19115:2003 extensions; OGC KML (Keyhole Markup Language); OpenGIS Geography Markup Language (GML) Encoding Standard; and many more. These OGC standards and supporting documents are available to the public at no cost. The OGC also coordinates its activities with other international standards organisations, such as ISO/TC 211.

Dublin Core Metadata Initiative

The Dublin Core Metadata Initiative (DCMI) (DCMI, 2014c) is an open organisation that began as an informal group of volunteers in 1995. Dublin Core takes its name from the invitational OCLC/NCSA Metadata Workshop that originated the initiative and took place in Dublin, Ohio. The DCMI is engaged in the development of simple interoperable metadata standards to facilitate description, search, sharing, and management of information and resources. The Dublin Core metadata standard (Dublin Core Metadata Element Set, Version 1.1) is a small and simple vocabulary of fifteen elements (see Table 2.1) that can be used to effectively describe and document a wide range of information and resources (DCMI, 2014b). It is designed to provide generic descriptions that are application-independent and are usable in combination with terms from other compatible vocabularies.

Table 2.1: Dublin Core Metadata Element Set, Version 1.1 (DCMI, 2014b).

Aston University

Illustration removed for copyright restrictions

The Dublin Core Element Set version 1.0 was defined in 1998 and standardized in ISO 15836:2009 in February 2009. The Dublin Core standard can be represented in many formats, but is typically implemented using XML.

Federal Geographic Data Committee (FGDC)

In the 1980s, the US Office of Management and Budget (OMB) initiated two studies to assess the scale and growth of digital cartographic activities in the Federal government. The first study, conducted in 1980, identified the scope of Federal digital cartographic activities and assessed the next course of action in this evolving field. The second study was conducted in 1982 and its results revealed the following: "*1) there was a substantial duplication of effort in the Federal community, which was expected to increase; 2) there was* *a lack of prescribed standards; and, 3) there was inadequate interagency coordination*" (FGDC, 2006). In 1983, following completion of the studies, OMB established the Federal Interagency Coordinating Committee on Digital Cartography (FICCDC) to coordinate digital cartographic activities among Federal agencies. On 19th October 1990, following a two-day "Forum on Spatial Data Coordination" carried out in December 1989 to discuss and formulate the future of coordination, OMB formally established the Federal Geographic Data Committee (FGDC).

The Federal Geographic Data Committee (FGDC) is "*an interagency committee that promotes the coordinated development, use, sharing, and dissemination of geospatial data on a national basis*" (FGDC, 2014a). It consists of representatives from the Executive Office of the President, and Cabinet level and independent Federal agencies. One of the tasks of the FGDC is to coordinate the development of the National Spatial Data Infrastructure (NSDI) – the nationwide data publishing effort designed to enable the development and sharing of digital geographic information resources (FGDC, 2014b). As part of this activity, the FGDC is working on development of new or adoption of the existing geospatial data standards for implementing the NSDI. The standards are only developed if no equivalent voluntary consensus standards exist. The FGDC-endorsed standards include: Content Standard for Digital Geospatial Metadata; Spatial Data Transfer Standard (SDTS); National Vegetation Classification Standard; Geospatial Positioning Accuracy Standards; Coastal and Marine Ecological Classification Standard (CMECS); and many more. In September 2010, the FGDC endorsed 64 non-Federally authored standards, including ISO metadata standards, giving them equal status as FGDC developed standards. With the adoption of the ISO metadata standards, the FGDC is now encouraging the federal agencies to transit to ISO metadata when possible. It is expected that the transition to ISO metadata will be occurring over the next few years.

CEN/TC 287: Geographic Information

The European Committee for Standardization (CEN) (CEN, 2014) is a non-profit organisation which was officially founded in 1974 under Belgian law. CEN develops the majority of the European Standards (ENs) and technical specifications and is officially recognised by the European Union and by the European Free Trade Association (EFTA). It is formed of 30 members from the national standards bodies of the European Union countries and EFTA countries.

In 1991, CEN formed the Technical Committee 287 (CEN/TC 287) to define a structured set of standards for geospatial data (Litwin and Rossa, 2011). CEN/TC 287 standards are aimed at defining, describing, structuring, interrogating, updating, codifying, transforming and transferring geospatial data and metadata. The main objective of CEN/TC 287 is to facilitate the development and usage of geographical information in Europe which will be achieved by:

- "*adopting the ISO/TC 211 standards series as CEN standards;*
- *developing and maintaining standards, specifications and profiles of standards;*
- *developing technical guidance and best practice documentation;*
- *collaborating with other standards related initiatives; and*
- *educating the user community and promoting the use of standards for geographic information*" (CEN/TC287, 2010).

In June 1991, CEN and ISO signed the Vienna Agreement (ISO and SEN, 2001) primarily to avoid potentially conflicting duplication of geospatial data standards. At present, most of the CEN/TC 287 published standards are directly mapped to official ISO standards for geospatial information.

International standards are essential in enabling geospatial data accessibility and interoperability, yet standards do not always ensure data quality as they cannot consider all the possible data uses and applications (Caprioli *et al.*, 2003). For this reason, the knowledge gathered from international standards needs to be combined with users' perspectives on geospatial data quality to define essential geospatial data quality indicators.

2.4.3 Geospatial Data Quality as Fitness for Use (External Quality)

As described in the previous section, international organisations, initiatives, and working groups such as ISO, OGC, INSPIRE, and many more, are actively working to establish, improve and extend geospatial data and metadata standards. These efforts have significantly improved metadata interoperability, but the increasing choice of metadata standards poses a number of unresolved questions: *Which standards are best to follow? How much metadata should be provided? How do providers make metadata 'useful' and not just 'usable'?* (Gahegan, 2005; Longhorn, 2005; Comber *et al.*, 2007b; Goodchild, 2009; Brown *et al.*, 2013). By documenting dataset characteristics such as ownership, legal constraints, format and some quality metrics, geospatial metadata can go some way towards helping users assess whether data is actually suitable for their intended use. Due to the fact that metadata standards are mostly focused on data *production* rather than potential data *use* and *application*, a typical metadata document is not however sufficient to effectively communicate this fitness for purpose to users from a variety of domains and expertise levels (Boin and Hunter, 2007; Comber *et al.*, 2007b; Goodchild, 2009; Brown *et al.*, 2013).

Over the past decade, many researchers and scholars have therefore actively attempted to address the challenge of communicating geospatial data 'fitness-for-purpose' information. Devillers *et al.* (2002; 2005; 2007) and Bédard *et al.* (2004) expended considerable effort towards developing the Multidimensional User Manual (MUM) (see Figure 2.10) – a working prototype system to "*compare data specifications as provided by data producers, with the* *needs, as expressed by users*" and to give users visual indicators of data suitability (Devillers *et al.*, 2002, p. 50). Their research was driven by the risks of misuse of geospatial data and potential data abuse by non-expert users. As stated by Goodchild (1995, p. 421), "*GIS is its own worst enemy: by inviting people to find new uses for data, it also invites them to be irresponsible in their use*". The MUM system organised metadata records in a multidimensional database for easy query and evaluation, and used the ISO 19113:2002 standard as a measure of geospatial data quality. The quality evaluation process in MUM involved users selecting the required area of the dataset and then setting the metadata parameters they are interested in. When parameters were selected, the system compared the selected metadata elements with the ISO standards and displayed a subsequent quality evaluation using quality indicators. The system was designed to provide a variety of visual indicators to report data quality to users which included traffic lights, colours, numbers, scale bars, speed meters, and smiley faces. Unfortunately, the MUM prototype developers do not provide details on how the quality parameters are set by users, how the evaluation process is implemented or how the quality values are measured against the ISO standard. Furthermore, while the MUM system was developed to address 'fitness for use', the researchers did not carry out user evaluations of the developed prototype. As a result, the development of the MUM system did not provide any insight into which quality indicators are most suitable and useful, and what parameters are most appropriate for assessing 'fitness for use' of geospatial data from a user's perspective. The system can only be considered as proof of the feasibility of using a quality dashboard with a multidimensional database of metadata records.

Aston University

Illustration removed for copyright restrictions

Figure 2.10: The MUM prototype for communicating geospatial data quality (Devillers *et al.*, 2005p, 213).

Boin (2008) and Boin and Hunter (2006; 2007) conducted thorough studies to elicit user requirements in terms of spatial data fitness for use evaluation. The results indicated that geospatial data users are more interested in data content rather than the metadata records because metadata records are too complex and are hard to understand. Consequently, to ensure richer user-focused information, the authors suggested adding more quality information to the descriptions of the data content, putting emphasis on data suitability and reliability, and allowing for user opinions.

In their papers, Comber *et al.* (2007a; 2007b) summarised the results of a workshop on "*Activating Metadata: The role of metadata in effective spatial data exploitation*" held by the National Institute for Environmental eScience (NIEeS) in Cambridge, UK, 6th – 7th of July 2005. The workshop gathered together different communities, which included data users, data producers, experts involved in creating metadata standards, and data mediators, and was focused on exploring how metadata can be made more '*useful'* rather than just '*usable'* for datasets' fitness for use evaluation. As argued by the workshop participants, data producers cannot predict all possible data applications, therefore there is a need for more user-focused metadata – "*information that* [actually] *helps the user assess the usefulness of a dataset relative to their problem*" (Comber *et al.*, 2007a, p. 2). Subsequently, it was recommended that metadata records be enriched with: references to relevant literature (citation information); less formal opinions from the data producers; expert opinions of data quality; and user feedback regarding previous data use. Goodchild (2009) also proposed a more "*user-centric approach*" to geospatial metadata, suggesting the introduction of informal voluntary user feedback and commentaries. Recent reviews, however, suggest that these recommendations have not yet been put into practice, with no practical means for collating and searching user-focused metadata, and many of the metadata records that are available being incomplete (Goodchild, 2012; Ellul *et al.*, 2013).

2.5 Decision Support Systems

The large part of this research focus on the design and implementation of a decision support system (DSS) for geospatial dataset discovery, intercomparison and selection, this section provides a review of the literature on decision making, DSS and visual information seeking using starfield displays. Section 2.5.1 provides an introduction to decision making processes and describes the concepts involved in solving selection decision problems. Section 2.5.2 offers an overview of the DSS research field; it discusses DSS development and application areas and highlights current challenges faced by DSS research. Finally, Section 2.5.3 describes use of starfield displays for exploration of large information databases and presents the seminal FilmFinder prototype visualisation system which supports interactive information retrieval and dynamic queries.

2.5.1 Introduction to Decision Making

Decision making is an inherent part of human life. On a daily basis humans face situations in which they have to make selection decisions, often in an uncertain environment based on imprecise and incomplete information (Martinez *et al.*, 2010). As argued by Van Schaik (1988, p. 4), "*decision-making is solving a problem*" and the problem is said to exist if "*someone is in doubt as to which choice is best to remove his dissatisfaction with his present state, where he can identify: one or more outcomes that he desires, two or more unequally efficient or effective courses of action, and an environment containing factors that affect the outcomes*"; that "*someone*" is a *decision-maker*.

Decision making implies that there are a number of alternatives to be considered, with the objective being to select the best option that fits a decision-maker's goals, preferences, desires, and values (Jankowski, 1995; Sharma, 2009; Martinez *et al.*, 2010). Decision making involves several considerations, such as the benefits gained from making the right choice, the costs and risks involved, and losses resulting from making the wrong decision (Alexander, 2012). Decision making is a complex process and is composed of different phases, such as information gathering, analysis and selection which, when followed, lead the decision-maker to a suitable choice from a set of possible options, optimising the decision-maker's utility (Evangelos, 2000; Teisman, 2000). Jankowski (1995) describes four stages of a structured approach to the decision making process: *problem definition*; *search for alternatives and selection criteria*; *evaluation of alternatives*; and *selection of alternatives*. *Problem definition* refers to a difference between the present state and a desired state that is formulated as a problem that requires a decision. *Search for alternatives and selection criteria* denotes identification of the feasible alternatives (potential problem solutions) and criteria for their evaluation. *Evaluation of alternatives* refers to the impact assessment of each alternative based on the identified evaluation criteria. Finally, the *selection of alternatives* stage refers to ordering of alternatives from the most preferable to the least preferable and consequent selection of the best option or selection of a group of options for further evaluation.

Information processing leading to a suitable choice varies depending on the task complexity and number of available alternatives (Payne, 1976). When faced with two alternatives, decision-makers typically employ a search strategy and search for an equal amount of information on each alternative in order to support selection of the best option. In contrast, when faced with multiple, complex alternatives, decision-makers typically employ a strategy of eliminating some of the alternatives as quickly as possible based on available partial information and then engaging in a search for additional information for the remaining alternatives (Payne, 1976). To make an appropriate selection decision, decision makers often seek more information than is required which leads to: (a) a delay in the decision due to additional time required to collect and process information; (b) information overload which affects ability to make a decision and decreases the quality of the decision made; (c) selective use of information to support a predetermined solution; (d) mental fatigue which affects quality and speed of work; and (e) decision fatigue which may lead to careless decisions or even decision paralysis (Malhotra, 1984; Lumsden, 2008). While an increased amount of available information on alternatives increases the confidence of a decision-maker, excessive information will accordingly increase variability and uncertainty of decisions and decrease the quality of the choices made (Slovic and Lichtenstein, 1971; Payne, 1976). The amount of information required to make a decision depends on the attractiveness of the alternatives to the decision-maker and the diversity of information gathered about the alternatives (Zeleny, 1982). There is less need for additional information if the alternatives are sufficiently divergent and if information about the alternatives is consistent.

Solving a decision problem can lead a decision-maker to a good or a bad decision. The traditional view has been that decision quality is reflected in the decision outcomes – that is, "*a good decision is one that produces positive outcomes*" (Higgins, 2000, 1218). When making a selection decision, a good choice is the alternative whose value or utility of outcomes is judged to be more beneficial than the available alternatives. The costs of obtaining the desired outcomes can also influence whether a decision is perceived as good (Higgins, 2000). The outcome benefits have to be weighed against the costs of obtaining the desired outcomes – if the costs are too high, the optimal alternative might not be selected. Both outcome benefits and outcome costs, therefore, contribute to a decision being considered good. Nevertheless, outcomes do not necessarily reflect the quality of a decision made – a good decision can bring about a bad outcome and a bad decision can result in a positive outcome (Higgins, 2000; Lumsden, 2008). Decision quality is reflected in the following conditions: (a) the selected alternative meets the objectives and values identified during the problem definition stage; (b) the final decision is "worthwhile" in that it meets the objectives, minimising costs, energy, and side effects; and (c) indirect advantages and disadvantages are considered when making a selection decision (Higgins, 2000; Lumsden, 2008).

2.5.2 Overview of Research on Decision Support Systems

Decision making is considered one of the most critical activities for human efficacy (Shirgaonkar *et al.*, 2010). Unsurprisingly, decision making is central to many scientific disciplines such as engineering, psychology, operations research, artificial intelligence and many more (Vroom and Yetton, 1973; Martinez *et al.*, 2010) and has led to the development of systems to support the process. The concepts involved in decision support systems (DSS) were first discussed in the early 70s by Scott Morton under the term 'management decision systems' (Sprague, 1980). The term 'decision support system' itself first appeared in a paper by Scott Morton in 1971, although some scholars argue that the DSS field began in 1965 with the acceptance of Scott Morton's PhD topic "*Using a Computer to Support the Decision-Making of a Manager*" by the Harvard Business School (Arnott and Pervan, 2005). In the 1980s, user organisations, information systems vendors, and researchers engaged in discussions of a new "era" in information described by Sprague (1980, p. 23) as a "*DSS Movement*". The DSS Movement was characterised by events and mechanisms such as systems development in organisations, hardware and software developments, publishing activities to report DSS experiences and research, and conferences to provide a forum for the exchange of DSS ideas among interested parties (Sprague, 1980). Since then, over the past 30 years, the DSS field has explored the use of every kind of technology to support DSS development and progression, including: spreadsheets; databases; networks; hypermedia; expert systems; visual programming; intelligent agents; neural networks; and many more (Beynon *et al.*, 2002). Computer-based DSS have been increasingly developed to support decision-makers in application areas such as production and operations, marketing and logistics, and management information systems (Eom and Kim, 2006).

There are numerous definitions of DSS. Arnott and Pervan (2008, p. 657), for instance, define DSS as "*the area of the information systems (IS) discipline that is focused on supporting and improving managerial decision-making*". (Uran and Janssen, 2003, p. 512) argue that "*a DSS implies a computer program that:*

- *assists individuals or groups of individuals in their decision process;*
- *supports rather than replaces judgements of individuals; and*
- *improves the effectiveness rather than the efficiency of a decision process*".

A widely quoted definition of DSS is one proposed by Sprague and Carlson in 1982 which articulates that DSS are "*computer-based systems that help decision makers confront ill-structured problems through direct interaction with data and analysis models*" (Lyons and Stuth, 1992, p. 124).

Initial DSS were primarily designed to support individual decision-makers but, with development of new technologies and the advent of the Web, DSS applications expanded to supporting teams, workgroups and groups of organisations (Shim *et al.*, 2002; Bharati and Chaudhury, 2004). Modern DSS provide their users or groups of users with a broad range of capabilities and facilitate a wide variety of decision tasks including information gathering and analysis, model building, sensitivity analysis, collaboration, alternative evaluation, and decision implementation (Bhargava *et al.*, 2007). While DSS are often developed and used to support ad hoc analyses, increasingly DSS technologies are being integrated into business processes and information systems (Bhargava *et al.*, 2007).

DSS is not a homogenous field. Throughout the history of the field, a number of fundamentally different approaches to DSS have had a period of popularity in both research and practice (Arnott and Pervan, 2005). Various types of DSS have been established, including, but not limited to: Personal Decision Support Systems; Group Support Systems; Negotiation Support Systems; Intelligent Decision Support Systems; Data Warehouses; Knowledge Management-Based DSS; Enterprise Reporting and Analysis Systems; and Spatial Decision Support Systems. Each of these DSS types utilise a variety of technologies, support diverse types of users, and represent different methods of support, system scales, levels of investment, and potential organisational impacts. Personal Decision Support Systems (PDSS) are usually small-scale systems that are developed to support decision tasks of one manager, or a small number of independent managers (Arnott, 2008). Group Support Systems (GSS) facilitate effective work of groups of users; this type of DSS utilises a combination of communication and DSS technologies (Fan and Shen, 2011). Negotiation Support Systems (NSS) primarily focus on supporting negotiation between opposing parties (Arnott and Pervan, 2012). Intelligent Decision Support Systems (IDSS) apply artificial intelligence techniques to decision support (Guerlain *et al.*, 2000). Knowledge Management-Based DSS (KMDSS) are systems that support decision making by aiding knowledge storage, retrieval, transfer and application (Arnott and Pervan, 2005). Data Warehousing (DW) are systems that provide large-scale data infrastructures to empower decision-makers with information that allows them to make decisions based on solid facts (Nemati *et al.*, 2002). Enterprise Reporting and Analysis Systems (ERASs) are enterprise-scale systems that include executive information systems (EISs), online analytical processing systems (OLAP), corporate performance management systems (CPM), business intelligence (BI), and, more recently, business analytics (BA) (Arnott and Pervan, 2012). Finally, Spatial Decision Support Systems (SDSS) are systems designed to support decision-makers in solving complex, semi-structured decision problems that have a spatial reference (Rinner, 2003).

With technological developments and the increasing availability and use of spatial data, SDSS are becoming increasingly popular in decision making processes (Uran and Janssen, 2003). Despite ongoing discussions, to date there is no general agreement on the definition of SDSS, with scholars usually listing general characteristics that apply to SDSS. This thesis will adopt the statement by Densham (1991, p. 405) that SDSS are "*explicitly designed to provide the user with a decision-making environment that enables the analysis of geographical information to be carried out in a flexible manner*". SDSS include wide areas of applications such as water management, crop management, urban planning, environmental planning, recycling, and many more. Due to the nature of complex spatial problems, SDSS often provide capabilities and functions that:

- support input of spatial data;
- allow representation of the complex spatial data relations and structures;
- provide analytical techniques specific to spatial and geospatial analysis; and
- output the results in a variety of spatial forms such as maps (Densham, 1991).

Additionally, SDSS typically include databases that integrate a variety of spatial data which needs to be selected prior to data analysis and decision making. SDSS do not, however, offer functionality to support users in searching and selecting spatial or geospatial data. Unfortunately, common geospatial data portals and clearninghouses do not offer decision support functions.

Over two decades ago, Angehrn and Lüthi (1990, p. 27) articulated that "*human-computer interaction remains a central issue in the DSS domain, and further research is needed to realise a high level of human-machine cooperation in problem solving and decision making*". To date, unfortunately, lack of practical relevance of DSS to their end-users still remains a major problem. Uran and Janssen (2003), for example, conducted a study in a search for explanations or reasons for success or failure of SDSS by systematically comparing five representative SDSS examples. The results of their study revealed that in all of the evaluated systems, contact with the decision process was lost during the SDSS development stage. Researchers found strong indications that users are not always able to adopt systems into use as intended or expected by developers. Consequently, the produced systems lack practical relevance and usefulness to their intended users. The authors argue that, to provide users with relevant and useful tools, there is a need for a closer link between developers and users during the SDSS development stage. In their critical analyses of the nature and state of DSS research, Arnott and Pervan (2005) demonstrated that almost 90% of DSS research has failed to identify the principal clients and approximately 60% failed to identify the DSS users. Furthermore, only 10% of reviewed DSS research demonstrated high or very high practical relevance, with approximately half of DSS research being regarded as having low or no practical impact. In their follow-up evaluation study, Arnott and Pervan (2008) again demonstrated extremely low relevance of DSS research to end-users. Their survey revealed that, overall, only 10.1% of DSS research is regarded as having high or very high practical relevance, while 49.2% of DSS research is regarded as either having low practical relevance or none at all. Consequently, the authors conclude that "*the relative lack of exposure of academics to contemporary professional practice is a particular problem for DSS*" (Arnott and Pervan, 2008, p. 661).

2.5.3 Visual Information Seeking using Starfield Displays

In 1992, Ahlberg *et al.* (1992) opened a discussion about *dynamic queries* to support direct manipulation of large databases of information. The authors argued that exploration of large information spaces is a challenging task, particularly for naïve users, and there is a need for easy to use, quick and powerful query methods for database information retrieval (Ahlberg *et al.*, 1992). To address the challenge, the authors proposed to employ direct manipulation interfaces which support:

- "*Continuous visual representation of objects and actions of interest;*
- *Physical actions or labelled button presses instead of complex syntax;*
- *Rapid, incremental, reversible operations whose impact on the object of interest is immediately visible; and*
- *Layered or spiral approach to learning that permits usage with minimal knowledge*" (Ahlberg *et al.*, 1992, p. 619).

To test the proposed solutions and to explore new approaches to visual information seeking (VIS), Ahlberg and Shneiderrnan (1994b) developed FilmFinder (see Figure 2.11 and Figure 2.12) – a prototype DSS application which supports interactive information retrieval and dynamic queries and which is now considered seminal in the field. The key concept behind the FilmFinder application is exploration of large information spaces using visual analytic techniques to overcome the issue of information "flood". The prototype application supports database browsing via rapid filtering to reduce result sets, progressive refinement of search parameters, continuous reformulation of goals, and visual scanning to identify relevant results (Ahlberg and Shneiderrnan, 1994a) . As articulated by the creators (1994a; 1994b), the FilmFinder environment is designed to encourage incremental and exploratory search.

The FilmFinder application, as its name implies, allows users to explore a large film database. The search results are represented as a starfield display – a two-dimensional scatterplot to structure result sets which supports zooming to reduce clutter – where each film is represented by a point of light of different colours (see Figure 2.11). The position of a film in the scatterplot is determined by when it was produced and a measure of popularity – x-axes coordinates represent the year of film production and y-axes coordinates represent user ratings between 0 and 9. This visualisation technique allows users to rapidly overview query results as queries are formulated to discard old or unpopular films (Ahlberg and Shneiderrnan, 1994b) . The search parameters can be progressively altered by manipulating a set of widgets (sliders, buttons and check boxes) to produce a complex Boolean query. Users can overview the whole results set, zoom in on desired items, and get details on demand (see Figure 2.12). A starfield display converts a large textual database into a single comprehensible and manipulable display.

Aston University

Illustration removed for copyright restrictions

Figure 2.11: The FilmFinder application (visualisation of results set in a starfield display) (Ahlberg and Shneiderrnan, 1994a, p. 479).

Aston University

Illustration removed for copyright restrictions

Figure 2.12: The FilmFinder application (obtaining detailed information about a particular film) (Waloszek, 2013).

Building on experiences from developing the FilmFinder prototype, Shneiderman (1996) proposes that, to facilitate effective processing of large databases of information, data visualisation systems should support seven tasks, namely:

- *“overview: gain an overview of the entire collection;*
- *zoom : zoom in on items of interest;*
- *filter: filter out uninteresting items;*
- *details-on-demand: select an item or group and get details when needed;*

- *relate: view relationships among items;*
- *history: keep a history of actions to support undo, replay, and progressive refinement; and*
- *extract: allow extraction of sub-collections and of the query parameters*" (Shneiderman, 1996, p. 337).

In recent years, the work of Ahlberg and Shneiderrnan (1994a; 1994b) has started to regain interest amongst HCI researchers and scholars with a number of prototype systems being developed and evaluated to explore new applications of starfield displays. Dunlop and Davidson (2000), for instance, explored the use of starfield displays for visualisation of large data sets on monochrome palmtops. The results of their evaluation study revealed that starfield displays work well and are simple to use even on small low-resolution black-and-white screens. Building on the data visualisation and filtering concepts of FilmFinder, Silva *et al.* (2003) developed EVA2D visual information seeking (VIS) environment for exploring a very large digital library of books. Various tests of the visualisation environment demonstrated acceptable performance for a collection of approximately 20,000 books. EVA2D was used for one semester by selected groups of students, faculty and library staff. Students and staff who were able to seek information and find relevant books by just clicking on neighbouring areas in the starfield display. Most recently, Sadana and Stasko (2014) explored use of dynamic scatterplot displays on tablet computers and successfully implemented an interactive scatterplot prototype application on an Apple iPad tablet. The authors have not yet evaluated the prototype system to validate its usability and effectiveness.

As will be demonstrated in Chapter 7, the research on decision making, DSS and scatterplot displays described in this section has been applied in the implementation of a DSS for geospatial dataset discovery, intercomparison and selection.

2.6 GEO Label

This section introduces the concept of a GEO label – a concept that was initially proposed by the Science and Technology Committee (STC) of the Group on Earth Observation (GEO). Section 2.6.1 describes the initial proposals that started the initiatives to conceptualise and define the notion of a GEO label. Section 2.6.2 reviews previous research attempts to formalise the concept of a GEO label and define its role within the Global Earth Observation System of Systems (GEOSS).

2.6.1 Why a GEO Label?

The Global Earth Observation System of Systems (GEOSS) is a distributed 'system of systems' which is being constructed by the Group on Earth Observation (GEO) to "*provide* *decision-support tools to a wide variety of users*" (GEO, 2014c). GEOSS allows users world-wide to discover and access geospatial data and models which can be used to support decision-making across a range of scientific domains. GEOSS supports policy-makers, resource managers, scientific researchers and many other experts and decision makers in their daily work with Earth Observation (EO) data, across a range of Societal Benefit Areas (SBA), namely: Health; Disasters; Weather; Energy; Water; Climate; Agriculture; Ecology; and Biodiversity (Béquignon *et al.*, 2010). Given that the GEOSS is estimated to contain more than 28 million dataset records, and is constantly growing, choices faced when selecting a dataset can (depending on usage domain) be quite daunting. With such a great choice of datasets comes the problem of data quality assessment and dataset selection decision-making. To tackle this challenge, in 2009 the GEO Science and Technology Committee (STC) proposed to establish a GEO label – a label "*related to the scientific relevance, quality, acceptance and societal needs for activities in support of GEOSS as an attractive incentive for involvement of the S&T communities*" (GEO, 2011b). The STC suggested that the development of such a label could significantly improve user recognition of the quality of geospatial datasets and that its use could help promote trust in datasets that carry the established GEO label (ST-09-02, 2010). In 2009, the activity of establishing a GEO label (activity 2b of the STC Road Map) was assigned to the GEO Task ST-09-02: Promoting Awareness and Benefits of GEO (GEO, 2011a). Several ST-09-02 meetings held in 2009 and 2010 revealed that there was no general agreement on the role that the GEO label should fulfil. Consequently, in March 2010, ST-09-02 asked the STC for further guidance on defining the GEO label and its role in GEOSS. In response to this call, the STC produced a short document (EGIDA, 2011, Appendix I) outlining more detailed GEO label requirements. As specified by the STC, the GEO label should comprise two functions: objective labelling (quality, reliability) and subjective labelling (relevance, usability). The criteria for the objective labelling could include: registration in the GCI; metadata completeness; indication of the time frame of commitment; availability of contact information; and compliance to quality standards. The subjective labelling could include a user rating system provided as part of the GCI. Both GEO label functions should be voluntary and self-assessed and should either be provided as two separate labels or, if combined in one label, should possibly be categorised into different ranks.

In 2010 – 2011, the European Commission, as part of its Seventh Framework Programme for Research and Technological Development (FP7), financed two projects (GeoViQua and EGIDA) to support the STC activities, with both projects adopting the responsibilities for defining the GEO label concept. The EGIDA project started on 1st of September 2010 and the GeoViQua project kicked off on 1st of February 2011. Both projects were expected to collaborate with the ST-09-02 Task team, contributing their research and findings towards development of the GEO label concept. In 2012, the ID-03 GEO Task (GEO, 2014b) took over the ST-09-02 officially adopting the responsibility for the development of the GEO label.

2.6.2 Previous Attempts to Define a 'Voluntary GEO Label'

In September 2010, the ST-09-02 Task team drafted a short initial report (ST-09-02, 2010) to the STC outlining their views on the concept of the GEO label and its role in GEOSS. As described in the report, the STC envisioned the GEO label which would:

- "*encourage scientists, researchers, and others to contribute their data and systems to GEOSS by offering an accepted voluntary label that provides recognition that their contribution is valued by the GEO community;*
- *differentiate components, data and products delivered through GEOSS and provide a "trusted brand" to GEOSS users; member governments may base their decisions on data/products of such contributions; and*
- *highlight the importance of GEOSS to those previously unaware they were reliant on this initiative for their data or product*" (ST-09-02, 2010, p. 4).

Based on these initial STC requirements and the clarifications provided in March 2010 (see previous section), the ST-09-02 Task team proposed that the GEO label should cover a broad range of criteria, combine objective and subjective components in a distinguishable way, and be based on:

1. "*an objective assessment measuring quality, reliability, accessibility, interoperability, etc.;*
2. *a subjective assessment scaling relevance, usability, etc.; and*
3. *a combined assessment objectively weighing the match between the entity and somewhat subjective user needs published in the GEOSS User Requirement Registry*" (ST-09-02, 2010, p. 4).

As a practical solution to the challenge, the ST-09-02 Task team proposed a voluntary tri-faceted label that would convey 1) a dataset's quality, 2) a dataset's relevance to user needs and 3) a dataset's scientific relevance. Due to the subjective nature of geospatial data quality, the ST-09-02 Task team proposed to use metadata completeness as a potential measure of dataset quality. The team suggested the use of a list of elements published by the US Global Climate Research Program (USGCRP) in a workshop report (USGCRP, 1999) as a criteria for metadata completeness evaluation. To evaluate datasets' relevance to user needs, the ST-09-02 Task team proposed to embed a 5-star rating system in the GEOSS Common Infrastructure (GCI) supporting subjective quality evaluation. Finally, to assess datasets' scientific relevance, they suggested using the GEO User Requirement Registry (URR) which collects community input on user needs, applications and requirements for observations and product. The report however did not provide any details on how the URR would be used in practice to measure scientific relevance for a given dataset.

As a potential graphical visualisation of the tri-faceted label, the ST-09-02 Task team proposed a star design with three star arms representing quality, user needs and scientific relevance (see Figure 2.13). These initial proposals, however, were not taken beyond the draft document, and neither the proposed GEO label function nor the tri-faceted star design materialised into physical GEO label systems.

Aston University

Illustration removed for copyright restrictions

Figure 2.13: Preliminary GEO label design proposed by the ST-09-02 Task team (ST-09-02, 2010, p. 8).

As already mentioned, the EGIDA project adopted the responsibility for developing and promoting the GEO label in GEOSS in September 2010. To support the ST-09-02 activities, the EGIDA project proposed to:

- "*further develop the draft concept for a GEO label and ensure that this concept is acceptable for major Earth Observation data providers;*
- *present the fully developed draft concept to the STC and, if necessary, adapt it according to guidance from the STC;*
- *support the STC in getting GEO-wide acceptance of the GEO label; and*
- *support the STC in implementation of the Label by applying the concept to a number of carefully selected services, observations, information, and data sets*" (EGIDA, 2011, p. 6).

The outcome of the project's work on the GEO label were published in their deliverable D.3.2 Proposal for a GEO Label. As described in the document, the EGIDA project members were highly sceptical about the voluntary and self-assessed nature of the GEO label as proposed by the STC. They argued that self-assessment would demote trust in the GEOSS datasets and voluntary application could lead to a large portion of the GEOSS datasets not being labelled.

A review on labelling programmes undertaken by EGIDA led to a proposal for a commercial approach to the GEO label. As a result, the Beyond Sustainability members of EGIDA submitted a Eurostars Proposal (EGIDA, 2011, Appendix IV) to establish a new project with a focus on developing an independent data certification body. It is, however, unknown whether this initiative was taken further and whether the proposed project received any funding form Eurostars or other sources.

The EGIDA project deliverable also discussed some potential benefits of introducing the GEO label in GEOSS, including:

- producer recognition for contributions to GEOSS;
- forward traceability (traceability of data usage) which would be very valuable for data producers;
- promotion of data sharing and signalling of data availability;
- transfer of information to users which would promote datasets' trustworthiness;
- help in communicating metadata information in an easy visual way; and
- the first step in the direction of a data certification.

The main outcomes of the EGIDA project's work on the GEO label were summarised as follows:

1. "*The goals of the GEO label have changed since inception and should be revisited.*
2. *A user survey regarding the GEO label is urgently required (this is being led by GeoViQua).*
3. *Voluntary labeling will likely not bring about desired results.*
4. *An external certification body is needed that can independently apply such valuation on data (internal or GEO valuation will not be accepted by data providers).*
5. *A user rating system should be implemented.*
6. *A document is required summarizing all GEO label related efforts to date (this document)*" (EGIDA, 2011, p. 16).

Although the EGIDA project provided some useful insights on possible GEO label functions and outlined the challenges that have to be considered when integrating the GEO label in GEOSS, no concrete definition of the GEO label function was formalised and no practical tools developed to support the GEO label.

This thesis describes the GEO label research conducted by the author as part of the GeoViQua project. The approach taken to define the notion of the GEO label and to provide practical tools to support the theory are described in the subsequent chapters of this thesis.

Chapter 3 Method

Since the importance of external data quality is widely recognised in the GIS community (see Section 2.4.3), it is surprising that there has been very little empirical research into user perspectives on geospatial data quality and fitness for use evaluation. User needs have somewhat been neglected when developing new solutions for improved quality evaluation and dataset selection. To address this issue, the main focus of this research was to understand how geospatial data users select the datasets to use, what the reasons for their decisions are, and what mechanisms could improve their experience. Consequently, this research adopted an iterative UCD approach in order to provide solutions that are tailored to geospatial data users' needs and that are likely to garner user acceptance once deployed.

UCD is "*a design philosophy and approach that places users at the centre of the design process from the stages of planning and designing the system requirements to implementing and testing the product*" (Baek *et al.*, 2008, p. 660). The main characteristic of UCD is that it attempts to optimise a product around user needs, abilities and desires, rather than forcing the users to change their behaviour to suit the product. UCD recognises that end user involvement in the design of tools is critical to the success of the developed solutions. In UCD, it is important to involve the potential users in the design process at an early stage and then continuously review and refine user requirements of the technology being developed (Baek *et al.*, 2008).

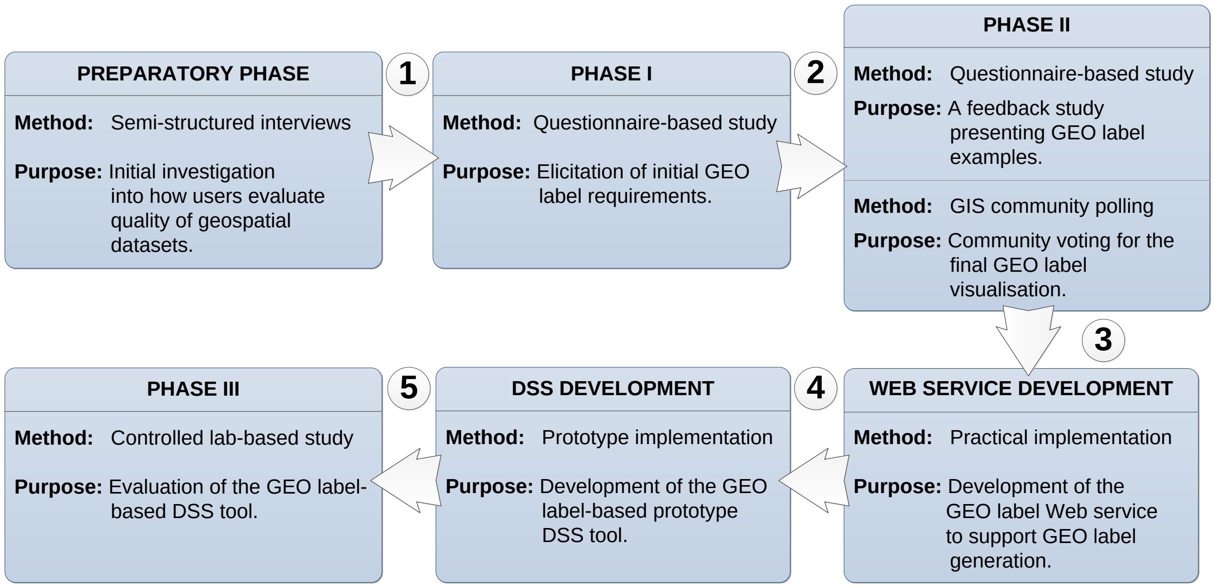


Figure 3.1: Phases of GEO label research.

Utilising various tried-and-tested UCD methods of collecting and analysing research data, this research comprised six main phases of exploration, development, evaluation and validation, with each phase building upon the knowledge gathered in the previous phases (see Figure 3.1). It should be noted that, while there are guidelines and techniques available for conducting UCD, each method will vary depending on the problem at hand.

The preparatory phase (see Chapter 4) was conducted using a series of semi-structured face-to-face and telephone interviews with geospatial data expert users and producers. The intention was to uncover initial information about dataset selection, including their use and production within representative application areas, in order to *inform* further research phases. The semi-structured interview technique is widely used in UCD at the beginning of the design process to gain an initial understanding of user needs and perspectives on the technology being developed (Abras *et al.*, 2004).

Phase I (see Chapter 5) was conducted via a comprehensive online questionnaire-based survey to solicit initial geospatial data producers' and users' views on the concept of a GEO label and the role it should serve. Questionnaire-based surveys can be successfully applied at an early stage of research to collect data related to the needs and expectations of stakeholders (Bevan, 2003; Abras *et al.*, 2004). This method was most appropriate to facilitate the collection of data from a large number of geospatial data users and producers from diverse GIS communities across the world.

Phase II (see Chapter 6) focused on the iterative design of the GEO label graphical representation. A comprehensive questionnaire-based study was conducted to solicit geospatial data producers' and users' views on the proposed GEO label visualisations. Questionnaire-based studies can be effectively applied in the evaluation of design alternatives in the early stages of the design cycle (Abras *et al.*, 2004). Similar to Phase I, this technique was used to collect a large sample of data. Following the questionnaire-based study, semi-formal feedback and recommendations from GIS peers and GeoViQua project partners were used to inform some of the GEO label design revisions (e.g., selections of colours, new icons design, etc.). This ensured geospatial professionals' involvement in all the stages of GEO label design. Due to the fact that two diverse designs were potential candidates for the GEO label visualisation, GIS community polling was conducted to identify the final GEO label representation. The intention was to obtain views from a large number of GIS professionals and identify community-dictated graphical representation.

The Web service development stage (see Chapter 7) focused on the implementation of a GEO label service to support the generation of dynamic GEO label representations. This stage was conducted via rapid prototyping and development, and resulted in a stable and fully-functional RESTful Web service. As part of the interoperability testing and validation, the service was integrated into a number of real-world GIS applications (see Section 10.2). The DSS development phase (see Chapter 8) focused on the design and implementation of a GEO label-based dataset discovery and intercomparison tool (GEO LINC). This development phase also followed a rapid prototyping approach. While the GEO LINC prototype represented a proof of concept, it was important to provide a functional system to allow for the evaluation of system interactivity and simulation of a real dataset selection experience.

Finally, Phase III (see Chapter 9) was conducted via a controlled lab-based human subject study to evaluate the usability and effectiveness of the prototype GEO label-based tool. A lab-based study approach was selected because it can be successfully applied to elicit rich qualitative data and offer a high level of experimental control and repeatability.

Chapter 4 Initial Investigation: Soliciting User and Expert Views on Geospatial Data Quality

This chapter presents the results of the preparatory phase – an initial investigation that was conducted to (a) elicit high level user requirements regarding geospatial data quality and trustworthiness assessment and evaluation, and subsequent dataset selection and (b) identify the informational aspects of geospatial datasets upon which users rely when assessing dataset quality and trustworthiness.

Section 4.1 describes the initial investigation process – essentially, a series of semi-structured interviews with geospatial data expert users and producers. Section 4.2 discusses the results of the initial study and describes 11 geospatial data informational aspects that were identified as important for evaluation of geospatial datasets' quality and trustworthiness. The limitations of the initial investigation are discussed in section 4.3. Finally, section 4.4 concludes with a discussion of the parallels between study findings and research on trust conducted in the field of B2C e-Commerce.

4.1 Study Process

The GEO label research comprising this thesis commenced with an initial investigation into how users and producers of geospatial data evaluate quality and trustworthiness of datasets. Using a series of face-to-face and telephone interviews the intention was not to elicit specific requirements for the GEO label, but rather to uncover initial information about dataset selection, use and production within representative application areas in order to *inform* further research into the design and development of the GEO label. The interviews were relatively informal, and discussion was guided by the following set of high-level questions or prompts:

1. *Please describe a current area of your work in which you use external data sources.*
2. *What data do you use in your work, and where does it come from?*
3. *How do you choose which datasets to use in your work? What are the reasons for your decisions?*
4. *Are you aware of any data certificates or seals in selecting your data? Do you look for specific certificates or meta-information in a data set you use? How do you know whether to trust the data?*
5. *Does the data you use come with sufficient supporting information to allow you to make an informed judgement about which one(s) to choose? How much information do you need?*

Based on interviewee responses to the above prompts/questions, follow-up and clarification discussion ensued as appropriate to generate a rich set of qualitative information regarding dataset production, selection and use. Where interviewees used community- or domain-specific jargon, they were asked for further explanation to eliminate any misunderstandings. The interviews were directed to capture sufficient contextually rich, qualitative information to allow distillation of specific information regarding the informational aspects that are perceived as significant in terms of determining dataset trustworthiness and dataset quality assessment for fitness for purpose-based dataset selection.

Representative interviewees – including geospatial data users, researchers, data archivists, academics and data producers – were identified and contacted to participate in the initial interviews. The diversity of interviewees supported the elicitation of a broad and inclusive picture of user needs as they relate to quality assessment of geospatial datasets. A total of 6 participants were recruited for telephone or face-to-face interviews with the author; each interview took between 30-60 minutes.

Table 4.1 profiles the six interviewees.

Table 4.1: Profiles of initial investigation interviewees.

Interviewee 1 is a **data archivist** who works as part of a science network of people, organisations, and, most importantly, observation platforms, that performs Long-Term Ecological Research (LTER).
Interviewee 2 is a **researcher data user** who is a part of a group that is working on projects to monitor forests and the tropics or, specifically, changes in the forest cover or tree cover in the tropics.
Interviewee 3 is a **land use researcher** who works for a government department that covers areas such as: the natural environment; biodiversity; plants and animals; sustainable development and the green economy; food, farming and fisheries; animal health and welfare; environmental protection and pollution control; and rural communities and issues.
Interviewee 4 is a **climate forecaster** who works on climate forecasting for protected areas using climate data which he pre-processes himself to get more descriptive data for his own needs. Interviewee 4 does not use a lot of external data.
Interviewee 5 is primarily a **data provider** who typically takes low level data (typically oceanography-related) and works it up into higher levels to arrive at some physical product.
Interviewee 6 is an **academic researcher** in earth and environmental sciences who uses external data sources "*across the board*".

The six interviews were audio-recorded to allow for detailed and accurate data capture as well as in-depth post-interview analysis; the interviewer also took written notes during the course of each interview. Data saturation refers to the point at which consulting additional participants would not have provided new information or identified new themes in the data (Guest *et al.*, 2006; Francis *et al.*, 2010); data saturation points are specific to each study but 73% of thematic discovery can occur from as few as six interviews (Guest *et al.*, 2006). In this study, despite the diversity of interviewees, data saturation occurred after completing the six interviews. The occurrence of data saturation was evident from the repetition of the themes as the interviews progressed and from the fact that the sixth interview did not reveal any new themes. Additionally, data from a further 12 interviews conducted by other GeoViQua partners was used to validate the information collected first hand. As such, after the data analysis was completed, the interview notes from other partners (no full transcripts were available) were carefully reviewed to identify any additional themes. The notes did not reveal new themes or requirements, which further confirmed the validity of the interview results and the occurrence of data saturation in this study.

Verbatim transcripts of the six interview recordings (see Appendix A.2) were generated to support detailed data analysis. A first pass of analysis was conducted in order to derive user stories – that is, very high-level informal statements of requirements that capture what users want to achieve. User stories typically follow the template: "*As a <role>, I want <goal/desire> so that <benefit>*". These user stories (see Appendix A.3) helped identify high level requirements for a GEO label. The transcripts were further analysed to identify, in greater detail, the informational facets of importance to users when assessing dataset fitness-for-purpose and to derive detailed user requirements that relate specifically to quality and trustworthiness assessment of datasets for the purpose of making dataset selection decisions (see Appendix A.4). The remainder of this chapter discusses the identified informational facets as they relate to the design of the GEO label.

4.2 Study Findings

The analysis of the interview transcripts identified that geospatial data users highly value good quality metadata records. The study participants stated that complete and well-documented metadata records are essential in the assessment of geospatial data quality and trustworthiness. Core metadata defined in ISO and Dublin Core standards must, according to the majority of the interviewees, be provided with geospatial datasets to enable comparative evaluation of dataset quality and trustworthiness. The study revealed the importance of dataset provenance and licensing information when assessing whether to trust a dataset based on its fitness for purpose. Data users confirmed that provenance information is usually incomplete and licensing information is normally missing from the metadata records of datasets. Dataset users are also interested in soft knowledge about data quality – i.e., data providers' comments on (a) the overall quality of a dataset, (b) known data errors, (c) potential data use, and (d) any other information that can help in the assessment of fitness for use of datasets. Also important when selecting a quality dataset are peer recommendations and reviews: dataset users are keen to be able to obtain feedback from their peers and are willing to accept peer recommendations when trying to select the most appropriate dataset for their given needs. The study results revealed the importance of citation information when assessing whether a dataset is fit for purpose, yet there was general consensus that citation information is, unfortunately, hard to acquire. It was discovered that, when selecting a dataset, users typically seek information about dataset providers and, in particular, value the availability of valid contact details for providers. Finally, study findings indicated that having side-by-side dataset and metadata comparison functionality would make the dataset selection process much easier for users. Each of these informational aspects is explored in more detail below.

4.2.1 Metadata Completeness and Compliance with Standards

The survey results revealed that metadata records play a very important role in the dataset discovery and trust-based selection process. Users of geospatial data heavily rely on information provided in the metadata records to assess a dataset's fitness for use, and to evaluate its quality and trustworthiness. For instance, the data archivist (interviewee 1) explained that his organisation "*should at least have metadata for all data that* [they] *can find, so that* [they] *know what exists*". The land use researcher (interviewee 3) stressed that "*what* *you need is good metadata which will tell you all about how it was collected, when it was, who did it, who to contact*". Unfortunately, at present, users find metadata records are typically incomplete with a lot of essential data omitted. As the data archivist stated, "*I don't think that metadata as it's provided at present contains enough information about the data to make it useful*". When discussing his previous workplace, the researcher data user (interviewee 2) recalled occasions when *"customers provided* [their organisation] *with data which* [they] *should use for some kind of data processing or analysis and there was a lot of metadata missing*". The climate forecaster (interviewee 4) also raised the metadata completeness issue, arguing that some datasets "*don't come with quality statements* [and] *are not particularly well described*". While users recognise that "*now there is more and more pressure for good metadata and standards in metadata*" (interviewee 3), users often "*have to manually inspect every single dataset before* [they] *start understanding whether the metadata adequately describes the data*" (interviewee 1).

Despite the standardisation bodies' work towards establishing core metadata elements and enforcing good metadata practices, dataset providers do not always follow metadata standards and leave metadata records incomplete. Nevertheless, it was discovered that data providers typically agreed on the importance of metadata and metadata standards. The data archivist commented that "*if data sources are standardised and automated then one can obviously do a bit more with it*", and that his organisation "*support*[s] *the ISO family of spatial metadata standards* [and] *Dublin Core for non-spatial datasets*" (interviewee 1). The data producer (interviewee 5) explained that his organisation is currently using *ad hoc* standards to include information which seems most relevant, but they *"are trying to move a bit more towards using some of the ISO standards for the discovery data*". The interviewees commented that lack of complete and well-documented metadata records particularly affects historical datasets and data that was collected as part of smaller scale projects or collected by individual researchers. As a result, users suffer from insufficient metadata records which makes the dataset selection process more difficult and time consuming.

The interviewed dataset users stated that, when searching for geospatial data or evaluating data quality, as a minimum they require information on coverage, licensing, methodology, uncertainty, resolution, source, acquisition date, pre-processing steps, and accuracy (see Table 4.2 for full metadata elements list). Although all these metadata elements are, in fact, defined as core in the existing Dublin Core and ISO 19115:2003 metadata standards (see section 2.4.2.2), the interviewees stated that the metadata records that they typically encounter are far from complete and are usually missing many of these core elements. These findings highlight the importance of metadata standards at the same time as illustrating that geospatial metadata still typically lacks metadata quality control.

Table 4.2: Metadata elements considered by interviewees when selecting datasets to use.

Metadata Element	Metadata Element	Metadata Element
Title	Subject	Author/Responsible party
Associated party or parties (co-authors, etc.)	Topic	Category
Keywords	Abstract	Language
Character set	Scale	Temporal scale
Attribute	Format version	Data representation
Spatial reference system	Bounding coordinates	Coverage (main dimensions of the data)
Spatial coverage	Geographic coverage	Temporal coverage
Taxonomic coverage	Metadata standard	Metadata creation date
Custodian	Custodian's reference	Legal usage information
Restrictions	Caveats	Licensing
Usage rights	Data set owner(s)	Associated party or parties (co-authors, etc.)
Access control	Access rules	Contact
Methods (methodology)	Physical structure	Technical data set parameters
Attribute(s)	Online distribution	Online resource
Sum elevation	Uncertainty	Resolution
Spatial resolution	Temporal resolution	Precision
Source	Acquisition date (year and time)	The pre-processing steps
Accuracy	Horizontal accuracy	Vertical accuracy
Absolute accuracy	Projection	Units
Bias gain	Cloud cover	

4.2.2 Metadata Visualisation and Comparison

The interviews revealed a need for more sophisticated tools for visualisation of metadata records, with users commenting that, at present, metadata records are not only incomplete but are also typically not easy to examine and assimilate. As the land use researcher explained:

> "*Having just a little view of the attribute table…a description of what is in* [it]*, what each attribute is and what is actually in your dataset, that's really handy. So you can, rather than downloading it, have a look at it … quickly to see what's in it*" (interviewee 3).

Metadata records are typically supplied as complex XML documents, making it far from easy to examine and assimilate them. Non-expert dataset users would appear to suffer the most from not being able to absorb and understand all of the information recorded in metadata. Considering the aforementioned importance of metadata in terms of the perceived fitness-for-use and trustworthiness of datasets, effective visualisation methods for metadata records need to be investigated to support users in data quality evaluation and trust-based decision-making.

The interview results highlighted another important aspect of metadata visualisation – that is, the ability to *compare* metadata records across candidate datasets. The dataset user interviewees indicated that side-by-side visualisation of all available metadata elements would allow them to compare geospatial datasets more effectively in order to determine which to trust. The side-by-side visualisation of metadata would be most beneficial when comparing datasets that are very similar and so the differences are hard to initially identify. For instance, the land use researcher argued that:

> "*there are so many datasets that probably do very much the same things but one might have an extra field in it that is crucial, or one might be at a different scale, or one has errors in it, or different units too, all sort of things like that. So it would be handy to sort of say 'oh that weather station data is much better than this weather station data'*" (interviewee 3).

The interviewees stressed that selection of a suitable dataset can be a very difficult task, especially when there are several that potentially fit the purpose. A side-by-side visualisation functionality would support and simplify data searches, the selection decision-making process, and data quality evaluation. Metadata comparison would be particularly invaluable for less-experienced or non-expert users who find it hard to manually inspect datasets to assess their fitness-for-use and trustworthiness.

4.2.3 Peer Review/Recommendation

As described in Section 2.1.3, e-Commerce has utilised peer review to engender customer trust in vendors and their services. At present, a significant number of commercial websites use commenting and rating functionality to facilitate user feedback on products and services that they offer. In particular, auction-based websites, such as e-Bay, heavily rely on user feedback and ratings because peer reviews validate the trustworthiness of vendors and the quality of their products. Surprisingly, geospatial data portals did not at the time of writing provide peer review functionality to allow dataset users to comment on and rate geospatial data products. On the basis of the interviews, it would appear that geospatial dataset users are very keen on having peer review/recommendation facilities available to them.

Peer review in this study appeared as one of the most prominent themes, with four out of six interviewees indicating its strong influence on the data selection process. The interviewed users of geospatial datasets stated that they heavily rely on peer recommendations when selecting a dataset to use. They contact their peers to elicit suggestions on what datasets are most suitable for their purpose, are of good quality, and are trustworthy. The land use researcher (interviewee 3), for instance, commented that, when evaluating data quality and trustworthiness, she would talk *"to other people that have used it"* because she trusts her peers. The dataset provider agreed that his organisation "*certainly accept*[s] *recommendations from other people*" (interviewee 5) and would also "*welcome*" any comments on the data that they produce:

> "*we have a help desk* [and] *if we get comments back we normally try do something about them if there is anything that's a problem*" (interviewee 5).

The dataset provider also positively welcomed "*the idea of a community of practice who could rate* [their] *reputation in terms of the datasets* [they] *provide*" (interviewee 5). This indicates that geospatial dataset *producers* are also interested in having their datasets peer reviewed; having user feedback on datasets that they produce would allow data producers to identify and resolve any issues within their datasets, and afford them an opportunity to respond to users' comments.

The climate forecaster, when discussing one of the datasets he was currently using, said "*one of my colleagues told me about this dataset that wasn't only one result, but also had uncertainty*" (interviewee 4) which further indicates the importance of peer advice. Also the data archivist confirmed the importance of peer recommendations and commented that his organisation "*have defined a framework for quality assessment that relies on peer review of some kind, almost "crowd sourcing" the assessment of the data by way of likes or dislikes, or frequency of use, or citation but* [they] *haven't implemented that yet*" (interviewee 1). The land use researcher actively supported the idea of "*people being able to comment on metadata, to have user feedback*" (interviewee 3). She argued that:

> "*Sometimes you'll look for a dataset, and you know people have used it, and you want to speak to people that, you know, you don't want to find them out yourself… And then having this active community where the person that publishes that data is responsible for what they are publishing. So then they can take the comments on board and comment back. I think it needs something like 'I used it for this', 'it's good for this but maybe not for this'*".

Peers typically provide valuable feedback on what datasets they used, what these datasets were good for, potential problems with the datasets and other potentially useful information. Having this sort of feedback available for every geospatial dataset would facilitate improved dataset selection and quality and trustworthiness evaluation. Peer review functionality, as proposed by the land research user, could also potentially be used to validate metadata records, highlight any inaccuracies or even fill in any informational gaps. Such feedback functionality would allow dataset users and geospatial data experts to provide their comments on datasets and flag any limitations or problems associated with the datasets.

Although the interviewees illustrated the importance of peer review, they qualified the value of such peer recommendations by indicating that they would not accept recommendations from just anyone – they mostly rely on known peers whose opinion they respect and trust. This suggests that any future peer review functionality for geospatial datasets would have to be controlled in order to enhance reliability and exclude the possibility of malicious actions.

4.2.4 Rating of Datasets

In line with having functionality to provide comments on datasets, the interviewed users expressed their interest in being afforded the ability to *rate* the quality and trustworthiness of datasets. Users indicated that having something similar to 5-star ratings assigned to datasets would make quality-based dataset selection processes much simpler. The land use researcher provided an example of a data provider that uses quality indicators and star ratings to convey dataset quality information:

> "*they have these three or four little indicators at the bottom of their datasets that say how good the spatial coverage is, how accurate the dataset is… It is all on their website … so that is really, really good*" (interviewee 3).

It is recognised, however, that a single quantitative rating assigned to a dataset would not on its own provide enough information to *meaningfully* and *contextually* evaluate the dataset's quality and trustworthiness. For instance, a 1 star rating would not supply any valuable information regarding *why* the dataset is considered to be of poor quality. Furthermore, whilst a dataset might be rated poorly relative to one context of use, its rating might be significantly higher for another context (perhaps one more closely aligned to its intended purpose). It is therefore suggested that to be meaningful and of value, quantitative ratings would need to be justified with detailed explanations regarding why a particular rating was being given to a dataset. This further suggests that rating functionality would perhaps be best combined with peer review/recommendation facilities wherein ratings would act as visual indicators of overall quality but user comments and reviews would provide the underlying rationale to substantiate and contextualise the rating.

4.2.5 Expert Review

In line with having peer reviews and recommendations, it was suggested that domain *experts* could provide valuable and highly informed judgements of dataset quality and trustworthiness for the benefit of more general users. The data archivist (interviewee 1) described a new project in which he is involved where domain experts' reviews are used to describe datasets and their quality. The project consists of seven or ten themes, which "*are similar almost to the Societal Benefits Areas that they use in GEOSS*" (see section 2.6.1 for more information on the Societal Benefit Areas), and where "*funded theme conveners and scientists that work on those themes describe the datasets that are available and most useful and so on*". The interviewee argued that "*not all datasets are useful or have equal value*" and therefore there could be considerable benefit in "*having some kind of value judgement by experts, on top of, let's say, more common requests for data supporting a given theme*". The interviewee further explained that his organisation has come up with three categories of peer review statements – (1) syntactic, (2) schematic and (3) semantic – to describe whether datasets are (1) syntactically correct and interoperable, (2) comply with a widely accepted schema, and (3) fit for a particular purpose.

This clearly demonstrates that effort is being made to provide the geospatial user community with expert value judgements of dataset quality. Nevertheless, as discussed in sections 2.4.3 and 4.2.3, neither peer nor expert review functionality is yet available for public consumption. Domain experts' value judgements could potentially provide invaluable information about dataset quality, errors, suggested dataset use, areas of applications, etc. to a broader consumer audience.

4.2.6 Reputation of Data Providers

The reputation of dataset providers was identified as one of the key factors in dataset selection. The interviewed users typically rely on data from producers that they already know or those who have a very good reputation in the community. The researcher data user strongly supported the data provider reputation theme and, when discussing his former job, acknowledged that dataset choices were *"driven by reputation of the data provider*" (interviewee 2). The data provider's identity alludes to how trustworthy the data is perceived to be by users because "*knowing who provides the data* [is] *the measure of how you could treat it afterwards*" (interviewee 2). Furthermore, the reputation of a dataset provider is seen as "*the substitute for any quality seal*" (interviewee 2). The study results also showed that large international organisations and governmental data providers are accepted by data users as more reputable:

> "*When it comes to landsat, the sources you have are huge national or international organisations which provide the data. And they apply certain standards...quality standards… I would assume them to be very professional and reliable*" (interviewee 2).

The land use researcher further confirmed that large-scale data producers are seen as more reputable, noting that "*there are probably organisations that you do just trust*" (interviewee 3). The academic researcher also agreed that data from national repositories and governmental data producers "*tend to be pretty uniform standards*" (interviewee 6). It follows, therefore, that if current dataset users are more likely to select data from producers that already have established a good reputation, smaller and unknown data producers are likely to experience much lower dataset demand. As argued by the researcher data user, "*if* [data producers] *have no reputation at all or low reputation you will act accordingly*" (interviewee 2). This is similar to e-Commerce, as discussed in Section 2.1.3, where new smaller vendors are initially much less trusted by customers, making it harder to secure online sales until they have established a positive reputation.

The interviewees indicated the importance, when assessing the trustworthiness of data providers, of documentation supplied by data providers with their datasets. Well-organised and easily accessible documentation engenders user trust in both data providers and their datasets. For instance, the data provider argued:

> "[Producer X] *has a very good reputation. And that is strongly backed up by when you look at their documentation and such, you can see they've considered a vast number of things with great deal of care. And the more you dig into it it's very clear that they are very good at what they do*" (interviewee 5).

The study results also revealed the importance of availability of contact information to obtain additional details about the dataset, for instance, to find out the approach that the data provider followed to generate the data or to clarify what licence applies to the dataset. As articulated by the land user researcher, "*having a good contact, someone that* […] *actually replies* [is] *really handy*" (interviewee 3).

Data users will typically have more confidence in data providers that invest resources into documenting their data and responding to user queries and will trust such data providers above others. A facility to support rating of dataset providers would allow emerging providers to gain, via word of mouth and excellent service provision, the reputation to support their successful entry into the GEO community. Similarly, it would encourage established providers to maintain a level of service over time.

4.2.7 Soft Knowledge

As already noted in Section 4.2.1, metadata records provided with geospatial data are often incomplete and are even sometimes ambiguous. The interviewed dataset providers commented that there are cases when data quality measures cannot be recorded in standard metadata records. For instance, a provider might be aware of problems with a particular satellite, such as higher uncertainty in specific areas of the image it takes; in such cases, they can provide some *soft knowledge* about the quality of the data, including information which they think may be relevant to potential users. As explained by the data provider (interviewee 5):

> "*for example, we have a sensor that has a particular range. We know it is pretty good in the middle of the range but at the edges we had seen indications that it is not so good. But we don't have any way to actually* […] *prove* […] *that. So all we can do is warn people that we believe it to be less good, they should be more cautious of it. But we don't actually have a quantitative estimate* […] *we can just point towards data signal noise and say 'clearly it is not going to be as good here'. So it is almost like a relative estimate*".

Interviewee 5 stated that his organisation follows some metadata standards such as NetCDF Climate and Forecast (CF) Metadata Conventions and ISO standards, but they also "*try to include information about what* [they] *think is good and bad in the dataset on general terms*". Although his organisation is now *"trying to move a bit more towards using some of the ISO standards for the discovery data*", they continue "*incorporating stuff that other people are suggesting would be good to have in there*". Interviewee 5 also argued that the organisation he works for cannot predict all possible uses of their datasets in a long term, but they "*certainly support any use that people can sensibly make of* [their data]".

In tandem, the users who were interviewed stressed the importance of data producers' comments and recommendations relative to the datasets they provide. The climate forecaster (interviewee 4) argued:

> "*Often a dataset is derived* [for] *some particular purpose and it might be good for that purpose but it is not necessarily good for other users. And often the* [person] *that creates the dataset might know that better than the average user"* and noted that it would be useful to, at the very least, have "*some kind of soft knowledge of what are the boundaries, when should I trust the dataset, when should I not trust the dataset*".

Interviewee 4 further emphasised the importance of soft knowledge about data uncertainty and error estimates:

> "*the* [most important] *thing is the uncertainty* […] *Maybe the one who derived the dataset doesn't know the exact uncertainty themselves...*[but] *at least if they could include the soft knowledge about uncertainty, that would also help*".

It is clear that both producers and users of geospatial data appreciate the importance of 'soft knowledge' as provided by dataset producers. Although, as outlined in section 2.4.2.2, the standardised metadata records can provide key information about datasets, there exist cases when additional, producer-derived information can be more relevant and valuable to potential data users. In particular, descriptive estimates of dataset quality could be particularly valuable to non-expert data users who are not accustomed to reading and using complex XML documents.

4.2.8 Citation Information

The interviews highlighted that the majority of users also base dataset quality evaluation on dataset citation information. When making dataset selections, users are interested in accessing publications where data quality checks are reported; consequently, journal articles that describe dataset use and evaluation are considered to be very important in dataset quality assessment. For instance, the land use researcher argued that "*when you get data from academic journals, that's good, because you can read exactly how they collected* [it]*, why they collected* [it] *and you get the sort of analysis after seeing all the errors*" (interviewee 3). The data provider stated that, when assessing quality and trustworthiness of datasets, their organisation "*tend*[s] *to look for reports on quality and such like, so that is more textual documentation*" which can typically be found on data producers' websites and documents repositories. As discussed by the data provider, availability of well-organised textual documentation may also affect users' perceptions of data providers' credibility and establish providers' reputation in the community.

The climate forecaster, when discussing one of the datasets that he used, said it "*was the first one that I was aware of that gave this relatively high resolution climate data. And it* [ha]*s also been used in several other papers*" (interviewee 4). Interviewee 2 also mentioned that *"there are loads of publications"* on the data that his organisation uses. This clearly indicates the extent to which citation information may affect dataset quality and trustworthiness evaluation and dataset selection decisions. Unfortunately, as stated by the data archivist, citation information is often missing, which complicates dataset discovery and fitness for use evaluation.

4.2.9 Quantitative Quality Information

In line with *qualitative* quality information discussed above, the interviewees stressed the importance of *quantitative* quality information in evaluation of dataset quality, trustworthiness and fitness-for-use. Four out of the six interviewees listed a number of quantitative measures which they consider when selecting a dataset to use, namely: spatial and temporal resolution; spatial and temporal scale; precision; geometric correctness; horizontal, vertical and absolute accuracy; precision; error estimates; and uncertainty. The interviewees expressed particular interest in availability of uncertainty estimates, especially at the pixel level. The climate forecaster (interviewee 4), when describing the quality aspects that influence his dataset selection decisions, argued that "*uncertainty is the most important*" but conceded that he is still likely to use a dataset that has no uncertainty information available because uncertainty measures are usually missing.

The data provider (interviewee 5) further supported the importance of uncertainty measures in dataset quality evaluation:

> "*satellites* […] *measuring what is in the oceans* […] *can do fairly well out in the open ocean but as you are getting towards the shore line there is all kind of rubbish washing out of rivers and things that messes up the processing quite a lot. So we would expect higher uncertainty there, so you should treat those data with more caution*".

This data provider also explained that some of the projects he is involved in *"are actually trying to derive those* [uncertainty] *measurements*" and he is very much interested *"in actually associating the uncertainty measurement*" with data. Without knowing the uncertainty estimates, and without carrying out "*uncertainty propagation to estimate what the final error is*", it is not possible to identify "*how good end products are*". For the datasets he produces, the interviewee attempts (as already mentioned) to provide at least "*a relative estimate*" – soft knowledge – of uncertainty and errors in cases when it is not possible to provide quantitative evidence; he noted that *both* quantitative and qualitative quality descriptions are important in evaluation of datasets' quality and trustworthiness, describing a different perspective on each:

> "*I guess I would see uncertainty as a somewhat mathematical estimation of how good or rather how accurate a dataset can be, whereas quality, I would probably say, is possibly more subjective. And we certainly have an interest in both of those angles.*" (interviewee 5).

Quantitative quality information describes the internal quality of datasets and is particularly important to expert users. Quantitative measures can determine how data will be used and how trustworthy it is; for instance, data with high uncertainty and error estimates should not be used in high-risk and high-impact areas of application. Providing the GIS community with both objective and subjective quality information could ensure that users with different dataset quality requirements could make an informed dataset selection decision.

4.2.10 Data Provenance and Method

The interview results highlighted that significant importance is placed on provenance information by geospatial data users when evaluating quality and trustworthiness. Five out of the six interviewees indicated interest in availability of provenance information and identified the following provenance elements as essential in data quality and trustworthiness evaluation:

- original dataset provider;
- party (parties) who has subsequently processed the dataset;
- party (parties) who has used the dataset before;
- method adopted for dataset data collection;
- how a dataset was derived and on what it is based;
- dataset harvesting pathway;
- the purpose for which a dataset was originally collected; and
- dataset processing log.

The method adopted for data collection plays an important role in evaluation of data quality and trustworthiness. For instance, the data archivist (interviewee 1) stressed the importance of methodological information in long-term data preservation:

> "*the methodology is probably also important in the metadata because if we are going to preserve data for the very long term, in 50 years from now the people who generated that dataset are not going to be around anymore"*.

The results revealed that information on the equipment that was used for data gathering and any problems that occurred during the gathering process can help data users in quality and trustworthiness evaluation. Interviewees stressed the importance, in terms of fitness for use evaluation, of knowing the purpose for which the data was initially gathered "*because you could have someone collecting data for a purpose which misses out what you really need*" (interviewee 3).

The interviewed users commented on the importance of knowing about the data processing steps and availability of a processing log which would allow users to trace the events and transformations in the life of a dataset. For instance, the researcher data user (interviewee 2) explained:

> "*We would definitely need what pre-processing steps have been applied and which methods* […] *we would want everything concerning the processing or pre-processing*".

Additionally, as indicated by the academic researcher (interviewee 6), access to the raw data is also important so that researchers *“can go back to the root of that processing* ‘*cause that processing isn’t always helpful*” and can then do things “*a little bit better* [to] *improve* […] *results subsequently*”.

Having all the information on processing steps and algorithms applied to data would allow users to measure data accuracy and better evaluate its quality, fitness-for-use and trustworthiness. Processing information would also support reversal of any processing if necessary/possible to improve quality of end products.

Unfortunately, however, according to the interviewees dataset provenance information is, at present, not usually available to data users. As the data archivist explained (interviewee 1):

> “*the path that the metadata follows to arrive at your doorstep is lost. You only ever, usually, have a portal that you harvested it from, and maybe the starting point but you don’t have the intermediaries. So one other thing we are looking at is to improve the way in which we store maybe a pathway where each portal that touches a metadata record has an element to describe who they are and where they are located and who to contact at the end of the metadata record, so that one can trace the path*”.

This claim is very interesting because, although geospatial data portals and clearinghouses are *designed* to provide a gateway to datasets and enable better dataset discovery, in practice this valuable information on the source of datasets is usually lost. Interviewee 1 further argued that “*if you can cut out a very long harvesting route by going directly to another source then maybe one should be able to do so*”. The academic researcher (interviewee 6) observed that acquiring processing information is “*going to be very difficult ‘cause it’s been programmed through the numerous people who did not necessarily document”* their processing. Incomplete metadata records supplied with datasets can also contribute to the loss of processing information.

4.2.11 Licensing

ISO 19115:2003 and Dublin Core metadata standards include restrictions and licensing information as core metadata elements for all geospatial datasets (Figure 4.1 and Figure 4.2). Despite this, data producers often fail to provide such information, with a number of interviewees pointing out that licensing information is nearly always missing.

Aston University

Illustration removed for copyright restrictions

Figure 4.1: Dublin Core Metadata Element Set, Version 1.1 – Rights element (DCMI, 2014b).

Aston University

Illustration removed for copyright restrictions

Figure 4.2: ISO 19115:2003(E) – Constraint information (ISO/TC211, 2003, p. 21).

The researcher data user (interviewee 2) observed that although geospatial data generally has usage limitations, the restrictions information is often not provided with the geospatial datasets and commented that "*a lot of people are not aware that geo data actually often has limitations on the way you are allowed to use it. This is less true for those data that you can download but if it is provided by via other sources it often has certain licence limitations and I think a lot of people have no clue about this*".

4.3 Study Limitations

The main purpose of this study was to gain an initial insight into factors that may affect geospatial dataset selection decisions in order to direct further research work. Although this initial investigative study is not without its limitations, as noted below, it has been successful in terms of eliciting rich, qualitative data to begin to illustrate issues surrounding dataset selection in the GIS domain. Obviously, the study has been conducted with a very small sample of users/producers. Whilst it could be argued that this may have influenced the results, and rendered them less than globally representative, data saturation was observed and the findings from these six interviews validated against those of the additional 12 interviews conducted by project partners. As such, it is felt that the findings represent a reliable overview of domain stakeholders' opinions. Although the interviews were semi-structured to optimally retain focus whilst supporting flexibility of investigation, it is recognised that the set of questions selected to guide the interviews could have potentially influenced interviewees' thinking about data quality and trustworthiness and the issues important therein; for instance, interviewees were specifically asked about the importance of supporting documentation in dataset selection and this may not have been something that they would intuitively have considered if unguided. That said, every effort was taken to mitigate against influencing interviewees' responses by keeping the questions as generic as possible to allow for new, interviewee-led topics of discussion to emerge. Examination of the interview transcripts would suggest these study limitation have been successfully minimised as far as was feasibly possible.

By selecting expert geospatial data users and producers it is recognised that an elite bias has been introduced to the study sample. This decision was made, however, on the basis that it was felt that only geospatial data experts could provide sufficiently deep insight into current challenges regarding geospatial data and help to identify important factors that influence dataset selection decisions based on their expertise and experience. Despite the fact that the study involved a variety of experts from different domains and countries, there nevertheless remain limitations in terms of the types of experts involved – three out of six study participants were researchers – which could have biased the results in terms of identified geospatial data quality and trust themes. The interviews only involved one geospatial data provider and no large-scale data producers, limiting insight into producers' views on geospatial data quality. Although the results from this study are deemed acceptable for an initial investigation to direct future work in this area, the requirements elicited from these findings were taken further and explored in a more structured quantitate study, as reported later in this thesis.

4.4 Summary and Conclusions

The results from this study lend empirical evidence to support the observation of direct parallels, as discussed at a theoretical level in Chapter 2, between geospatial data quality and trust themes and well-defined trust models and trust triggers that are already used extensively in B2C e-Commerce to engender consumer trust and increase user willingness to engage in online transactions. Compliance of geospatial data and metadata with international standards directly relates to *vertical* and *institution-based* trust (see discussion in Section 2.3.3) and is comparable to use of B2C e-Commerce trust seals (see Section 2.1.3). Peer reviews and recommendations on quality of geospatial datasets together with ratings of datasets relate to *horizontal* trust (see discussion in Section 2.3.3) and reflect consumer testimonials that are widely used by online vendors to engender consumer trust in the products and/or services that a vendor provides (see Section 2.1.3). Expert reviews also relate to *horizontal* trust but are different from consumer testimonials; hence this informational aspect does not have an exact counterpart in terms of an e-Commerce trust trigger. Producer reputation relates to *credibility* and *benevolence* dimensions of trust (see discussion in Section 2.3.5) and reflects the e-Commerce branding trust trigger (see Section 2.1.3). The B2C e-Commerce trust trigger relating to alternative channels of communication is also relevant here since geospatial data users are highly interested in the availability of dataset producers' contact information. Informal producer comments (soft knowledge) also relate to producer *credibility* and *benevolence*, although this informational aspect does not have a corresponding e-Commerce trust trigger. Availability of citations information relates to both *horizontal* and *vertical* dimensions of trust – *horizontal* trust is supported by descriptions of the dataset quality and recommendations on the dataset's use supplied in scientific papers, while *vertical* trust is trust in a journal or a conference where the document was published. Availability of quantitative quality information and licensing information links to *institution-based* trust since it indicates that the dataset adheres to some international standard. Geospatial dataset provenance and the method(s) adopted for data collection and processing relate to *vertical* and *technological* dimensions of trust, respectively – i.e., trust in the organisation that produced the dataset and technologies that were used to collect and process the data to produce the dataset. Citations information, quantitative quality information, data provenance and licensing do not have corresponding B2C e-Commerce trust triggers.

The interviews with GEO data users and producers were very enlightening in terms of identifying their informational dependence when selecting datasets for use within particular contexts. Discussion during the preparatory phase of this research suggested that a GEO label would best serve a drill-down function whereby, at the top level, it visually represents the *availability* of specific informational elements for its associated dataset and, thereafter, permits users to click the label to drill down into and interrogate the detail for each informational element. Based on the interviewees' responses, 11 GEO label-appropriate themes were extrapolated (as described in Section 4.2) and, from these, 8 facets were further identified as potential candidates for inclusion in the GEO label (see Figure 4.3), namely:

- the reputation of the dataset producer;
- producer comments on the dataset quality;
- dataset's compliance with international standards;
- community advice;
- dataset ratings;
- expert value judgments;
- links to dataset citations; and
- side-by-side metadata records comparison.

The metadata completeness theme, including availability of quantitative quality information, licensing information, and provenance information, has been incorporated into the compliance with international standards facet.

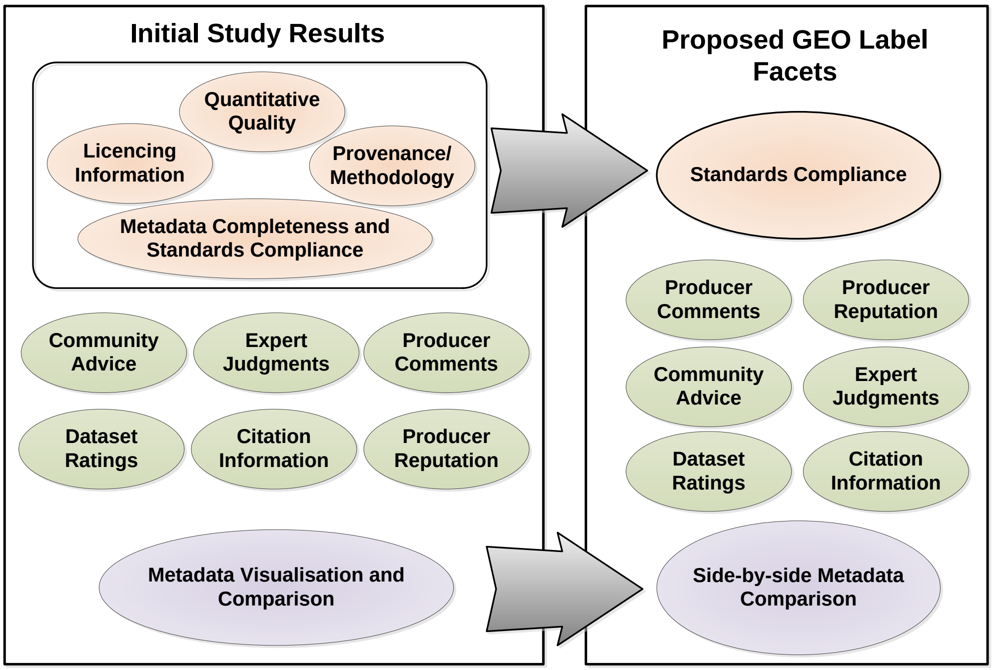


Figure 4.3: GEO label facets derived from the initial interviews results.

Having identified candidate informational/functional foci of the GEO label, a subsequent study was designed and administered to investigate users' and producers' opinions on the concept of a label, and to elicit their assessment of the role that such a label should assume (e.g., whether or not there was general agreement that a GEO label should provide a drill-down interrogation facility covering all 8 of the candidate facets). This study, forming the first main phase of this research agenda, is documented in the next chapter of this thesis.

Chapter 5 Phase I: Soliciting User and Expert Views on a GEO Label

This chapter presents the first main phase of this research study which was conducted to solicit geospatial data producers' and users' views on the concept of a GEO label and the role it should serve. Section 5.1 outlines the mechanism by which this phase of the research was conducted. Section 5.2 presents and discusses the study results. The study limitations are discussed in section 5.3. Finally, section 5.4 presents three prototypic graphical GEO label representations which were derived and informed by the results of this and the initial study (see previous chapter) as potential means to convey availability of geospatial dataset quality information.

5.1 Soliciting Opinion on a GEO Label

In order to conduct an initial investigation into both geospatial data producers' and users' views on the concept of a GEO label and the role it should serve, an online questionnaire-based survey was conducted. The questionnaire comprised five parts, A to E, each of which is described below (a full copy of the questionnaire is included in Appendix B). On commencing the questionnaire, respondents were first asked to identify themselves as primarily either users or producers of geospatial data. Depending on their self-identified role, respondents were then presented with an appropriately tailored version of Section A. The user-oriented version of Section A (see Appendix B, Section A – Data User) focused on eliciting information about the degree to which users have a choice of dataset to use and what typically impacts or influences this choice. The primary purpose of this section was to verify the results of the initial interviews and to validate the informational aspects that were identified as important when making a dataset selection decision. It was also important to ascertain whether users actually have choice in geospatial data and what sources of data they consider. To address these questions, this section comprised a series of questions asking respondents:

a) to classify themselves as belonging to one or more user categories (e.g., private sector data user, and/or governmental data user, and/or researcher data user, etc.);
b) to identify the data sources they rely on for their work;
c) to reflect on the importance of different informational aspects about datasets when selecting a dataset to use; and
d) to illustrate their awareness of any certificates or seals that apply to geospatial data.

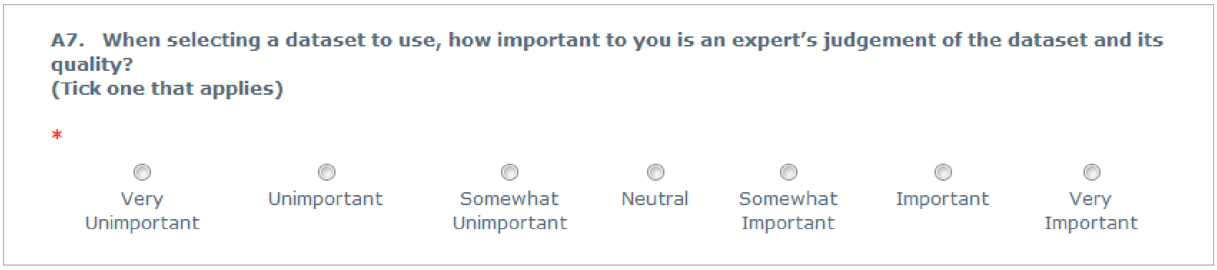


Figure 5.1: Example Likert scale to gauge opinion about the importance of different informational aspects.

Using a 7-point Likert scale, ranging from 'very unimportant' (1) to 'very important' (7) (see Figure 5.1 for an example), dataset users were asked to reflect on how important they felt the availability of each of the following informational aspects (as derived from the preparatory phase research) associated with datasets is in terms of selecting a dataset to use:

- information about the reputation of the dataset provider;
- producer comments on the dataset quality ('soft knowledge', i.e., subjective and informal statements provided by the creator or provider of the dataset);
- dataset's compliance with international standards;
- community advice and recommendations;
- expert value judgments of the dataset and its quality;
- links to dataset citations (e.g., journal articles or other publications where the dataset has been used); and
- side-by-side metadata records comparison.

As can be noted, dataset users were only asked to reflect on the importance of 7 informational aspects. This is because at the time of conducting the study dataset rating functionality was not supported by the geospatial data portals and clearinghouses.

These questions were specifically designed to validate the results of the initial interviews and to identify the informational aspects that are considered most influential when making a dataset selection decision. The most important/influential informational aspects could then be considered as candidate GEO label functions. The 'Dataset ratings' informational aspect (see Section 4.4) was intentionally omitted because practical implementations of such functionality are not currently provided within the GIS community; although indicated as desirable (see Section 4.2.4), it seemed unreasonable to ask about the in-practice importance of an aspect which is not yet currently available.

The data producer-oriented version of Section A (see Appendix B, Section A – Data Producer) was naturally focused on the production rather than selection of datasets. It comprised a series of questions asking respondents:

a) to classify themselves as belonging to one or more producer categories (e.g., GEO committee member, and/or data producer not part of GEOSS, and/or private sector data producer, and/or governmental data producer, and/or research/academic data producer, etc.);
b) about the type of data they produce and about international standards their datasets observe; and
c) about their awareness of any certificates or seals that apply to geospatial data.

It should be noted at this point that, although users and producers were asked about their awareness of any geospatial data certificates or seals, at present, no such publicly accessible certifications exist. While there are many active geospatial data and metadata standardisation initiatives (see Section 2.4.2.2), technically, such activities and the resulting standards are not certification programmes. The main purpose of this question, thus, was to identify any certification programmes that are, perhaps, internal to respondents' work organisations.

The remaining sections of the questionnaire were common to all respondents. Section B comprised a small number of questions to gather information about users' and producers' initial views on the role that a GEO label should serve. Respondents were asked if and why geospatial data or metadata records would benefit from the application of certification programme(s). They were also asked to comment on the role they would want a GEO label to serve and identify whether the presence of such a label would influence their dataset selection decisions.

Section C focused on gauging respondents' awareness and opinion of commonplace e-Commerce rating systems and certification programmes. The main purpose of this section was to bring respondents' attention to the mechanisms that are already widely used in e-

Commerce for communicating the quality of products and trustworthiness of e-vendors. The examples presented in this section were kept as generic as possible to minimise the influence on respondents' views on the applicability of such mechanisms to geospatial data. Respondents were asked what review or rating systems they use in their everyday life and, if used, to specify the frequency of use. Respondents were then presented with screenshots of two review and rating systems: one taken from an eBay products listing page (see Figure 5.2; also see Appendix B, Section C) with some of the listed products carrying a Top-rated Seller label, and the other taken from a TripAdvisor hotel listing page (see Figure 5.3; also see Appendix B, Section C) with all the listed hotels having user ratings and reviews.

Aston University

Illustration removed for copyright restrictions

Figure 5.2: eBay product listing page used to solicit information about respondents' awareness and opinion of e-Commerce review/rating systems.

Aston University

Illustration removed for copyright restrictions

Figure 5.3: TripAdvisor hotel listing page used to solicit information about respondents' awareness and opinion of e-Commerce review/rating systems.

Table 5.1: Trust questions related to the top-rated seller label.

1. The Top-rated seller label encourages me to trust the vendor more than I would otherwise.
2. The Top-rated seller label would have a strong negative effect on my intention to purchase.
3. I would be more likely to purchase a product from a vendor that carries the Top-rated seller label.
4. A vendor that carries the Top-rated seller label provides products of high quality.
5. A vendor that carries the Top-rated seller label has a good reputation.
6. A vendor that carries the Top-rated seller label is more reliable than vendors that do not carry such a label.
7. A vendor that carries the Top-rated seller label has my best interests at heart.

Table 5.2: Trust questions related to the TripAdvisor hotel reviews.

1. A positive review encourages me to trust a hotel more than I would otherwise.
2. A positive review would have a strong negative effect on my intention to book a hotel.
3. I would be more likely to book a hotel that has positive reviews.
4. The hotels that have positive reviews are of high quality.
5. A hotel that has good reviews has a good reputation.
6. A hotel that has good reviews is more reliable than hotels that have bad reviews.
7. Good reviews do not promote trust towards a hotel.
8. I would be less likely to book a hotel that has positive reviews.

Using a 5-point Likert scale, ranging from 'strongly disagree' (1) to 'strongly agree (5), respondents were asked to reflect on their level of agreement with a series of statements about both these systems (see Table 5.1 and Table 5.2; also see Appendix B, Section C).

Both positive and negative statements were used to try and ensure that respondents provided properly-considered ratings. As can be seen from Table 5.1 and Table 5.2, the statements covered issues of trust, purchase intention, product quality, reputation, and reliability. These statements were intended to encourage respondents to think of the type of functionality that is already offered by commonplace review and rating systems, and the effect they have on consumer decisions about the products to which they are applied.

Section D was designed to explore respondents' reaction to the concept of a GEO label that is clickable and supports a drill-down function such that, if the label was clicked, users would be navigated to a GEO label certification programme homepage which would provide users with information regarding the GEO label itself and offer a number of tools to help users to assess a dataset's fitness for use. Similar to Section C, the main aim of this section was to introduce respondents to the mechanisms that are successfully used in e-Commerce (drill-down function in this case) and could potentially be adopted in the GIS domain to communicate geospatial data quality information. To encourage respondents to think about how a GEO label could benefit from a drill-down function, and to provide some tangible examples of how such functionality is used in practice in other domains, this section presented some example e-Commerce seals of approval that use click-to-verify functionality (see Figure 5.4; also see Appendix B, Section D). It was anticipated that respondents would recognise at least some of the presented seals of approval and that these examples would help them to better understand the proposed drill-down functionality for a GEO label.

Aston University

Illustration removed for copyright restrictions

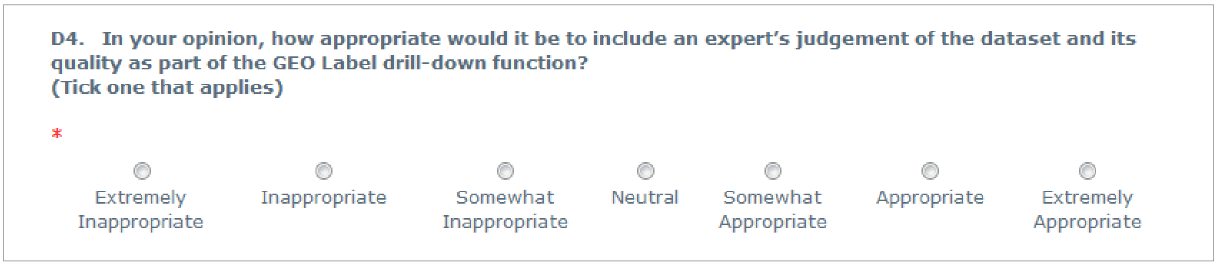
Figure 5.4: Examples of e-Commerce seals of approval that use click-to-verify functionality.

Respondents were asked whether they had ever come across any e-Commerce seals that use click-to-verify functionality, including those presented as examples. Those who indicated their awareness of the seals of approval were then asked: (a) whether they ever clicked on any seals to verify that they were genuine; (b) the reason why they did or did not click on the approval seals for verification; and (c) whether they thought that a GEO label should fulfil a drill-down function and have click-to-verify functionality.

The subsequent part of Section D explored both dataset users' and producers' opinions on eight potential facets that could be included as part of the GEO label drill-down functionality. Based on the findings from the interviews conducted in the initial investigative study, it was anticipated that GEO label tools could potentially:

- provide functionality to access experts' judgements about the value and quality of datasets;
- provide access to information on datasets' compliance with international standards;
- enable users to comment on datasets' quality, as well as to access comments provided by other community members;
- enable users to rate datasets' quality, as well as to access ratings provided by other community members;
- potentially rate the reputation of the providers that supplied datasets;
- give access or links to the documents, scientific papers and reports where datasets have been cited;

- provide access to ‘soft knowledge’ (subjective and informal statements) about datasets’ quality as provided by the creator or provider of the dataset; and
- provide a facility to visualise several metadata records side-by-side to enable better dataset comparisons.


D4. In your opinion, how appropriate would it be to include an expert’s judgement of the dataset and its quality as part of the GEO Label drill-down function?
(Tick one that applies)

*

Extremely Inappropriate | Inappropriate | Somewhat Inappropriate | Neutral | Somewhat Appropriate | Appropriate | Extremely Appropriate

Figure 5.5: Example question regarding perceived appropriateness of including proposed facets as part of the GEO label function.

Using a 7-point Likert scale, ranging from ‘extremely inappropriate’ (1) to ‘extremely appropriate’ (7), respondents were therefore asked to rate how appropriate they felt it would be to include the facility to interrogate (and, where applicable, contribute to) each of the informational aspects listed above as part of the GEO label drill-down function (see Figure 5.5; also see Appendix B, Section D). Unlike the examination of the features users *currently* rely on when selecting a dataset (Section A), here respondents were presented with *all* eight informational aspects, including ‘dataset ratings’. This section was essentially exploring users’ and producers’ opinions as to the appropriateness of including, in a GEO label for use by all, the informational aspects elucidated from the preparatory study. It was particularly interesting to investigate the level of correlation between (a) users’ reflection on what was of importance to them personally as users of geospatial data and (b) what all respondents (users and producers alike) thought more generally about a global GEO label function.

Finally, in Section E, respondents were asked to indicate and explain their preferences for the role of a GEO label – specifically, whether it should stand as a certification seal or a drill-down interrogation facility; respondents were additionally given the option to suggest any other function they felt a GEO label could potentially serve. Respondents were also asked whether they believe a GEO label should combine multiple functions – that is, whether they thought the label should combine several tools to access producer profile information and producer comments about dataset quality, access peer and expert reviews, access dataset citation information, visualise several metadata records side-by-side, provide quality assurance, and, by combining all these functions and tools, represent an all-in-one quality indicator. Armed with the information presented in the previous sections, respondents would be able to make an informed decision on the most appropriate function that a GEO label could fulfil. Consequently, this section focused on eliciting respondents’ concluding views on a GEO label and its role in communicating geospatial data quality.

The questionnaire was constructed and administered using the QuestionPro (QuestionPro, 2014) online-survey software. To inform potential respondents about the GEO label questionnaire, a small leaflet (see Appendix C) was produced that was distributed at the GEO Plenary session in Istanbul, Turkey on 16th of November 2012. Emails were also sent to a number of professionals from key organisations such as NASA, ESA and EPA, academics and researchers that work in the GIS field, and other GIS professionals, asking them to complete the questionnaire and to, wherever possible, circulate the questionnaire more broadly within their network of contacts. The goal was to target a variety of user and producer groups, ranging from individual researchers to experts who work in large-scale organisations. A total of 87 valid questionnaire responses were received: 57 from self-identified dataset users and 30 from self-identified dataset producers. The questionnaire was accessed and completed from a number of countries including Austria, Spain, Germany, Slovenia, Greece, United States, Netherlands, France, Italy, United Kingdom, Belgium, Switzerland, China, Bolivia, and Estonia – suggesting that it was successful in soliciting the opinions of experts from a variety of user groups and cultures around the world.

5.2 Results and Discussion

As noted, of the 87 valid responses received, 57 (66%) were from self-identified geospatial data users and 30 (34%) were from self-identified geospatial data producers.

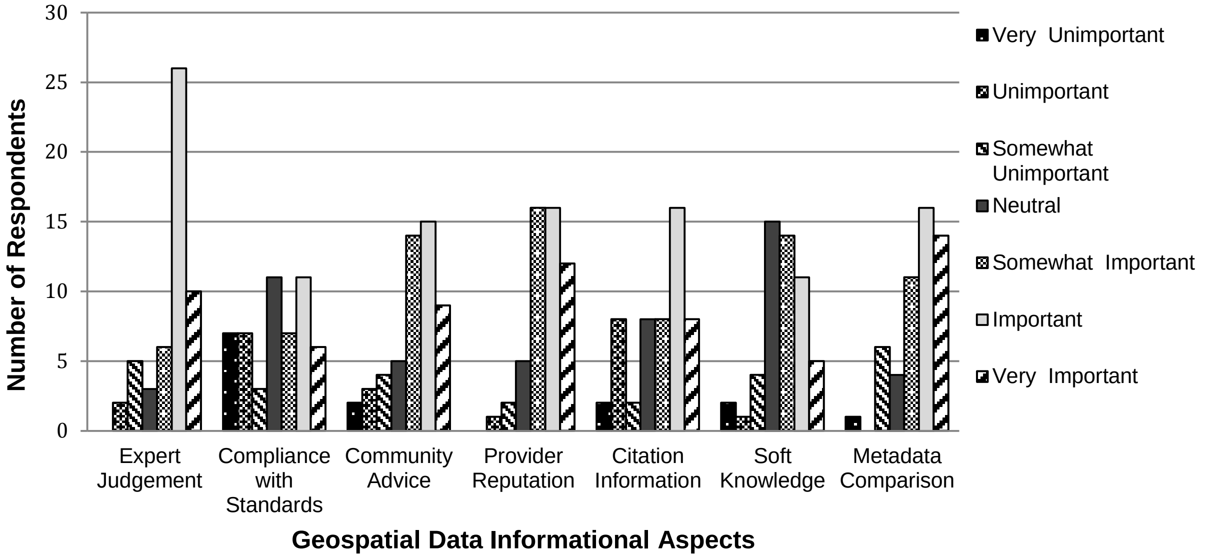
Table 5.3: Respondents' self-identified user and producer classifications.

User Type	Number of Respondents	Producer Type	Number of Respondents
Group on Earth Observations committee member	8	Group on Earth Observations (GEO) committee member	5
Private sector data user	2	Data producer with some dataset(s) in the Global Earth Observation System of Systems (GEOSS) Common Infrastructure	13
Governmental data user	15	Data producer not part of Global Earth Observation System of Systems (GEOSS)	7
Researcher data user	**40**	Private sector data producer	3
Academic data user	27	Governmental data producer	13
Other	2	Research/academic data producer	**16**
		Other	5

Amongst the former, 8 respondents identified themselves as Group on Earth Observations committee members, 2 as private sector data users, 15 as governmental data users, 40 as researcher data users, 27 as academic data users, and one respondent identified his/her user type as other, namely a manager within projects that use data (see Table 5.3; note that the users were free to self-identify as belonging to one or more categories). Among the data producers, 5 identified themselves as a Group on Earth Observations (GEO) committee member, 13 as a data producer with some dataset(s) in the Global Earth Observation System of Systems (GEOSS) Common Infrastructure, 7 as a data producer not part of Global Earth Observation System of Systems (GEOSS), 3 as a private sector data producer, 13 as a governmental data producer, 16 as a research/academic data producer, and 5 identified themselves as other (see Table 5.3; again, producers were free to self-identify as belonging to one or more categories).

5.2.1 Dataset Selection by Users

The majority of data users (91%) stated that they have a choice of dataset for their work. Of these, 86% stated that there is more than one data provider supplying the types of datasets they need. When asked to list the external data providers from whom respondents source their data, they indicated a broad spectrum of data providers including large governmental data producers such as NASA, NOAA, ESA, WorldClim, Ordnance Survey, etc., repositories that are largely free to the academic community, local universities, "*data sets from research papers*", and internet websites. Of the data users that have a choice of data, 85% confirmed that they considered metadata records when making a dataset selection. When asked to describe the types of metadata they typically consider, they indicated a broad spectrum ranging from literally any/all metadata that is available to specific metadata aspects including, but not limited to, information about data resolution, accuracy, and lineage, license information, information about the software that processed the data, supported standards, the purpose for which the dataset was initially generated, missing values within the dataset, the cost of obtaining the dataset, author information, and domain-specific elements such as, for example, "*field measurements of vegetation properties*". These findings suggest that users of geospatial data consider a wide variety of metadata elements and indicate the importance of the availability of complete metadata records for effective dataset selection (recall metadata completeness discussion in Section 4.2.1).



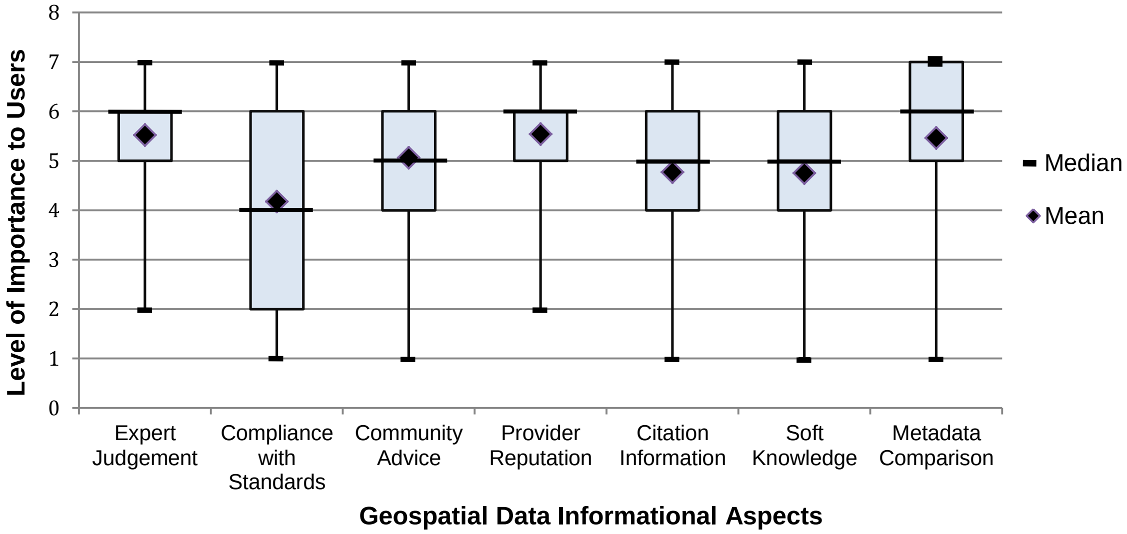


Figure 5.6: Data user attributed importance of informational aspects associated with geospatial datasets shown as bar chart and box plot respectively.

Figure 5.6 shows a summary of the importance ratings attributed by dataset users (who have a choice in dataset for their work activities) to the various informational aspects associated with datasets. With more than 70% of users rating them as important at some level, expert judgment, community advice, dataset provider reputation, and support for metadata comparison seem to be considered universally important across users. In contrast, the importance of compliance with standards and citation information was not generally agreed by users. This is somewhat surprising given that these are the more metric-based or easily validated measures of quality. Although the majority of users rated soft knowledge as important at some level, the agreement on this was not as widespread as the earlier interviews suggested might have been the case.

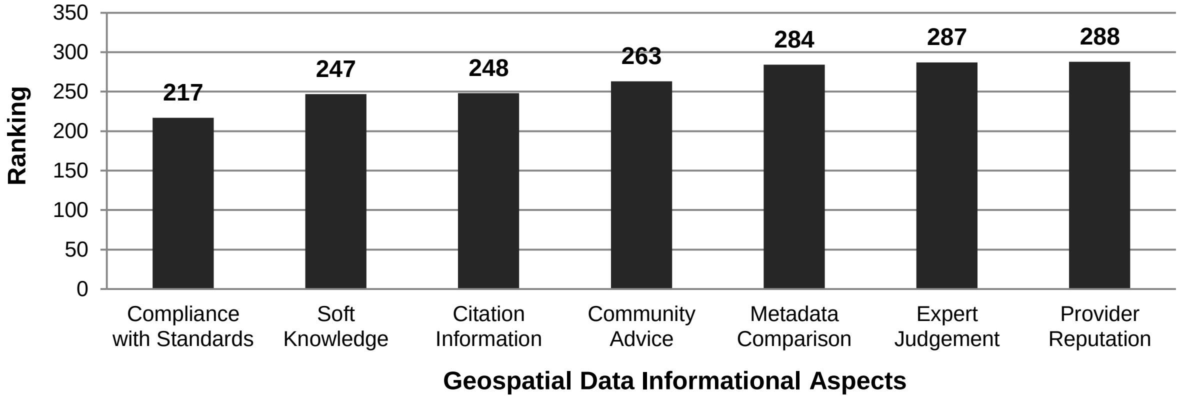


Figure 5.7: Ranking of overall importance of informational aspects according to dataset users.

Figure 5.7 shows an overall ranking of the importance of the 7 informational aspects (as noted previously, the 'dataset ratings' aspect was not included in the assessment of importance when considering what users *currently* rely on due to its as yet unavailability). To calculate this, a ranking was applied (per question) to each Likert scale item ranging from 1 for 'extremely unimportant'/'very useless' to 7 for 'extremely important'/'very useful'; the number of responses for each item was then multiplied by its corresponding rank and the ranks were added to arrive at an overall rank for each informational aspect. As can be seen, dataset users place more importance on the more subjective metadata aspects, with compliance with standards being ranked lowest overall.

5.2.2 Dataset Production

As already noted, 30 study respondents (34%) were self-identified as geospatial data producers. When asked about the data they produce, they listed a range of data types, including: climate data; environmental monitoring and observational data (natural hazards, hydrology, natural resources, etc.); weather forecasts; climate predictions; observations of the atmosphere, oceans, and hydrology; data about species occurrence, status and distribution; topographic maps and databases; informatics products; and socio-economic information.

The majority (93%) of data producers indicated that they provide metadata records with datasets they produce. They listed a range of metadata elements, including: ISO/FGDC/DIF metadata elements (all with translation between formats); all INSPIRE mandatory fields; standard NWS (National Weather Service) reporting; "*metadata for discovery and for assessing provenance, quality, etc.*"; "*station site, instrumentation and measuring parameters and their changes during the period of usage*"; "*Common Data Index, information on instruments, data and methodologies used*"; and "*quality, extent, units, lineage*". One of the data providers stated that "*the amount of detail and support depends on a defined 'level of service'*". Two data providers suggested that they also provide 'soft knowledge' metadata, described by them as taking the form of "*several MS Word and PDF documents describing* *the data, Excel spreadsheets describing the single instruments*", and other "*descriptive documentation*".

When asked about the data and/or metadata standards they support, 87% of data producers stated that they observe international data or metadata standards, listing ISO 19115/19139/19119, FGDC, INSPIRE regulations on metadata, Dublin Core, Darwin Core, WMO, NASA DIF, and netCDF metadata protocol as the main standards that they support. Of the data producers who do not support any international standards, only two provided some explanatory comments stating that they do not yet support metadata standards because "*the need has not arisen yet*".

5.2.3 User and Producer Awareness of Geospatial Data Certifications

When asked about awareness of any certificates or seals that certify geospatial datasets or metadata records, 15 study respondents (17%) indicated awareness of any such certificates. Two study respondents listed ISO 9000 – a family of quality management standards which includes ISO 9001:2008, "*the only standard in the ISO 9000 family which an organisation can become certified to*" (The British Assessment Bureau, 2014). One of the data users specified 5 STAR OPENDATA (5 Star Open Data, 2012) – a 5 star re-usability indicator for Public Sector Information (PSI), ranging from 'very poor' to 'perfect'. One data producer explained a certification framework that is used for the data he produces:

> "*As a US NWS*[2] *field office, we provide certified climate records (F6, CF6) for certain sites to NCDC*[3]*. We also collect/log various COOP*[4] *forms B-91, data tapes, and certain electronic transmissions and forward these to NCDC for archival. In turn, the NCDC at Asheville NC conducts data QA/QC involving such things as data completeness, representativeness, and nearest-neighbour comparisons to these datasets … and publishes these data as 'official' under a form of NOAA*[5]*/NCDC certification*".

Interestingly, 7 study respondents (47% of the respondents who indicated awareness) listed various international data standards as examples of such certificates, including ISO 19100 (series of Geographic Information standards), ISO 19115, INSPIRE metadata implementation rules, OGC (Open Geospatial Consortium), and OWS (OGC Web Services). One of the data users provided a list of several international standards and initiatives, including *"metadata standards like those in use by ESA, NASA etc., e.g. the new GEOMS, statements of compliance with metrology standards like BIPM, GUM,* [the] *GEO QA4EO*

[2] National Weather Service.
[3] National Climatic Data Center
[4] Cooperative Observer Program
[5] National Oceanic and Atmospheric Administration

framework, HDF or NetCDF formalism (compliance mandatory for a list of data archives), [and] *various ISO*". International standards are not, technically, certification programmes but are clearly being viewed by some as serving this role. All 15 study respondents who indicated awareness of certificates or seals agreed that these "certifications", i.e., international standards, are useful to the user community. Four data producers explained why they consider these "certifications" valuable to the GIS community (in some way addressing the subsequent section of the questionnaire focused on respondents' views towards certification programmes for geospatial data or metadata records (see Section 5.2.4):

a) the certifications "*help the user to find the fit-for-purpose datasets*";
b) the certifications "*provide some kind of guarantee and standard procedures of claiming*";
c) "*in our case,* [the certifications] *ensure to the community the quality of our productive processes and existence of documentation about all the process steps, indicators of quality control, etc.*"; and
d) "*Local NWS office or subsequent NCDC publication of observations are routinely used and trusted by partner agencies, industrial users, academics, and others. NCDC 'certified' observations are regularly used in legal cases to establish weather or hydrologic conditions at the scene*".

5.2.4 User and Producer Views on Certification of Geospatial Data

When asked their initial views as to whether geospatial data or metadata records would benefit from the application of certification programme(s), the opinions of users and producers were reasonably concordant: 49% of users and 57% of producers agreed that certification would be beneficial; only 12% of users and 13% of producers disagreed with the benefit of certification programmes; and 39% of users and 30% of producers were not sure. Many respondents agreed that data certification "*would help to improve general quality of data* […], *would also help the user to know the limitations of the data, and if done in a standard, nonbiased manner, would be a useful comparison among similar datasets*". One of the data users pointed out that "*the amount of geospatial data without proper metadata is overwhelming; establishing the quality of data by trial and error takes too much time to benefit from the increasing availability of data*". Respondents argued that certification would "*ensure that a minimum amount of information is captured*", "*it would encourage data providers to follow a common framework of data formats or metadata provision*", and would ensure consistency of data quality. Furthermore, respondents suggested that certification would help to "*identify datasets from authoritative sources and distinguish those datasets from similar data of uncertain quality*". Respondents also confirmed the previous findings on importance of provenance and licensing information, stating that certification could ensure that metadata "*provides a record of the history of the processing, which tends to be neglected if there is a long path towards the final product. It may help with IP issues and similar questions. Some datasets are provided with restrictions on them which can be difficult to sort out after the event. Resolving them becomes a paper chase*".

Although geospatial data certification was viewed positively, some of the study respondents indicated that certification will never outweigh other important data characteristics such as data content, citation information and peers' recommendations. Generally, users appear to view geospatial data certification as a type of formal data quality control (e.g., certification of a dataset's conformance to a defined level of uncertainty, accuracy, resolution, etc., data/metadata interoperability, etc.) and control over metadata completeness; data content, peer review and citations do not appear to be considered as potential certification metrics. A study respondent highlighted that the "*most important quality indicator is actually ratings by users (partly equivalent to peer review)*". Another respondent further argued:

> "*In general it would be good to know whether fundamental metadata is attached to the data in a form that is straightforward to use. This might be helpful when browsing the data, but when it comes down to the choice of what to use for your application I do not see how this would circumvent the process of looking through metadata, searching the literature and talking to other users*".

Consistent with the initial interview results, these findings indicate the value of producer-supplied metadata records as well as more subjective information to support dataset quality evaluation.

In contrast, those respondents who disagreed with certification stated that "*certification is an extra effort* […and] *it seems unlikely asking for more extra effort will improve the current* [situation] *on a broad basis*". Data producers stated that they "*have in-house quality checks and provide information on these* [and the] *procedures are ISO-certified already*". ISO, OGC, IETF, OASIS, W3C, etc. also have their own conformance schemes, consequently "*there are already too many labels/certificate/etcetera in the world, and it is VERY difficult to distinguish the really useful/neutral labels with industry 'self-labelling' things*". The cost of certification also raised a concern because the "*regimented bureaucracy of* [an] *approval process may be costly and difficult for data providers to implement*" and "*if you need too much time to produce this certification, the updating of the data can be compromised*". Also certification could potentially "*reduce the number of datasets made available (due to publishers not wanting to go through the certification process – it is hard enough to encourage filling in of Metadata records)*". This would not be desirable because certification would "*make the data much more expensive and hard to obtain*".

Finally, a number of study respondents did not believe that certification of geospatial data quality is feasible because "*quality is in the eye of the beholder, so that is not certifiable*". It was argued that "*at best you can certify whether the producer followed recommended practice of documentation and stewardship*" and it was suggested that "*data is used for a range of purposes, and data sets not passing the certification process could sometimes be better suitable for* [one's] *purpose than certified ones*".

5.2.5 Initial Perspectives on a GEO Label

When asked about what role they would want a GEO label to serve, respondents provided a variety of comments and suggestions. Some respondents were unsure about, or were not in favour of, a GEO label in general because "*quality is subjective*" and "*different data sets might have different needs*". Others suggested that a GEO label could act as some kind of quality indicator or could "*certify that the data have correct minimum metadata information and are ready to be used*". It was noted that a GEO label could act as an "*incentive for producers to provide metadata*", "*assuring metadata contents are trustworthy, properly filled, and quality control accepted*". It was also suggested that a GEO label "*should be a gold seal of approval by a third party indicating that the data or data sets meet all of the usability, quality and performance criteria*".

A number of respondents proposed that a GEO label "*should give an overview of the data set which can be grasped with one glance on the GEO label*" and provide "*an easy visual indication that data* [has] *met some quality standards,* […] *is complete, useful, etc.*" and would serve as "*an immediate vision that it can be a trusted data source"*. As proposed by study respondents, a GEO label could potentially:

a) indicate compliance with standards such as ISO, Inspire, QA4EO, etc.;
b) provide identification of standard-compliant metadata format;
c) mark that certain Quality Control (QC) procedures are in place;
d) give information about the most relevant metadata;
e) indicate metadata completeness (also containing methodology and uncertainty information);
f) indicate frequency-of-use (data popularity);
g) provide user feedback and ratings;
h) act as an endorsement by relevant experts and user communities;
i) inform users that information on the quality of the data is available;
j) inform users about what the dataset can be useful for;
k) indicate usability levels, certifications, and the main communities using the data;
l) rank the quality of the datasets;
m) describe how the data was created; and

n) provide information on data availability and affordability.

As a caveat to all the proposed functions it might serve, it was stressed that "*the label would only be useful if it is easy to use and understand*".

5.2.6 GEO Label as a Certification Seal

When asked whether the presence of a GEO label would, if provided with datasets to *certify* their quality/trustworthiness, influence respondents' dataset selection decisions, 48% of respondents stated that its presence *would* influence their decision, 22% stated that it would depend on whether they had previously used the data, and 11% stated that a GEO label would not influence their selection. When asked to provide a brief explanation to support their answers, a number of respondents argued that the importance of a GEO label would largely depend on the *availability* of datasets that fit the task at hand. Data users stated that at present there is not much choice of datasets that fit their needs and they are "*often lucky to find the data needed*" at all. It was also highlighted that "*a lot of the data* [that users] *deal with is only available from one source, as such there is often limited choice made when collecting it*". A data user suggested that, if faced with no alternative, he "*would use the data although the quality info is not provided*". Another data user noted that "*if there* [is] *more than one similar dataset to choose from, and* [he is] *not personally aware of the differences, then* [having a GEO label] *would be useful*" in supporting a selection decision.

Other study respondents argued that a GEO label would influence dataset selection decisions because users "*would hopefully trust the ones who certified the dataset"* and because they felt that such "*data sets would become de facto standards*". A data producer argued:

> "*such a label could assist in selecting one of* [a] *few best options from a large range of possibly very different data sets. If/where this happens, the label might also develop into an incentive for producers to implement minimum standards (metadata, QC)*".

Although respondents generally agreed that they would prioritise certified data, they also noted that they would also spend time locating non-certified data from trusted sources. It was further suggested that users "*are looking first for good and reliable data sources*" and "*would have confidence in using data from trustworthy sources*" – implying that the source rather than the certification was significant. In line with the preparatory findings reported in the previous chapter, these results indicate that the reputation of data providers plays an important role in data quality evaluation and is often viewed as a 'quality seal' in itself. Also in support of the interview findings, many study respondents suggested that even if provided with assurance seals for a dataset, users would still want to visualise data within the dataset and talk to or read evaluations from other users of the data.

Respondents indicated that, in cases where users are already familiar with a dataset or source, the presence of a GEO label would not affect their perception of the dataset quality. In contrast, for those users who have never encountered a dataset, the *“GEO Label* [has a potential to] *represent a warranty of the data*”.

Finally, a number of study respondents stated that geospatial data cannot be usefully certified because its quality and trustworthiness depends entirely on the intended use. It was suggested that “*a dataset that suits one person's needs may be completely inappropriate for someone else”* and it was felt that *“this would be nearly impossible to capture*”. Respondents argued that they are more concerned with whether a dataset is suitable for their particular requirements than whether or not it has a seal of approval; they indicated that they would rather check the data themselves than trust the assessment to someone else. The issue of certification costs was again raised, with respondents wondering “*who will pay for the sustainable maintenance of any accreditation scheme? The banks?*”.

User Viewpoint:

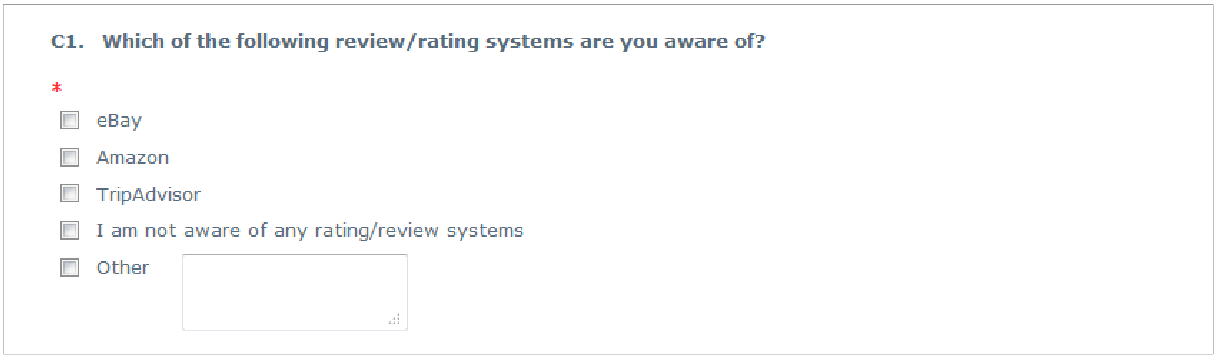
“*The problem is that you cannot certify quality and trustworthiness if you do not know about the intended use of the data. I could see this would be helpful if the user's profile matches a typical user of the data – but there are many cases where people (successfully) use data for purposes other than what was intended. I do not see how the certification would help here.*”

Producer Viewpoint:

“*So how are you going to measure this? What makes something quality data? Why would I trust it? What value does this GEO label provide unless the process is totally transparent, as least as rigorous as peer-review of papers, and open. How likely is any of that!*”

5.2.7 GEO Label as a Review or Rating System

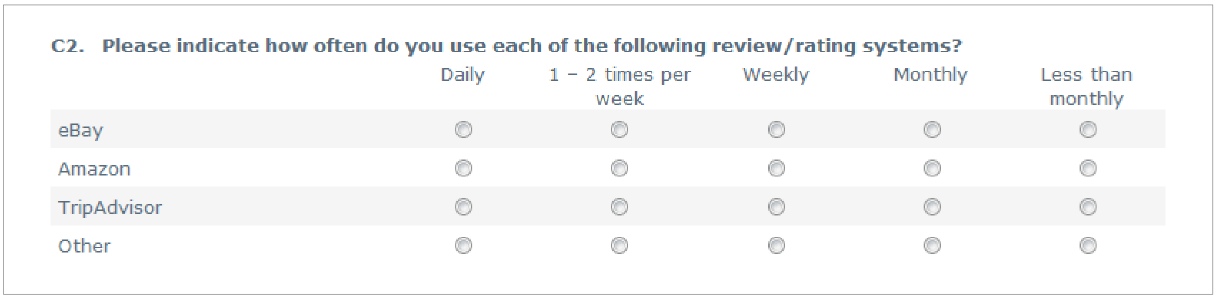
To first observe respondents’ general awareness of review/rating systems, they were asked to identify everyday e-Commerce-type review/rating systems with which they were familiar (see Figure 5.8 and Figure 5.9; also see Appendix B, Section C). Surprisingly, 23% of respondents indicated that they are not aware of any such systems; the remaining 77% indicated familiarity with eBay, Amazon, TripAdvisor, hotels.com, booking.com, and even Google maps search results. Of these, the majority of respondents stated that they use such systems less than once per month.


C1. Which of the following review/rating systems are you aware of?

*

- [ ] eBay
- [ ] Amazon
- [ ] TripAdvisor
- [ ] I am not aware of any rating/review systems
- [ ] Other

Figure 5.8: Example of questions designed to elicit general awareness of review and rating systems.


C2. Please indicate how often do you use each of the following review/rating systems?

	Daily	1 – 2 times per week	Weekly	Monthly	Less than monthly
eBay	○	○	○	○	○
Amazon	○	○	○	○	○
TripAdvisor	○	○	○	○	○
Other	○	○	○	○	○

Figure 5.9: Example question to assess respondents' frequency of use of review and rating systems.

When presented with a screenshot of an eBay product listing page incorporating a Top-rated seller label (see Figure 5.10; also see Appendix B, Section C), respondents showed a positive attitude towards its presence: 63% agreed that they would be encouraged to trust the vendor more than they would a vendor without the label and its associated meaning; 63% agreed that they would be more likely to purchase a product from a vendor that carries the Top-rated seller label; 56% agreed that a vendor that carries the Top-rated seller label has a good reputation; and 47% agreed that a vendor that carries the Top-rated seller label is more reliable than vendors that do not carry such a label. In contrast, the Top-rated seller label was not generally viewed as an indicator of product quality nor was it seen to carry assurance that the vendor has the consumers' best interests at heart. Reaction to a screenshot of TripAdvisor hotel reviews and ratings (see Figure 5.11; also see Appendix B, Section C) followed a similar pattern.

Aston University

Illustration removed for copyright restrictions

Figure 5.10: A screenshot of an eBay product listing page incorporating a Top-rated seller label used to elicit awareness and reaction to the labelling.

Aston University

Illustration removed for copyright restrictions

Figure 5.11: A screenshot of a TripAdvisor hotel listing page incorporating reviews and ratings used to elicit awareness and reaction to the rating presentation.

After presenting respondents with these examples, respondents were asked to state whether they believe a GEO label should support a similar role. Nearly half of the respondents (49%) agreed that a GEO label should fulfil an assurance and review/rating function; 34% recorded no opinion; and 16% disagreed. When asked to provide a brief explanation to support their answer, a number of respondents agreed that user ratings systems are very useful, they provide more insight into the data, and *"while always subjective in nature,* [are] *a valuable tool for deciding on data relevance and utility*". Respondents observed that datasets recommended by many users with similar needs could be seen as more trustworthy and relevant. Furthermore, it was also suggested that introducing a review and rating system for GEO datasets would add *"value to the community of users and give GEO an identity and make a real contribution to the understanding of the planet Earth*".

Producer Viewpoint:

"*I think it strongly depends on who is providing the label and how it's achieved. I'm highly suspicious of seller ratings as they can be 'farmed', but will be more likely to trust a large number of positive user/independent reviews (a small number will make me distrust the reviewers as family/friends). A user based review system is therefore more appealing to me than a certification programme.*"

Although study respondents agreed that labels and user ratings might initially attract their attention, they stressed the importance of actual comments as opposed to simple star-based ratings. They suggested that they would consider free text user comments to ultimately arrive at their own, independent "*judgement about the usefulness of the rating*". One data user suggested that user feedback and review functionality would actually be more valuable and relevant with respect to geospatial data than any form of label similar to the 'Top rated seller'. He argued that it is the products (geospatial datasets) that must be reviewed and not the dataset providers because "*the data provision procedure (IT aspects, etc.) is certainly important but secondary with respect to the information/data*" itself. A data producer noted that, if dataset ratings and user feedback are to be provided as part of the GEO label function, these subjective quality measures should not be combined with objective quality aspects.

Producer Viewpoint:

"*I'd be more interested in actual user comments than a simple star rating. Given the wide diversity of potential data uses (unlike things reviewed on Amazon or trip advisor) a star rating is not very meaningful. A five star data set for an atmospheric modeller could be a one star data set for a terrestrial ecologist.*"

In contrast, many respondents appeared to be unsure whether rating and review systems could be usefully applied to geospatial data. It was argued that "*rating systems are too one-dimensional for scientific data*", that they are "*too subjective*", and that e-Commerce-led review and rating systems are not viable comparable examples because "*scientific datasets are more complex than a hotel and their quality is much more difficult to assess*". Respondents once again highlighted the use case-dependent nature of geospatial data quality, arguing that "*data sets serve different goals so for one purpose a dataset may be very good while for another it may be very poor, so then you cannot attach a single label, it all depends on the intended purpose*". A data user argued that "*just because someone 'highly rated' a dataset does not necessarily mean that it is of good quality"*; he also aptly commented that "*there are many aspects that one should take in to account regarding the user background, which would partially help in determining how the user perceived and rated* [the dataset]". Some respondents proposed that "*the label should be limited to reflecting that the data are documented with a specific level of metadata…* [and] *independently, the GCI should pursue hosting a ratings system to solicit feedback from users on each data set, acting as intermediary between user and provider, and sharing experiences of users*".

User Viewpoint:

"*I think it could support a reviewing system. But using tripAdvisor/booking.com as* [an] *example, the rating is usually relative to the expectation of the hotels. I would expect a 5 star hotel with an average rating of 6 still to be better for many things than a 1 star hotel with rating 8, as you pay a lot more. Rating of data sets would similarly depend on the expectation of the data set. But would a high resolution data set be better than a low resolution data set, even if the first has been downscaled from the second one with a simple and naive method? Many users would say yes, but the high resolution might give a false indication of precision. The second question is whether users would be able to recognize a good data set. I can see if a hotel room is dirty and the receptionist is impolite, but I might need to do my own cross-validation to be able to correctly rate a data set*".

Several study respondents indicated strong opinion that a GEO label should be a certification seal and not a review/rating system. They argued that the GEO label "*should be formally assigned by a data provider that certifies this dataset has the backing of the government agency that supplied it*" and should not include user reviews and ratings. As a data user stated, “*the GEO label should in my opinion be a certification of the completeness of the metadata, not the popularity among the users*”. Another data user also argued that a GEO label should not fulfil a general user review function, but should instead support “*expert evaluation - not rating by the masses*”.

The limited availability of suitable datasets was again referenced when a number of respondents argued that, unlike hotels or other electronically retailed products for which there is normally a wide choice of alternative products/vendors, there is often little choice when it comes to GEO datasets and so criteria other than ratings become more important when selecting a dataset to use. Indeed, respondents indicated that they would often have to use data of low quality in cases when there is no better choice; in such cases, user feedback and evaluation would not have any impact on the dataset selection decision.

5.2.8 GEO Label for Drill-Down Interrogation

Prompted by some examples (see Figure 5.12), respondents were asked about their awareness of e-Commerce seals and associated click-to-verify functionality.

Aston University

Illustration removed for copyright restrictions

Figure 5.12: Example e-Commerce seals presented to study respondents.

Three quarters (75%) of respondents stated that they had come across e-Commerce seals before; of these, 46% had clicked on an approval seal to verify if it was genuine, stating that that they “*clicked to check that they were legitimate and how they proved that they were legitimate*”. In general, respondents explained that, in accessing the seal’s webpage, they were trying to reassure themselves about the security of the transaction, especially in terms of credit card use. A few, however, clicked “*to check what criteria are used*”, “*to see briefly what it was, to understand it better*”, or “*mostly out of curiosity*”.

Three study respondents, who had never clicked on an approval seal, suggested that they “*only used trusted sources* [that they] *knew from previous experience*” and acknowledged that most of the time, they purchase products on the basis of already-established trust in a brand or producer, meaning that they largely only see the seals on the pages of vendors they already trust. Rather candidly, four respondents admitted that they never clicked on a seal as a result of naivety because they had never “*had a negative experience with fake seals*”; they commented that “*seeing the approval seal was enough*” and that they have generally taken them at face value. Other reasons for not investigating approval seals included a lack of belief in such systems, lack of consideration (i.e., the seals were not considered important or, worse, weren’t even thought about at all), and lack of time. Finally, four respondents stated that they were not previously aware of the click-to-verify functionality, with one responding indicating that he would consider this more closely from now on!

When asked whether a GEO label should fulfil a drill-down function and have click-to-verify functionality, more than half of respondents (54%) agreed that a GEO label should fulfil such a function. Respondents were then asked to rate their perception of the appropriateness of different candidate data interrogation functions/foci that could be provided within a drill-down GEO label. Their responses are shown in Figure 5.13.

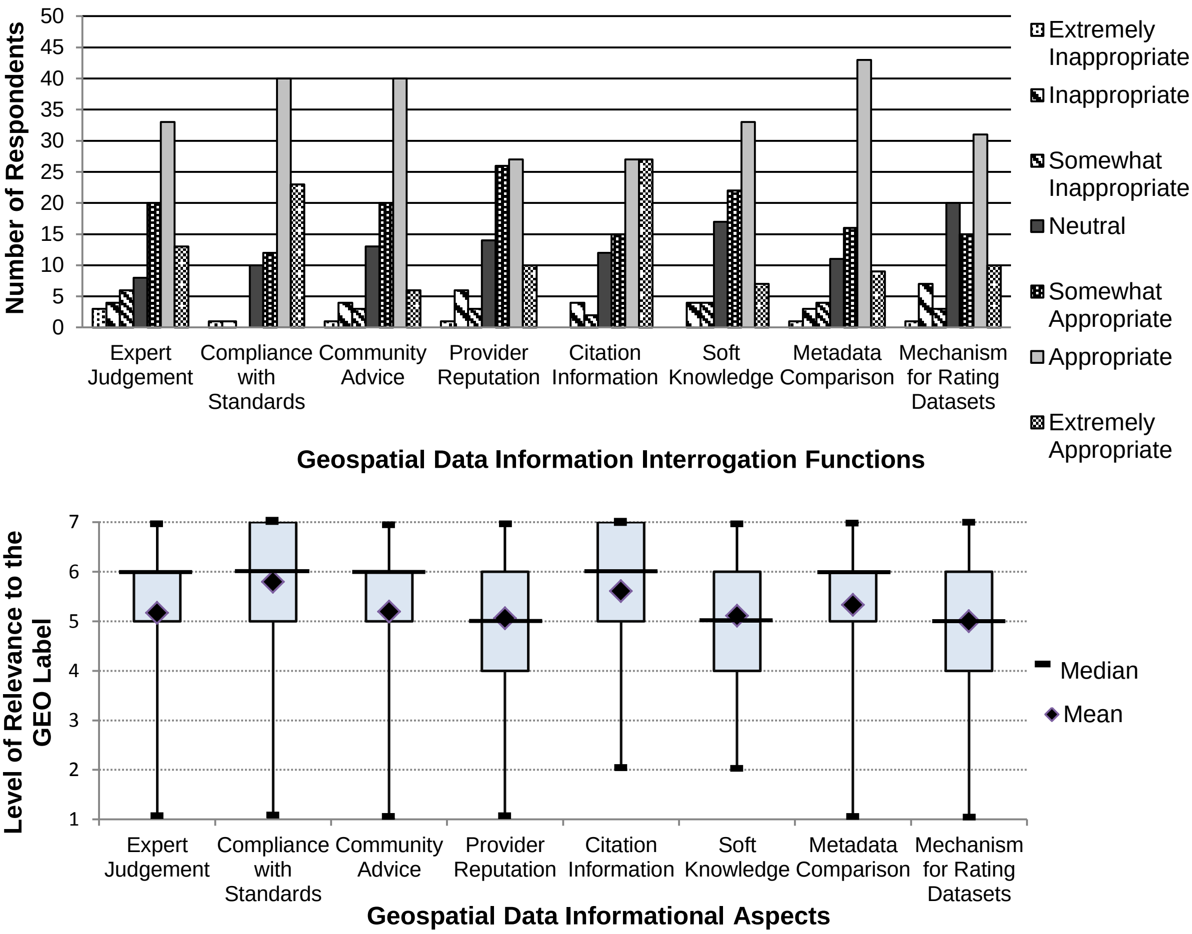


Figure 5.13: Attributed appropriateness of candidate information interrogation functions/foci within a GEO label shown as bar and box plots.

The Spearman's Rank Correlation Coefficient was used to discover the strength of correlation between users’ *personal* perceptions of the importance of geospatial data informational aspects (when personally selecting datasets as discussed previously) and their assessment of the *general* appropriateness of inclusion of associated informational interrogation functionality within a GEO label drill-down function. (It should be noted that only dataset users' responses were included here because only this group was asked to rate importance of informational aspects for dataset selection.) Data analysis showed significant correlation for citation information ($r(50)=0.44$, $p=0.001$), soft knowledge ($r(50)=0.664$, $p=0.001$), and metadata comparison ($r(50)=0.509$, $p=0.001$). Compliance with standards ($r(50)=0.37$, $p=0.005$), community advice ($r(50)=0.353$, $p=0.01$), and expert judgement ($r(50)=0.323$, $p=0.025$) showed much weaker correlation. Finally, no significant correlation ($r(50)=0.226$, $p=0.10$) was discovered between the perceived importance of reputation of dataset providers and its value within a GEO label function. On this basis, some level of variance can be seen between what users rely on personally when selecting datasets and their professional opinion in terms of what should be included within a GEO label drill-down function for dataset representation. Whereas in section 5.2.1 there was a clear emphasis on more subjective metadata, we are now perhaps witnessing some element of concern about whether it is *possible* to effectively or appropriately support such metadata in an interrogation facility.

Respondents were asked to identify any additional functions/foci that they felt could usefully be provided within a drill-down GEO label. Their suggestions included:

a) indication of level of interest in the data set – a popularity indicator (e.g., how many users acquired and used the dataset);
b) information on the purpose of creation of the dataset;
c) example uses of the dataset;
d) complete provenance (lineage) information, if not provided in the metadata;
e) full processing history;
f) information on stability of the set over the years;
g) links to other relevant or similar data (e.g., data gathered as part of the same campaign, same geographic area, same time);
h) information on known gaps or shortcomings of the data;
i) automated checks (syntax, schema or structure, existence of the data link, etc.);
j) access to provider responses to community members' comments about a dataset;
k) discussion forums about the different datasets and/or providers; and
l) knowledge and information on licenses (free and open) – that is, a clear description of the ownership and use that can be made of the data, including commercial and quality (likely errors) restrictions.

Study respondents cautioned that a GEO label should supply information regarding "*who awarded the label* [and] *which criteria were used*". A data provider highlighted the importance of 'soft knowledge' and provided US National Weather Service COOP B-91 reports as a practical example: he commented that these reports had previously included tremendously useful observer remarks which have, unfortunately, "*since vanished* […] *due to automation or ha*[ve] *been erased from the historical record due to digitization*".

5.2.9 Preference for GEO Label Role

Respondents were asked to identify their preference(s) for one or more of the following as the role a GEO label should serve:

a) certification seal;
b) drill down interrogation facility;
c) other; and/or
d) don't know.

The majority of respondents (57%) indicated preference for a drill-down interrogation facility, with a large number additionally and/or alternatively stating preference for a certification seal. Where respondents suggested alternative functionality, they proposed that a GEO label represent a popularity indicator, user feedback facility, or a metadata completeness seal that would include a template of metadata elements that are required to be filled in. A data user argued that "*there are common metadata fields that should be included in EVERY data set, that are missing in part in almost every dataset available*" and that a standardised template would solve this issue.

When asked to provide a brief explanation in support of their stated preference(s), the respondents who selected only the certification role argued that certification would be "*very useful to verify data quality*" and could provide "*a quick way of verifying the usability of the dataset*". It was also suggested that because "*seals are simple and easier to notice"* they would therefore serve as effective *"means of noting quality of products to other users whether professional or non-professional*".

Respondents whose preference was solely the drill-down function argued that such a facility might be more flexible and complete compared to certification programs which "*usually apply to the existence of documentation accompanying the data, not to the quality of this documentation*". Respondents suggested that providing more information about datasets in this way would be the *"best way to provide the information, whether it is metadata from the data provider or comments/evaluations by users*". Respondents argued that dataset users should be able to arrive at their own assessment of data quality and, since quality is dependent on personal perception and intended use, they considered a drill-down function as the most convenient way to support such quality assessment. They reiterated that it may not be possible to determine quality without knowing the intended context of use and so a drill-down function to allow interrogation of dataset-related information would be more meaningful as opposed to some abstract assertion of quality. One data producer considered that information from providers and user reviews available via a drill-down facility were "*perhaps the most useful way to provide more information on a dataset*", adding that citations would also serve as "*a useful indication of what the data are suitable for*" but expressing concern about expert judgements and certification, based on who is purporting to be expert and on what basis the quality is being certified.

Respondents who were equally in favour of both certification and drill-down roles argued that, combined together, both functions would provide users with the information necessary to establish confidence or trust in a dataset, especially when a user is faced with an unfamiliar data platform. They highlighted that "*a certificate implies well-defined and independent controls*" and "*can be useful to state that certain minimum requirements were followed*" whilst "*the drill down interrogation facility with rating could provide the user with lots of additional information*" and potentially support *"online comparison of essential items*". It was further suggested that the presence of a certification seal might encourage users to select those datasets for inspection and, thereafter, a drill-down functionality would allow users to assess whether the data "*is worth working with*".

In contrast, a number of respondents were "*not sure how to integrate all these aspects in a simple form*" arguing that "*the one-label-for-all-solution might be a bit optimistic and clutter everything more than necessary*".

Finally, respondents who did not indicate any preference expressed doubt about the fundamental usefulness of the GEO label (regardless of form and function) or did not understand the concept of 'drill-down functionality'.

When asked whether a GEO label should combine multiple functions (e.g., data ratings, reviews, quality assurance, etc.) and thereby represent an all-in-one quality indicator, the majority of respondents (66%) agreed, often on the basis that a single label of this nature would support initial decision-making and that it would circumvent the problem of having too many labels to deal with. One respondent considered this option for a GEO label as being an *"easily consultable*" option which would provide dataset users *"with a one-stop source of quality assurance and reliability*". Respondents who thought an all-in-one function was inappropriate argued that it would be impossible to provide all the information in one place, with many of them commenting that expert judgement, community advice, data ratings, quality assurance, etc., are different concepts and, as such, cannot be combined into one function and represented via a single label.

5.3 Study Limitations

Although this study has returned rich information about a previously unexplored concept, it is recognised that it is not without its limitations. As with the initial investigation, the majority of study respondents were research-based data users or producers, and only 34% of all respondents represented the geospatial data producer group. Although every effort was made to target a variety of user and producer groups, researchers and data users were more willing to participate in the study. That said, while it would be valuable to elicit broader views of geospatial data producers, geospatial data users are the primary consumers of geospatial data and metadata records with unique knowledge of geospatial data selection requirements. Comments and feedback from producer respondents provided a sufficient understanding of producer perspectives on usefulness and feasibility of the potential GEO label functions and their impact on geospatial dataset production and availability. The fact that the majority of user respondents were from the research community allowed acquiring deeper insight into issues of geospatial data quality and quality evaluation; benefits and weaknesses of existing international standards; feasibility and potential effectiveness of certification programmes for geospatial data and metadata; and mechanisms that can be adopted to support geospatial data quality evaluation. Research data users were not only able to express their own views and opinions but also discuss issues that affect geospatial data community as a whole. Overall, it is therefore felt that the results elicited in this study represent a reliable overview of geospatial domain stakeholders' opinions.

Another *potential* limitation of the study is that users and producers were considered as two distinct groups. Although done to allow comparison of attitudes across these geospatial data roles, some of the respondents stated that they had to make a difficult choice between identifying themselves as a 'user' or 'producer' because they consider themselves as both. This might be one of the reasons why there was no obvious, strong distinction between the responses from these two groups. It would be interesting to elicit views of 'pure' data producers (if such a group exist) to determine if there is greater polarity in opinion between producers and users than this study would suggest is the case; that said, many professionals working in the GIS domain do assume multiple roles (consuming and producing datasets) and so it is anticipated that the results do reflect target users.

It could be argued that the questionnaire should have provided respondents with concrete examples of how a proposed GEO label would work. This approach was, however, intentionally *avoided* in favour of use of more abstract questions to prevent leading respondents and thereby to encourage more freedom of thought. That said, it is acknowledged that some respondents may have been somewhat restricted in response by their own imagination and/or prior experience. For instance, from some of the respondents' comments, it would appear that they *assumed* that the intention was to combine expert judgement, community advice and 'soft knowledge' into one aggregated "sign" – a concept about which they held a negative opinion and which was subsequently reflected their negative attitude towards an 'all-in-one' GEO label function. This clearly reinforced the need to, in line with the overall research strategy (see Chapter 1) adopted by this project, base further investigation into the concept of a GEO label on the use of actual GEO label *prototypes*; instead, however, of doing this based on unfounded assumptions, the results of this phase of the research were used to establish user-led prototypes for the further investigation (see Chapter 6), hence placing stakeholders and their opinions at the centre of the investigation at all stages.

5.4 Summary and Conclusions

Overall, the results of this study show that users and producers of geospatial datasets appear to have generally very positive attitudes towards the development and introduction of a GEO label. Interestingly, there was no distinct difference between user and producer views on the role that a GEO label should serve; it is hoped that this is because both groups equally recognise the challenges of assessing quality of geospatial data and that producers may be willing to place themselves under greater scrutiny to better support informed dataset selection amongst their clients, i.e., users. The study illustrates that geospatial dataset users rely heavily on metadata records when assessing dataset fitness for use, and reiterates the problems associated with the current lack of uniform availability of quality-associated information despite ongoing standardisation efforts. For these reasons, many respondents agreed that a GEO label could potentially fulfil a certification or assurance seal function and be used to impose higher standards on provision of metadata records.

Respondents demonstrated positive attitudes towards the concept of a GEO label that provides some sort of rating and review facilities, seeing this as appropriate support for more subjective metadata recording and assessment for datasets. **The majority of users and producers strongly supported the notion of a GEO label providing an all-in-one drill-down interrogation facility that would combine expert value judgements, community advice, links to citation information, side-by-side visualisation of metadata records, etc.** This suggests that, in order to make an informed dataset selection decision, respondents require and would appreciate as much information as possible presented in one place in a format that allows for easy comparison.

The results of this study confirmed the importance of all 8 of the facets presented to study respondents, namely:

- dataset producer information;
- producer comments on the dataset quality;
- dataset's compliance with international standards;

- user feedback (community advice);
- dataset ratings;
- expert value judgments;
- links to dataset citations; and
- side-by-side metadata records comparison.

On further reflection, however, since side-by-side metadata visualisation would require at least two datasets and does not represent an informational facet of a single dataset alone, it was decided not to include this function in the GEO label visualisation itself. Instead, this level of metadata comparison functionality could and should be provided as a separate facility when the GEO label is developed and deployed in the GEOSS (or equivalent).

As part its work dedicated to the development of quantitative measures of dataset quality GeoViQua has defined an extended version of a standard ISO 19115:2003 compliant metadata record to include quantitative quality information (<gvq:dataQualityInfo> metadata element). This, together with the fact that both the preparatory interviewees (see section 4.2.9) and respondents to this phase of study indicated that both objective (quantitative) *and* subjective quality information was important in the assessment of dataset fitness-for-use, indicates the value of inclusion of an additional facet – 'dataset quality information' – in a GEO label.

Based on the study findings, three GEO label examples were developed (see Figure 5.17, Figure 5.18 and Figure 5.19). These were prototypic graphic representations (i.e., static images) which could potentially be used to convey availability of spatial dataset quality information. These GEO label visualisations combined the 8 identified and confirmed informational aspects, namely:

a) dataset producer information;
b) producer comments on the dataset quality;
c) dataset's compliance with international standards;
d) user feedback (community advice);
e) user ratings of the dataset;
f) expert reviews (expert value judgments);
g) dataset citations; and
h) quantitative dataset quality information.

Each informational facet was designed to show whether the information it represents is '*available*', '*not available*' or '*only available at a higher level*' for the dataset with which the GEO label is associated. The design of prototype example 1 (see Figure 5.17) was influenced by the GEOSS Societal Benefits logo (blue circles + white icons) (see Figure 5.14). Colour variations to convey information availability were utilised as follows: blue background with white icon represented 'information is available' state; white background with icon outline represented 'information is not available' state; and white background with blue icon represented 'information is available only at a higher level' state. Potential graphical representations of each informational aspect were explored to identify the icons and symbols that are already successfully used in other domains. A 'writing hand' icon was selected to represent the *producer profile* informational aspect. While there is no commonly accepted icon to represent such concept, it was felt that this graphic could successfully symbolise data and metadata creation. To support facet recognition via visual similarity with the *producer profile* facet, a 'writing hand in a speech bubble' icon was designed to represent the *producer comments* facet. The *compliance with standards* facet incorporated an 'ISO+' icon, with the plus sign indicating a variety of standards. Taking into consideration that the ISO standards are widely known within the GIS community, it was anticipated that users will intuitively associate this icon with standards. The 'speech bubble' and 'star' icons are commonly accepted in e-Commerce to represent user feedback and ratings respectively; consequently, these icons were adopted for *user comments* and *user ratings* facets. A 'magnifying glass with a sheet of paper' icon was designed to represent the *expert review* facet; although the magnifying glass symbol is commonly used to indicate search facilities, it is also often employed to symbolise *review*. An 'open book' icon was designed to represent the *citations information* facet. Since a book symbol is commonly used to indicate bibliography, it was anticipated that users would be able to associate this symbol with citations. Finally, a 'Q' icon was used to represent the *quality information* facet. While a 'quality seal' icon could be utilised for this facet, such symbology could potentially be interpreted as certification and mislead the label users. Short description labels were assigned to each facet to enable better recognition and recall of the informational aspects presented in the label.

The design of prototype example 2 (see Figure 5.18) was influenced by TripAdvisor, Amazon and eBay review and rating labels (see Figure 5.15 for an example) that are widely known and recognised in e-Commerce. This prototype example communicated the quality information using three different techniques: six graphical representations were used to convey availability of *producer profile*, *producer comments*, *quality information*, *expert reviews*, *citations information* and *user feedback*; 5 stars were used to communicate the average *user rating*; and name of the standard to which the dataset complies was used to represent the *compliance with standards* informational aspect.

Finally, the design of prototype example 3 (see Figure 5.19) was *based* on and adapted from an early proposal by the ST-09-02 in their Draft GEO Label Concept paper (ST-09-02, 2010) (see Figure 5.16). This prototype example was designed as a multifaceted star and did not include any iconic representations. Textual labels and star arm locations were employed to convey which informational aspect the star arm represents. Star arm variations to convey information availability were used as follows: gray star arm with black outline represented 'information is available' state; transparent star arm represented 'information is not available state; and transparent star arm with black arrow represented 'information is available only at a higher level' state.

Aston University

Illustration removed for copyright restrictions

Figure 5.14: The GEOSS Societal Benefit Areas logo (GEO, 2014c).

Figure 5.15: An example TripAdvisor review.

Aston University

Illustration removed for copyright restrictions

Figure 5.16: Preliminary GEO label design proposed by the ST-09-02 Task team (Plag, 2012, p. 3).

Aston University

Illustration removed for copyright restrictions

Figure 5.17: GEO label example 1. **Figure 5.18: GEO label example 2.** **Figure 5.19: GEO label example 3.**

To evaluate the comparative effectiveness of, and user preference for, the three prototype visualisations, a third user study was conducted, also administered as an online questionnaire. The study and its results are presented in the following chapter.

Chapter 6 Phase II: Soliciting User and Expert Views on GEO Label Prototypes

This chapter presents studies that were conducted as part of the second main phase of this GEO label research. A questionnaire-based study was designed to solicit geospatial data producers' and users' views on the proposed GEO label visualisations. The main aim of this study was to evaluate the effectiveness of, or potential issues with, the proposed GEO label designs and to arrive (if possible) at a final, community-supported GEO label representation. Due to the fact that the overall results of the questionnaire-based study did not show strong preference for any of the proposed GEO label visualisations, the GEO label designs were modified, adapted and improved in line with geospatial experts' feedback and recommendations. To arrive at a definitive user-accepted GEO label visualisation specification, the geospatial community was further polled regarding the final GEO label design.

Section 6.1 of this chapter outlines the structure and the distribution process of the questionnaire-based study conducted as part of this phase of the research. The results of the questionnaire-based study and study summary are presented in sections 6.2 and 6.3, respectively. Section 6.4 then describes evolution of the GEO label visualisations based on experts' feedback, recommendations and community voting for the final GEO label design. Section 6.5 provides a final and definitive specification for the GEO label visualisation. Finally, study limitations and conclusions are presented in sections 6.6 and 6.7, respectively.

6.1 Soliciting Opinion on the Geo Label Prototype Designs

A second online questionnaire-based study (see Appendix D) was again administered to geospatial data users and producers. This questionnaire comprised six sections, A to F, each of which is described below.

Section A consisted of a small number of questions to gather background information about the respondents and their requirements when selecting geospatial data to use. Here respondents were first asked to identify themselves as one of the following: primarily data users; primarily data producers; or equally data users and producers. The respondents were then asked: to pick one or more statements from a set which best describes their dataset user or producer type; to identify the type of organisation they work for; to indicate how long have they been working with geospatial data; and to approximate the percentage of their time they spend working directly with geospatial data. In this section, the respondents were also asked whether they have a choice of dataset to use, what clearinghouses they use, if any, and whether they find dataset selection a challenging task.

Section B was designed to explore respondents' levels of understanding of the proposed GEO label facets (informational elements) and the information the facets convey about the datasets they represent. This section first presented a short description of: how quality information can affect dataset selection decisions; what informational aspects may be considered when evaluating a dataset's quality; at a high level, the proposed GEO label and what it represents; and how the GEO label could be used to support dataset selection. After the introduction, the icons representing the eight GEO label information facets (see Figure 6.1) were presented to the respondents who were asked to describe what geospatial dataset informational aspect they believed each icon represented (see Figure 6.2 for an example; also see Appendix D, Section B).

Figure 6.1: GEO label informational facets presented to study respondents.

The intention was to assess respondents' unguided initial interpretation of the meaning of each icon and thereby identify the effectiveness and intuitiveness of the graphical representations. After soliciting respondents' *initial* and unguided opinions as to what they believed the icons represented, the actual intended meaning of each icon was explained (see Figure 6.3 for an example; also see Appendix D, Section B). Armed with this information, the respondents were then asked to rate the intuitiveness of each icon and to provide any comments or suggestions on each icon's representation and indicate possible improvements.

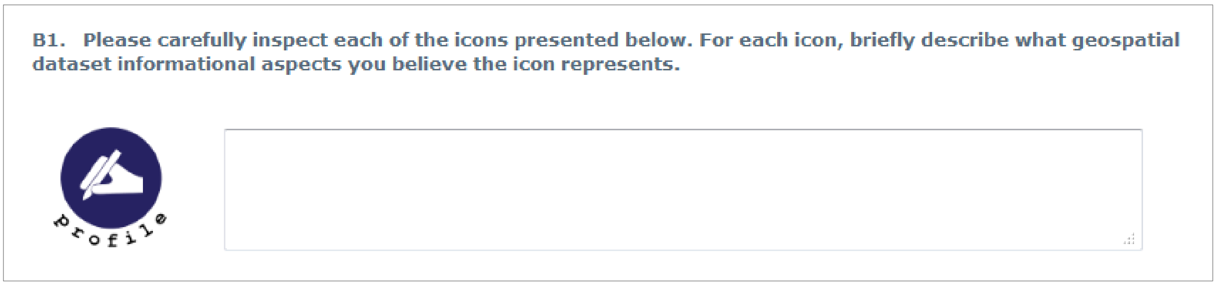


Figure 6.2: Question B1 (see Appendix D) asking respondents to identify the meaning of the icon.

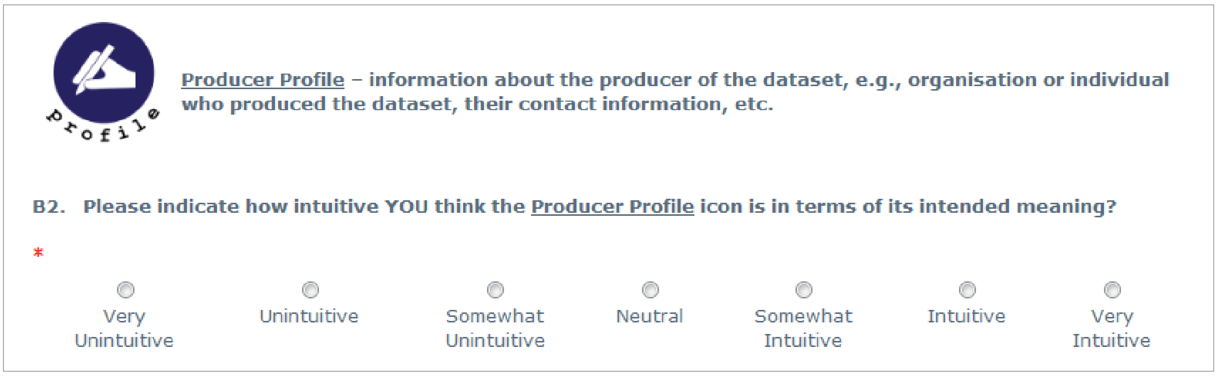


Figure 6.3: Question B2 (see Appendix D) providing the intended meaning of the icon and asking respondents to grade its intuitiveness relative to that meaning.

Once the intuitiveness of the meaning of the icons themselves had been assessed, the section then evaluated the effectiveness of the variations for each icon in terms of conveying the *availability* of quality information for a given dataset. As before, the availability variations for each icon were first presented to the respondents without explaining their intended meaning and respondents were asked to interpret their meaning without guidance (see Figure 6.4 for an example; also see Appendix D, Section B). The intended availability meaning was then indicated (see Figure 6.5 for an example; also see Appendix D, Section B), and respondents were asked to consider how intuitive they thought each availability representation was. Respondents were also welcomed to provide any comments and suggestions regarding the icons' variations and their perceived meaning.

Section C was focused on gauging respondents' opinions on the effectiveness of the proposed GEO label prototypes – that is, the complete labels rather than the component parts (as studied up until this point) – at conveying availability of dataset quality information. This section presented three GEO label designs, two of which (examples 1 and 2 in Figure 5.17 and Figure 5.18) used collective visualisations of the individual icons examined in section B, and one (example 3 in Figure 5.19) used an alternative visualisation without iconic representation of the facets (drawing on ideas from the skeleton ST-09-02 proposal). The prototype designs were presented to respondents separately. Each GEO label example was first presented together with a series of both true and false statements about quality information availability for a fictitious dataset that the label was supposed to represent; respondents were asked to indicate those availability statements that *they* believed were correct (see Figure 6.6, Figure 6.7 and Figure 6.8; also see Appendix D, Section C). This exercise was designed to assess the ease of interpretation or intuitiveness of the proposed visualisations as well as ensure that participants were familiar with each GEO label representation before proceeding to more demanding dataset ranking scenarios (as described later).

Following this, each GEO label example was examined within the context of a dataset selection scenario: respondents were presented with a scenario identifying key dataset qualities that were flagged as important for a given selection decision. They were then provided with 5 different mock-ups (based on the given GEO label design) that conveyed availability of different quality information for 5 fictitious datasets, and were asked to examine the mock-up GEO labels and rank them (as representatives of their underlying datasets) in order of fitness-for-use based on a match between the specified scenario and the information the labels conveyed (see Figure 6.9 for an example; also see Appendix D, Section C). After each ranking scenario, respondents were asked to indicate how effective the GEO label designs were at supporting, and how easy it was to use them for, dataset selection (see Figure 6.10). At the end of the section, respondents were asked to rank the three different GEO label prototypes in order of preference and to provide some justification regarding their indicated preferences (see Figure 6.11).

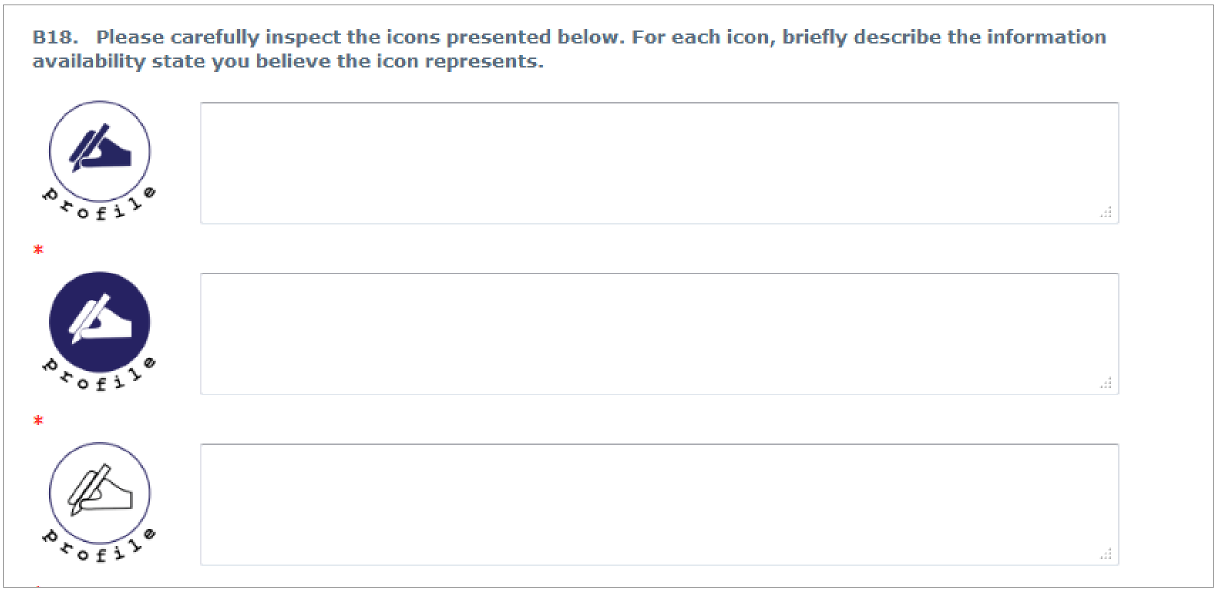
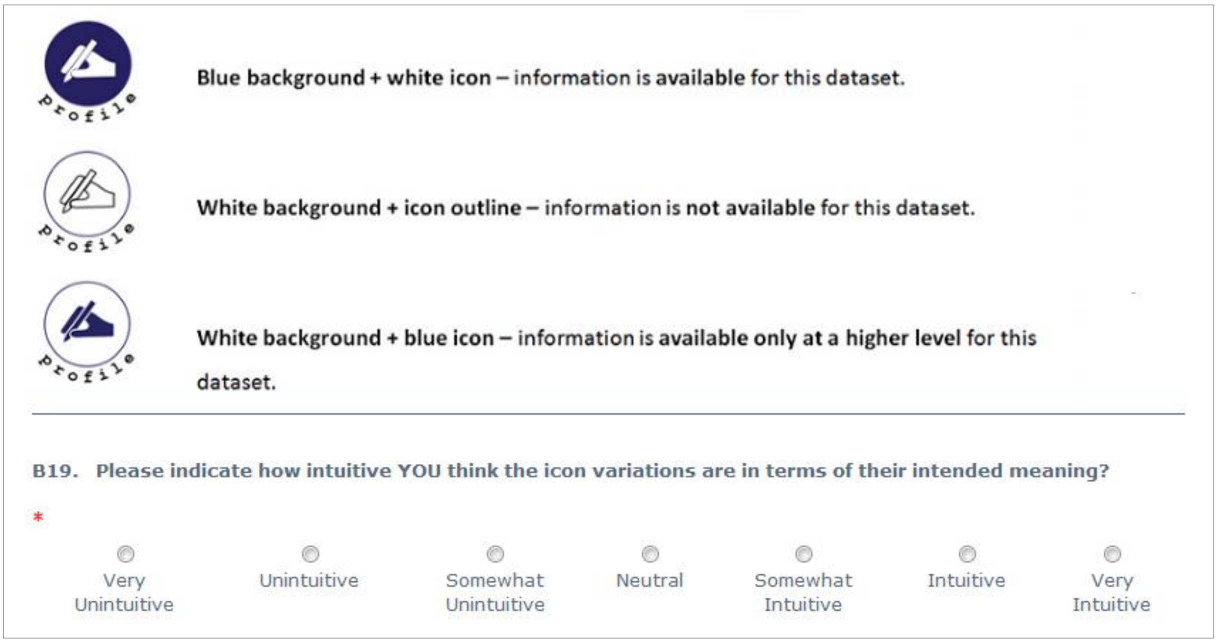


Figure 6.4: Question B18 (see Appendix D) asking respondents to identify the availability-related meaning of the icon variations.


Blue background + white icon – information is available for this dataset.

White background + icon outline – information is not available for this dataset.

White background + blue icon – information is available only at a higher level for this dataset.

B19. Please indicate how intuitive YOU think the icon variations are in terms of their intended meaning?

*

Very Unintuitive	Unintuitive	Somewhat Unintuitive	Neutral	Somewhat Intuitive	Intuitive	Very Intuitive
○	○	○	○	○	○	○

Figure 6.5: Question B19 (see Appendix D) providing the intended meaning of the icon variations and asking respondents to grade their intuitiveness relative to that meaning.


Please carefully inspect the label and proceed to the questions below.

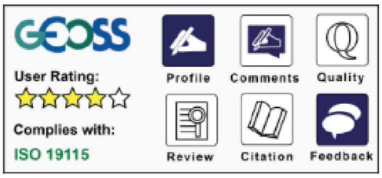
C1. Based on the GEO label provided, please select the statements about the dataset that you believe are TRUE.
(Tick all that apply)

*

- ☐ Expert review is available
- ☐ User rating is only available at a higher level
- ☐ Information about the dataset's compliance with international standards is available
- ☐ Quality information is not available
- ☐ User feedback is not available
- ☐ Producer comments are only available at a higher level
- ☐ Citation information is available
- ☐ Producer profile information is not available
- ☐ User feedback is available

Figure 6.6: GEO label example 1 presented with a mixed set of true and false availability statements from which the respondents were required to identify the true statements.

Please carefully inspect the label and proceed to the questions below.



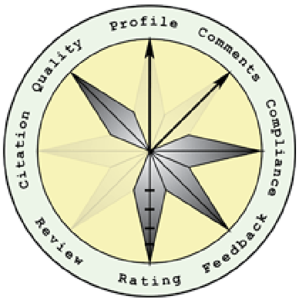
C11. Based on the GEO label provided, please select the statements about the dataset that you believe are TRUE.
(Tick all that apply)

*

- [ ] Producer comments are not available
- [ ] User feedback is available
- [ ] User rating is available
- [ ] Citation information is only available at a higher level
- [ ] Producer comments are only available at a higher level
- [ ] Expert review is not available
- [ ] Quality information is not available
- [ ] The dataset complies with ISO standard
- [ ] Producer profile information is only available at a higher level
- [ ] The dataset has a rating of 3

Figure 6.7: GEO label example 2 with a mixed set of true and false availability statements from which the respondents were required to identify the true statements.

Please again carefully inspect the label and proceed to the questions below.

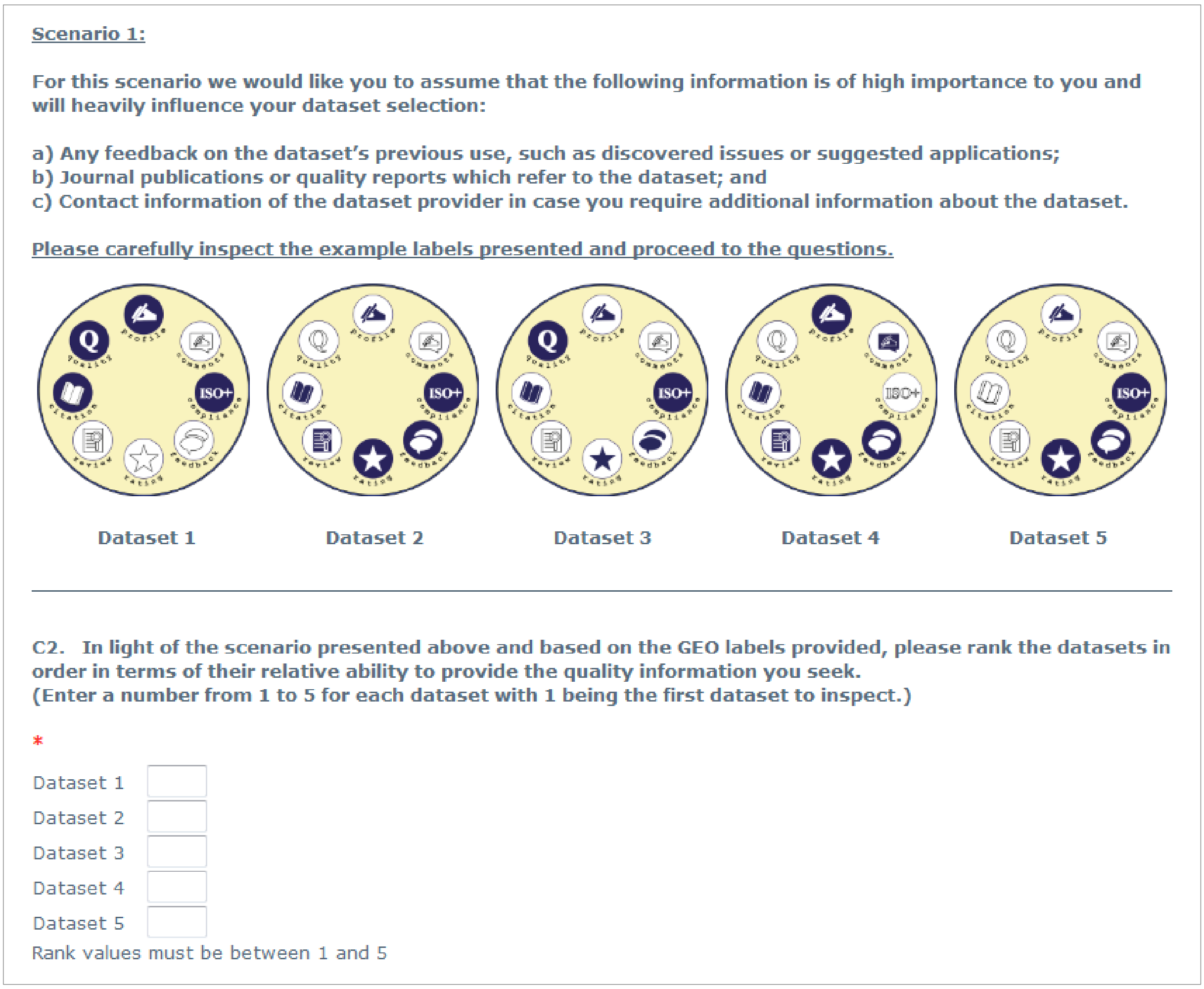


C20. Based on the GEO label provided, please select the statements about the dataset that you believe are TRUE.
(Tick all that apply)

*

- [ ] Citation information is available
- [ ] Information about the dataset's compliance with international standards is not available
- [ ] Expert review is not available
- [ ] Producer comments are only available at a higher level
- [ ] User feedback is available
- [ ] Citation information is only available at a higher level
- [ ] Quality information is only available at a higher level
- [ ] User rating is not available
- [ ] Producer profile information is only available at a higher level
- [ ] Expert review is available

Figure 6.8: GEO label example 3 with a mixed set of true and false availability statements from which the respondents were required to identify the true statements.



Scenario 1:

For this scenario we would like you to assume that the following information is of high importance to you and will heavily influence your dataset selection:

a) Any feedback on the dataset's previous use, such as discovered issues or suggested applications;
b) Journal publications or quality reports which refer to the dataset; and
c) Contact information of the dataset provider in case you require additional information about the dataset.

Please carefully inspect the example labels presented and proceed to the questions.

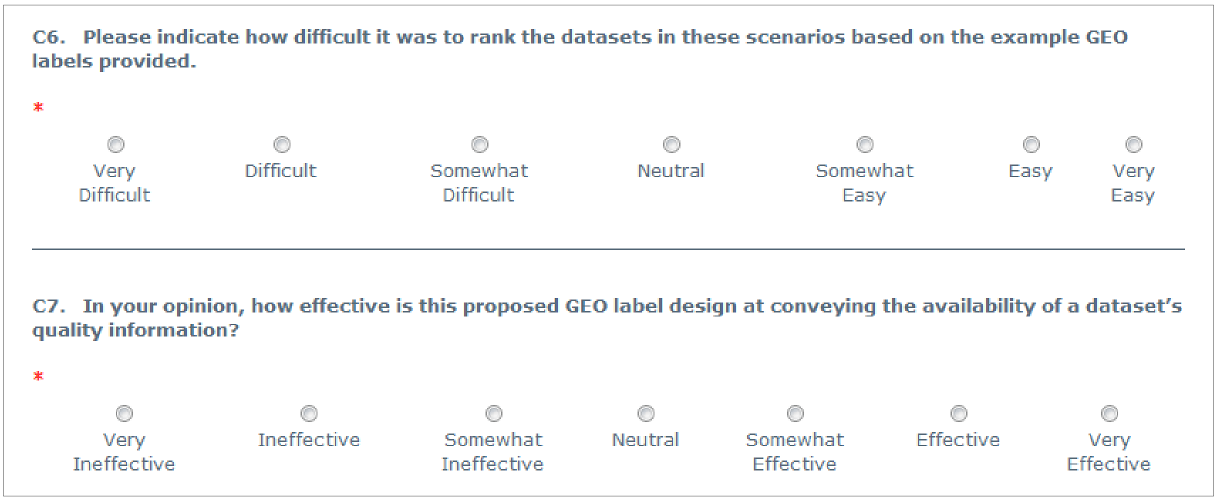
Dataset 1 | Dataset 2 | Dataset 3 | Dataset 4 | Dataset 5

C2. In light of the scenario presented above and based on the GEO labels provided, please rank the datasets in order in terms of their relative ability to provide the quality information you seek.
(Enter a number from 1 to 5 for each dataset with 1 being the first dataset to inspect.)

*

Dataset 1
Dataset 2
Dataset 3
Dataset 4
Dataset 5
Rank values must be between 1 and 5

Figure 6.9: GEO label example 1 presented within the context of a dataset selection scenario where respondents were asked to rank the datasets according to best fit for purpose on the basis of the representative GEO labels.



C6. Please indicate how difficult it was to rank the datasets in these scenarios based on the example GEO labels provided.

*

Very Difficult | Difficult | Somewhat Difficult | Neutral | Somewhat Easy | Easy | Very Easy

C7. In your opinion, how effective is this proposed GEO label design at conveying the availability of a dataset's quality information?

*

Very Ineffective | Ineffective | Somewhat Ineffective | Neutral | Somewhat Effective | Effective | Very Effective

Figure 6.10: Questions C6 and C7 (see Appendix D) in which respondents were asked to reflect on the use of ranking datasets according to the GEO labels for each and to indicate the label effectiveness, respectively.

Please review the three GEO label examples presented in this study.

Example 1 Example 2 Example 3

C26. After reviewing the examples presented here, please rank the GEO label examples in order of YOUR preference (with 1 being your most preferred version).

*

GEO label – Example 1 [ ]

GEO label – Example 2 [ ]

GEO label – Example 3 [ ]

Rank values must be between 1 and 3

Please provide a brief explanation for your answer:

Figure 6.11: Question C26 (see Appendix D) in which respondents were asked the rank the three prototype GEO labels in order of preference.

Section D consisted of a small number of questions to gather respondents' opinions on the informational aspects presented in the GEO label examples. Here respondents were asked to identify the informational aspects that they consider important and relevant to the GEO label, to indicate the aspects that they considered to be redundant and should not be included in the GEO label, and to describe any additional aspects that they felt may be relevant to the GEO label function. This was done to (a) further validate findings from the previous phases of this research, and (b) to identify whether, when used in reality within the context of a proposed GEO label, respondents felt differently to the proposed informational elements than had been suggested would be the case in previous phases of this investigation.

Section E consisted of a small number of questions designed to gather respondents' opinions on the use of 'GEO' branding in the GEO label. The main aim of this section was to investigate whether the presence of branding affects perceived credibility and trustworthiness of the label and the information it conveys. As discussed in Section 2.6, the concept of a GEO label was initially proposed by GEO to support the recognition of the GEO activities and promote GEOSS; consequently, 'GEO' was an obvious branding to use as part of the evaluation examples. In this section, respondents were presented with two GEO label examples (see Figure 6.12), one with ‘GEO’ branding and one without any branding. Using a 5-point Likert scale, ranging from ‘strongly disagree’ (1) to ‘strongly agree (5), respondents were asked to reflect on their level of agreement with a series of statements about the perceived trustworthiness of the two examples (see Figure 6.13 for examples of the statements). Both positive and negative statements were used to try and ensure that respondents provided properly-considered ratings. Respondents were ultimately asked to indicate whether they would prefer a GEO label with ‘GEO’ branding or a label without such branding; they were asked to justify their choice.

Figure 6.12: Examples of a GEO label with and without branding (see Appendix D).

E1. Presence of the GEO branding encourages me to trust the GEO label more than I would otherwise. (Tick one that applies)

*

○ Strongly Disagree ○ Disagree ○ Neither Agree Nor Disagree ○ Agree ○ Strongly Agree

E2. A GEO label that carries the GEO branding is more reliable than a label that has no branding. (Tick one that applies)

*

○ Strongly Disagree ○ Disagree ○ Neither Agree Nor Disagree ○ Agree ○ Strongly Agree

Figure 6.13: Questions E1 and E2 (see Appendix D) asking respondents to reflect on the impact of GEO branding in terms of their interpretation of the label.

Finally, in section F, the respondents were asked to provide any further comments and suggestions on the GEO label and the proposed GEO label designs.

As with the previous GEO label questionnaire, this questionnaire was constructed and administered using the QuestionPro (QuestionPro, 2014) online-survey software. To inform potential respondents about the questionnaire, emails were sent to a number of professionals, academics and researchers who work in the GIS field, asking them to complete the study. The questionnaire was accessed and completed from a number of countries including Spain, Germany, United States, Netherlands, France, Italy, United Kingdom and China.

6.2 Study Results and Discussion

This section presents and discusses the results of this questionnaire-based study. Very detailed commentaries provided by the study respondents supported very rich qualitative data analysis which was of significant benefit to the ongoing design efforts, especially in terms of gauging potential community acceptance of a future GEO label.

6.2.1 Background Information

A total of 26 valid questionnaire responses were received, 10 from ‘primarily dataset users’, 3 from ‘primarily dataset producers’, and 13 from ‘equally data users and data producers’. In terms of the best description(s) of their dataset user and/or producer type (bearing in mind that the descriptions were not mutually exclusive and respondents were at liberty to select one or more description), 1 respondent self-identified as a Group on Earth Observations committee member, 3 as private sector data users or data producers, 4 as governmental data users or data producers, 13 as research data users or data producers, 12 as academic data users or data producers, and 1 as other: “*private company service provider*”. Over half of the respondents (14) indicated that they work in an academic institution. All of the study participants had at least two years of experience working with geospatial data or maps, with 10 respondents having worked with geospatial data for between 2 to 9 years. The results indicated that, in the context of respondents’ current positions, most respondents (10) only work directly with geospatial data 5%–20% of the time. When asked about choice of datasets to use in their typical work, 22 study participants (85%) stated that they do have a choice of data; of those, 12 participants indicated that they use data portals or clearinghouses for selecting datasets to use. Participants listed a number of data sources that they use including the GEOSS portal, NASA, ESA, NERC, National Snow and Ice Data Centre (NSIDC), INSPIRE, and even Google. Of the 22 respondents who have a choice of dataset, 17 indicated that they find selecting datasets to fit their needs a challenging task (detailed respondents’ profiles are provided in Appendix E).

6.2.2 Recognising the Intended Meaning of the GEO Label Facets

6.2.2.1 Producer Profile

Figure 6.14: Producer profile facet.

When asked to describe their interpretation of the meaning of the '*producer profile'* facet (see Figure 6.14), only one respondent was able to guestimate what the proposed icon represented, describing it as an indicator of "*who created the data or is commenting on it*". The majority of respondents did not appear to understand what informational aspect the icon represents, with some stating that the word '*profile*' suggests information about a dataset's "*general contents*", "*some overall metadata about a dataset*", or a dataset's "*basic metadata*" (e.g., title of the dataset, abstract, description, time period, resolution, etc.).

When the intended meaning of the facet was explained, many respondents suggested adding the word '*producer*' to the description label: they suggested this would better focus the profile to the producer rather than the dataset itself, and thereby eliminate the aforementioned confusion. The visual presentation of the label was identified as a potential problem because "*when these icons are made smaller, having the text in a curve will hamper clarity*".

Regarding the icon itself, there was no general agreement as to whether or not it is suitable to represent '*producer profile*', with three respondents stating that the 'writing hand' was perhaps more indicative of comments or commenting rather than a profile of a producer because "*it seems you are writing*" and so hints at "*editing more than producing*". In contrast, other respondents argued that it was acceptable because it could be intuitively linked to authorship. The difficulty in generating a perfect icon for producer profile information is perhaps best summed up by an academic data user who suggested that "*there is probably no widely-accepted icon for 'producer' that could be used here and the icon here is probably as good as any*".

6.2.2.2 Producer Comments

Figure 6.15: Producer comments facet.

As with the '*producer profile'* facet, the majority of respondents were unable to identify the intended meaning of the '*producer comments*' facet (see Figure 6.15). Sixteen respondents stated that they considered it to convey some kind of "*human-readable*", "*free text comments*", although they were "*not sure if this represents users specifically, or whether this also encompasses comments from the data producers*". Only three study respondents were able to correctly identify the intended meaning stating that the facet represents:

a) "*additional comments by data creators describing pitfalls and problems the data set has or things the user of this dataset should consider when using the data*";
b) "*qualitative comments from the producers, issues with the data set, assumptions behind generation or use of the data set*"; and
c) "*comments* [from the] *data provider possibly about potential use*".

After learning the intended meaning of the facet, many study participants again suggested adding the word 'producer' to the logo. They also raised the issue that it is "*somehow unclear what the difference is between 'comments' and 'feedback'*", with the result that producer comments could be misinterpreted as user feedback and vice versa.

Respondents did not provide many comments on the effectiveness of the actual facet icon itself, but they did recognise the connection between the previous '*producer profile*' and '*producer comments*' icons, arguing that informational facets associated with the producer were very clear and visibly connected once the key producer icon was known.

6.2.2.3 Compliance with Standards

Figure 6.16: Compliance with standards facet.

When asked to describe their interpretation of the intended meaning of the '*compliance with standards*' facet (see Figure 6.16), 25 respondents were able to identify that this facet represents compliance with some international standards; of these, 17 assumed compliance with various ISO standards. Only one respondent suggested that the facet indicates that data "*has some certification*". Although most respondents linked the representation to ISO standard compliance, they argued that there was no helpful information about *which* standard and an associated certification date. The '+' symbol was not perceived as intuitive and respondents were uncertain whether it indicated that the dataset complied with other, non-ISO standards as well. Interestingly, one respondent reacted rather negatively to this facet saying that it indicated that "*the data is standards compliant which means the structure will be complex"*, meaning that he would *"have to find software to read it - which might not exist!*".

When the intended meaning was explained, respondents argued that the explicit 'ISO' component of the symbol seems to place heavy or even sole focus on ISO compliance at the expense of other standards. Respondents again suggested that "*some reference to which version of ISO would be desirable*", as would a link to the corresponding ISO standard's webpage. Consequently, it was proposed to replace the symbol 'ISO+' with word 'Standard' or to "*introduce more symbols in the icon*" to indicate that "*standards are multiple and not only ISO*".

Some respondents more deeply questioned the inference of the facet: they questioned what 'compliance' actually means, who would *"police such claims"*, and "*to what must* [a dataset] *be compliant to get this*". It was argued that 'compliance with standards' may have limited meaning to some users who are neither aware of nor interested in ISO or other standards. Correctly interpreting the icon's meaning, one academic data user and producer argued it is impossible to achieve full compliance with ISO *and* other standards; another academic data user suggested that "*in reality, data and metadata can partially conform to many standards, so assessing the level of compliance can be difficult*". This respondent further suggested that "*the 5-star 'open data' rating proposed by Tim Berners-Lee could be adapted"* in which case our *'compliance with standards'* facet could *"become 'available in an open data format' or 'uses non-proprietary standards'"*.

More positively, some respondents argued that our '*compliance with standards'* facet "*is rather intuitive*" and is, in fact, "*the most intuitive of all the icons*" because ISO is well known, even if other standards are used. One respondent agreed that 'ISO+' is, itself, slightly misleading but conceded that it was clearly connected to standards and could not suggest a way to make the intention clearer.

6.2.2.4 User Feedback

Figure 6.17: User feedback facet.

When asked to describe their interpretation of the intended meaning of the '*user feedback*' facet (see Figure 6.17), the majority of respondents (18) correctly recognised that the facet represents availability of user feedback. Interestingly, however, a small number of participants assumed that this facet indicates "*an option to provide feedback*" rather than just to read existing feedback.

When the intended meaning was explained, respondents argued that the "*difference between 'feedback' and 'comments'* [as in 'producer comments'] *is not so clear*" and these two facets can be confused. Respondents proposed adding the word '*user*' to the label to avoid ambiguity. Regarding the actual icon itself, respondents recognised it as "*very intuitive"* based on its common use, especially in social networking environments.

6.2.2.5 User Ratings

Figure 6.18: User ratings facet.

When asked to describe their interpretation of the intended meaning of the '*user rating*' facet (see Figure 6.18), 18 respondents correctly described the facet as representing some sort of user rating, but some were unsure as to the source of the rating (e.g., user, supplier, peer-review, or that of an independent body). As with findings regarding the '*user feedback*' facet, several study respondents mistakenly assumed the ‘*user rating*’ facet linked to a facility for actively rating geospatial datasets as opposed to indicating the provision of user ratings.

After learning the intended meaning of the facet, respondents agreed that this is “*one of the most intuitive icons*” and that the representation is “*recognizable from hotel ratings, etc.*”. Despite the facet’s overall intuitiveness, some responses suggested that the filled-in star might be mistaken for the actual rating (i.e., 1 out of 5 stars) as opposed to an indication that ratings are available. In line with ‘*user feedback’*, respondents suggested adding the word ‘user’ to the description label to avoid ambiguity and highlight that the ratings are provided by data users. To further improve the facet, respondents also proposed that a connection between the ‘*user feedback*’ and ‘*user rating*’ facets should be visually indicated (via some kind of visual similarity). It was further suggested that colour coding should be used to better distinguish between facets related to users and producers.

In terms of the facet in general, some respondents demonstrated strong concerns regarding the criteria that users would use to rate the datasets because of the subjective nature of the perceived quality of a dataset based on the area of application. They suggested that user ratings should only be permitted if supported by feedback or a review, and that it would be important to distinguish between ratings provided by authoritative experts and novice users; this could possibly be resolved by enforcing user registration before allowing any feedback.

6.2.2.6 Expert Review

Figure 6.19: Expert review facet.

When presented with the ‘*expert review*’ facet (see Figure 6.19), six study respondents were able to correctly identify the facet’s intended meaning, describing it as: (a) “*expert opinion*”; (b) “*scientific review*”; (c) “*a formal review of a dataset, perhaps published in the literature*”; (d) “*expert review of the data, more expert-based than the above rating and feedback*”; (e) “*peer review… mainly specialized scientists*”; and (f) “*technical reviews of the data*”.

The remaining 20 respondents were unable to correctly identify the intended meaning of the icon, arguing that the facet is not intuitive. Six respondents identified the facet as representing user reviews (such as feedback and comments) – which led to their confusion with the comment- and feedback-related facets already discussed. According to respondents, the description label was misleading because it could variably be interpreted as referring to a singular review or as an invitation to submit a review (as opposed to the facility to read existing reviews). Interestingly, one of the respondents assumed that the facet conveys the “*results of comparison of the dataset to other geospatial datasets offering the* *same kind of data*". The remaining respondents were unable to interpret the intended facet's meaning, and therefore did not provide any specific descriptions.

After learning the intended meaning, six respondents suggested that the word 'expert' should appear in the facet's label; without it, it is "*not clear who supplied the review*" or *"it seems like an invitation*" to supply one. A private company service provider warned that some users/producers may not be familiar with the notion of expert review which he considered to be an academic concept; somewhat corroborating this academic/commercial distinction, an academic data user and producer argued that "*the magnifying glass is very intuitive*" and effectively conveys the intended meaning.

To improve this facet's recognition, a data producer proposed to visually link the user- and expert-review icons, stating that "*it'd be nice if this was somehow linked to the user reviews icon - e.g., a star with some indication of 'expertise' added. Seeing the two together would definitely help make it clear how they differed*". Although a valid suggestion, it was felt that introducing visual similarity between user and expert icons could lead to further confusion as to why there would be *three* feedback-related facets within a GEO label.

6.2.2.7 Citation Information

Figure 6.20: Citation information facet.

When presented with the '*citation information*' facet (see Figure 6.20), 9 respondents were able to correctly identify the facet's intended meaning, describing the facet as, for example:

a) a "*list of papers that used the dataset*";
b) "*links to papers that have cited the data*"; and
c) a "*list of publications (both peer-reviewed and grey literature) which describe the data set or the instruments and data processing steps used to create the dataset*".

Six respondents were unsure whether the facet was supposed to convey availability of references to "*paper*[s] *published in the literature that cite a dataset*" or whether it "*indicates the presence of a 'canonical citation' about a dataset*". Five respondents assumed that the facet provides information on "*how the data set can be referenced*". The remaining 6 respondents were unable to identify the intended meaning.

When presented with the intended meaning, respondents generally agreed that the label should read "citations" because it is otherwise unclear whether it is indicating how the dataset itself should be cited or where you can find those articles where it has been cited. Regarding the facet's icon, study respondents did not provide any comments or suggestions, this suggesting that the proposed icon is effective and appropriate and would only benefit from a slight labelling tweak.

6.2.2.8 Quality Information

Figure 6.21: Quality information facet.

When asked to identify the intended meaning of the ‘*quality information*’ facet (see Figure 6.21), only one respondent was able to correctly identify that it related to numerical measures of uncertainty. Three respondents assumed that it represents the ‘big five’ spatial data quality indicators (see section 2.4.2.1), namely “*positional accuracy, attribute accuracy, completeness, logical consistency, scale*”. The remaining respondents were unsure as to its intended meaning, but some guessed that it represents “*some quality indicator*” or “*some sort of overall quality flag*”, with one academic data user commenting that it “*probably represents the presence of some information about data quality, but is rather vague about the specifics*”.

After learning the intended meaning (that being, it represents the presence of quantitative quality information (such as uncertainty measures recorded in UncertML, errors, accuracy information, etc.), a data producer remarked that perhaps a more iconic represent would be better, with one respondent suggesting some kind of graph-based representation (e.g., normal distribution or error bars). An academic data user and producer observed that the ‘Q’ is only meaningful in languages where the word for ‘quality’ begins with a ‘Q’.

Regarding the facet’s description label, an academic data user and producer proposed that the label should be extended to include the word “information” for clarity, whilst other respondents recommended changing the description to *“quality control*”, “*quality assurance*” or “*quantitative (uncertainty)*”. While some study respondents admitted that the facet “*is intuitive*”, they highlighted that it does not clearly convey “*what kind of quality is offered*”.

6.2.3 Intuitiveness of the Proposed GEO Label Facets

Figure 6.22 and Table 6.1 present the results of respondents’ ratings of icons’ intuitiveness. As can be seen from Figure 6.22, the respondents rated *most* of the proposed GEO label icons as workably intuitive with producer profile, producer comments and expert review icons considered least intuitive.

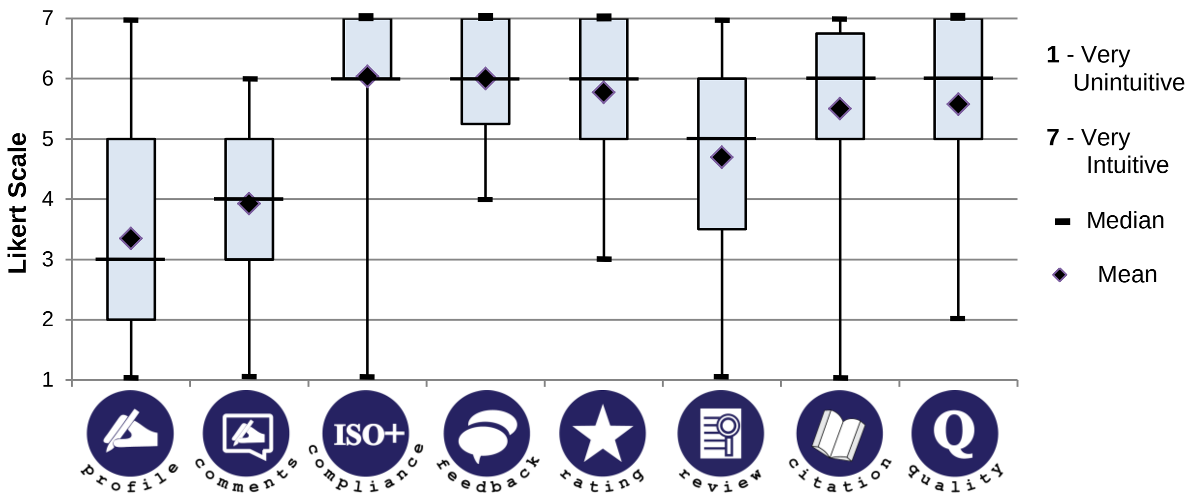


Figure 6.22: Respondent-attributed intuitiveness of the GEO label icons.

Table 6.1: Respondents ratings of intuitiveness of the GEO label icons.

Icon / Intuitiveness	profile	comments	ISO+ compliance	feedback	rating	review	citation	Q quality
1 / Very Unintuitive	2 7.7%	3 11.5%	1 3.9%	0 0.0%	0 0.0%	1 3.9%	1 3.9%	0 0.0%
2 / Unintuitive	**8** **30.8%**	3 11.5%	0 0.0%	0 0.0%	0 0.0%	2 7.7%	1 3.9%	1 3.9%
3 / Somewhat Unintuitive	6 23.0%	5 19.2%	0 0.0%	0 0.0%	1 3.9%	4 15.4%	2 7.7%	2 7.7%
4 / Neutral	2 7.7%	3 11.5%	1 3.9%	1 3.9%	3 11.5%	0 0.0%	1 3.9%	2 7.7%
5 / Somewhat Intuitive	6 23.0%	**6** **23.1%**	3 11.5%	6 23.1%	6 23.1%	11 42.3%	3 11.5%	5 19.2%
6 / Intuitive	1 3.9%	**6** **23.1%**	10 38.5%	**11** **42.3%**	7 26.9%	**6** **23.1%**	**11** **42.3%**	**8** **30.8%**
7 / Very Intuitive	1 3.9%	0 0.0%	**11** **42.3%**	8 30.8%	**9** **34.6%**	2 7.7%	7 26.9%	**8** **30.8%**
Mean	3.35	3.92	6.04	6.00	5.77	4.69	5.50	5.58
Standard Deviation	1.60	1.72	1.31	0.85	1.18	1.57	1.63	1.42

6.2.4 Conveying Information Availability through Icon Variations

Overall, respondents' comments on icon variations to convey information availability (see Figure 6.23) indicated that the variations were not as effective as was hoped. Respondents indicated that the intended meaning could only be estimated *"by process of elimination*" because the variations were not distinct enough.

Information is available | **Information is NOT available** | **Information is available at a higher level**

Figure 6.23: Icon variations to convey information availability.

When asked to describe their interpretation of the intended meaning of the '*available*' icon variation, 10 study respondents were able to correctly identify the intended meaning; 9 were unable to arrive at a conclusion regarding the meaning, arguing that the icon is '*not clear*'; the remaining 7 participants incorrectly interpreted the meaning, of which 4 participants assumed that the variation conveys the *degree* of information availability – i.e., "*full information present*".

When asked to describe their interpretation of the intended meaning of the '*information is not available*' icon variation, 18 respondents were able to correctly describe its intended meaning regarding information availability. Five respondents were unable to interpret the intended meaning, arguing that the icon's meaning was not at all clear, especially if viewed in isolation. A research data producer proposed placing an "X" over each underlying icon to make the variation more intuitive. The remaining three participants incorrectly described the availability state as either '*available*' or '*available at a higher level*'.

When asked to describe their interpretation of the intended meaning of the '*available only at a higher level*' icon variation, only 6 study participants were able to correctly identify its intended meaning. Of these, however, one respondent argued that he was only able to guess the intended meaning *"by process of elimination!*". A large number of respondents (11) were unable to state the intended availability state, arguing that they could not differentiate between this icon and the 'not available' icon without the availability of a legend. The remaining 9 respondents incorrectly described the intended availability, with 5 stating that it represents the '*available*' state and 4 identifying the state as indicating '*partial*' or '*low*' availability.

After learning the intended meaning of the three icon variations, respondents stated that the proposed variations are not sufficiently distinct to be effective at conveying degrees of information availability. They conceded that the '*available*' and '*not available*' variations "*were reasonably clear*" when presented together, but felt that the '*available at a higher level*' representation was very unclear and unintuitive. Even with the benefit of explanation, respondents found the semantics of the distinction between the '*available*' and '*available at a higher level*' icon variations hard to retain over time. Where respondents were able to guess the relative meanings, they commented that the icon meanings were not obvious, with some going so far as to reject the variations as *"just style choices"* which *"don't indicate any information on the existence or not of data*".

In terms of the colour alterations used in creating the icon variations, a number of respondents agreed that such colour saturation had helped them to identify information availability. They stated that "*the more coloured"* the icons, *"the more available"* the information was interpreted to be, confirming the choice of unfilled for no information and filled for available information and lauding the former colour scheme for '*not available*' as the most intuitive of all three variations.

In contrast, other respondents argued that the blue colour and its alterations were not effective, indicating that they could not see the link between icon fill and availability of information and that the *'available'* and *'available at higher level'* variations should not look as opposite/different as they do. Consequently, respondents proposed a number of alternative representations to improve effectiveness, suggesting to, for example:

a) use "*red/green/yellow*" or "*red, amber, green"* colour schemes;
b) *"use pale color*" or "*softer blue background*" to indicate availability at higher level;
c) leave the "*'not available' as it is"* and make "*'available' as* [the] *opposite of that, i.e. all dark, light lines* [and] *'available at higher level' like 'available' but lines dotted or colors shaded*"; and
d) "*use a red line through* [the icon] *to mean not available, or ghost it out, and various shading to indicate how full it is*".

In general, the use of a cross or line through the icon to indicate '*not available'* proved popular, with respondents indicating it would better convey the meaning because at present all the icons could be interpreted as indicating information was available.

User Viewpoint:

"*'Available' is pretty intuitive because of the use of lots of 'ink', indicating positivity. 'Not available' is only intuitive if you have 'available' to compare it with. A line through the icon would reinforce its negative nature. 'Available at a higher level' is always going to be a tricky concept to convey, since it is specific to this case. I think the icon here is probably the best that can be done, particularly if 'not available' is clarified.*"

Table 6.2 presents respondents' ratings of intuitiveness of the proposed icon variations. Overall, as a set, the icon variations were not particularly effective (M = 3.38, SD = 1.83, where 7 is '*very intuitive*'), and results indicated a substantial level of disagreement between respondents in terms of perceived intuitiveness.

Table 6.2: Respondents' ratings of intuitiveness of the proposed set of icon variations for information availability.

Intuitiveness	**Number of Respondents**	**Percentage of Respondents**
1/ Very Unintuitive	5	19.2%
2/ Unintuitive	6	23.1%
3/ Somewhat Unintuitive	3	11.5%
4/ Neutral	1	3.9%
5/ Somewhat Intuitive	**9**	**34.6%**
6/ Intuitive	1	3.9%
7/ Very Intuitive	1	3.9%

6.2.5 Conveying Information Availability through Star Arm Variations

Section 6.2.4 discussed respondents' reactions to the availability representation based on the iconic GEO label prototypes (examples 1 and 2, see Figure 5.17 and Figure 5.18). The effectiveness of the star-arm variations, as used in example 3 (see Figure 5.19) to convey the three information availability states (see Figure 6.24), was also investigated. When presented with these variations, the majority of respondents were able to correctly recognise the intended meaning of each, with overall results indicating that all three star arm visualisations were intuitive. It should be noted, however, that this study did not adopt any counterbalancing and the proposed examples were presented in same sequence to all respondents. Consequently, since the star design was always presented to the respondents last, there is a clear scope for learning to influence responses at this point.

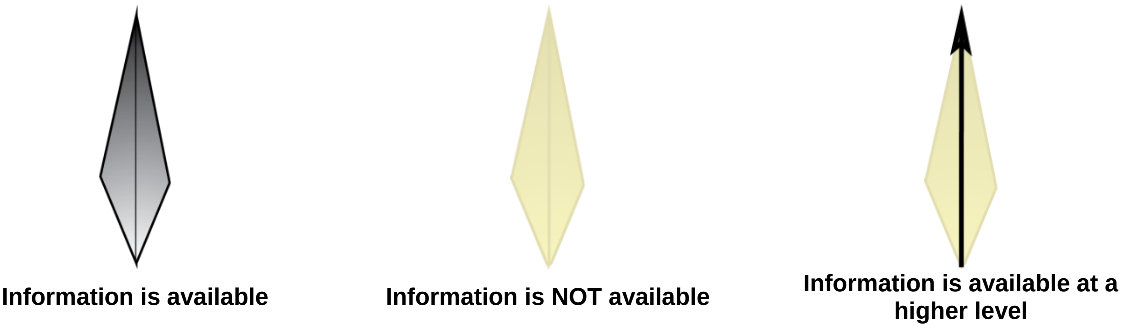


Figure 6.24: Star arm variations to convey information availability.

When presented with the '*available*' star arm variation, 22 respondents were able to correctly interpret the variation's intended meaning. Three respondents were unable to describe the intended state, arguing that the visualisation is "*not intuitive*". The remaining respondent identified the variation as indication that "*there is information available only at a higher level*".

The '*not available*' icon variation proved to be equally intuitive, with 22 respondents being able to correctly interpret its intended meaning.

When presented with the '*available only at a higher level*' star arm variation, 21 respondents were able to correctly interpret the variation's intended meaning. One respondent was partially correct in his interpretation that "*there is information available*", but four respondents were unable to describe the intended availability state.

Table 6.3 shows respondents' ratings of the intuitiveness of the proposed star arm variations. As findings indicate, this visualisation garnered greater levels of agreement across the respondents, with all three star arm variations perceived as intuitive (M = 5.12, SD = 1.75, where 7 is '*very intuitive*'). As previously noted, such level of agreement could potentially result from learning effect.

Table 6.3: Respondents' ratings of the intuitiveness of the proposed star arm variations.

Intuitiveness	Number of Respondents	Percentage of Respondents
1/ Very Unintuitive	2	7.7%
2/ Unintuitive	2	7.7%
3/ Somewhat Unintuitive	1	3.9%
4/ Neutral	0	0.0%
5/ Somewhat Intuitive	4	15.4%
6/ Intuitive	**15**	**57.7%**
7/ Very Intuitive	2	7.7%

While the vast majority of respondents were able to correctly interpret the intended availability states of all three star arm variations, the majority of follow-up comments discussing the intuitiveness of the star arm availability variations were negative in tone. This is somewhat unsurprising, considering that, by the time respondents were presented with the star design, they have already benefited from learning the meanings of three availability states from two previous examples. Bearing this learning effect in mind, respondents' comments and suggestions were carefully considered to elicit more representative data on the effectiveness and intuitiveness of the star arm variations.

Respondents argued that the arrow that indicates '*available at a higher level*' can be confusing and "*is not intuitive, since it could be pointing up, down or sideways in the final icon*". Respondents also suggested that the text provided with the star arms is too small and too curved to be practicably readable, further commenting that this impedes identification of which information element is available. Respondents suggested using only colour variations[6]

[6] It should be noted that colour-only based distinction in icon design is considered poor design practice and excludes informational access for those with colour vision deficiencies or reduced contrast sensitivity.

to convey availability, and remove the arrow from the design altogether. For instance, it was proposed using a paler colour for information at a higher level rather than introduce another arrow. Only two respondents, both governmental data users and producers, provided positive comments about the proposed star arm visualisations. They stated that these variations are "*much easier*" and "*much more intuitive than the filled symbols*" studied previously.

With hindsight, it can be suggested that the better levels of interpretation (and associated intuitiveness ratings) returned for the star arm variations are likely a consequence of learning effect brought about by respondents having previously considered the associated semantics in relation to the icon-based graphics; as such, the more accurate interpretation in this instance was not likely to be a true reflection of their acceptance of the visualisations (as realised in the accompanying comments) or their stand-alone intuitiveness.

6.2.6 Using GEO Labels for Identifying Information Availability

As discussed in section 6.1, after exploring the intuitiveness of the icons and availability graphics used within each GEO label example, each complete label was then presented together with a series of both true and false statements about quality information availability for a fictitious dataset that the label was supposed to represent; respondents were asked to indicate those availability statements which they believed were *true* (see Figure 6.6, Figure 6.7 and Figure 6.8).

Table 6.4: Percentage correct selection of true availability statements and total incorrect selections.

			
Total Statements:	9 (3 true, 6 false)	10 (6 true, 4 false)	10 (4 true, 6 false)
% True Statements Correctly Selected:	87.2% (68 out of 78 possible)	80.8% (126 out of 156 possible)	89.4% (93 out of 104 possible)
% Total Statements Incorrectly Selected:	17.3% (27 out of 156 possible)	25.0% (26 out of 104 possible)	13.4% (21 out of 156 possible)

Table 6.4 shows the percentage of possible true statements that were correctly selected, as well as the percentage of erroneous selections made. The percentage was calculated out of the total number of true/false statements for each example multiplied by the number of respondents, for instance, if the first example was accompanied with 3 true statements then

the total number of possible selections was (3 * 26) 78. As can be seen, the results show a very similar pattern for all three GEO label examples, indicating that all three examples – regardless of form – were equivalently effective in terms of conveying information availability. Whilst 80%+ is not a poor accuracy rate, it was considered advantageous to try and improve on this for the final GEO label proposal.

6.2.7 GEO Label Example 1

Figure 6.25: GEO Label Design 1.

As described in section 6.1, each GEO label example was examined within the context of a dataset selection scenario: respondents were presented with a dataset use scenario and 5 different mock-up GEO labels, and were asked to rank the mock-ups (representing datasets) in order of fitness-for-use based on the scenario description. After completing the ranking exercises for GEO label example 1 (see Figure 6.25 for one mock-up example for this design format), respondents were asked to provide a brief explanation to support their ranking decisions. Although some respondents did provide positive comments regarding their use of the mock-ups in terms of comparing the datasets for use, the majority of respondents were somewhat negative about the GEO label design in terms of supporting such a ranking exercise. They argued that "*the representations are not that great for ranking*" and they "*found it too difficult to compare the labelling in a circular format*"; to this end, some commented that a more "*tabular format would have been much easier*". Many of the more specific comments regarding the visualisation of the label reiterated concerns already expressed by respondents when considering the facet icon and availability indication design as previously discussed. Respondents agreed that the visualisation needs further work "*to make it easier to get an overview*" of overall information availability for a given dataset. These comments and feedback indicate that, although the proposed visualisation may be considered fundamentally effective for conveying the availability of quality information for a given dataset, the design is too demanding to allow for efficient dataset comparison in terms of selection for further inspection.

Interestingly, despite the fact that the '*available at a higher level*' icon variation was not shown to be effective (as previously outlined), some study respondents admitted that this availability state influenced their selection decision, suggesting that once familiar with the concept, it proved useful when assessing datasets. In relation to data availability, a number of study participants commented that their ranking decisions were based on the presence of the blue colour, oftentimes simply picking the dataset with the most icons *"with background in blue*". This suggests that the saturation of the icons was the most engaging and dominant visual component of the labels and should therefore be taken into consideration when redesigning the icons and labels.

When asked to describe which aspects of the circular design were most effective/ineffective, respondents largely agreed that the proposed design is generally not sufficiently effective and is difficult to use for dataset selection. Their specific comments echoed the findings previously reported for each of the icons and availability variations, including comments about issues with reading the curved text and, more generally, the lack of intuitiveness of the colour coding and the availability representations. Being generally perceived as too complex, respondents agreed that the proposed circular GEO label needed to be simplified; they highlighted that there is no obvious ordering in terms of positioning of the facets around the circle. Regarding the number of facets, respondents argued that eight facets is too many for a label when using the label to compare datasets based on fitness-for-use; including 8 was perceived as imposing too much cognitive demand when trying to compare different labels, with respondents suggesting that the facets should either be limited to 5 or that nesting should be used to reveal more information on drill-down whilst reducing the complexity at the top level. Aside from the icon-related improvements already discussed, some respondents suggested the inclusion of a legend with the label to help interested users interpret the label's information.

6.2.8 GEO Label Example 2

Figure 6.26: GEO Label Design 2.

After ranking 5 fictitious datasets using the second of the GEO label designs (see Figure 6.26 for an example), 16 respondents provided brief comments to support their ranking decisions. Of these, five respondents attributed very high importance to the top-level availability of the 5-star user ratings, arguing that the easily visible user rating "*is the only part that tells you directly how good the data is"* as opposed to merely indicating the availability of quality information. These respondents noted that the user rating was the very first thing that caught their attention and therefore highly influenced their ranking decision – for example, "*my eyes were drawn to the yellow stars so this defined my ranking*". Although the star-based user rating was very positively accepted, respondents admitted that they would have liked to have been able to determine the reliability of the users providing the ratings, including determining whether the users had taken assumptions about the dataset into account when submitting their ratings.

In contrast, other respondents argued that "*the 'user rating' category is probably the least useful*" because geospatial data quality is use case dependent. As with the feedback on the previous example, respondents again noted that the presence of the blue colour had substantially influenced their ranking decision.

Overall, the findings suggest that respondents generally preferred the rectangular label format (see Figure 6.26) over the circular format discussed in Section 6.2.7, in part because of its similarity to other webpage-related labels. When asked to elaborate on the effectiveness/ineffectiveness of the rectangular design, respondents commented that this format was easier to work with, allowed for easier scanning of the facets, appeared more structured, was more compact (so used space more efficiently), and that the icons were more readable. They also considered this design to have more informational content, probably because the actual user rating and specific standards information was itemised at this top level – which they highlighted as very useful. That said, some respondents argued that the label should not include a 5-star-based rating, suggesting that "*users will focus too much on the yellow stars*" (as earlier comments did, indeed, suggest had been the case) and that its use could be misleading, with one respondent suggesting that example one supported "*a more fair comparison of aspects*". As before, comments regarding the facets and information availability presentation echoed those already discussed.

6.2.9 GEO Label Example 3

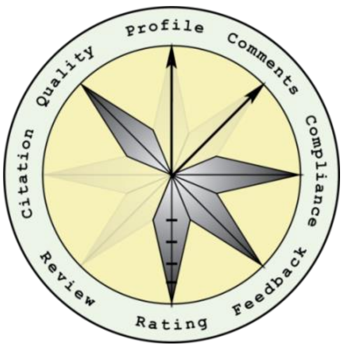


Figure 6.27: GEO Label Design 3.

After ranking 5 fictitious datasets using the third of the GEO label designs (see Figure 6.27 for an example), a number of study respondents commented rather negatively on the label design, arguing that the star label representation is too complicated, creating confusion and failing to deliver any useful information. Respondents suggested that it is "*hard to tell what the user rating is*" and, due to the very small font size used, the labels at the end of the star arms are unreadable – with some respondents going so far as to indicate they mainly considered "*the number of available items, not so much which these were*" on account of the lack of readability. Two study respondents stated that they based their ranking decision on "*the number of bold arrows*" and dark colours within the star arms, further demonstrating the effect of colour on perceived information availability. At best, respondents commented that "*it looks good for an overall picture of record completeness, but* [it is] *hard to deal with detail*".

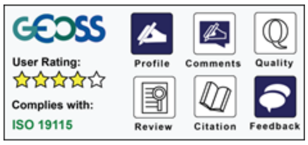
Despite a better response to the availability interpretability of the star arm variations (as discussed in section 6.2.5), no overall agreement was found on the effectiveness of the star arm-based proposed GEO label visualisation when used to compare dataset suitability for a given context of use. The respondents who felt this label design *was* effective at conveying information availability argued that overall availability information *"is available in a single glance, rather than many eye movements*" making it "*easier to see how much information is given*"; this benefit is, however, countered by the fact that other respondents considered it less intuitive than the other design and felt that the design made it hard to deal with the detail. It was generally perceived that, in terms of practical use, iconic representation of the facets (despite issues with the icons themselves) was more effective than the text-only representation in this design which made facet recognition very difficult, forcing respondents "*to read the labels all the time*" in order to identify the informational aspects that the star arms represented[7]. These findings indicate that, as confirmed in Section 6.2.5, the star arms are perhaps most effective (of the three designs studied) at conveying the three information availability states and, consequently, when combined in a GEO label, are good at delivering a global overview of information availability. The lack, however, of effective indication of what information each individual arm represents, i.e. lack of iconic representations, makes the star design hard to use in practice. Confirming this observation, some respondents went so far as to propose a combination of the star design with the icons to arrive at a more effective label visualisation. Combining the two design concepts together could potentially support both easy facet recognition and perceived information completeness. The combination of the star arms for information availability and the icons to indicate the facets themselves would seem to address many of the concerns previously discussed regarding both designs – it would seem to eliminate the weaknesses and play to the strengths of both designs.

6.2.10 Comparative Difficulty of Use of the GEO Label Designs

Table 6.5 and Figure 6.28 present respondents' ratings of the difficulty of use of the three GEO label designs tested. Overall, the results suggest that none of the proposed GEO label visualisations are particularly easy to use for ranking fictitious datasets. According to respondents' ratings, the circular GEO label design (example 1) was the most difficult to use for evaluating dataset fitness-for-use (M = 2.81, SD = 0.98, where 7 is '*very easy*'), with nearly 90% of respondents rating it as difficult to some degree. The rectangular design (example 2) was perceived as slightly easier to use (M = 3.42, SD = 1.21, where 7 is '*very easy*') than example 1, which is probably due to the label layout being similar to commonly used e-Commerce trust labels. Finally, although overall ratings indicated that the star design (example 3) was the easiest to use of the three proposed visualisations (M = 3.54, SD = 1.61, where 7 is '*very easy*'), the ratings were widely distributed across respondents, demonstrating no general agreement on the difficulty or ease of its use. It is also likely that its strength at representing information availability had an unduly strong influence on its perceived ease of use overall.

[7] These findings are more in line with established design principles.

Table 6.5: Respondents' ratings of the difficulty of use of the proposed GEO label designs, with majority rating highlighted in bold.

Label / Example Difficulty of Use	 Example 1	 Example 2	 Example 3
1 / Very Difficult	1 3.9%	0 0.0%	2 7.7%
2 / Difficult	9 34.6%	5 19.2%	6 23.1%
3 / Somewhat Difficult	**13** **50.0%**	**13** **50.0%**	**7** **26.9%**
4 / Neutral	0 0.0%	2 7.7%	2 7.7%
5 / Somewhat Easy	3 11.5%	4 15.4%	5 19.2%
6 / Easy	0 0.0%	2 7.7%	4 15.4%
7 / Very Easy	0 0.0%	0 0.0%	0 0.0%
Mean	2.81	3.42	3.54
Standard Deviation	0.98	1.21	1.61

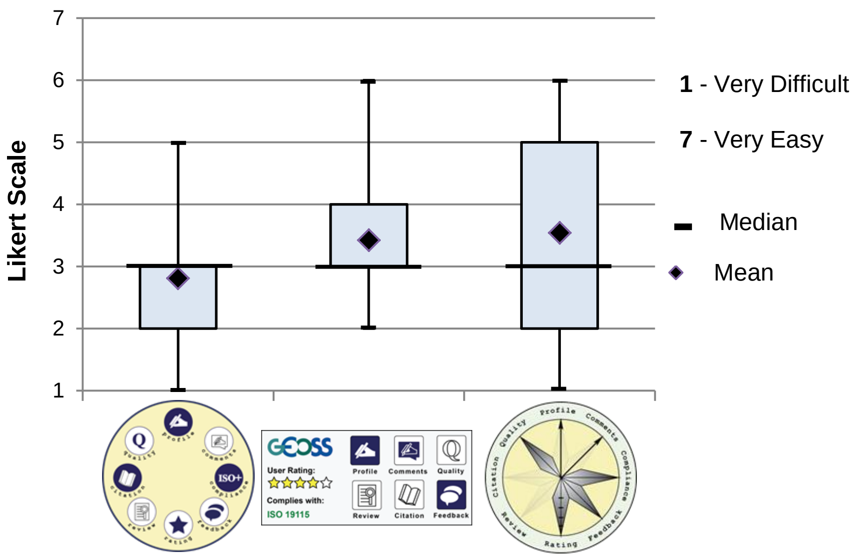
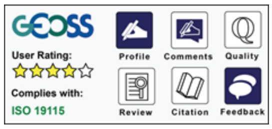


Figure 6.28: Respondent-attributed difficulty of use of the GEO label examples.

6.2.11 Comparative Effectiveness of the GEO Label Designs

Table 6.6: Respondents' ratings of the effectiveness of the proposed GEO label designs, with majority rating highlighted in bold.

Label Example / Effectiveness	 Example 1	 Example 2	 Example 3
1 / Very Ineffective	1 3.9%	0 0.0%	2 7.7%
2 / Ineffective	3 11.5%	2 7.7%	5 19.2%
3 / Somewhat Ineffective	8 30.8%	6 23.1%	4 15.4%
4 / Neutral	0 0.0%	3 11.5%	1 3.8%
5 / Somewhat Effective	**10** **38.5%**	7 26.9%	5 19.2%
6 / Effective	4 15.4%	**8** **30.8%**	**9** **34.6%**
7 / Very Effective	0 0.0%	0 0.0%	0 0.0%
Mean	4.04	4.50	4.12
Standard Deviation	1.48	1.36	1.82

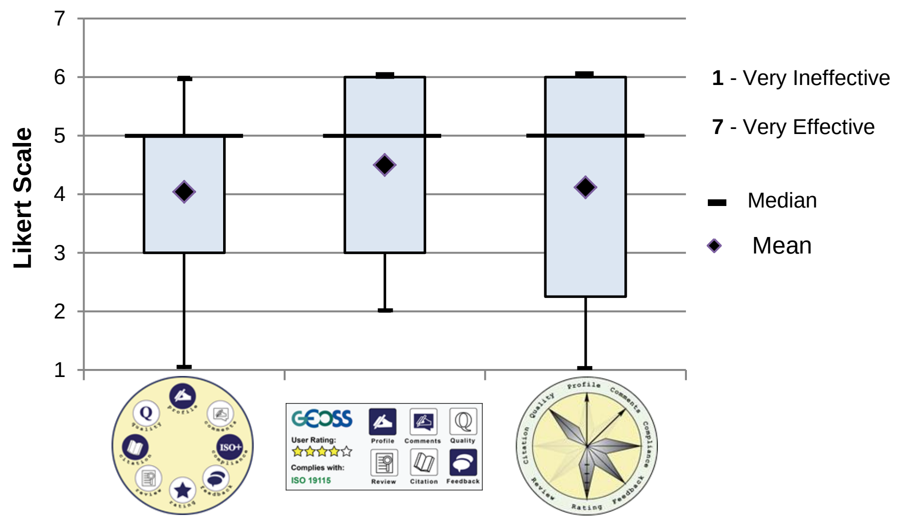


Figure 6.29: Respondent attributed effectiveness of the GEO label examples.

Table 6.6 and Figure 6.29 present respondents' ratings of the effectiveness of the proposed GEO label designs. As can be seen, whist perceived to be more effective than easy to use, the overall results suggest that none of the proposed designs were, as yet, sufficiently effective. Respondents' ratings indicated that the circular GEO label design (example 1) was perceived as the least effective of three proposed visualisations (M = 4.04, SD = 1.48, where

7 is ‘*very effective*’). The star design (example 3) was, as with ease of use, perceived slightly more favourably than example 1 (M = 4.12; SD = 1.82, where 7 is ‘*very effective*’); once again, however, respondents’ ratings were widely distributed for this design, demonstrating no general agreement on its perceived effectiveness. Although the rectangular GEO label design (example 2) had the highest rating for effectiveness overall (M = 4.50; SD = 1.36, where 7 is ‘*very effective*’), here too the ratings were widely distributed for this design, indicating no strong agreement among study respondents as to its effectiveness.

6.2.12 Respondents’ Ranking of the Proposed GEO Label Example

Having worked with each GEO label design for the purpose of ranking datasets according to scenarios of use, respondents were then asked to rank the 3 proposed GEO label designs in order of preference. Figure 6.30 shows the respondents’ rankings. Figure 6.31 shows the results in the form of overall weighted rankings; to calculate the weighted ranks, for each label the number of responses for each rank was multiplied by 3 for 1st choice, 2 for 2nd choice, and 1 for 3rd choice, the ranks were then added to arrive at an overall weighted rank for each GEO label example.

As can be seen, the circular GEO label design (example 1) was the least preferred (weighted rank 43) with only 2 respondents (7.7%) ranking it as their first choice. These results are in line with our previous findings, where respondents agreed that the circular example was the hardest to use and was the least effective at conveying information availability. The star design (example 3) was slightly better received (weighted rank 46) than the circular design, with 7 study respondents (26.9%) ranking it as their first choice. On the basis of previously-reported respondent feedback, preference for example 3 is likely to reflect that the fact that, although this GEO label design failed to effectively represent individual facets, it was most effective at conveying overall information availability. Finally, the rectangular design was the preferred design (weighted rank 67), with 17 study respondents (65.4%) ranking it as their first choice. Overall, this GEO label example received most positive comments, with respondents finding it easier to use because of perceived commonality with e-Commerce trust seals they have encountered before.

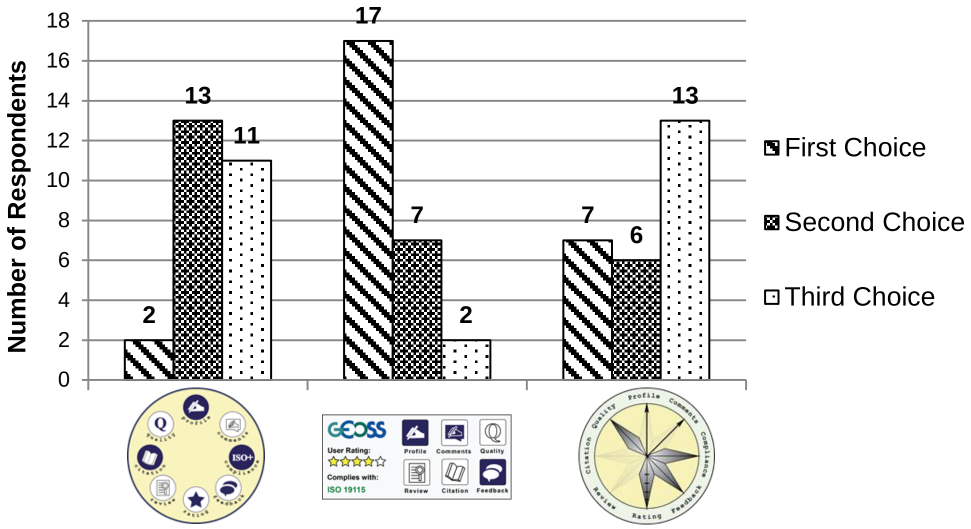


Figure 6.30: Respondent-attributed ranking of the GEO label examples in order of preference.

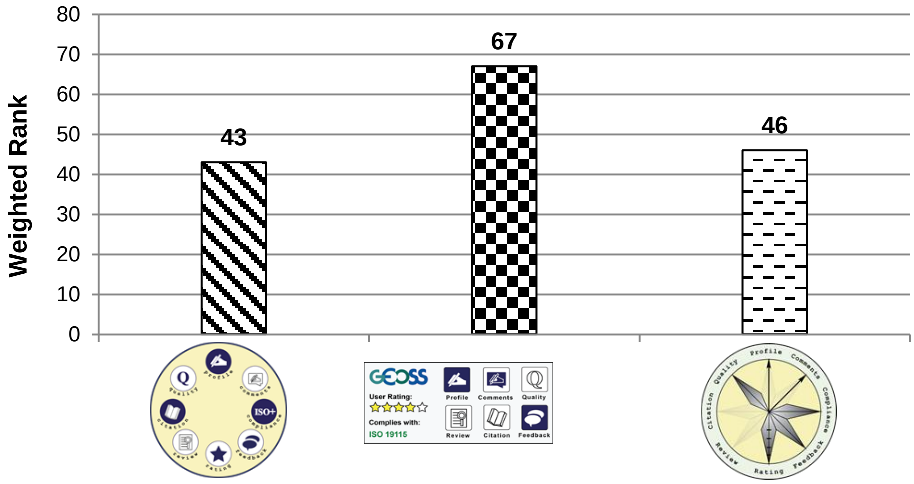


Figure 6.31: Weighted ranking of the GEO label examples in order of respondent preference.

6.2.13 GEO Label Facets

In Section D of the online survey, respondents were presented with a list of the 8 informational aspects included in the GEO label examples and asked to identify the aspects that *they* considered important and relevant to the GEO label function. Respondents were asked to provide comments to support their assessment. The results (see Figure 6.32) were in line with the previous two studies and indicated that all eight informational aspects were generally considered important and relevant to the GEO label function.

The informational aspects related to 'producer comments', 'user rating', and 'expert review' were perceived as being of the lowest importance to respondents. Attribution of low importance to these 'soft knowledge' related elements could be a consequence of the fact that such information, especially the likes of producer comments, is not currently typically provided with geospatial datasets (with users as yet mainly relying on formal metadata records for assessment and some users commenting that important producer comments should be included within the metadata itself). Additionally, some users were sceptical that, since they perceived producers as only being interested in selling their data, any producer comments could be biased. The relatively low importance attributed to user ratings would appear, on the basis of comments provided in the previous sections, to be a consequence of the fact that user ratings are not perceived to be of any value unless supported with textual feedback. Respondents' repeated concerns regarding how expertise would be determined with respect to expert reviews appears to have manifested in their low attribution of importance to this informational facet.

Producer profile and quantitative quality information were shown to be of highest importance to respondents. Indeed, this reflects the findings to date which have shown that users are more likely to acquire data from producers they know and those they trust. The same is true of the importance assigned to quantitative quality information: in addition to subjective quality information, users want to make their own judgements about data quality and its fitness for intended use on the basis of quantitative measurements of quality.

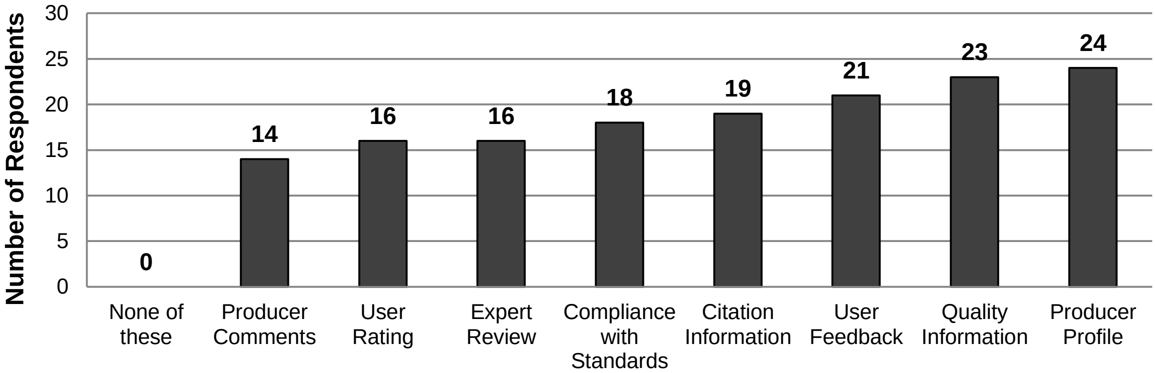


Figure 6.32: Respondent-attributed importance and relevance to the proposed eight informational GEO label aspects.

User Viewpoint:

"*I would primarily look for dataset quality information. I'd also want to know about the dataset producer so I could see if they are 'trusted' and find a contact point for questions. 'Producer comments' are probably relatively unlikely to be available but could be useful. User feedback could be very useful provided the user gives enough information to be useful, and for me to know whether their use case was comparable. Citations would be a highly-trusted way to find third-party information about the dataset.*

As I've discussed above, I don't think 'user rating' is very useful. There are unlikely to be enough users to provide a meaningful rating, meaning that results are probably going to be skewed, or not relevant to me. 'Expert review' appears to overlap with 'citation information', and could be merged.

'Compliance with standards' is only useful if this is defined more precisely, since there are so many potentially-relevant standards. A dataset could publish Dublin Core metadata, but this may be far too high-level for my needs. I would suggest borrowing from the Open Data community and using something like 'published in a non-proprietary format', or something similar that is easier to define."

Respondents were then asked to rank the 8 informational aspects in order of importance; the results of their rankings are shown in Figure 6.33. Figure 6.34 shows the results in the form of overall weighted rankings; to arrive at the weighted ranks, for each informational aspect the number of respondents who attributed each rank to the facet (1st choice, 2nd choice, … 7th choice, 8th choice and so on) was multiplied by a score for that rank (by 1 for '8th choice', 2 for '7th choice', …, 8 for '1st choice'), the results were then added. As can be seen, producer profile and quantitative quality information were again ranked as the most important aspects. When weighted, citation information and user rating facets proved to be of least importance to respondents. It is possible that the low importance being attributed to citation information is due to the fact that such information is not generally available with datasets at present and, as the preparatory research suggested, users currently find it challenging to locate publications on datasets in which they are interested. Possible causes of low importance attribution to user rating information have already been discussed in this section.

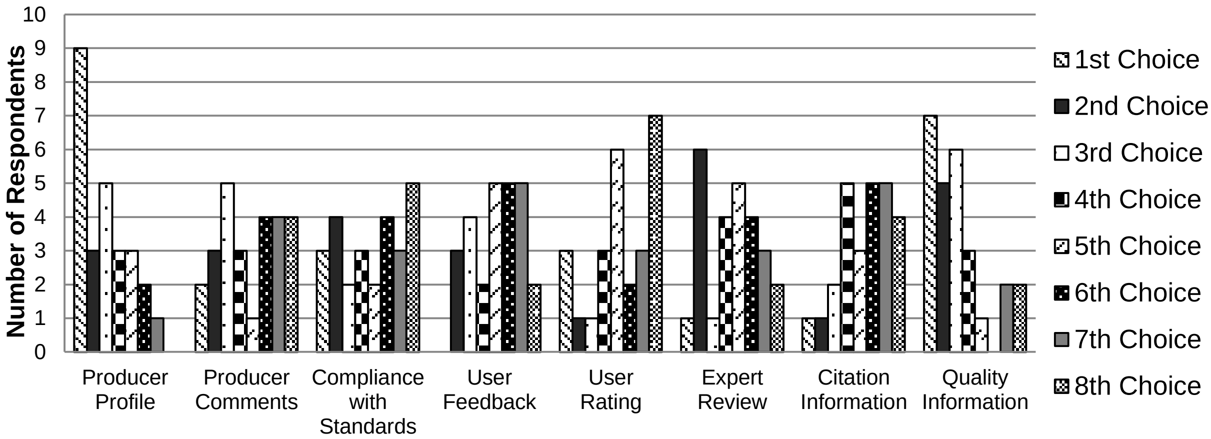


Figure 6.33: Respondent-ranked importance of the eight GEO label informational aspects.

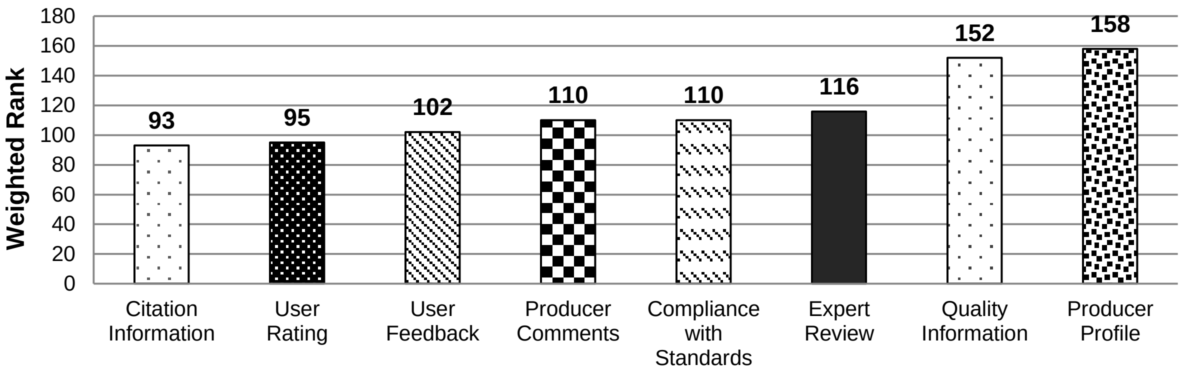


Figure 6.34: Weighted respondent ranking of importance of the eight GEO label informational aspects.

Respondents were then asked to indicate the informational aspects that they consider redundant, and which should not be included in the GEO label (see Figure 6.35). Overall, respondents did not indicate strong redundancy for any of the informational aspects; in accord with respondents' previously discussed reservations about the facet, only user rating received a relatively high number of redundancy votes (9 respondents, 34.6%). Consistent with their attributed importance levels, producer profile and quantitative quality information were not considered to be redundant by any of the respondents.

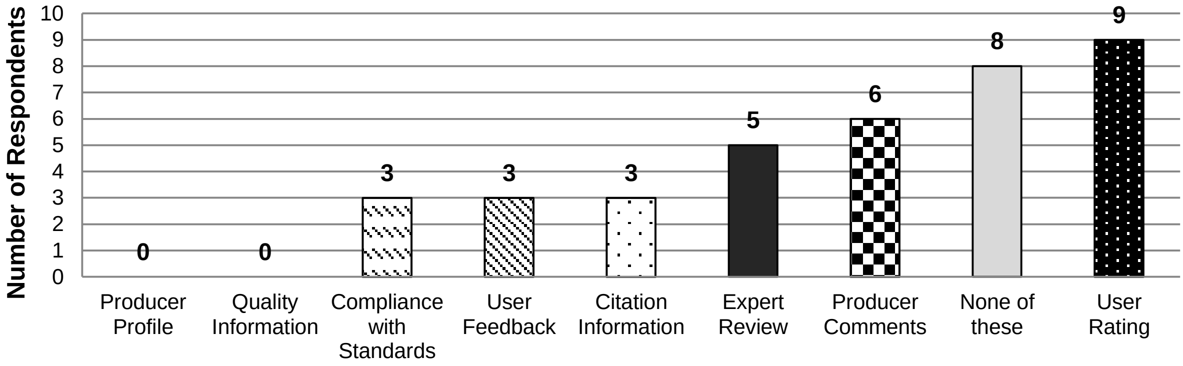


Figure 6.35: Respondent-attributed redundancy of the eight GEO label informational aspects.

When asked to describe any additional aspects that they felt may be relevant to the GEO label function, respondents typically listed specific metadata elements, namely: geographic extents; coverage; data format; feature types, i.e., vector (point, polyline, polygon), raster; and licensing information.

6.2.14 Using Branding in the GEO Label

In section E, respondents were presented with two versions of a GEO label – one using 'GEO' branding and one without the branding – and asked to indicate which version they prefer, supporting their selection with brief explanation. As can be seen from Figure 6.36, the majority of respondents (61.5%) preferred a GEO label with branding. Perhaps unsurprising (especially in light of known e-Commerce findings related to the importance of branding in promotion of trust in online environments (recall discussion in Section 2.1.3)), these results indicate that users are more likely to trust well-known and established sources. Respondents commented that the branding indicates the organisation which initiated the scheme and "*gives a 'traceability' to the label itself*".

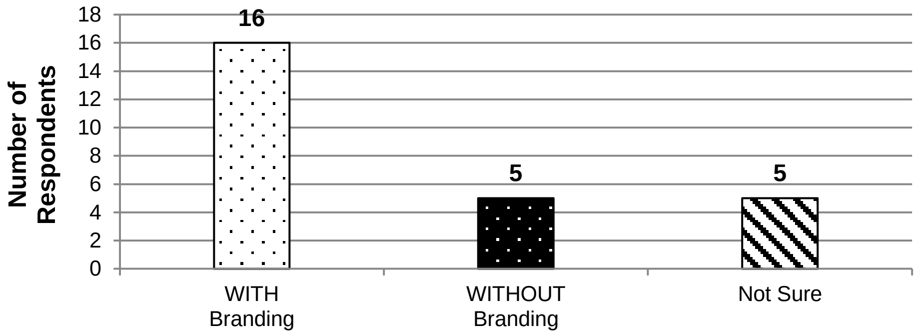


Figure 6.36: Respondents' opinion on the use of branding in the GEO label.

Respondents who preferred a label without the branding stated that without it "*it is easier to concentrate on the provided information*". Two study respondents also explained that because they "*don't know the GEO branding*" and "*have never actually used GEOSS to access data*" they were rather sceptical of the value of the brand. This again indicates that consumers put more trust in well-established brands with which they are already familiar.

6.3 Study Results Summary and Conclusions

Overall, the study results indicated that, unfortunately, none of the proposed GEO label visualisations were as yet sufficiently effective to stand as the final GEO label design. Nevertheless, respondents' feedback provided rich information on which basis to identify essential GEO label design modifications and improvements and derive user-defined GEO label requirements.

6.3.1 GEO Label Facets

One of the aims of this study was to identify whether, when included as part of a GEO label, the eight informational aspects identified in the previous studies would still be perceived as important and relevant to the GEO label role. With producer information and quantitative quality information being clearly identified as the most important GEO label facets, the study results indicated that the GEO label visualisation should comprise:

a) dataset producer information;
b) producer comments on the dataset quality;
c) dataset's compliance with international standards;
d) user feedback;
e) user ratings of the dataset;
f) expert reviews;
g) dataset citations; and
h) dataset quality information.

6.3.2 GEO Label Icons

The study results indicated that icons are absolutely essential to achieve effective representation of individual GEO label facets. Respondents did, however, propose a number of modifications to improve the proposed GEO label icons – as summarised in Table 6.7 below:

Table 6.7: Recommended improvements to facet icons.

Icon	Recommendations for Improvement
	May require modification because the writing hand symbol suggests commenting or editing.
 	With above change implemented, the producer profile and producer comments icons must retain visual similarity to support facet recognition.
	May require modification because the ISO symbol strongly suggests the relationship to the ISO standards.
	User feedback icon should be modified to indicate the relationship to users.
	User ratings icon should only convey availability of user ratings and not the average user rating directly; given the use-case dependent nature of geospatial data quality, provision of an actual average rating within the label itself could mislead or falsely direct selection decisions.
 	As with the producer icons, the user feedback and user rating icons should include visual similarity to indicate connection and support better facet recognition (in particular, the meaning of the ratings icon).
	May require modification because the icon was not perceived as intuitive by non-academic users.
	Results suggest the citation information icon does not require modification and is intuitive in its current form.
	Quantitative quality information icon needs modified to better reflect the statistical nature of the information and to remove linguistic dependency if possible.

6.3.3 GEO Label Facets Labelling

The study results indicated that it is important to include effective labels with facet icons to fully convey the intended meaning of the GEO label informational components, with study respondents relying heavily on the labels in order to interpret the facet icons. Whilst the aforementioned improvements to the icon designs should hopefully alleviate reliance on labelling, it is important to take on board that, when the facet labels are curved or too small, the text becomes unusable. In practical terms, when integrated in the GEOSS or other geospatial data portal, based on the size of the label as a whole, the facet labels could easily become too small to remain readable. A suggested solution to this is to provide a legend (or key) to itemise the facets; this could prove particularly useful for first-time users who are not familiar with the GEO label concept, as well as acting as an on-demand reminder for users who are familiar with the GEO label. Alternatively, or indeed in addition, a dynamic hover-over functionality could be utilised to display facet labels.

6.3.4 Final GEO Label Visualisation Layout

The study results indicated that the circular GEO label design (example 1) was the least effective and the most difficult to use, especially in terms of conveying information availability and in supporting side-by-side comparison of GEO labels. Despite its shortfalls, however, respondents did react positively to the provision of the facet icons in this GEO label representation – noting their support for easy facet recognition and recall. Conversely, although the star-based GEO label design (example 3) effectively represented information availability, it failed to successfully represent the individual informational facets. Consequently, study respondents proposed to combine these two GEO label visualisations to arrive at a final GEO label design.

The rectangular GEO label visualisation (example 2) was generally favoured by respondents on the basis of its similarity to common e-Commerce features that they had previously encountered. Although this label design was considered the most intuitive and the easiest to use for dataset selection, it should be remembered that respondents came to this visualisation with the benefit of learning from using the first visualisation, a fact that will undoubtedly have positively skewed their opinion of this design. A caveat to bear in mind when considering this design is that respondents indicated that the upfront average user rating was very influential, potentially to the detriment of the other GEO label facets. Respondents agreed that, considering the subjective nature of geospatial data quality, such a prominent average rating could be misleading; some went so far as to suggest it should not, therefore, be included as part of the GEO label function.

Overall, respondents did not indicate strong acceptance of any of the proposed GEO label visualisations as presented. Consequently, it was concluded that the final user-dictated graphical GEO label representation should *either* be a hybrid of two of the tested prototype designs (the circular and star-based designs) *or* should adopt a modified version of the rectangular design, comprising the 8 informational aspects but solely conveying information availability (i.e., changing the user rating meaning). On this basis, as will be discussed in the following sections, two modified GEO label designs were developed and further evaluated in order to arrive at the final GEO label representation.

6.3.5 Representation of Information Availability

Study respondents were strongly in favour of the intuitive use of colour/colour boldness to indicate information availability; the proposed additional use of arrows to convey information availability was not accepted (see Section 6.2.9). Consequently, respondents suggested the use of white icon fill for 'not available', a dark icon fill colour for 'available', and a light or graduated icon fill to indicate information 'available at a higher level'. Going one step further, respondents actually indicated that they felt the 'available at a higher level' information state was redundant, suggesting that it should be omitted altogether to avoid overcomplicating GEO label visualisation; they argued that it makes it hard to process the GEO label information if they are required to consider too many availability options – being forced to do so largely makes the GEO label unusable.

6.3.6 GEO Label Colour Scheme

The study respondents argued that existing arrangements of the GEO label facets did not clearly indicate any organisation, grouping, or relationships between facets (see Section 6.2.7). To address this issue, the facets needed be reordered to use adjacency to indicate relationships; it was also felt that use of different colour schemes could further strengthen visual grouping of the GEO label facets according to contextual relationships. The following four groups that convey related information were identified:

- producer profile and producer comments facets both relate to dataset producer;
- compliance with standards and quantitative quality information facets relate to more objective quality information;
- user comments and user ratings facets relate to data users and their feedback; and
- expert review and citation information facets both relate to expert feedback/reports on data quality.

6.3.7 Use of Branding in the GEO Label

The study results indicate that the final design of the GEO label *must* include branding to gain better user acceptance. As discussed in Section 6.2.14, the 'GEO' branding would encourage users to trust the GEO label and would make the label more recognisable.

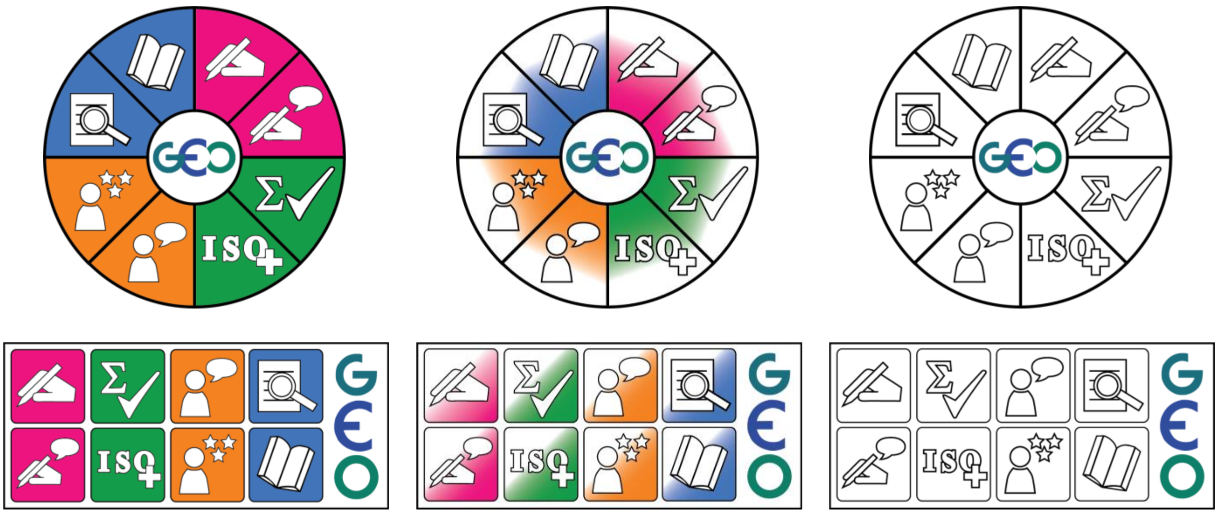
6.4 GEO Label Representation Evolution

To move closer to the development of a user-accepted graphical GEO label representation, a further iteration of the adopted user-centred design approach was undertaken whereby the GEO label designs were adapted and improved in line with geospatial experts' feedback and recommendations. Firstly, on the basis of the findings reported above, two enhanced designs of the circular and rectangular GEO label representations were created (see Figure 6.37), where the circular design represented a hybrid of two of the previously-tested prototype designs. To evaluate these enhanced representations and obtain timely experts' feedback, these GEO label designs were presented in two poster sessions at the GISRUK 2013 conference in Liverpool, UK, and at the EGU 2013 conference in Vienna, Austria. When presented with the design proposals, geospatial data experts positively responded to both circular and rectangular GEO label visualisations. No comments or suggestions were received for facet or icon modifications. Such results indicated that these versions of the GEO label visualisations were on the right pathway to the final GEO label design.

Information is available | **Information is available at a higher level.** | **Information is not available**

Figure 6.37: Enhanced sketches of the GEO label representations based on Phase II study outcomes.

While enhanced GEO label designs were accepted positively at the conferences, it was felt that the proposed icons and colour scheme could be further improved to reduce any unnecessary clutter and make facets more distinct. Following several design iterations which were focused on refining the above GEO label visualisation proposals, the GEO label designs shown in Figure 6.38 were generated. These updated graphical representations incorporated simplified and harmonised facet icons and also adopted a brighter colour scheme to ensure label's effectiveness even when scaled to a very small size.



Information is available | **Information is available at a higher level** | **Information is not available**

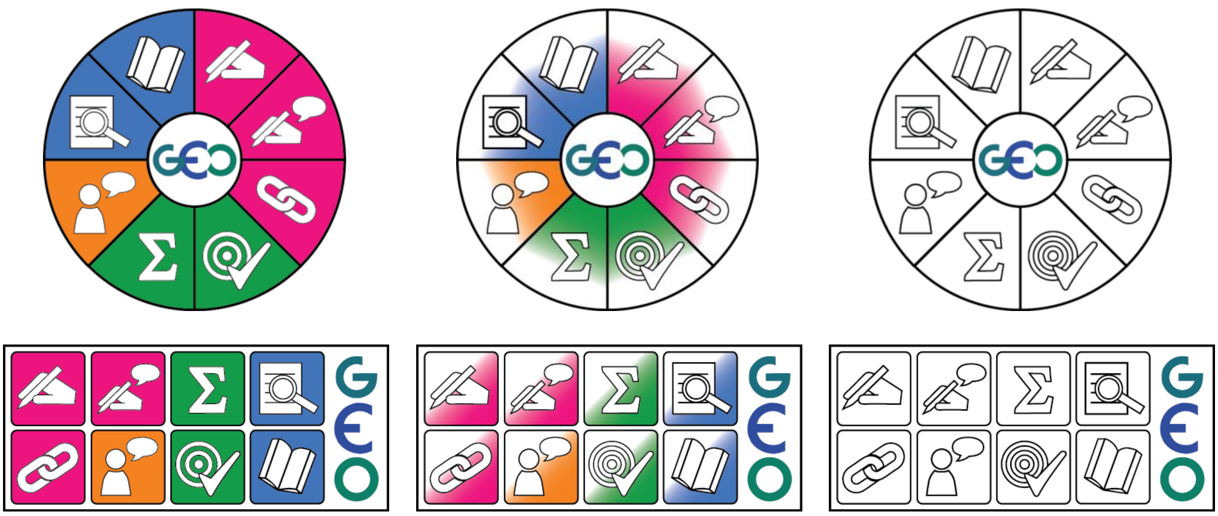
Figure 6.38: Updated GEO label representations with harmonised icons and improved colour scheme.

To gather further feedback and recommendations from the geospatial data community on the refined GEO label graphical representations, a GEO label informational website was produced which summarised the GEO label research and developments allowing visitors to send feedback and vote for the final GEO label layout. The website was made publicly available at http://www.geolabel.info in June 2013 with an objective to raise the community's awareness of the GEO label developments and also to open an active discussion about the GEO label concept and its graphical representation (the process and results of the community voting are outlined in Sections 6.4.2 and 6.4.1, respectively).

The proposed GEO label representations were also presented to the GeoViQua consortium for discussion and evaluation. The geospatial data producer members of the consortium raised a serious concern that the proposed visualisations lacked provenance and lineage information. Data producers argued that producer-related quality information was underrepresented in the label, arguing that such information is vital for geospatial data users to make an informed dataset selection decision. Despite the attempt to convey provenance information through the producer profile, producer comments and citations information facets, it was obviously insufficient to clearly represent provenance information. To better convey availability of provenance and lineage information via the GEO label, data producers asked that an additional facet was established to solely represent lineage information. After a careful consideration of the producers' appeal, it was decided to indeed introduce a new *lineage information* facet. This decision was not, however, purely based on the producers' feedback; the initial user interviews (see Section 4.2.10) had previously indicated the significant importance of provenance information, and this was confirmed in the Phase I study where respondents identified provenance as a potential GEO label function/facet (see Sections 5.2.5 and 5.2.8). When developing the prototype GEO label visualisations up to this point, it was assumed that provenance information would be adequately conveyed via the producer profile, producer comments and citations information facets and, as such, adding a separate facet to solely represent provenance information was not considered.

Phase II study outcomes indicated that 8 label facets were already pushing user-acceptance limits in terms of how much information users can process at once (see Section 6.2.7); for this reason, introducing an additional 9th facet was not considered desirable and it was felt that an additional facet would potentially decrease the label's overall effectiveness. Consequently, it was necessary to review Phase I and Phase II study results to identify which, if any, of the existing facets could be combined or removed from the GEO label representation to make room for the *lineage information* facet. Upon reflection, the Phase II study results indicated relatively low importance and perceived redundancy of the *user ratings* facet (see Section 6.2.13), with study respondents demonstrating strong concerns regarding the criteria that users would apply to rate the datasets. Taking into consideration these concerns and the fact that user ratings should only be permitted if supported by feedback or a review, it was decided to combine user ratings and user feedback into a sole *user feedback* GEO label facet. Drilldown and a dynamic hover-over functionality could then be utilised to query more detailed information from the facet, including but not limited to total number of feedback posts, total number of ratings, average rating, details of individual user feedback posts, etc. Combining *user feedback* and *user ratings* facets therefore freed up space for the new *lineage information* facet, retaining the total of 8 facets overall.

The GEO label facet icons all underwent some modifications and improvements. To represent a new lineage information facet, a ‘chain’ symbol was used to visually indicate a chain of processing steps to which a dataset has been subjected. The combined user feedback facet retained the former ‘person with a speech bubble’ icon; it was decided not to add any star-rating representation to the symbol to avoid unnecessary clutter. Regarding the standards compliance facet, the GeoViQua consortium raised concerns about using an official ISO trademark in the GEO label. Further investigation revealed that the International Organization for Standardization owns the registered trademarks for the "ISO" abbreviation and the graphical logo and only ISO members and ISO technical committees (TCs) are permitted to use ISO trademarks (ISO, 2013). Consequently, it was decided to adopt a more generic symbol for this facet and a ‘target with a tick’ icon was designed to represent compliance with standards. To avoid user confusion due to visual similarity with this new standards compliance icon, the ‘tick’ symbol was removed from the quality information facet. The resulting revised GEO label visualisations are presented in Figure 6.39 and Table 6.8.



Information is available | **Information is available at a higher level** | **Information is not available**

Figure 6.39: Revised GEO label representations with combined user feedback and a new lineage information facet.

Table 6.8: Revision stages of the GEO label facet icons.

Initial Facet Icon	Enhanced Facet Icon	Revised Facet Icon	Facet description
			***Producer profile*** facet conveys availability of information about the producer of the dataset.
			***Producer comments*** facet conveys availability of any informal comments about the dataset quality as provided by the dataset producer.
Not previously defined	Not previously defined		***Lineage information*** facet conveys availability of lineage/provenance information.
			***Standards Compliance*** facet conveys availability of information about dataset's compliance with international standards.
			***Quality information*** facet conveys availability of formal quality measures of the dataset.
			***User feedback*** facet conveys availability of feedback, comments and ratings provided by the users of the dataset.
		Removed from the GEO label	***User Ratings*** facet conveys availability of ratings provided by the users of the dataset.
			***Expert reviews*** facet conveys availability of domain experts' comments on dataset quality.
			***Citations information*** facet conveys availability of citations where the dataset was used and cited.

6.4.1 Voting for the Final GEO Label Design

As described in Section 6.3, the overall results of the Phase II questionnaire-based study did not show strong preference for any of the proposed GEO label visualisations and indicated that both the rectangular design and an updated circular design could be potential candidates for the final GEO label representation. Consequently, to arrive at a definitive user-accepted GEO label visualisation specification, the geospatial community was polled regarding the final GEO label design; the voting process was conducted on-site at two scientific conferences and online via the www.geolabel.info website. To collect on-site community votes, a small voting leaflet was produced (see Appendix F) which was distributed at the GISRUK 2013 conference in Liverpool, UK between 3rd and 5th of April 2013, and at the EGU 2013 conference in Vienna, Austria between 7th and 12th of April 2013.

Following this, links to an online voting page were constructed (see Appendix G) and administered using a Google drive form (Google Drive, 2014) to collect responses from a wider community of geospatial data experts. To inform potential voters about the GEO label voting web-form, emails were sent to a number of professionals from key organisations such as NASA, ESA, EPA, the GEO Secretariat, academics and researchers that work in the GIS field, and other GIS professionals, asking them to vote and to, wherever possible, circulate the voting web-form's URL more broadly within their network of contacts. The voting web-page was also advertised on the www.geoviqua.org website.

6.4.2 Voting Results

Due to the fact that the GEO label representations was still undergoing iterative design and development during the course of community voting (see section 6.4), three subsets of voting results were collected (see Table 6.9, Table 6.10 and Table 6.11).

Table 6.9: Results of the on-site voting for the initial GEO label designs.

	 Circular Design	 Rectangular Design	Neither Design
On-site Votes	**58.3%**	41.7%	0.0%

Table 6.10: Results of the online voting for the updated GEO label designs.

	 Circular Design	 Rectangular Design	Neither Design
Online Votes	**65.0%**	25.0%	10.0%

Table 6.11: Results of the online voting for the finalised GEO label designs.

	 Circular Design	 Rectangular Design	Neither Design
Online Votes	**85.7%**	14.3%	0.0%

As can be seen from Tables 6.9 – 6.11, despite the fact that the icons, colour scheme and the facets were varied during the voting window, the circular GEO label representation remained the most favoured by the geospatial community. As can be seen from Table 6.9, on-site voters were less polarised in their preference for one design over the other; it is anticipated that this may be because these voters were provided with fairly limited information about the GEO label and its intended purpose upon which to make a fully informed design selection. In contrast, the online voters had the advantage of access, via the website, to more detailed information about the GEO label and its underlying studies; furthermore, some of the online voters had participated in the earlier GEO label studies.

When voting for a GEO label design, the on-site participants were asked to elaborate on what influenced their selection decision. The voters who selected the circular layout explained that it utilises available space more efficiently and the clock-like design makes it much easier to navigate the facets. These results indicate that a hybrid of the initial circular and the star designs is, indeed, more effective at conveying information availability than the initial design proposals. While the initial circular design only used a small fraction of the label space to convey information availability (i.e., only icons was signalling the availability state), the hybrid design effectively utilised the whole area to communicate information to users. With added visual partitioning to separate the facets, the clock-like layout became more apparent making facet navigation easier and more intuitive. As a result, the participants argued that the arrangement of the circular facets allows them to better capture information availability at a glance, something which they felt was much harder to achieve with the rectangular layout. It was also suggested that the circular arrangement appears more harmonised and actually looks like a label. The majority of the respondents who voted for the rectangular layout explained that the linear arrangement of facets allows for larger icons, making it easier to distinguish facet symbols. Some of the voters simply stated that they generally dislike circular designs and prefer to work with linear shapes.

The participants who voted online were also required to provide feedback on their selection decision. The voters who selected the circular design stated that this layout feels more "*inclusive*" and is very simple and easy to understand. It was also suggested that having directionality in the clock-like position of facets will allow easy browsing and comparison of a large number of dataset labels. While the questionnaire results indicated that the initial circular layout was neither intuitive nor effective, the applied modifications seemed to alleviate these issues. Other voters did not provide direct comments about the label layouts, but instead suggested softening the colour scheme, replacing an ISO symbol with something more generic (this relates to the first of the voted-on label designs), and providing more examples of real labels that are based on real geospatial datasets. The minority who preferred the rectangular GEO label design perceived this design as more dynamic and extensible; respondents who voted for this layout suggested that, if needed in the future, an additional facet could easily be added to a rectangular label. It was also argued that the rectangular layout is more suitable for websites because it does not produce a lot of wasted space and the facets are bigger and easier to click. Voters who did not select any of the GEO label designs (a clear minority) argued that the label is too complex for a simple logo and "*will create unnecessary confusion and diminish the overall effectiveness of GEO*".

The experts' comments collected from on-site and online voting suggested that the circular GEO label design is more effective at conveying information availability. The directional position of the facets provides immediate visual feedback on what information is available and supports better facet recognition and recall. Consequently, it is believed that the circular design will better support cognitive processing of a large number of dataset labels at once. Feedback provided to support the rectangular label layout did not strongly indicate its effectiveness at conveying information availability. Although it was argued that the rectangular GEO label design offers larger facet icons, this would be of greatest advantage for novice label users rather than expert GEO label users who, it is anticipated, will primarily rely on the position and colour of given facets rather than their iconic representations. Although, as proposed by voters, extension of the rectangular label would perhaps be easier, having too many label facets in total would diminish the effectiveness of the GEO label as a whole, and so the current total of 8 facets is an advisable maximum for the label.

On the basis of voting outcomes, it was concluded that the final GEO label visualisation should adopt the circular layout. The following section provides, therefore, the final and definitive specification for the circular GEO label.

6.5 Final Specification for the GEO Label Representation

On the basis of the research outcomes outlined above, a final and definitive specification for the GEO label visualisation was created, which is presented in the following sections.

6.5.1 GEO Label Facets

The final GEO label should comprise 8 informational aspects, namely producer profile, producer comments, lineage information, standards compliance, quality information, user feedback, expert review, and citations information. These informational aspects should be represented via the following 8 user-accepted graphical visualisations (Table 6.12).

Table 6.12: Graphical representations and descriptions of the GEO label informational aspects.

Facet Icon	Facet Description
	***Producer profile*** facet conveys availability of information about the producer of the dataset, e.g., organisation or individual who produced the dataset, their contact information, etc.
	***Producer comments*** facet conveys availability of any informal comments about the dataset quality as provided by the dataset producer, e.g., any identified problems, suggested use, etc.
	***Lineage information*** facet conveys availability of lineage/provenance information, e.g., processing applied to data and number of process steps.
	***Standards Compliance*** facet conveys availability of information about dataset's compliance with international standards, e.g., compliance with ISO 19115, Dublin Core, etc.
	***Quality information*** facet conveys availability of formal quality measures of the dataset, e.g., uncertainty measures recorded in UncertML, errors, accuracy information, etc.
	***User feedback*** facet conveys availability of feedback, comments and ratings provided by the users of the dataset, e.g., general comments on dataset quality, identified problems, suggested use for the dataset, etc.
	***Expert reviews*** facet conveys availability of domain experts' comments on dataset quality, e.g., results of formal quality checks, expert suggestions on the dataset applications, etc.
	***Citations information*** facet conveys availability of citations where the dataset was used and cited, e.g., formal reports on dataset quality checks, journal articles, etc.

6.5.2 Representation of Information Availability

To convey the availability of quality information for a given dataset, each informational facet can represent one of three availability states: '*available*'; '*not available*'; and '*available only at a higher level*' (to indicate that information is not immediately available for the dataset, but is available for a parent dataset). These three information availability states should be expressed by varying the appearance of the facet icons as shown in Table 6.13.

Table 6.13: Graphical representations and descriptions of the GEO label availability states.

Facet Appearance	Availability State Description
	Fully filled-in background + white icon with black outline – indicates that ***information is available*** for this dataset.
	White background + white icon with black outline – indicates ***that information is not available*** for this dataset (at any level).
	Partially filled-in background + white icon with black outline – indicates that ***information is available only at a higher level*** for this dataset.

6.5.3 GEO Label Colour Scheme

The GEO label representation should visually convey organisation, grouping, or relationships between facets. This should be achieved via the use of four different colours to indicate relationships between the facets, as outlined in Table 6.14.

Table 6.14: The GEO label colour scheme for grouping of related informational aspects.

Colour	Colour HEX	Colour Group Description
	#ED1E7F	Bright pink – indicates producer-related information (producer profile, producer comments and lineage information facets).
	#0F9B48	Bright green – indicates formal quality information (standards compliance and quality information facets).
	#F38020	Bright orange – indicates data user-related information (user feedback).
	#4274B9	Blue – indicates formal reviews information (expert reviews and citations information facets)

6.5.4 Final GEO Label Representation

Each of the above requirements should be combined into graphical GEO label representation variants, as shown in Figure 6.40, Figure 6.41 and Figure 6.42. Some additional examples of the GEO label visualisations can be seen in Figure 6.43.

Figure 6.40: Final GEO label design (information is available).

Figure 6.41: Final GEO label design (information is available at a higher level).

Figure 6.42: Final GEO label design (information is not available).

Figure 6.43: Example GEO labels with information availability variations.

6.6 Study Limitations

Although once again returning rich data, and resulting in a user-led GEO label specification in which there is confidence, the Phase II studies are not without their limitations. As with both the initial investigation and Phase I study, the majority of the questionnaire-based study participants for this study phase were researchers, with over half of the respondents working for academic institutions. Consequently, the preferred function of the GEO label and its design were highly influenced by the scientific GIS community. That said, in attempts to ensure that the final GEO label addresses the needs of a wide spectrum of geospatial data users and producers, the proposed GEO label function and graphical representations were presented for discussion and feedback to a variety of stakeholders, ranging from non-expert data users to large-scale data producers. Throughout all the GEO label development stages, various GeoViQua consortium meetings, scientific conferences, and workshops were used to gather as much feedback and comments on the proposed GEO label as possible from as broad a representation of the stakeholder community as possible. It is felt that, despite the majority participation from the research domain, the results and therefore outputs of this research are representative both of a truly user-centred approach to the design of the GEO label and of the opinions and requirements of the community as a whole.

One of the main limitations of the Phase II questionnaire-based study, and therefore its associated findings, is the order in which the proposed designs were presented to the study participants; the proposed examples were presented in same sequence to all respondents (example 1, example 2 and example 3) and, as such, there was clear scope for learning to influence responses from one example to the next. The study did not adopt any counterbalancing for two main reasons: a) the online-survey software available for the research did not provide functionality to randomise the study sections; and b) based on the drop-off rate witnessed for the Phase I questionnaire (61%), it was anticipated that many respondents might leave the survey before completing all the sections, defeating attempts to achieve full counterbalancing of results even if it were possible. While it is recognised that full counterbalancing would produce more reliable results, available resources did not permit to construct a counterbalanced questionnaire. To counter the recognised effects of learning on the study results, a strong emphasis was placed on analysing textual feedback and recommendations but with an awareness of the impact of learning. As discussed in Section 6.2.5, the final results did indicate that study participants were affected by the order in which the examples were presented to them; yet, this learning effect was taken into account when designing the updated versions of the GEO label representations.

In regards to the GEO label design voting, once again it is anticipated that a large portion of votes came from the scientific GIS community. The first round of votes was collected on-site at scientific conferences, and so the majority of respondents in that round of voting are anticipated to be from the research domain; no demographic data was collected for the online voters in order to minimise the number of questions and ensure a higher response rate and, as such, it is impossible to say which communities were represented by online voters.

6.7 Summary and Conclusions

This chapter presented the studies that were conducted as part of the second main phase of the GEO label research to solicit geospatial data producers' and users' views on proposed GEO label visualisations and to arrive at a final, community-supported GEO label representation. The findings of the questionnaire-based study indicated that the final user-dictated graphical GEO label representation should *either* be a hybrid of two of the tested prototype designs (the circular and star-based designs) *or* should adopt a modified version of the rectangular design, comprising the 8 informational aspects but solely conveying information availability. Due to the fact that the overall study did not show strong preference for any of the proposed GEO label visualisations, the GEO label designs were modified, adapted and improved in line with geospatial experts' feedback and recommendations. Following the GEO label design modifications and improvements, the geospatial community was polled to arrive at a definitive user-accepted GEO label visualisation specification. The voting results indicated that the final user-defined GEO label representation should adopt a circular layout.

The following chapter of this thesis describes the implementation of a GEO label Web service developed to support use of the graphical GEO label defined in Section 6.5.

Chapter 7 GEO Label Web Service Development

To support use of the graphical GEO label outlined in Section 6.5.4, a PHP Web service was then developed to generate GEO label representations for datasets by combining producer metadata (from standard catalogues or other published locations) with structured user feedback. This chapter describes implementation of this GEO label service and its Application Programming Interface (API).

Section 7.1 provides an introduction to the purpose of the GEO label service and discusses metadata models that support it. Section 7.2 explains the method adopted for evaluation of information availability in supplied producer and feedback documents; the section describes use of an external XPath configuration file to ensure service interoperability and consistency. The output formats supported by the service are outlined in Section 7.3. The service API, together with the resources offered by the service, is presented in Section 7.4. Sections 7.5 and 7.6 describe supported dynamic hover-over and drilldown functionalities, respectively. The technical architecture of the service is presented in Section 7.7: this section outlines use of Object-Oriented (OO) PHP and the open-source PHP micro-framework Silex in the service implementation. Section 7.8 briefly introduces the GEO label service website which provides documentation for the service API and offers simple service demo pages. Finally, a chapter summary is provided in Section 7.9; here, practical and scientific implications are discussed.

7.1 GEO Label Service Introduction

The GEO label service has been developed as a stand-alone Web-based server-side application, exposed via a publicly available RESTful API. Representational State Transfer

(REST) is an abstract architectural style that constrains the implementing application to adopt a stateless client-server model with a uniform interface, meaning that “resources” made available by an application are represented by a Uniform Resource Identifier (URI) with a communication protocol that defines methods for accessing and modifying the state of these resources. A prime example of a system implementing this architecture with the Hypetext Transfer Protocol (HTTP) used for communication is the World Wide Web, where clients use HTTP method verbs to inform a server how to process their requests for a resource’s URI, e.g., GET for the retrieval of information and POST for accepting data (commonly used when creating new resources). The emergent properties of the World Wide Web are highly desirable for any software architecture: decoupled, reliable, scalable, and performant. Furthermore, clients and servers within the World Wide Web are implemented in a variety of programming languages and deployed across countless architectural permutations, yet are completely interoperable due to adhering to the HTTP standard in their communication, forming a unified system-of-systems with a self-descriptive interface for consuming resources on a global scale. A GEO label service with these qualities allows for wider deployment within service-oriented architectures, easier integration and reuse across the GIS software ecosystem. The service has been deployed on one of the GeoViQua servers and is available live at http://www.geolabel.net. Although its internal implementation can be subject to changes and upgrades, the API should remain stable for the foreseeable future.

The GEO label service is designed to dynamically process producer metadata and feedback XML documents for a given dataset and, based on evaluated information availability, build a clickable SVG (Scalable Vector Graphic) GEO label representation for that dataset. The service accepts encoded URLs of publicly available metadata documents or metadata XML files as part of an HTTP GET request, or locally-available files uploaded through a POST request, and applies XPath and XSLT mappings to transform the supplied XML documents into SVG representations. The service is underpinned by two metadata XML-based quality models that were developed by the GeoViQua project. The first is the Producer Quality Model (PQM) (Bastin *et al.*, 2012) that extends ISO 19115:2003 (ISO/TC211, 2003), ISO 19115-2:2009 (ISO/TC211, 2009) and ISO 19157:2013 (ISO/TC211, 2013), adding means to report publications, discovered issues, reference datasets used for quality evaluation, traceability, and statistical summaries of quantified uncertainty. This model introduces elements to record qualitative and quantitative quality information, and to identify resources (i.e., geospatial datasets) in order to relate metadata in hierarchical or other ways. The second is the User Quality Model (UQM) (Bastin *et al.*, 2012; Broek *et al.*, 2013), developed to enable application of ‘customer’ reviews to datasets which span a variety of user expertise levels, thematic, temporal and spatial domains. This model re-uses a few ISO quality and metadata elements, and elements of the PQM, but is far less strictly bound to existing ISO schemas. Each item of feedback references a particular target, and may contain: numerical ratings, with text justification; user comments and reports of usage and problems identified; citation of publications; tags to assist with topic-based search and linking; supplementary quality reports; and information on any spatial, temporal or thematic foci of the feedback. These two models aim to fill significant perceived gaps identified by users and producers of geospatial data (see Section 4.2 for more details), such as the formalisation of the soft knowledge quality parameters (e.g., discovered issues, publications, lineage), the standardisation of statistical quality metrics, and the ability to collect feedback from users to support the more 'user-centric' metadata. Although the services primarily rely on the GeoViQua quality models, an external XPath configuration file which is used for determining whether information is available (see Section 7.2) can be adapted to support any XML-based metadata models.

The service has been deployed on one of the GeoViQua servers and is available live at http://www.geolabel.net.

7.2 XPath Transformations

The GEO label service uses a set of XPath 1.0 expressions to determine whether the required information for a dataset is available in supplied producer and feedback XML metadata documents. To ensure service interoperability and consistency, the implementation uses an external JSON configuration file which defines XPath transformation rules and other transformation templates (see Figure 7.2 for an example).

The XPaths transformations configuration file contains the following information:

- a set of Boolean XPath expressions which are used to determine information availability; when evaluated, these XPath expressions will return either true or false;
- text templates to support hover-over functionality;
- a set of XPath expressions to retrieve hover-over text; when evaluated, these XPath expressions will return relevant hover-over text (e.g., a name of the dataset producer); and
- templates for the drilldown URLs.

Figure 7.1 provides an example XML extract from an ISO 19115:2003 metadata document showing dataset producer information. As can be noted, producer information is recorded in <gmd:contact><gmd:CI_ResponsibleParty> XML elements and the producer's organisation name, which is used in the producer profile GEO label hover-over summary, is recorded in <gmd:organisationName><gco:CharacterString> elements. Figure 7.2 provides an example of XPath transformations and templates for the producer profile GEO label facet. When evaluating information availability for a given GEO label facet, the corresponding `availabilityPath` XPath transformation is applied to the dataset's metadata XML document. In this example, the producer profile XPath transformation checks whether `gmd:contact/gmd:CI_ResponsibleParty` XML elements contain any textual information. If the elements exist and contain some text, the XPath transformation returns TRUE. The facet's hover-over summary text can be obtained by applying the `"hoverover":"text":` XPath transformation. In this example, `organizationNamePath` transformation locates `gmd:organisationName/gco:CharacterString` elements and returns the extracted producer organisation name, if found. The drilldown URL for a given facet is constructed using the `"drilldown":"url":` template; in the case of the producer profile facet, the `facet=producer_profile` URL argument is used for constructing the URL of the producer profile HTML drilldown page.

```
[…]
<gmd:contact>
  <gmd:CI_ResponsibleParty>
    <gmd:individualName>
       <gco:CharacterString>Example Producer Name</gco:CharacterString>
    </gmd:individualName>
    <gmd:organisationName>
      <gco:CharacterString>Food and Agriculture Organization</gco:CharacterString>
    </gmd:organisationName>
  […]
    <gmd:role>
      <gmd:CI_RoleCode […] codeListValue="originator"/>
    </gmd:role>
  </gmd:CI_ResponsibleParty>
</gmd:contact>
[…]
```

Figure 7.1: An example of ISO 19115:2003 XML metadata document showing producer information.

```
[…]
"producerProfile":{
      "availabilityPath":"boolean(normalize-space(string(//*[local-
name()='contact']/*[local-name()='CI_ResponsibleParty'] | //*[local-
name()='ptcontac']/*[local-name()='cntinfo'] | //*[local-
name()='pointOfContact']/*[local-name()='CI_ResponsibleParty'])))",
      "hoverover":{
              "facetName":"Producer Profile",
              "template":"Organisation name: %s.",
              "text":{
                   "organizationNamePath": "normalize-space(string(//*[local-
name()='contact']/*[local-name()='CI_ResponsibleParty']/*[local-
name()='organisationName'] | //*[local-name()='ptcontac']/*[local-
name()='cntinfo']//*[local-name()='cntorg'] | //*[local-
name()='pointOfContact']/*[local-name()='CI_ResponsibleParty']/*[local-
name()='organisationName']))"
              }},
      "drilldown":{
            "url":"%s?metadata=%s&facet=producer_profile"
      }
}
[…]
```

Figure 7.2: An example of XPath transformations and templates for the producer profile GEO label facet.

The transformer_1.1.0.json configuration file supports ISO19115, FGDC and GeoViQua-derived PQM and UQM XML metadata models; it can, however, be adapted to support any other XML-based metadata models. The latest transformer_1.1.0.json file is publicly available on GitHub at http://geoviqua.github.io/geolabel/mappings/transformer_1.1.0.json. Older versions and additional configuration file examples can be found on GitHub at http://geoviqua.github.io/geolabel/.

7.3 Output Formats

Depending on the requested resource, the GEO label service offers three output formats for a given dataset: SVG, JSON and styled HTML Web pages. The GEO label representations are returned in an SVG format which not only allows for better image scaling but also offers interactivity which is essential for supporting hover-over and drilldown GEO label functions. The service also offers a more lightweight JSON format which encodes core GEO label information as a textual JSON representation (see Figure 7.3). The returned JSON file consists of the following information:

- dataset identification number;
- facets' availability encodings (0 – information is not available, 1 – information is available, 2 – information is available at a higher level);
- facets' hover over text; and
- facets' drilldown links.

```
{"datasetIdentifier":"c0dc2fd0-88fd-11da-a88f-000d939bc5d8",
"facets":{
     "producerProfile":{
          "availability":1,
          "organisationName":"JRC"},
     "producerComments":{
          "availability":1,
          "supplementalInformation":"The GVM unit delivers products and services
                   to the various DGs of the European Commission, Space Agencies,
                   the scientific community at large and other users.",
          "knownProblems":"Legend issues with South East Asia"},
     "lineage":{
          "availability":1,"
          processStepCount":3},
     "standardsComplaince":{
          "availability":1,
          "standardName":"ISO 19115:2003\/19139", "standardVersion":"1.0"},
     "qualityInformation":{
          "availability":1,
          "scopeLevel":"dataset level"},
     "userFeedback":{
          "availability":0,
          "feedbacksCount":0, "ratingsCount":0, "feedbacksAverageRating":0},
     "expertReview":{
          "availability":0,
          "expertReviewsCount":0, "expertRatingsCount":0,
          "expertAverageRating":0},
     "citations":{
          "availability":1,
          "citationsCount":5
}}}
```


Figure 7.3: GEO label JSON summary.

Such JSON summary information can, for instance, be used to construct a GEO label representation using JavaScript client-side technology.

The drilldown resource (see Section 7.4.3) returns a styled Web page in an HTML format.

7.4 GEO Label Service API

The GEO label API provides a simple interface for generating GEO label representations and accessing styled drilldown HTML pages. The API uses GET and POST functionality of the HTTP protocol to create and retrieve GEO label resources. In this case, a resource can be either a GEO label representation or an HTML page. The following sections describe the GEO label API endpoints – the addresses or connection points to the web service – that return various GEO label resources.

7.4.1 GEO Label SVG Resources

The dynamic GEO label SVG resource can be obtained via the /geolabel endpoint (available at http://www.geolabel.net/api/v1/geolabel) using HTTP GET and POST methods. HTTP

GET method should be used when the producer and feedback XML documents are accessible via publicly available URLs.

Table 7.1 provides a list of parameters that can be specified as part of the GET request. The metadata, feedback, parent_metadata, and parent_feedback parameters can be used to specify encoded URLs of producer metadata, feedback, parent dataset's metadata, and parent dataset's feedback XMLs respectively. While the metadata and feedback parameters are optional, at least one of these parameters must be provided to obtain a GEO label SVG. The size parameter can be used to specify the diameter of the returned GEO label SVG in pixels.

Table 7.1: GET /geolabel resource parameters for obtaining SVG GEO label representations.

Parameters	Data Type	Description
metadata* optional	URL	Encoded URL of producer metadata document. Metadata document must be ISO19115, ISO19157, FGDC or GVQ-PQM compliant.
feedback* optional	URL	Encoded URL of feedback document. Feedback document must be GVQ-UQM compliant.
parent_metadata optional	URL	Encoded URL of parent dataset's producer metadata document. Metadata document must be ISO19115, ISO19157, FGDC or GVQ-PQM compliant.
parent_feedback optional	URL	Encoded URL of parent dataset's feedback document. Feedback document must be GVQ-UQM compliant.
size optional	int	Required size of the GEO label SVG in pixels. If not specified, the default size is 200x200 pixels.
* Although optional, at least one of these parameters must be provided.		

When the producer and feedback XML documents are not available via publicly exposed URLs, the /geolabel endpoint (available at http://www.geolabel.net/api/v1/geolabel) can be used to upload locally-available XML files through a POST request.

Table 7.2 provides a list of parameters that can be specified as part of the POST request. The metadata, feedback, parent_metadata, and parent_feedback parameters can be used to upload producer metadata, feedback, parent dataset's metadata, and parent dataset's feedback XML files respectively. While the metadata and feedback parameters are optional, at least one of these parameters must be provided to obtain a GEO label SVG. Similar to the GET request, the size parameter can be specified to define the diameter of the returned GEO label SVG in pixels.

Table 7.2: POST /geolabel resource parameters for obtaining SVG GEO label representations.

Parameters	Data Type	Description
metadata optional	File	Producer XML metadata document. Metadata document must be ISO19115, ISO19157, FGDC or GVQ-PQM compliant.
feedback optional	File	Feedback XML document. Feedback document must be GVQ-UQM compliant.
parent_metadata optional	File	Parent dataset's producer XML metadata document. Metadata document must be ISO19115, ISO19157, FGDC or GVQ-PQM compliant.
parent_feedback optional	File	Parent dataset's feedback XML document. Feedback document must be GVQ-UQM compliant.
size optional	int	Required size of the GEO label SVG in pixels. If not specified, the default size is 200x200 pixels.
* Although optional, at least one of these parameters must be provided.		

7.4.2 GEO Label JSON resources

The GEO label JSON resource can be obtained via the /facets endpoint (available at http://www.geolabel.net/api/v1/facets) using HTTP GET and POST methods. GET method should be used when the producer and feedback XML documents are accessible via publicly available URLs.

Table 7.3 details the metadata and feedback parameters that can be used to specify encoded URLs of producer metadata and feedback XMLs respectively. While these parameters are optional, at least one parameter must be provided to obtain a GEO label JSON representation.

Table 7.3: GET /facets resource parameters for obtaining JSON GEO label representations.

Parameters	Data Type	Description
metadata* optional	URL	Encoded URL of producer metadata document. Metadata document must be ISO19115, ISO19157, FGDC or GVQ-PQM compliant.
feedback* optional	URL	Encoded URL of feedback document. Feedback document must be GVQ-UQM compliant.
* Although optional, at least one of these parameters must be provided.		

The /facets endpoint (available at http://www.geolabel.net/api/v1/facets) can also be used to upload locally-available XML files through a POST request.

Table 7.4 details the metadata and feedback parameters that can be used to upload producer metadata and feedback XML files respectively. While these parameters are optional, at least one parameter must be provided to obtain a GEO label JSON representation.

Table 7.4: POST /facets resource parameters for obtaining JSON GEO label representations.

Parameters	Data Type	Description
metadata* optional	File	Producer XML metadata document. Metadata document must be ISO19115, ISO19157, FGDC or GVQ-PQM compliant.
feedback* optional	File	Feedback XML document. Feedback document must be GVQ-UQM compliant.
* Although optional, at least one of these parameters must be provided.		

7.4.3 GEO Label Drilldown Resource

Styled HTML representation of the supplied producer and feedback XML documents can be obtained via /drilldown endpoint (available at http://www.geolabel.net/api/v1/drilldown) using HTTP GET method.

Table 7.5 lists the parameters that can be specified as part of the GET request. The metadata and feedback parameters can be used to specify encoded URLs of producer metadata and feedback XMLs respectively; at least one of these parameters must be provided to obtain a GEO label SVG. The facet parameter must be provided as part of the GET request to specify the informational aspect for which the detailed information should be returned.

Table 7.5: GET /drilldown resource parameters for obtaining drilldown HTML pages.

Parameters	Data Type	Description
metadata optional	URL	Encoded URL of producer XML metadata document. Metadata document must be ISO19115, ISO19157, FGDC or GVQ-PQM compliant.
feedback optional	URL	Encoded URL of feedback XML document. Feedback document must be GVQ-UQM compliant.
facet required	string	GEO label facet identifier. Allowed values: ▪ producer_profile ▪ producer_comments ▪ lineage_information ▪ standards_compliance ▪ quality_information ▪ user_feedback ▪ expert_review ▪ citations_information

7.5 Hover-Over Functionality

Generated GEO labels offer dynamic hover-over functionality for obtaining quick summary information. Hovering over an individual facet in the GEO label displays a summary of the information related to the facet for the associated dataset – e.g., producer name, producer comments, the name of the standard to which the dataset complies, etc. (see Figure 7.4). Table 7.6 provides a list of hover-over text templates for each GEO label facet.

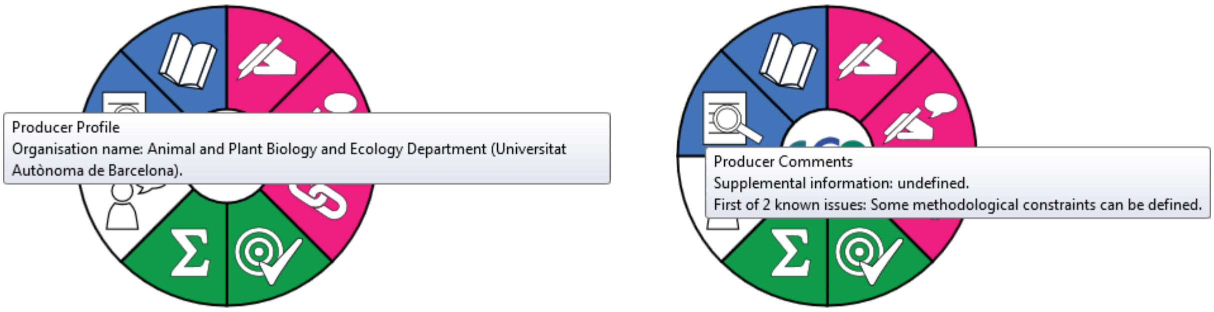


Figure 7.4: Examples of producer profile and producer comments hover-over functionalities.

Table 7.6: GEO label hover-over text templates.

Facet	Hover-over text template	Facet	Hover-over text template
	Producer profile Organisation name: […].		Quality information Quality information scope: […].
	Producer comments Supplemental information: […]. First of […] known problems: […].		User feedback Number of feedbacks: […]. Average rating: […] ([…] ratings).
	Lineage information Number of process steps: […].		Expert reviews Number of reviews: […]. Average rating: […] ([…] ratings).
	Standards Compliance Standard name: […], version […].		Citations information Number of citations: […].

To integrate the hover-over function into GEO label SVG representations, a `<title>` element was used which is widely supported in modern browsers. When an SVG GEO label representation is constructed, a relevant `<title>` element is added to every label facet. The title is then shown as a tooltip text when the mouse pointer moves over the facet. Below is an example of the citations information title element:

```
<title>Citations Information Number of citations: 5.</title>
```

7.6 Drilldown Functionality

The drilldown GEO label function is designed to provide detailed structured information extracted from the associated dataset's metadata record when a facet is clicked. The GEO label service API is used to transform producer metadata and feedback XML documents into styled structured HTML pages (see Section 7.4.3). Figure 7.5 provides an example of a citations information summary page that was generated using the GEO label drilldown function.

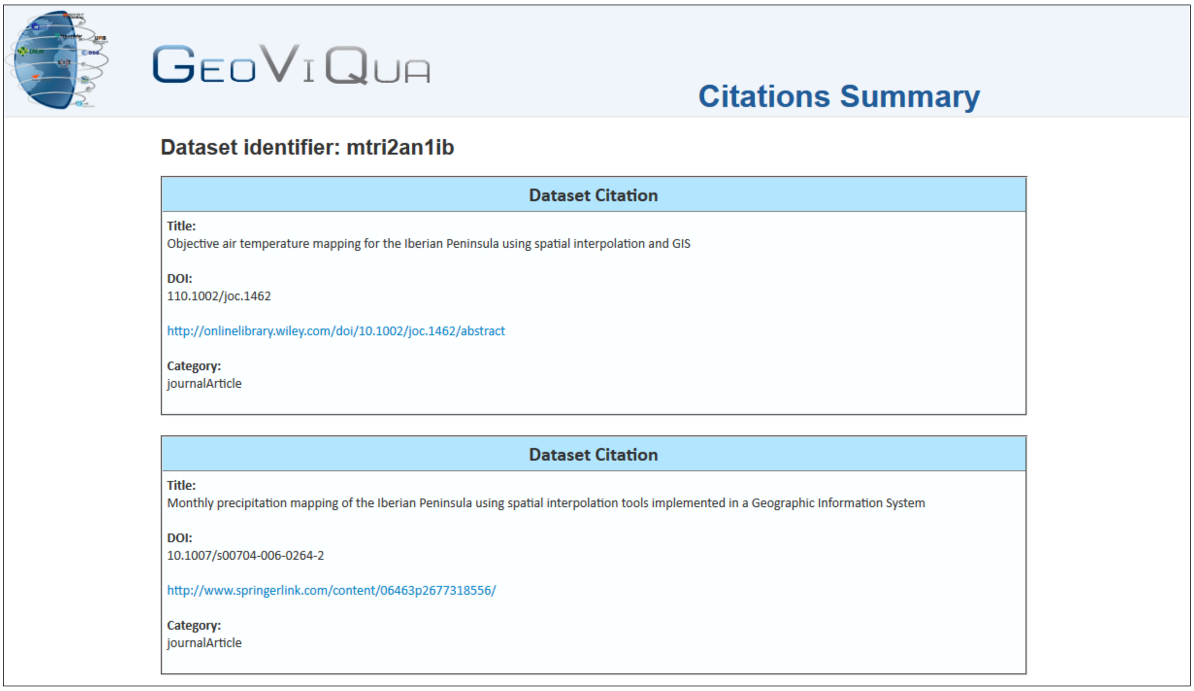


Figure 7.5: Example of citations information drilldown page.

In the GEO label SVG representation, the drilldown functionality is integrated using anchor `<a>` HTML element and XLink (XML Linking Language) `href` attribute. Each GEO label SVG facet contains an external link to its corresponding drilldown page. Below is an example of the citations information drilldown link:

```
<a xlink:href="http://www.geolabel.net/api/v1/drilldown?metadata
=[URL]&facet=citations_information" target="_blank">
```

7.7 GEO Label Service Implementation

To enable the rapid prototyping and development of a RESTful web-service, Object-Oriented (OO) PHP and the open-source PHP micro-framework Silex (Sensio Labs, 2011) were used to implement the GEO label service. The choice of using a framework was an important decision early on: since Silex is comprised of components used in an industry-standard enterprise-level framework called Symfony2 – also developed by Sensio Labs – it allowed for development to be focused on business logic central to the GEO label API, rather than implementation details specific to how a RESTful API should behave and perform correctly. Additionally, Silex encourages code to be written to the PSR-0 (class autoloading) and PSR-2 (coding style) standards defined by the PHP Framework Interop Group. Given the open-source nature of this research project, Silex's lightweight nature and the embracing of these standards was crucial to ensure that other developers outside the project could reuse components of the API in their own code (such as generating a GEO label SVG) or host and maintain the service themselves.

The core functionality of the GEO label service comprises five PHP classes, namely: `XMLProcessor`, `MappingsProcessor`, `SVG`, `SVGParser`, and `Drilldown`.

Figure 7.6 provides an example of the sequence of interactions among the GEO label service objects when an HTTP GET request is made to obtain an SVG GEO label representation. Here, the Silex controller first sends a call to the `XMLProcessor` object to fetch and parse metadata XMLs from the URLs supplied in the request query string. The `XMLProcessor` object combines the obtained XMLs into a single XML DOMDocument object which is then returned to the controller. A sequence of calls is then made to the `MappingsProcessor` object to obtain availability, hover-over and drilldown arrays. Once obtained, these arrays of information are passed to the `SVGParser` object which builds an SVG String using the components from the `SVG` object. Finally, the controller transforms the SVG String into an image/svg+xml format and returns the resulting SVG 1.1 object in an HTTP response.

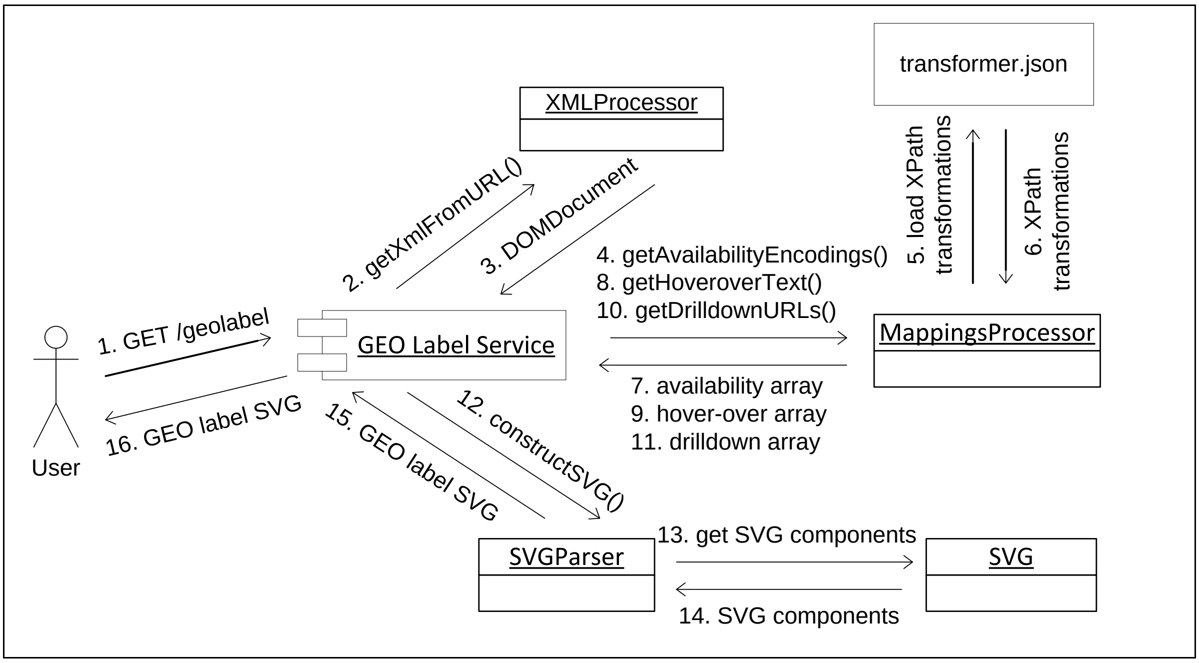


Figure 7.6: Interaction among the GEO label service objects when a GET GEO label request is received.

`XMLProcessor` is a small class that provides functionality for loading XML documents from publicly accessible URLs, validating these documents, and converting them into PHP DOMDocument objects for easier XML manipulation and processing. This class also offers functionality for joining two XML documents, in this case producer and feedback metadata XMLs, into a single DOMDocument object to allow for more efficient evaluation of information availability.

The `MappingsProcessor` class is responsible for loading XPath transformations from an external configuration file (discussed in section 7.2) and applying these transformations to the combined producer and feedback metadata XML to evaluate information availability. The class provides four main functions: `getAvailabilityEncodings`, `getHoveroverText`, `getDrilldownURLs`, and `getJsonDatasetSummary`. The `getAvailabilityEncodings` function accepts a combined producer and feedback metadata XML DOMDocument and a combined parent dataset's producer and feedback metadata XML DOMDocument and returns an associative array of information availability where facet names act as keys and integers 0, 1, and 2 specify the values for the '*not available*', '*available'*, and '*available at a higher level*' availability states respectively. To obtain the availability encodings of each facet, this function passes the supplied XML documents and an appropriate XPath transformation to the `evaluateAvailability` function. The `evaluateAvailability` function first applies the XPath transformation to the dataset's XML and if the transformation evaluates to TRUE the function returns integer 1 indicating that the information is available for the dataset. If the XPath transformation evaluates to FALSE and no parent XML is supplied, the function returns integer 0 indicating that the information is not available. If the parent XML is available, the XPath transformation is applied to the parent XML document and integer 2 is returned in case where the transformation evaluates to TRUE or 0 is returned otherwise.

Using XPath expressions and templates from the configuration file, the `getHoveroverText` function provides functionality for extracting snippets of information from the supplied XML documents and constructing hover-over text for each GEO label facet. The function returns an associative array of hover-over text where facet names act as keys and hover-over texts specify the values. In cases where producer and/or feedback XML documents are supplied via publicly accessible URLs, the GEO label can support drilldown functionality (discussed in Section 7.6). The `getDrilldownURLs` function is then used to construct drilldown URLs for each GEO label facet. As with the previously described functions, this function returns an associative array of drilldown URLs where facet names act as keys and drilldown URLs specify the values. Finally, the `getJsonDatasetSummary` function is responsible for producing a JSON summary (see Section 7.3) which contains all the essential GEO label information extracted from the supplied XML documents.

The `SVG` class provides essential components for constructing dynamic SVG GEO label representations; as such, each of the eight GEO label facets are represented as 3 separate SVG components to convey different availability states (available, not available, and available at a higher level) (see Figure 7.7). These facets' SVGs are stored as simple Strings that encode the facets' shape coordinates, outline colour, outline width, fill colour, etc. (see Figure 7.8 for an example encoding). These String representations can be concatenated to construct a required GEO label graphic.

available   not available   available at a higher level

Figure 7.7: The SVG representations of the quality information facet conveying 3 availability states.

```
private $quality_information_available = '
<path fill="#0F9B49" stroke="#000000" stroke-width="2" stroke-miterlimit="10"
      d="M98.128,152.0211- 59.795,59.794  c24.018,24.019,55.496,36.027,86.974,
      36.027V163.28C115.47, 163.28, 105.633,159.527,98.128,152.021z"/>
       […]
';
private $quality_information_not_available = '
<path fill="#FFFFFF" stroke="#000000" stroke-width="2" stroke-miterlimit="10"
      d="M98.128, 152.0211-59.795,59.794 c24.018,24.019,55.496,36.027,
      86.974,36.027V163.28C115.47, 163.28,105.633,159.527,98.128,152.021z"/>
       […]
';
private $quality_information_higher_level = '
<linearGradient id="SVGID_5_" gradientUnits="userSpaceOnUse" x1="77.8032"
      y1="238.4434" x2="115.0769" y2="158.5097">
       […]
';
```

Figure 7.8: An example PHP script for storing the quality information SVGs as Strings.

The `SVG` class also provides various helper functions that return such SVG components as: the root element which contains version, namespaces, label ID, size, and other essential SVG attributes; SVG elements for grouping various components, for example, to define a clickable area; the SVG representation of the GEO branding; and the GEO label facets. Figure 7.9 presents an example of a `getFacet` function which accepts a facet name, availability state (as an integer number), hover-over text, and a drilldown URL and, based on the supplied arguments, produces a dynamic SVG component – a GEO label facet – with embedded hover-over text and a link to the appropriate drilldown URL.

```
public function getFacet($class, $availability, $drilldownText, $drilldownURL){
  // prepare URL for XML
  $drilldownURL = str_replace('&', '&amp;', $drilldownURL);
  […]
  $facetSVG = '';
  switch($class){
    case 'producer_profile':
    if($availability === 0){ $facetSVG = $this->producer_profile_not_available; }
    […]
   $anchorOpeningTag = '<a xlink:href="'.$drilldownURL.'" target="_blank">';
   $anchorClosingTag = '</a>';
   if(empty($drilldownURL)){ $anchorOpeningTag = ''; $anchorClosingTag = ''; }
   return '<g class="'.$class.'" id="'.$class.'">
           <title>'.$drilldownText.'</title>'.
           $anchorOpeningTag . $facetSVG . $anchorClosingTag . '</g>';
   }
```

Figure 7.9: Function `getFacet` for obtaining a dynamic SVG GEO label facet.

The `SVGParser` class provides functionality for constructing an integrated SVG GEO label representation from the components provided by the `SVG` class. The class is lightweight and only consists of one function – `constructSVG` – which accepts a required label size in pixels and three associative arrays of (1) information availability, (2) hover-over text, and (3) drilldown URLs generated using the `MappingsProcessor` class and, based on the supplied arguments, builds a String representation of the GEO label SVG (see Figure 7.10).

```
public function constructSVG($availabilityArray, $hoveroverTextArray,
$drilldownURLsArray, $size){
      […]
      $labelSVG = $this->svg->getHeader($size);
      $labelSVG .= $this->svg->getFacet('producer_profile',
                                        $availabilityArray['producerProfile'],
                                        $hoveroverTextArray['producerProfile'],
                                        $drilldownURLsArray['producerProfile']);
      […]
      $labelSVG .= $this->svg->getBranding();
      $labelSVG .= $this->svg->getFooter();
      return $labelSVG;
}
```

Figure 7.10: Function `constructSVG` for constructing an integrated SVG GEO label representation.

Finally, the `Drilldown` class provides one main function – `getDrilldown` – for generating the GEO label drilldown pages (discussed in Section 7.6). The function accepts producer XML, feedback XML and XSL documents in a DOMDocument format and, by applying the XSL Transformations (XSLT) to the supplied XMLs, generates a structured HTML page (see example code in Figure 7.11).

```
public function getDrilldown($producerXML, $feedbackXML, $xsl){
      […]
      $gvqXML = $xmlProcessor->joinXMLDoms($this->updateNamespaces($producerXML),
                                     $this->updateNamespaces($feedbackXML));
      […]
      $xsltProcessor = new XsltProcessor();
      $xsltProcessor->importStylesheet($xsl);
      // transform the XML into HTML using the XSL file
      if($html = $xsltProcessor->transformToXML($gvqXML)) { return $html; }
      else {
      // If no document is supplied, return an empty styled page by default
           $dom = new DOMDocument('1.0', 'UTF-8');
           return $xsltProcessor->transformToXML($dom);
      }
}
```

Figure 7.11: Function `getDrilldown` for generating GEO label drilldown pages.

The Silex micro-framework is utilised to define routes and the controllers containing business logic that should be called when a particular request matches an HTTP verb and resource, adhering to RESTful principles. Figure 7.12 provides an example of a GET route that matches an HTTP request to obtain an SVG GEO label representation.

```
$app->get('/api/v1/geolabel', function(Request $request) use ($app) {
       $metadataURL = $request->query->get('metadata');
       $feedbackURL = $request->query->get('feedback');
       $parentMetadataURL = $request->query->get('parent_metadata');
       $parentFeedbackURL = $request->query->get('parent_feedback');
       $size = $request->query->get('size');
       […]
       $svgParser = new SVGParser();
       $svg = $svgParser->constructSVG($availabilityArray, $hoveroverTextArray,
       $drilldownURLsArray, $size);
       return new Response($svg, 200, array('Content-Type' => 'image/svg+xml'));
});
```

Figure 7.12: A GET route that matches an HTTP request for obtaining an SVG GEO label representation.

Here the controller gets the metadata and feedback URLs from the query string included in the HTTP request along with other necessary parameters. These XML documents are then fetched and parsed by the `XMLProcessor` (omitted), with the resulting mapping used in the generation of a GEO label SVG.

7.8 GEO Label Service Website

To promote geospatial community awareness of the GEO label and the service capabilities, a GEO label service website was developed using simple HTML pages and a Bootstrap framework (version 2.3.2). The website has been deployed as part of the GEO label service and was made available online at http://www.geolabel.net/home.html in July 2013. The website provides documentation for the service API and also offers simple demo pages that can be used to test-drive the service functionality. The API demo pages offer capability to generate sample GEO label SVGs by supplying metadata URLs or uploading metadata XML documents and also to view the GEO labels for the example metadata documents (see Figure 7.13).

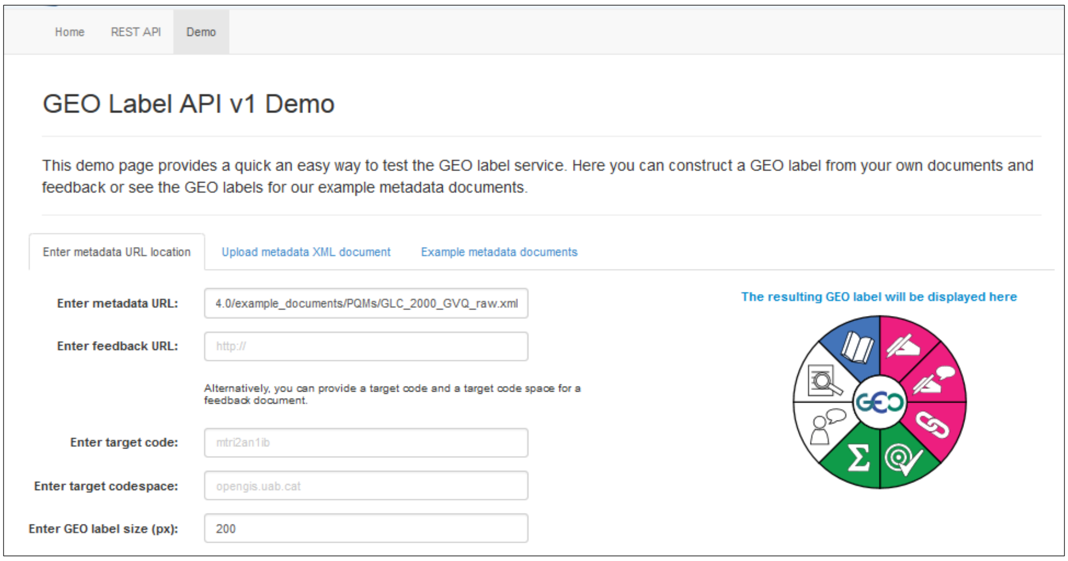


Figure 7.13: GEO label API demo page.

7.9 Summary and Conclusions

The GEO label web service was initially developed as a proof of concept to demonstrate that practical implementations of the GEO label are possible. From there, the prototype has evolved into a stable stand-alone Web-based server-side application which, as will be shown in the following chapters, is now being actively used in the GIS domain to integrate the GEO label into geospatial data portals and applications.

As discussed in Section 2.6.2, the EGIDA project suggested that (a) it is unlikely that a voluntary label would be able to fulfil community requirements and deliver desired results, and (b) that an external certification body is needed to independently evaluate the quality of geospatial datasets. The practical implementation of the GEO label has, however, demonstrated that it is possible to provide a dynamic voluntary quality and trust label without having to establish dedicated standardisation bodies or certification programmes. In addition, this practical GEO label implementation not only fulfils the needs of the geospatial community (as identified in this research), but also reflects the STC's initial vision that a GEO label should comprise two functions: objective labelling (quality, reliability); and subjective labelling (relevance, usability) (see Section 2.6.2). Within the developed GEO label, the objective labelling is supported by producer metadata, expert reviews and citations information and the subjective labelling is based on user feedback and dataset ratings. As demonstrated by this practical implementation, producer metadata documents can, in practice, be effectively combined with user feedback to generate an integrated visualisation of a user-focused summary of geospatial dataset quality and trustworthiness.

The practical implementation has additionally confirmed the feasibility of not only the drilldown GEO label function for obtaining detailed dataset information, but also the hover-over function for viewing a quick quality summary. The SVG format of the GEO label representation allows for integration of the essential dataset quality information and ensures label interactivity. Regarding the technological side of the GEO label implementation, the GEO label has been realised as a RESTful web-service with a publicly available API which means that it is technology agnostic and interoperable. Essentially, since every programming language has HTTP libraries available, as long as a GIS tool can send an HTTP request it can embed a GEO label directly within itself. The service is reusable and, because it is a stand-alone Web-based application, it is possible to scale it up with future growth in demand or if more functionality is required. The GEO label service API is so lightweight and well documented that it is a low barrier to entry. The key advantage of the service is in its interoperability – it allows for the GEO label to be integrated within any GIS application that supports HTTP requests.

Chapter 8 GEO Label-Based Dataset Intercomparison and Decision Support System

As discussed in Section 2.6.1, with the growth in availability of geospatial data, exploration of large multi-attribute geospatial data databases is becoming increasingly challenging, particularly for less-experienced geospatial data users. Studies conducted as part of this research (see Chapters 3, 4, and 5) demonstrated that dataset users consider many different aspects when selecting a dataset to use; the studies also revealed that users often have to manually inspect the dataset data and metadata to decide on fitness for intended purpose. Although lack of complete metadata records still presents a major barrier to effective evaluation of geospatial data quality and trustworthiness, it is anticipated that, with the development of new tools to support more efficient recording of data quality, including automated metadata generation, and provision of feedback tools for geospatial datasets, the generation of more descriptive and informative metadata documents is now possible. To effectively assess and intercompare large quantities of such complex geospatial dataset metadata records while avoiding information overload, geospatial data users will require innovative dataset discovery tools and decision support systems. While geospatial data portals and clearinghouses do offer search facilities to retrieve individual datasets (e.g., search by region, keywords, type of data, date when data was collected), and this allows a search to be filtered according to a potentially complex set of analysis requirements, search-by-quality is not currently available.

To explore novel approaches of visualising metadata records and intercomparing large numbers of datasets, a GEO label-based dataset discovery tool has been designed and developed as a prototype online system that supports geospatial dataset intercomparison and selection. Built on the principles of starfield displays (see Section 2.5.3), the system represents geospatial datasets' metadata records as GEO labels and allows dataset filtering based on the informational aspects' availability recorded in eight GEO label facets. This novel approach of visualising metadata records allows for a more efficient evaluation of datasets' fitness for purpose and enables 'at a glance' intercomparison of large numbers of datasets, which is not currently possible when using traditional dataset cataloguing systems and data discovery portals.

The remainder of this chapter describes the interface design, functionality and implementation of the prototype GEO label-based dataset discovery and intercomparison tool – GEO LINC (GEO Label INterComparison tool). Section 8.1 overviews the interface design and functionality of the prototype tool: it describes functionality provided to support (a) searching geospatial datasets, (b) dynamic filtering of search results, (c) obtaining detailed information about a dataset, and (d) highlighting datasets of interests for later reference and dataset intercomparison. Implementation of the GEO LINC tool is outlined in Section 8.2; the section describes the server-side and client-side technologies used in the development of the prototype system. Finally, a chapter summary is provided in Section 8.3.

8.1 Interface Design and Functionality

The interface design and functionality of the GEO LINC tool were influenced by two data search and exploration systems – the FilmFinder application (see Figure 8.1) and the EuroGEOSS brokering tool (see Figure 8.2). The FilmFinder application was an appropriate choice for the design inspiration because it enables visualisation of thousands of entities within a single search space, visualises data as simple graphics, and offers a dynamic querying system to filter search results. The EuroGEOSS brokering tool, on the other hand, is designed specifically for geospatial data discovery and incorporates specific geospatial data query elements. Unlike common geospatial data catalogues, clearinghouses and portals (e.g., GeoNetwork-based catalogues, GEOSS data portal (http://www.geoportal.org/web/guest/geo_home_stp), ESA data portal (https://earth.esa.int/web/guest/data-access/browse-data-products), NASA data portal (http://gcmd.nasa.gov/KeywordSearch/Home.do?Portal=NASA&MetadataType=0), etc.), the EuroGEOSS brokering tool provides the look and feel of an integrated standalone desktop application rather than a website. Due to this key difference, the interface design of the EuroGEOSS brokering tool was selected as a model when developing a prototype decision support tool for geospatial dataset discovery and selection based on GEO-labels.

Aston University

Illustration removed for copyright restrictions

Figure 8.1: The FilmFinder application for exploring film databases (Waloszek, 2013).

Aston University

Illustration removed for copyright restrictions

Figure 8.2: The EuroGEOSS geospatial data brokering tool (GEO, 2014a).

The FilmFinder application influenced the visualisation and filtering of the search results (see Figure 8.1). Similar to FilmFinder, the search results area of the GEO label-based tool resembles a starfield display (interactive scatterplot) with the difference being that this tool does not employ the use of a scale on the x- and y-axes to order the search results. The use of sliders to dynamically filter search results was also influenced by the FilmFinder application.

In terms of the search results ordering, there were several options for utilising the search area x- and y-axes – for instance, the results could be ordered by information availability (number of informational aspects available) on x-axes and by popularity or expert acceptance (user or expert ratings) on y-axes (i.e., most popular datasets would appear at the top of the search area). As discussed in Section 6.2.8, the results of the Phase II study revealed that dataset ratings can be highly influential when selecting geospatial datasets; considering the subjective nature of geospatial data quality, such an influence was considered undesirable when developing a prototype decision support tool (i.e., it was not considered appropriate to base a visualisation axis on subjective ratings). Another option would be to order the results by either dataset creation (or updated) date or alphabetically by dataset title on x-axes and by information completeness on y-axes. In this case, the GEO labels with no information available would be at the bottom of the search area and the most complete GEO labels (i.e., with all 8 informational aspects available) would appear at the top. Alternatively, the results could be clustered (visually grouped together) by dataset producer and the resulting clusters could be ordered, for example, alphabetically by producer name.

While various search results ordering options could be implemented, after careful consideration, it was decided to focus on the overall feasibility and effectiveness of visualising a large number of geospatial datasets using the GEO labels. Additionally, the initial implementation of the GEO LiNC tool was aimed at facilitating the Phase III study to evaluate the effectiveness of the GEO label at supporting dataset intercomparison and selection. Considering a large number of filtering options provided in the initial prototype, it was anticipated that the use of x- and y-axes could potentially add another level of complexity increasing mental demand from Phase III study participants. The results in the search area are therefore ordered in a sequence in which they appear in the database. Although x- and y-axes are not utilised in the initial version of the GEO LiNC tool, the software design allows for easy extension to include the ordering functionality in the later versions. Some relatively minor modifications to the search results ordering algorithm would be required to employ the use of a scale on the x- and y-axes. The modifications would involve (a) getting the information required for the ordering from each search results GEO label (e.g., the number of informational aspects available and updated date), (b) calculating the x- and y-coordinates depending on the information extracted from the GEO label, and (c) rendering the GEO label graphic in the appropriate location of the search results area.

The geospatial data search parameters such as keywords, location, start and end dates, and access and use constraints (see Figure 8.2) were adopted from the EuroGEOSS brokering tool. In addition, the EuroGEOSS brokering tool influenced the overall look and feel of the GEO label-based tool, including use of tabs for separating different views and filtering options and use of OpenLayers JavaScript library (OpenLayers, 2014) for providing an interactive map.

The prototype tool has been developed to support four major functions to: (1) allow users to search geospatial datasets by defining initial search criteria; (2) offer an interactive and visual way of filtering metadata records that match initial user requirements; (3) facilitate obtaining detailed information about a dataset; and (4) enable highlighting of datasets of interests for later reference and dataset intercomparison. Each of these functions is described in more detail the following sections of this chapter.

8.1.1 Searching for Geospatial Datasets

The initial datasets search view (see Figure 8.3) provides an interface for discovering geospatial datasets by searching the metadata records available in a system catalogue. The *Query Constraints* tab is used to define the search criteria, allowing users to:

a) specify query keywords (e.g., cloud cover, precipitation, sea surface temperature, etc.);
b) specify required spatial coverage by either selecting a predefined location option or selecting a custom area using an interactive map;
c) specify required temporal coverage by selecting the start and end dates; and
d) select dataset access and use constraints.

After the search constraints have been defined and the search query has been submitted by clicking the *Search* button, the system returns the GEO label representations of all the datasets that match the defined search criteria (see Figure 8.4).

Figure 8.4 provides an example of the visualisation of 351 dataset metadata records (each represented via their corresponding GEO label) within a single query results area. Even with such a large number of search results being displayed at once, it is still possible to not only identify at-a-glance the most complete metadata records but also assess overall metadata completeness in the catalogue.

In the current version of the prototype tool, the zoom and pan control (see Figure 8.4, top right corner of the *Search Results* area) is not functional but has been added as a placeholder for future development.

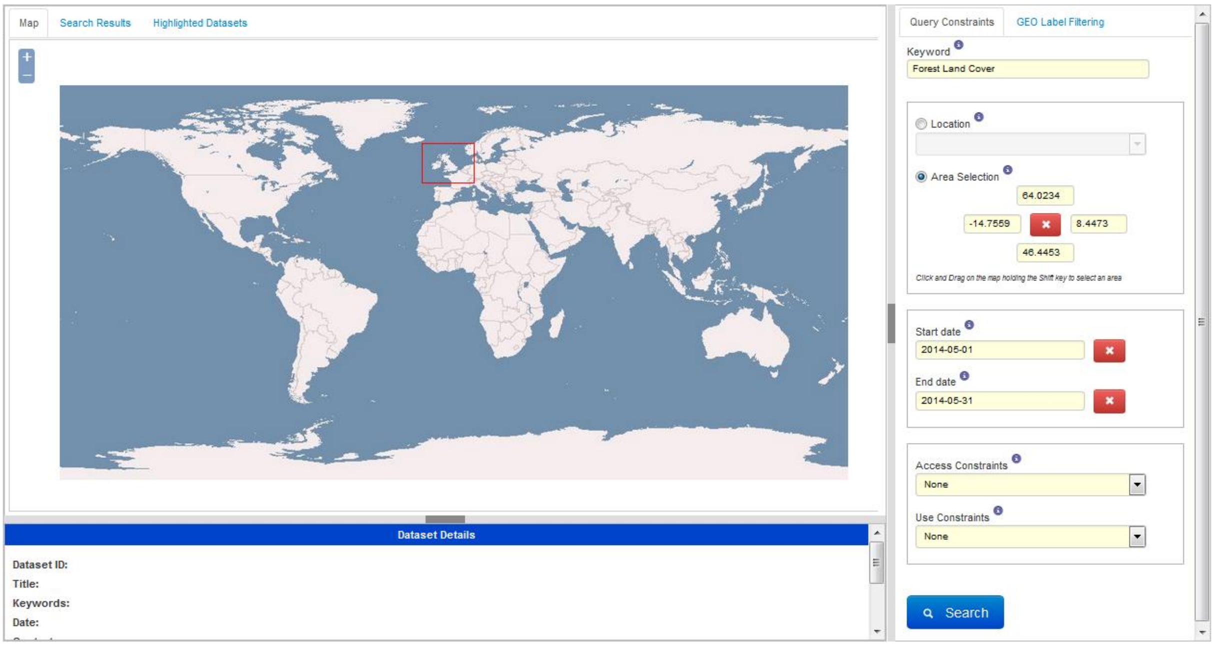


Figure 8.3: GEO LINC tool – initial dataset discovery view.

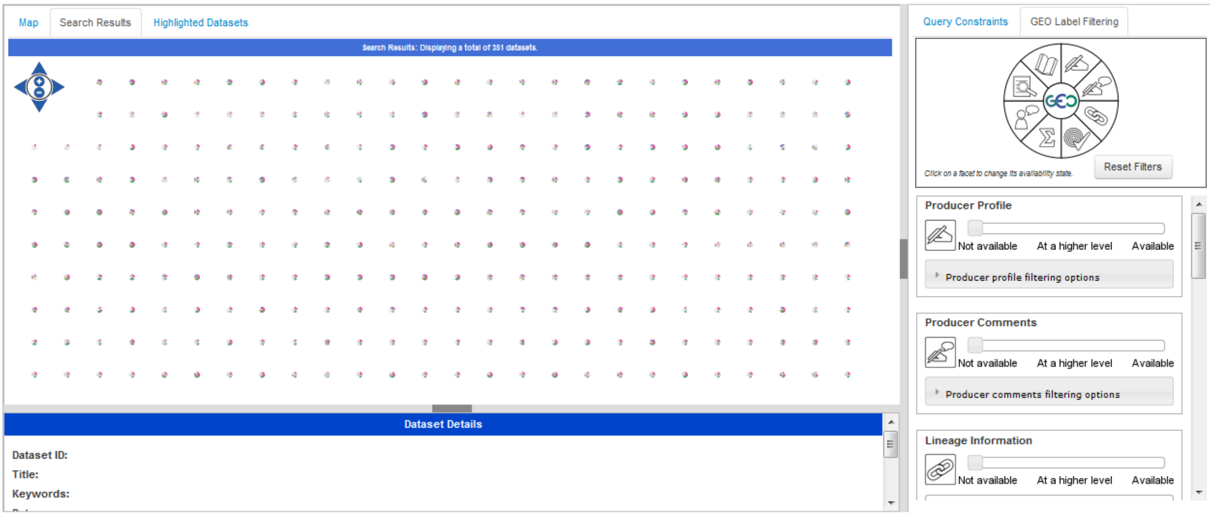


Figure 8.4: GEO LINC tool – search results view.

8.1.2 Filtering Search Results

When search results are displayed, the *GEO Label Filtering* tab allows a user to apply information availability filtering based on informational aspects' availability recorded in eight GEO label facets. Using either the interactive clickable label (located at the top of the *GEO Label Filtering* tab) or eight facet sliders (see Figure 8.5), users can select one of three availability states ('*available*', '*not available*' or '*available only at a higher level*') for each GEO label facet. For instance, a user might only be interested in datasets that have producer information and producer comments immediately available (i.e., available for the dataset itself and not just for its parent dataset); he/she would set producer profile and producer comments availability to '*available*' to filter out all the datasets that do not contain this information. Figure 8.6 presents an example with producer profile and producer comments filtering being set to the '*available'* state. As can be noted from the example, when facet filtering is applied the GEO labels that do not match the specified availability state are removed from the search results (leaving gaps in the starfield display). Such interactive filtering enables users to quickly narrow down their search to a manageable number of datasets that they can then inspect in more detail.

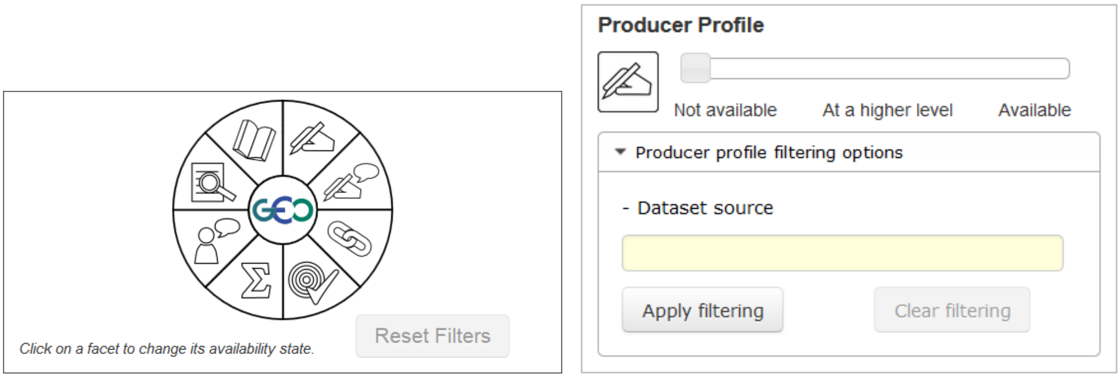


Figure 8.5: The GEO LINC tool – (left – interactive clickable label, right –producer profile facet slider and dataset source filtering).

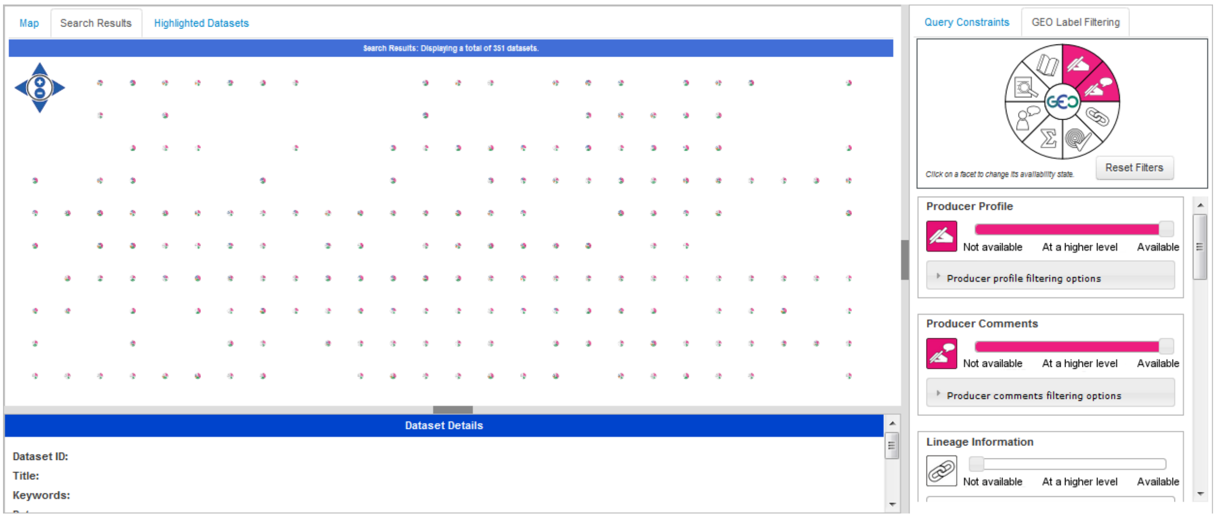


Figure 8.6: The GEO LINC tool – producer profile and producer comments filters applied.

The tool also offers *additional* filtering options for every GEO label informational aspect (see Figure 8.7). These additional filtering options allow users to specify:

a) dataset source, i.e., name of the dataset producer;
b) producer comments' type (supplemental information, known problems, or both supplemental information and known problems);
c) minimum and maximum number of process steps that have been applied to the data;
d) name of the metadata standard to which the dataset complies;
e) quality information scope (dataset or pixel level);
f) average user rating and minimum number of user feedbacks;
g) average expert rating and minimum number of expert reviews; and
h) minimum number of citations which refer to the dataset.

Unlike the information availability filtering described above, when applied the additional filtering options do not remove the GEO labels that do not match the specified criteria. These filtering options alter the GEO labels' size to indicate most relevant datasets. Figure 8.7 presents an example of additional filtering being applied. As can be noted from the example, the datasets' label representations that match the specified filtering criteria are larger in size.

Figure 8.7: The GEO LINC tool – dataset source and comments type filters applied.

8.1.3 Obtaining Detailed Information about a Dataset

The tool also allows users to inspect detailed information about the datasets of interest. For instance, the dataset's ID can be inspected by hovering over its GEO label representation and full information about a dataset can be accessed by clicking on its GEO label representation (see Figure 8.8). When selected (clicked on), the GEO label is highlighted and the dataset's title, abstract, producer details, link to a full metadata record, etc. are displayed in the *Dataset Details* section of the discovery tool.

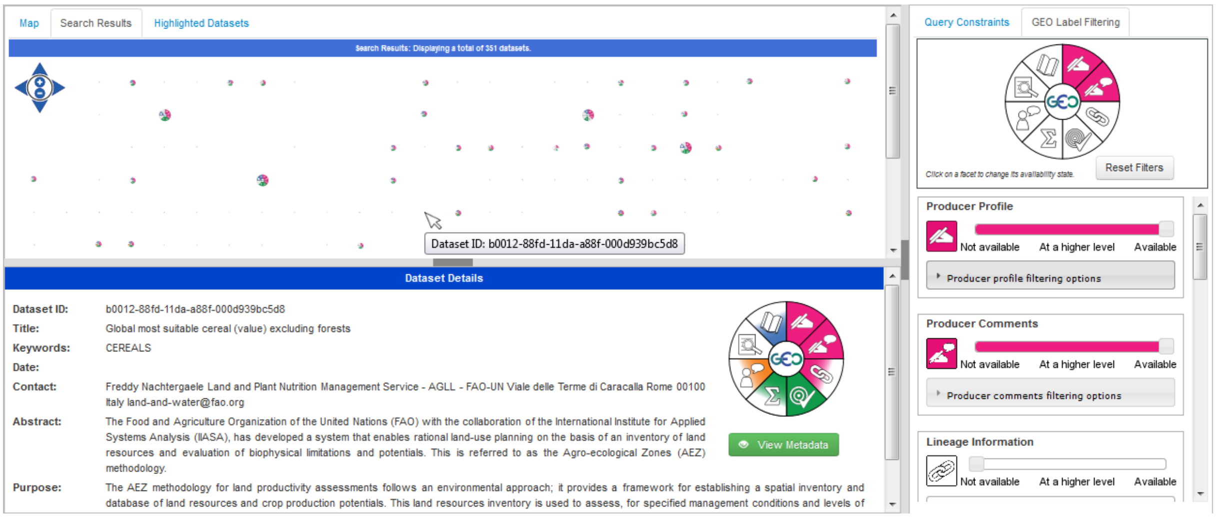


Figure 8.8: The GEO LINC tool – dataset details are displayed.

The *Dataset Details* section also provides an enlarged representation of the selected dataset GEO label which can be used to obtain further details about the dataset it represents (see Figure 8.8). Hovering over each facet of the enlarged dataset label displays a facet summary – e.g., name of the dataset producer, producer comments, number of process steps applied to the data, etc. (see Section 7.5 for more information on hover-over functionality). The enlarged dataset label additionally supports the drill-down GEO label function (see Section

7.6 for more information on drilldown functionality) – i.e., when a facet is clicked, styled structured information extracted from the dataset's metadata record is displayed in a new browser window. Figure 8.9 shows a user feedback summary displayed after the *user feedback* facet was clicked.

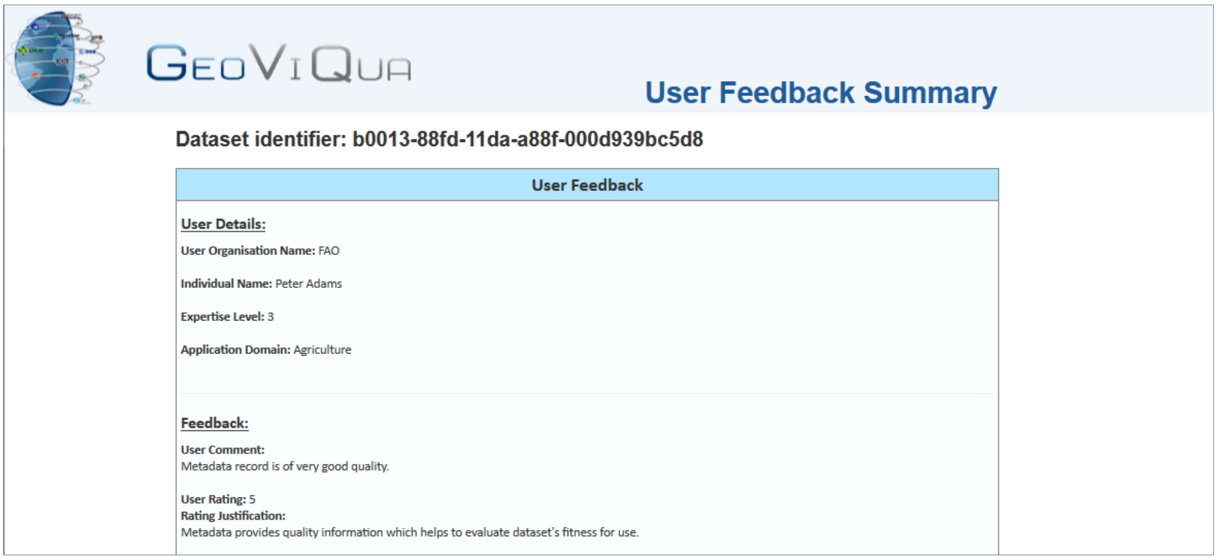


Figure 8.9: The GEO LINC tool drilldown functionality – user feedback summary.

8.1.4 Highlighting Datasets of Interest

Specific datasets of interest can be highlighted for later reference by right-clicking on their GEO label representations and selecting the *Highlight* option (see Figure 8.10). An outer glow is applied to the highlighted GEO labels for a visual distinction.

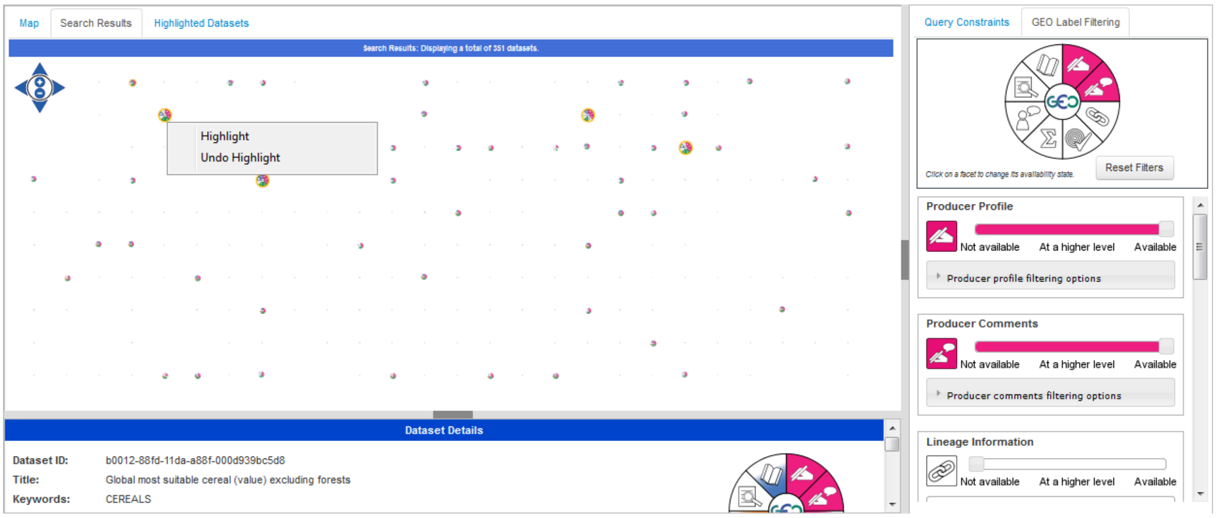


Figure 8.10: The GEO LINC tool – highlighting a dataset of interest.

When a dataset is highlighted, its detailed information is displayed in the *Highlighted Datasets* tab (see Figure 8.11). This allows users to keep track of datasets that closely match their requirements and also supports datasets intercomparison. The highlighted datasets can be removed from the highlighted list at any time by either clicking on a *Remove from List* button in the *Highlighted Datasets* tab or by right-clicking the dataset's label representation in the *Search Results* tab and selecting *Undo Highlight*.

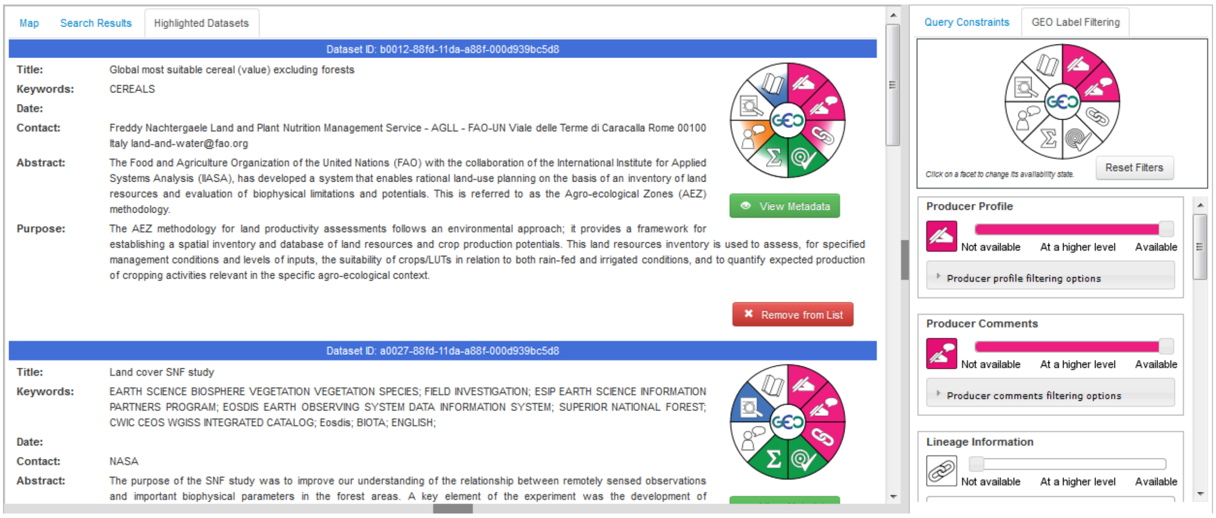


Figure 8.11: The GEO LINC tool – highlighted datasets.

8.2 Implementation

The GEO label-based dataset discovery tool has been implemented as a standalone, Web-based system using both client-side and server-side technologies. On the client-side, HTML, CSS and JavaScript have been used to provide an interactive responsive UI and, on the server-side, PHP and MySQL have been used to store, retrieve and process system data. Most of the GEO label-based tool functionality has been provided using the client-side technologies to minimise communication with the server and hence allow for improved interactivity by immediately responding to users' actions. The following sections describe the server-side and the client-side implementations in more detail.

8.2.1 Server-Side Implementation

The main role of the server-side implementation is to provide a geospatial data catalogue with functionality for storing, retrieving and processing of XML metadata records that describe geospatial datasets. To ensure rapid processing of search queries, the system catalogue has been developed using (a) a MySQL database that stores key dataset information and (b) a file system that stores physical XML metadata records. The database is fairly simple and consists of one `query_constraints` table which contains the following fields: `dataset_id`; `keywords`; `start_date`; `end_date`; `latitude_north`; `latitude_south`; `longitude_west`; `longitude_east`; `access_constraints`; and `use_constraints`. The unique dataset identifiers stored in the database (`dataset_id` field) are used as the filenames for the corresponding physical XML documents in the file system; for instance, a database record with a dataset ID `a0001-88fd-11da-a88f-000d939bc5d8` matches XML file `a0001-88fd-11da-a88f-000d939bc5d8.xml` stored in the file system directory. The unique dataset identifiers are also recorded in each XML metadata document in the `gmd:fileIdentifier` element. In case of traditional geospatial data portals and clearinghouses, the data catalogues are often represented by either lookup tables (e.g., XML or Json files) or file systems of all the available metadata records: for the latter, powerful commercial XML processors are used to dynamically parse all the catalogued XML documents. While it is appreciated that these approaches represent commercial standards, for the purpose of the prototype development the catalogue implementation was kept as simple as possible to minimise the development time while ensuring acceptable prototype performance.

The functionality of the server-side implementation is provided via three main PHP scripts: `process_search_request.php`; `process_details_request.php`; and `facet.php`. The `process_search_request.php` script provides functionality for acquiring JSON summaries (see Section X) of all the geospatial datasets in the system catalogue that match the search query. The script accepts a POST request with the parameters that correspond to the database fields in the `query_constraints` table (dataset id, keywords, etc.) and returns a JSON object with an array of JSON dataset summaries that can be transformed into equivalent SVG GEO label representations. When a POST request is passed to the script, the query parameters are extracted from the PHP $_POST variables (see Figure 8.12) and are used to construct a MySQL query for obtaining the unique identifiers (dataset IDs) of the datasets and their parent datasets that match the query (see Figure 8.13).

```
[...]
// -------- Get data from POST variables ------------
$keyword = $_POST['keyword'];
$startDate = $_POST['startDate'];
$endDate = $_POST['endDate'];
$accessConstraints = $_POST['accessConstraints'];
$useConstraints = $_POST['useConstraints'];
[...]
```

Figure 8.12: A PHP script for extracting search query parameters from the PHP $_POST variable.

```
[...]
$db = new mysqli($hostname, $user, $password, $database);
if($db->connect_errno > 0){
    die('Unable to connect to database [' . $db->connect_error . ']');
}
$keywordsString = "keywords LIKE '%%'";
       $keywordsString = implode(' OR ', $keywordItems);
[...]
$sql = "SELECT dataset_id, parent_id FROM query_constraints
             WHERE access_constraints LIKE '%$accessConstraints%'
             AND use_constraints LIKE '%$useConstraints%'
             AND (" . $keywordsString . ")";
[...]
```

Figure 8.13: A PHP script for obtaining the datasets that match user query from a MySQL database.

If successful, the query returns a resource with a list of dataset IDs and their parent IDs which are then used to load the corresponding metadata XML documents from the file system. Using the `getJsonDatasetSummary` function of the `xmlProcessor` PHP class, which has been developed as part of the GEO label service implementation (see Section 7.7), each dataset XML together with its parent XML are processed to produce a GEO label JSON summary that encodes information availability, hover-over text and drilldown URLs for every GEO label facet. While a series of GET requests to the GEO label service could be used to obtain the GEO label JSON summaries, this approach would significantly decrease the overall system performance. Consequently, relevant GEO label service PHP classes and scripts have been directly integrated as part of the GEO label-based tool functionality. When all the dataset XML documents are processed, the resulting summaries of all the datasets are combined into a single JSON String and returned to the client-side application (see Figure 8.14).

```
[...]
$jsonResponseString = '{"dataset": [';
[...]
// Get location of the metadata XML file
$metadataURL = 'metadata_records/' . $row['dataset_id'] . '.xml';
$parentURL = 'metadata_records/' . $row['parent_id'] . '.xml';
// Load metadata XML file
$metadataXML = new DOMDocument();
$metadataXML->load($metadataURL);
$parentXML = new DOMDocument();
$parentXML->load($parentURL);
$xmlProcessor = new xmlProcessor();
$json = $xmlProcessor->getJsonDatasetSummary($metadataXML, null, $parentXML);
[...]
$jsonResponseString .=  $json;
[...]
$jsonResponseString .= ' ]}';
echo $jsonResponseString;
[...]
```

Figure 8.14: A PHP script for producing a JSON summary of all the geospatial datasets that match user query.

The `process_details_request.php` script provides functionality for obtaining core information about a specific dataset in the system catalogue. The script accepts a POST request with a unique ID of the dataset of interest and returns a JSON String that contains the dataset title, keywords, creation date, producer contact details, abstract, and purpose. When a POST request is passed to the script, the dataset ID is extracted from the $_POST variable and is used to load the corresponding metadata XML document from the file system. A series of XPath expressions (see Figure 8.15) are then applied to the metadata XML document to extract required information. The resulting transformations are stored in an associative array which is transformed into JSON and returned to the client-side application (see Figure 8.16).

```
[...]
// XPath expressions:
$fileIdentifierXPath = '//*[local-name()=\'fileIdentifier\']/*[local-
name()=\'CharacterString\']';

$title = '//*[local-name()=\'title\']/*[local-name()=\'CharacterString\']';

$abstract = '//*[local-name()=\'abstract\']/*[local-name()=\'CharacterString\']';

$purpose = '//*[local-name()=\'identificationInfo\']//*[local-
name()=\'purpose\']/*[local-name()=\'CharacterString\']';

$date = '//*[local-name()=\'CI_Date\']/*[local-name()=\'date\']/*[local-
name()=\'Date\']';

$keywords = '//*[local-name()=\'descriptiveKeywords\']/*[local-
name()=\'MD_Keywords\']/*[local-name()=\'keyword\']/*[local-
name()=\'CharacterString\']';

$producerProfileXpath = '//*[local-name()=\'contact\']/*[local-
name()=\'CI_ResponsibleParty\'] | //*[local-name()=\'ptcontac\']/*[local-
name()=\'cntinfo\'] | //*[local-name()=\'pointOfContact\']/*[local-
name()=\'CI_ResponsibleParty\']';
[...]
```

Figure 8.15: An example of XPath expressions for extracting dataset details.

```
[...]
$summaryArray = array(
  'fileIdentifier' => getFirstNode($xml, $fileIdentifierXPath),
  'title' => getFirstNode($xml, $title),
  'abstract' => getFirstNode($xml, $abstract),
  'purpose' => getFirstNode($xml, $purpose),
  'date' => getFirstNode($xml, $date),
  'keywords' => getFirstNode($xml, $keywords),
  'producer' => getFirstNode($xml, $producerProfileXpath));
[...]
$json = json_encode($summaryArray);
if(!empty($json)){
       $jsonResponseString .=  $json;
}
$jsonResponseString .= ' }';
echo $jsonResponseString;
[...]
```

Figure 8.16: A PHP script for extracting dataset details.

Finally, the `facet.php` script provides functionality to support the drilldown GEO label function (see Section 7.6). This script reuses functionality of the `Drilldown` PHP class developed for the GEO label service (see Section 7.7).

8.2.2 Client-Side Implementation

The main role of the client-side implementation is to provide an interactive UI for searching, filtering and interrogating geospatial datasets (see Section 8.1.1, 8.1.2, 8.1.3, and 8.1.4 for more information about the corresponding interface design). The interface and functionality of the client-side implementation have been developed using HTML, CSS, JavaScript, JQuery, JQueryUI and JSON technologies. HTML has been used to define the structure and content of the tool web page; CSS has been used to define the style of content and interface components; and JavaScript has been used to support interactivity and to manipulate the web page via the DOM (Document Object Model) to create a rich web application. Bootstrap (Bootstrap, 2014) – an open source HTML, CSS and JavaScript framework – has been integrated as part of the development to utilise some of its HTML components (e.g., tabs, buttons, input boxes, dropdown lists, etc.), obtain customisable template CSS documents, and add basic JavaScript interactivity to the HTML components. The jQuery library (http://jquery.com/) had been utilised to provide a more powerful and intuitive DOM traversal and manipulation, event handling, and HTTP Ajax (Asynchronous JavaScript and XML) requests. The jQueryUI library (The jQuery Foundation, 2014) provided a set of UI effects, widgets and interactions to create a desktop-like application rather than a web page. Finally, the JSON format has been used to exchange data between the client-side and server-side applications through AJAX requests. The adjustable frames layout of the tool has been implemented using an open source jQuery UI Layout plug-in (http://layout.jquery-dev.net/index.cfm) in combination with jQuery UI functionality and custom CSS files.

The interface of the GEO label-based dataset discovery tool was constructed using one `index_tabs.html` HTML page which incorporates all the UI components and includes references to external CSS and JavaScript files. The styling and interactivity of the interface components is provided via two main custom JavaScript files: `dynamic-elements-styling.js` and `frames-layout.js`. When the tool is loaded, the `dynamic-elements-styling.js` script is used to initialise all the dynamic tool components; as such, the script adjusts the size of the frames depending on the screen size. sets content visibility, and enables appropriate input components and buttons. The script also provides functionality for resetting all the components, such as input boxes, star ratings, results area, etc. to their initial state and for dynamically showing and hiding the content when different content tabs are selected. The `frames-layout.js` script uses the jQuery UI Layout plug-in to set the layout of the content frames – it defines the position of the frames, their minimum and maximum size, and behaviour.

The core functionality of the client-side GEO label-based tool implementation is provided via four custom JavaScript files: `search-request.js`; `svg_facets.js`; `jquery_sliders.js`; and `filtering.js`. The `search-request.js` script provides functionality for constructing a dataset search query, sending the query to the server-side application, and converting the query results into dynamic SVG GEO label representations. When the *Search* button is clicked (see Figure 8.3), the script extracts values from the *Query Constraints* form, constructs a query string from the extracted values and sends an Ajax

POST request to the `process_search_request.php` server-side script (see Figure 8.17).

```
[...]
var keyword = $("input#keyword-autocomplete").val();
var startDate = $("input#start-date").val();
var endDate = $("input#end-date").val();
var accessConstraints = $("select#access-constraints").val();
var useConstraints = $("select#use-constraints").val();

var dataString = 'keyword='+ keyword + '&startDate=' + startDate + '&endDate=' +
endDate + '&accessConstraints=' + accessConstraints + '&useConstraints=' +
useConstraints;
[...]
// Send a POST request
$.ajax({
  type: "POST",
  url: "php/process_search_request.php",
  data: dataString,
  success: function(data){
    if(isJson(data)){
      var JSONObject = JSON.parse(data);
[...]
```

Figure 8.17: Example JavaScript for constructing and sending Ajax POST request.

If the search query is successful, the script performs a number of calculations to determine: the area available for displaying all the search results (this depends on the screen size); available area per label to define the size of the GEO labels and margins between the labels to allow for size alteration when the results are filtered (see Section 8.1.2); and the maximum number of GEO labels per results row (see Figure 8.18).

```
[...]
//1) get screen height and width
//2) available area = height * width
//3) area per label = available area / number of datasets
//4) label scale = area per label / (250 * 250)
var availableArea = searchAreaHeight * serachAreaWidth;
var areaPerLabel = parseInt(availableArea / JSONObject.dataset.length, 10);
var scale = parseFloat(areaPerLabel / 47000).toFixed(2);
var xOffset = parseInt((250 * scale) + 35, 10);
var yOffset = xOffset;
var maxLabelsPerRow = parseInt(serachAreaWidth / xOffset, 10);
[...]
```

Figure 8.18: Example JavaScript for determining distribution of the search results.

The GEO label descriptions (information availability encodings, hover-over text and drilldown URLs) of each dataset are then iteratively extracted from the JSON object returned by the server-side application (see Figure 8.19). Using functionality provided in the `svg_facets.js` script (see below), these descriptions are transformed into dynamic SVG DOM components which are then combined into a complete SVG GEO label representation.

When generated, each SVG is assigned a simple unique system identifier (an integer between 0 and the number of datasets returned by the query) which is used for dataset filtering purposes. Combined with the component name, this unique identifier is recorded in the `id` attribute of every SVG component: for instance, if a dataset was assigned ID '`1`' then the root SVG component will have ID `geolabel_1`, the producer profile facet will have ID `producer_profile_1`, the GEO branding component will have ID `branding_group_1`, and so on (see Figure 8.20). These unique component identifiers are used to obtain required information from the DOM elements, for example, the facet's availability encoding or hover-over text. The generated GEO label SVG representations are assigned x and y positions and are nested within a container SVG which is embedded into the `index_tabs.html` page and acts as a search results scatterplot (starfield) display. On completion, the *Search Results* area of the tool is updated and the resulting GEO label SVG representations are displayed to user.

```
[...]
// Process all JSON dataset objects and build GEO label representations
for (var i = 0; i < JSONObject.dataset.length; i++) {
  [...]
  var datasetID = JSONObject.dataset[i].datasetIdentifier;
  var parentID = JSONObject.dataset[i].parentIdentifier;
  var producerProfileAvailability = JSONObject.dataset[i].facets.
    producerProfile.availability;
  [...]
  var organisationName = JSONObject.dataset[i].facets.
    producerProfile.organisationName;
  var supplementalInformation = JSONObject.dataset[i].facets.
    produerComments.supplementalInformation;
 [...]
```

Figure 8.19: An example JavaScript for extracting dataset information from a JSON object.

```
[...]
for (var i = 0; i < JSONObject.dataset.length; i++) {
  [...]
  // Set GEO label svg
  var labelSVG = document.createElementNS ("http://www.w3.org/2000/svg" , "svg");
  labelSVG.setAttributeNS(null, "id", "geolabel_" + i);
  labelSVG.setAttributeNS(null, "class", "dataset_geolabel");
  labelSVG.setAttributeNS(null, "dataset_id", datasetID);
  labelSVG.setAttributeNS(null, "parent_id", parentID);
  labelSVG.setAttributeNS(null, "title", "Dataset ID: " + datasetID);
  [...]
 // Create producer profile facet
  var producerProfileGroup = document.createElementNS
("http://www.w3.org/2000/svg", "g");
  producerProfileGroup.setAttributeNS(null,"id", "producer_profile_"+i);
  producerProfileGroup.setAttributeNS(null, "producer_profile_name",
organisationName);
  // Check availability and generate appropriate facet
  if(producerProfileAvailability == 0){
    getProducerProfileNotAvailable(producerProfileGroup);
    producerProfileGroup.setAttributeNS(null, "availability", 0);
  }
  else if(producerProfileAvailability == 1){
    getProducerProfileAvailable(producerProfileGroup);
    producerProfileGroup.setAttributeNS(null, "availability", 1);
  }
[...]
```

Figure 8.20: An example JavaScript for generating a GEO label SVG from JSON descriptions.

The `svg_facets.js` script contains a number of JavaScript functions for generating simple SVG shapes (e.g., rectangles, circles, paths, polygons, etc.), linear gradient elements to produce the 'at a higher level' half-filled facets effect, and filter elements to produce an outer glow effect when a label is selected or highlighted (see Figure 8.21). Using these helper functions, the script provides functionality for constructing the SVG GEO label facets (see Figure 8.22).

Unlike the server-side implementation, where the GEO label SVG facets and shapes are represented as Strings, concatenated to produce a GEO label and returned in an SVG format (see Section 7.7), on the client-side all the SVG components must be dynamically generated as HTML DOM elements because JavaScript does not directly support conversion of String to SVG object.

```
[...]
// A helper function to create a path
function createPath(fill,stroke,stroke_width,stroke_miterlimit,d) {
	var path = document.createElementNS ("http://www.w3.org/2000/svg", "path");
	path.setAttributeNS(null,"fill",fill);
	path.setAttributeNS(null,"stroke",stroke);
	path.setAttributeNS(null,"stroke-width",stroke_width);
	path.setAttributeNS(null,"stroke-miterlimit",stroke_miterlimit);
	path.setAttributeNS(null,"d",d);
	return path;
}
[...]
 function getProducerLinearGradient(linearGradient){
	var stop1 = createStop (0.3, "stop-color:#FFFFFF");
	var stop2 = createStop (0.3417, "stop-color:#FCD8E9");
	var stop3 = createStop (0.3956, "stop-color:#F8ADD0");
	var stop4 = createStop (0.4537, "stop-color:#F586BA");
	var stop5 = createStop (0.5153, "stop-color:#F366A8");
	var stop6 = createStop (0.5815, "stop-color:#F14C99");
	var stop7 = createStop (0.6539, "stop-color:#EF378D");
	var stop8 = createStop (0.7357, "stop-color:#EE2985");
	var stop9 = createStop (0.835, "stop-color:#ED2180");
	var stop10 = createStop (1, "stop-color:#ED1E7F");

	linearGradient.appendChild(stop1);
	linearGradient.appendChild(stop2);
	linearGradient.appendChild(stop3);
	linearGradient.appendChild(stop4);
	linearGradient.appendChild(stop5);
	linearGradient.appendChild(stop6);
	linearGradient.appendChild(stop7);
	linearGradient.appendChild(stop8);
	linearGradient.appendChild(stop9);
	linearGradient.appendChild(stop10);

	return linearGradient;
}
[...]
// A helper function to create a filter
function createFilter(id, x, y, width, height, stdDeviation){
	var filter = document.createElementNS ("http://www.w3.org/2000/svg",
"filter");
	filter.setAttributeNS(null, "id", id);
	filter.setAttributeNS(null, "filterUnits", "userSpaceOnUse");
	filter.setAttributeNS(null, "x", x);
	filter.setAttributeNS(null, "y", y);
	filter.setAttributeNS(null, "width", width);
	filter.setAttributeNS(null, "height", height);

	var feGaussianBlur = document.createElementNS ("http://www.w3.org/2000/svg",
"feGaussianBlur");
	feGaussianBlur.setAttributeNS(null, "in", "SourceGraphic");
	feGaussianBlur.setAttributeNS(null, "stdDeviation", stdDeviation);
	filter.appendChild(feGaussianBlur);
	return filter;
}
[...]
```

Figure 8.21: An example of JavaScript functions for generating SVG elements.

```
[...]
// *****   PRODUCER PROFILE FUNCTIONS   ***** //
function getProducerProfileNotAvailable(parent_svg){
  // create paths
  var path1 = createPath("#FFFFFF", "#000000", 2, 10, "M152.178,97.822l59.796-
59.795
C187.958,14.008,156.478,2,125,2l0,84.563C134.837,86.563,144.674,90.316,152.178,97.
822z");
  var path2 = createPath("#FFFFFF", "#000000", 1.25, 10, "M154.14,31.925 c2.306-
0.358,5.972,1.384,5.972,1.384c-7.198,7.265-17.221,16.66-21.993,21.108c-1.67-0.165-
3.921-0.971-3.979-2.51 c-0.066-1.755,2.612-3.657,4.608-5.616c4.082-4.008,6.06-
6.125,10.018-10.091C150.775,34.184,152.558,32.167,154.14,31.925z");
  var path3 = createPath("#FFFFFF", "#000000", 1.25, 10, "M167.306,26.13
c1.821,1.689,2.298,2.29,4.055,3.965c-7.401,7.411-22.163,21.813-24.745,24.211c-
2.907,2.581-5.549,4.796-5.549,4.796 l-7.479,3.439l4.013-7.443c0,0,4.047-
3.986,6.425-6.279C151.787,41.333,159.915,33.34,167.306,26.13z");
  var path4 = createPath("#FFFFFF", "#000000", 1.25, 10, "M182.787,47.247l-
1.194,11.154l-25.367-0.137  c-4.976,0.13-9.948,0.983-9.535-3.671c0.214-
2.416,0.819-3.699,5.114-3.937h8.306c0,0-2.658-1.896-5.663-3.763 c1.758-
1.844,8.417-8.233,11.033-10.81C167.033,36.949,182.787,47.247,182.787,47.247z");

  // remove content
  parent_svg.innerHTML = '';
  // append new paths
  parent_svg.appendChild(path1);
  parent_svg.appendChild(path2);
  parent_svg.appendChild(path3);
  parent_svg.appendChild(path4);
}
[...]
```

Figure 8.22: An Example JavaScript function for constructing an SVG GEO label facet.

The `jquery_sliders.js` script provides functionality for applying information availability filtering based on the informational aspects' availability recorded in eight GEO label facets (see Section 8.1.2). The script contains eight JavaScript functions for initialising the facet sliders (one function per slider), defines the 'on-click' functionality of the interactive clickable filtering label located in the *GEO Label Filtering* tab (see Section 8.1.2), and provides functionality for setting search result GEO labels' visibility when the filters are applied.

When a value of a slider is changed, the slider's action listener saves the current value of the slider into an array and calls the `setLabelsVisibility` function (see Figure 8.23). The array consists of eight integers (with a value of 0 – not available, 1 – available, or 2 – available at a higher level) and is used to monitor the availability states of each GEO label filter for both sliders and the interactive GEO label. The `setLabelsVisibility` function (see Figure 8.24) obtains all the DOM children of the search results container SVG, i.e. all the search result GEO label SVGs, and iterates through the SVGs using unique system identifiers described earlier. For every dataset SVG, the function obtains the facets' DOM elements (producer profile, producer comments, etc.) and compares the value of the `availability` attribute in each facet with the corresponding filter value in the filter array. The `getMatch` helper function (see Figure 8.25) is used to compare the filter value with the facet availability; if the availability value recorded in the facet is equal or supersedes the filter value (for instance, 'available' state is better than 'available at a higher level') the function returns 1, otherwise it returns 0. The 'match' results for each facet are added together and if the sum is less than 8 (i.e., there is a mismatch in facets' availability values vs. filters), the `setLabelsVisibility` function hides the dataset SVG; otherwise, the dataset SVG is set to visible. The interactive clickable filtering label works in a similar way to the filtering sliders.

```
[...]
$(function() {
   $("#producer-slider").slider({
      range: "min",
      value: 0,
      min: 0,
      max: 2,
      step: 1,
      disabled: true,
      change: function( event, ui ) {
         // producer_profile group to append to
         var producer_profile = document.getElementById ("producer_profile");
         if(ui.value == "0"){
            $slidersVal[0] = 0;
            document.getElementById("producer-facet-img").src =
"img/facets/not_available/producer_not_available.png";
            getProducerProfileNotAvailable(producer_profile);
         }
         else if(ui.value == "1"){
            $slidersVal[0] = 1;
            document.getElementById("producer-facet-img").src =
"img/facets/higher_level/producer_higher_level.png";
            getProducerProfileHigherLevel(producer_profile, "SVGID_1_");
         }
         else if(ui.value == "2"){
            $slidersVal[0] = 2;
            document.getElementById("producer-facet-img").src =
"img/facets/available/producer_available.png";
            getProducerProfileAvailable(producer_profile);
         }
         // Hide or show labels depending on the current state
         setLabelsVisibility();
      }
   });
});
[...]
```

Figure 8.23: Example JavaScript function for setting up availability filtering sliders.

```
[...]
// Variable to keep the current value of the sliders
var $slidersVal = [0, 0, 0, 0, 0, 0, 0, 0];

function setLabelsVisibility(){
 // Iterate through all GEO labels
 var count = $("#zoom_pan_results_svg").children().length;
 for (var i = 0; i < count; i++) {
  var totalMatch =
  getMatch($slidersVal[0],$("#producer_profile_"+i).attr("availability")) +
  getMatch($slidersVal[1],$("#producer_comments_"+i).attr("availability")) +
  getMatch($slidersVal[2],$("#lineage_" + i).attr("availability")) +
  getMatch($slidersVal[3],$("#standards_compliance_"+i).attr("availability ")) +
  getMatch($slidersVal[4],$("#quality_information_"+i).attr("availability" )) +
  getMatch($slidersVal[5],$("#user_feedback_"+i).attr("availability")) +
  getMatch($slidersVal[6],$("#expert_review_"+i).attr("availability")) +
  getMatch($slidersVal[7],$("#citations_" + i).attr("availability"));
  if(totalMatch < 8){
    $("#geolabel_" + i).hide();
  }
  else{
    $("#geolabel_" + i).show();
  }
 }
}
[...]
```

Figure 8.24: The `setLabelsVisibility` function for filtering search results.

```
[...]
// Returns 1 if availability matches, or 0 otherwise
function getMatch(sliderVal, facetVal){
	if(sliderVal == 0){
		return 1;
	}
	else if(sliderVal == 1){
		if(facetVal == 1 || facetVal == 2){
			return 1;
		}
		else{
			return 0;
		}
	}
	else if(sliderVal == 2){
		if(facetVal == 1){
			return 1;
		}
		else{
			return 0;
		}
	}
}
[...]
```

Figure 8.25: The `getMatch` function for comparing filter value with facet's information availability.

The `filtering.js` script provides functionality to support additional filtering options for every GEO label informational aspect. The script defines the action listeners for all the *Filter* and *Reset* buttons in the *GEO Label Filtering* tab and provides JavaScript functions for applying and resetting the filters. Figure 8.26 provides an example of the producer profile filtering function. In this example, the script obtains the user-specified value from the `dataset-source-autocomplete` textbox and, if the textbox is not empty, iterates through the search result GEO labels. From each GEO label, the script obtains the values recorded in the `availability` and `producer_profile_name` attributes and, if information is available, compares the dataset producer name with the user input. If the facet's value matches the filter, the label SVG is 'up-scaled', otherwise the SVG is 'down-scaled'. The filtering function for the rest of the facets work in a very similar way by comparing DOM attribute values with the user filters and 'up-scaling' or 'down-scaling' the labels depending on the comparison results.

```
[...]
function filterProducer(){
	datasetSource = $("#dataset-source-autocomplete").val();
	if(datasetSource != ""){
		// Iterate through all GEO labels
		var count = $("#zoom_pan_results_svg").children().length;
		for (var i = 0; i < count; i++) {
			var availability = $("#producer_profile_" +
i).attr("availability");
			var producerName = $("#producer_profile_" +
i).attr("producer_profile_name");
			if(availability != 0 &&
producerName.toLowerCase().indexOf(datasetSource.toLowerCase()) != -1){
				// increase the size of the label
				upScaleLabel(i);
			}
			else{
				downScaleLabel(i);
			}
		}
		$("#dataset-source-autocomplete").prop('disabled', true);
		$("#filter-producer-btn").prop('disabled', true);
		$("#reset-producer-btn").prop('disabled', false);
	}
}
[...]
```

Figure 8.26: An example JavaScript function for applying additional filtering.

The filtering reset functions are very similar to the filtering functions, with the difference being that the GEO labels that match the filters are 'down-scaled' and those that do not match the filtering are 'up-scaled'. The reset functions essentially 'undo' the filters applied.

To increase or decrease the size of the search result GEO labels, the `filtering.js` script provides the `upScaleLabel` and `downScaleLabel` functions. The `upScaleLabel` function (see Figure 8.27) accepts the ID of the GEO label in question and, using this ID, obtains the current scale value (the original label size is 250^{x}250px hence the labels are scaled to achieve a required size) and x and y positions of the GEO label SVG from the size_scale, translate_x and translate_y DOM attributes. The function then increases the scale value by 0.015 and decreases the x and y values by 2. When the search result GEO labels are increased in size, their position slightly changes, therefore the x and y values have to be adjusted to avoid dislocation of labels from their original position.

```
[...]
// Helper function to resize the facet
function upScaleLabel(i){
  var currentScale = $("#size_group_" + i).attr("size_scale");
  var currentX = $("#size_group_" + i).attr("translate_x");
  var currentY = $("#size_group_" + i).attr("translate_y");
  var newScale = parseFloat(currentScale) + filterScale;
  var newX = parseFloat(currentX) - filterTransformX;
  var newY = parseFloat(currentY) - filterTransformY;
  if(parseFloat(newScale).toFixed(3) > parseFloat(minimumSize).toFixed(3)){
    newTransform = "translate(" + newX + " " + newY + ") scale(" + newScale + ")";
      $("#size_group_" + i).attr("translate_x", newX);
      $("#size_group_" + i).attr("translate_y", newY);
  }
  else{
      newTransform = "translate(" + currentX + " " + currentY + ") scale(" +
minimumSize + ")";
       }
       $("#size_group_" + i).attr("transform", newTransform);
       $("#size_group_" + i).attr("size_scale", newScale);
}
[...]
```

Figure 8.27: The upScaleLabel function for increasing the size of the GEO labels that match filtering.

The functionality of the downScaleLabel function is very similar to the upScaleLabel function described, with the difference being that the scale value is decreased by 0.015 and x and y position values are increased by 2.

The functionality for highlighting the datasets of interest is included as part of the search-request.js script. When a search result GEO label is clicked-on with the right mouse button, an event listener function is called (see Figure 8.28). The function locates the closest SVG root element to the selected target and extracts the GEO label ID from its id attribute. The extracted ID of the selected GEO label is saved in the $clickedLabelID global variable. Once the *Highlight* button in the pop-up menu is clicked, an event listener function assigned to the button (see Figure 8.29) obtains the dataset ID and verifies that the dataset is not already highlighted to avoid duplication. Using the getHighlightGlow function from the svg_facets.js script, the event listener adds a glow effect to the selected GEO label SVG and calls the getHighlightedLabelDetails helper function. The getHighlightedLabelDetails function sends an Ajax request to the server-side application to obtain dataset details and builds a fragment of HTML to display in the *Highlighted Datasets* tab. The HTML is then appended to the *Highlighted Datasets* and clickedLabelID is reset to null.

```
[...]
$(function() {
   $(".dataset_geolabel").mousedown(function(event) {
      if(event.which == 3){
        var targetLabel = $(event.target).closest("svg");
        $clickedLabelID = targetLabel.attr('id').replace("geolabel_", ""); } })
});
[...]
```

Figure 8.28: A JavaScript event listener function for highlighting a dataset with a right-click action.

```
[...]
var $clickedLabelID = null;
$(function() {
 $("#highlight").click(function(event) {
  if($clickedLabelID != null && !($("#highlight_glow_group_" +
                                  $clickedLabelID).length)){
   // Highlight the selected label by adding a 'glow' effect
   var highlightGlowGroup = document.createElementNS
                            ("http://www.w3.org/2000/svg", "g");
   highlightGlowGroup.setAttributeNS(null, "id", "highlight_glow_group_" +
                                                $clickedLabelID);
   getHighlightGlow(highlightGlowGroup);
   if($("#select_glow_group_" + $clickedLabelID).length){
    $("#select_glow_group_" + $clickedLabelID).after(highlightGlowGroup);
   }
   else{
    $("#size_group_" + $clickedLabelID).prepend(highlightGlowGroup);
   }
   // Add highlighted dataset into Highlighted Dataset tab
   var highlightedDiv = getHighlightedLabelDetails($clickedLabelID);
   $("#highlighted-datasets").append(highlightedDiv);
   $clickedLabelID = null; } })
});
[...]
```

Figure 8.29: A JavaScript event listener function for highlighting the dataset of interest.

The functionality for undoing a highlight action on a dataset is very simple (see Figure 8.30). When the *Undo Highlight* button in the pop-up menu is clicked, an action listener function checks the ID of the selected GEO label, removes the glow effect DOM element from the search results label SVG, removes the dataset description HTML DOM from the *Highlighted Datasets* tab, and resets the `clickedLabelID` variable to null.

```
[...]
$(function() {
      $("#undo-highlight").click(function(event) {
            if($clickedLabelID != null){
                  // Remove the highlight from the selected label
                  $("#highlight_glow_group_" + $clickedLabelID).remove();
                  $("#highlighted_item_" + $clickedLabelID).remove();

                  $clickedLabelID = null; } })
});
[...]
```

Figure 8.30: A JavaScript event listener function for removing highlighting from the dataset.

8.3 Summary and Conclusions

The prototype GEO label-based dataset discovery and intercomparison decision support tool has been developed to provide an innovative approach to visualising geospatial dataset metadata records and selecting datasets that fit user needs. The tool not only offers geospatial data search facilities that are common for real-world geospatial data portals and clearinghouses but goes beyond conventional methods of dataset discovery by providing a novel interactive technique for filtering and interrogating metadata records. The prototype tool effectively utilises the GEO label and transforms it into a powerful interrogation facility that allows dataset filtering based on the informational aspects' availability recorded in eight GEO label facets. By condensing large complex metadata records into simple, multi-faceted labels, the tool allows users to capture visual representations of the quality and possibly relevance of many datasets without having to inspect each dataset in great detail. Such a visualisation approach allows for hundreds of metadata records to be displayed at once within a single search space, giving users a sense of overall metadata completeness. At the same time, hover-over and drilldown capabilities allow users to access and inspect detailed information about the datasets that potentially fit their needs to make an informed dataset selection. The tool additionally offers functionality for highlighting the datasets of interest, which resembles the 'favourites' lists often used in e-Commerce recommender or comparison websites and is designed to further support more effective dataset intercomparison and informed dataset selection. Essentially, the prototype GEO label-based dataset discovery tool offers geospatial data search- and interrogation-by-quality functionality which is not currently provided by real-world geospatial data applications, portals and clearinghouses.

In terms of practical implications, the prototype implementation of the tool demonstrated that it is feasible to not only integrate the GEO label as part of a GIS application but to also utilise its SVG interactivity to develop an interactive decision support system. The SVG format allows the GEO label to internally store core dataset information which can be rapidly accessed and manipulated using simple HTML and JavaScript technologies. The fact that the GEO label SVGs can be constructed using both server-side and client-side technologies makes the GEO label integration even more flexible.

To evaluate the usability and effectiveness of the GEO label-based dataset discovery and decision support tool, a human-subject study has been conducted with geospatial data users and experts. The study design, procedure and results are discussed in the next chapter of this thesis.

Chapter 9 Phase III: User Evaluation of the GEO LINC Tool

To evaluate the usability and effectiveness of the design of the GEO LINC tool, a human-subject evaluation study has been conducted with geospatial data users and experts. This chapter presents and discusses the study design, procedure and results.

Section 9.1 describes the techniques applied to generate a set of test metadata XML records for use in the GEO LINC evaluation study. Section 9.2 outlines the procedures adopted to conduct a usability study of GEO LINC. The results are discussed in Section 9.3, and the study limitations and conclusions are presented in Sections 9.4 and 9.5, respectively.

9.1 Generating Test Study Data

Due to the fact that the GEO LINC tool was developed to support geospatial dataset discovery and intercomparison, it was necessary to generate a large set of test metadata records in order to conduct a meaningful evaluation study. As will be outlined in the next section, the evaluation study included three geospatial dataset selection scenarios (forest cover, agricultural land use and climate change); as such, three sets of metadata records were required. To ensure variety and sufficient number of metadata records, the GEOSS portal[8] and Food and Agriculture Organisation (FAO) GeoNetwork catalogue[9] were used to collect 351 suitable metadata records. Of these, 101 metadata records were collected using “forest”, “forest cover”, and “forestry” keywords to support the forest cover scenario, 130 were collected using “agriculture”, “agricultural land use”, “crops”, etc. keywords to support the agricultural land use scenario, and 120 metadata records were collected using “climate”,

[8] http://www.geoportal.org/web/guest/geo_home_stp
[9] http://www.fao.org/geonetwork/srv/en/main.home

“temperature”, “precipitation”, etc. keywords to support the climate change scenario. Spatial and temporal coverages of the datasets were not considered because more specific search queries were not returning enough datasets to collect a large-enough sample of test data. It was therefore decided to manually edit collected metadata records to suit the evaluation study scenarios.

Although all the collected metadata records were ISO19115:2003 compliant, most of the records were incomplete or inconsistent. It was foreseen that none of the ISO compliant metadata records would include citations information, user feedback and expert reviews since these elements are not supported by ISO standards, but the inspection of the acquired metadata records further confirmed the findings from the previous studies where geospatial data users repeatedly reported that the geospatial dataset metadata records are typically incomplete. The acquired metadata records at best contained producer profile information, ambiguous producer comments that were not adding any real value (e.g., generic information about the dataset provider), standards compliance information and lineage information that was often either empty or simply provided a ‘last updated’ date. Despite the fact that quality information can be recorded using standard ISO19115:2003 metadata documents, none of the collected metadata records contained any quality information.

While GEO LINC supports processing and visualisation of ISO19115:2003, FGDC and GeoViQua quality models compliant metadata records, the contents of the collected ISO metadata records were not sufficient to carry out thorough evaluation of the tool and its functionality. For instance, it would not be possible to evaluate the importance and usefulness of user feedback, expert reviews, and citations information. The GeoViQua quality models (see Section 7.1) that extend ISO19115:2003 and enable recording of rich information only recently became available and, at present, there are only a few geospatial datasets that have been fully described using these models. As a result, most of the test metadata records required manual editing to enrich them with: relevant producer comments; lineage information; quantitative quality information; user feedback; expert reviews; and citations information. Sample quantitative quality and lineage information was extracted from the example GeoViQua models-compliant metadata documents generated for the Global Land Cover[10] and Digital Climatic Atlas[11] datasets. A large number of fictitious user feedbacks, expert reviews and citations were generated specifically for the scenarios and allocated to metadata records based on their suitability. For instance, where metadata records were enhanced with detailed quality and lineage information, positive user and expert feedbacks were added to the records and, on the other hand, incomplete metadata records were accompanied with more negative feedbacks and lower star-ratings. Although

[10] http://schemas.geoviqua.org/GVQ/4.0/example_documents/PQMs/GLC_2000_GVQ_raw.xml
[11] http://schemas.geoviqua.org/GVQ/4.0/example_documents/PQMs/DigitalClimaticAtlas.xml

the test metadata records required heavy manual editing to provide rich dataset descriptions, it is anticipated that with the availability of GeoViQua quality models and GeoViQua feedback server[12] that supports feedback creation, the issues of metadata completeness will be addressed in the near future.

9.2 Soliciting Opinion on the GEO LINC Tool

To evaluate the prototype GEO LINC tool, a controlled human-subject study was conducted. The study was based on use of the 'think-aloud' protocol as described by Rubin and Chisnell (2008), where study participants were asked to 'think-aloud' – i.e., verbalise what they are looking at, thinking, and doing – as they completed their allocated tasks. If a participant remained quiet for a prolonged period of time, the evaluator (the author of this thesis) posed a question to obtain more information or to encourage the participant to continue verbalising his performance (hence, introducing a question-asking protocol to complement the think-aloud process). The study was carried out at two locations: half of the sessions (3 participants) took place at Aston University in an HCI lab and the other half (also 3 sessions) were carried out at participants' work places in an office environment. The sessions followed the same controlled procedures and replicated the same study environment and set-up. The sessions were carried out on a one-to-one basis with only a participant and an evaluator being present in the study room. The participants were provided with a laptop on which the prototype GEO LINC tool was installed, an additional large 21" screen, a keyboard and a mouse. A digital camera was focused on the screen of the computer, not the participant, to collect an audio-video (AV) recording of the participant's interaction with the system and his/her corresponding utterances for post-study analysis.

To recruit potential study participants, email invitations were sent to a number of personal contacts provided by the GeoViQua project partners. Due to restrictions in time and budget, only GIS professionals, academics and researchers who work in the UK were considered for the study. To ensure participants each had sufficient time to reflect on their participation and what would be asked of them before consenting to participate, and to minimise the time required for briefing the participants, a copy of the participant information sheet (see Appendix H.1), study materials (see Appendix H.3), and a consent form (see Appendix H.2) were emailed to all participants in advance. The participant information sheet included information about the purpose of the evaluation study, including what would be expected of them as participants, study location, duration of the study, information about the study investigators including their organisation, data collection method (AV recording), and how the collected data would be stored and used after the study.

[12] http://geoviqua.stcorp.nl/home.html

Table 9.1: Phase III consent form statements.

		Tick Box
1.	I confirm that I have read and understood the information sheet for the above study. I have had the opportunity to consider the information, ask questions and have had any questions answered satisfactorily.	
2.	I understand that my participation is voluntary and that I am free to withdraw at any time without giving any reason, without my legal rights being affected.	
3.	I agree to my interaction with the software being digitally audio-video recorded and later transcribed. I understand that such transcriptions will be anonymised such that I cannot be identified from the record.	
4.	I understand that data collected from me during this study will be available only to the study team, including the investigator and her supervisors.	
5.	I understand that data collected from me may be published in aggregated and/or anonymised form but that any publication will not contain any personal information that could identify me.	
6	I agree to take part in the above study.	

The consent form consisted of six statements to which the participants needed to agree by signing the form prior to engaging in the study (see Table 9.1). The evaluator made sure that all recruited participants fully understood the purpose and the procedures of the study before commencing their participation. The study materials document (see Appendix H.3) comprised seven sections, A to G, each of which is described below.

Section A consisted of a small number of questions to gather background information about the participants. Participants were first asked to identify themselves as one of the following: primarily data users; primarily data producers; or equally data users and producers. They were then asked: to pick one or more statements from a set which best describes their dataset user or producer type; to identify the type of organisation they work for; to indicate how long have they been working with geospatial data; and to approximate the percentage of their time they spend working directly with geospatial data. In this section, the participants were also asked whether they have a choice of dataset to use, what clearinghouses they use, if any, and whether they find dataset selection a challenging task.

Section B provided a brief introduction to the GEO label and its role. The section introduced the eight GEO label informational facets and explained the facet variations to convey information availability. GEO label examples presenting three different availability states were also presented.

Section C provided detailed information about the GEO LINC dataset discovery tool. This section acted as a user guide and provided instructions on how to use various tool functionalities to: (a) define initial search criteria; (b) filter search results; (c) obtain detailed information about a dataset; and (d) highlight a dataset of interest.

Section D consisted of three geospatial dataset selection exercises where study participants were required to use GEO LINC to complete a series of prescribed tasks. Each exercise provided a dataset selection scenario which participants needed to follow in order to select a dataset or a group of datasets that satisfied the prescribed scenario requirements (see Figure 9.1 for an example scenario). After completing each study scenario, participants were asked to write down the ID(s) of the dataset(s) that they selected and provide a brief explanation as to their dataset(s) selection decision. The evaluator would then reset the system ready for the next exercise.

Consider a scenario where you are searching our proposed dataset discovery tool for a dataset to use in your work. Your current task involves monitoring changes in global forest cover in the last two decades to identify where the changes occurred and to what degree (for instance, where were trees cut or newly planted). For this task, you require a dataset or a combination of datasets with global spatial coverage, temporal coverage between 1993 and 2013, and no access or use constraints.

We would also like you to assume that the following information is of high importance to you and will heavily influence your dataset selection:

a) Availability of contact information of the dataset provider in case you require additional information about the dataset. You are particularly interested in the datasets that are provided by the Joint Research Centre (JRC) because you have used its data in the past.
b) Availability of formal quality information such as uncertainty measures and dataset accuracy. For your current task quality information at a dataset level is acceptable.
c) Availability of information on datasets' compliance with international standards. You are specifically looking for the datasets that comply with ISO 19115 or at least FGDC standards.
d) Availability of experts' opinions on the dataset quality. If available, you would prefer the datasets that have been reviewed by at least 10 experts and received an average rating of 4 stars or higher.
e) Availability of reports on quality checks or journal publications which refer to the dataset. You would be particularly interested in the datasets that have been cited in 5 or more publications.

Figure 9.1: An example of a dataset selection exercise scenario.

Section E consisted of a small number of questions to solicit participants' feedback and opinions on the GEO label visualisation. Here, participants were asked to: (a) rate the effectiveness of the proposed GEO label design at conveying the availability of a dataset's quality information; (b) describe which aspects of the proposed GEO label design they found most effective/ineffective in conveying information availability; and (c) describe any modifications or improvements that they would apply to the proposed GEO label.

Section F consisted of a small number of questions to solicit participants' feedback and opinions on the prototype GEO LINC tool. In this section, participants were asked to: (a) rate the difficulty of completing the dataset selection exercises using GEO LINC; (b) describe the aspects of the system that they found the most challenging when comparing and selecting geospatial datasets; (c) rate the effectiveness of GEO LINC at supporting dataset intercomparison and selection; (d) describe the aspects of GEO LINC that they found most effective/ineffective; and (e) describe any modifications or improvements that they would apply to GEO LINC. Finally, Section G welcomed participants to leave any other comments or suggestions on the GEO label and the prototype dataset discovery tool.

Six participants were recruited for the evaluation study; their profiles are presented in Table 9.2.

Table 9.2: Profiles of Phase III study participants.

Participant 1 is an academic data user who works for an academic institution. He has been working with geospatial data for less than 2 years and, in the context of his current position, he spends approximately 20% of his time working directly with geospatial data. Participant 1 has no choice of datasets to use.
Participant 2 is a governmental data user who works for a municipal government organisation. He has been working with geospatial data for about 2 to 9 years and, in the context of his current position, he spends approximately 50% of his time working directly with geospatial data. Participant 2 has a choice of datasets to use and he uses http://data.gov.uk/ for selecting datasets. Participant 2 finds selecting datasets that fit his needs a challenging task because: • "*category* [and] *search terms are subjective and do not provide insight;* • [it is] *difficult to distinguish regional from national datasets when selecting* [spatial data]*;* and • *quality and date of data can be variable*".
Participant 3 is a private sector, research and academic data user and data producer who works for an academic institution. He has been working with geospatial data for more than 20 years and, in the context of his current position, he spends approximately 5% of his time working directly with geospatial data. Participant 3 has no choice of datasets to use, but uses The British Atmospheric Data Centre (BADC) to acquire datasets. Participant 3 finds selecting datasets that fit his needs a challenging task because "*much of the data is not well documented and* [is] *very hard to find*"*. "It is often not clear which portal to search in and data is often held in several portals adding to the confusion. Often* [the participant is] *also not 100% clear on what* [he] *want*[s] *(or* [he] *know*[s] *what* [he] *want*[s] *is not available so* [he is] *looking for something related).*"
Participant 4 is an academic data user who works for an academic institution. She has been working with geospatial data for about 2 to 9 years and, in the context of her current position, she spends approximately 10 – 15% of her time working directly with geospatial data. Selection of geospatial data is not applicable to the participant, hence she is not sure whether geospatial dataset selection is a challenging task. This participant's organisation is "*an end-user of datasets and therefore require* [datasets] *to fulfil a basic set of requirements: latitude/longitude/time and variable with error. It is very rudimentary, so* [the organisation is] *not trying to find data that has to be 'fit for purpose'*".
Participant 5 is a research data user and data producer who works for an academic institution. She has been working with geospatial data for about 2 to 9 years and, in the context of her current position, she spends approximately 70% of her time working directly with geospatial data. Participant 5 has a choice of datasets to use and she uses the following websites to acquire data: NASA's Ocean Colour website; NOAA's Comprehensive Large Array-data Stewardship System (*CLASS*) website; ESA's MeRCI and FTP sites with MeRCI/AATSR data; the Natural Environment Research Council (NERC) Earth Observation Data Acquisition and Analysis Service (NEODAAS) website; and Earthnet Online ESA (EOLISA). Participant 5 finds selecting datasets that fit her needs a challenging task because "[data portals are] *not user friendly for GIS applications* [, it is] *difficult to select fit for purpose data* [, and it takes] *too long to get a large number of datasets*".
Participant 6 is a research and academic data user who works for an academic institution. She has been working with geospatial data for about 2 to 9 years and, in the context of her current position, she spends approximately 5% of her time working directly with geospatial data. Participant 6 has a choice of datasets to use and she uses National Snow and Ice Data Center (NSIDC) and NASA Distributed Active Archive Centers (DAAC) (for MODIS Land Products) to acquire data. Participant 6 finds selecting datasets that fit her needs a challenging task because, while she "*tend*[s] *to know what's available through word-of-mouth, it's hard to know about the whole range of data sources available to choose from"*. "*The CEOS EO Handbook is an enormous help in this respect (for EO data anyway!)*".

9.3 Results and Discussion

As already noted, the evaluation study sessions adopted a ‘think-aloud’ protocol, complemented with question-asking where required, and were AV recorded for post-study data analysis. The AV recordings of the study sessions were transcribed and the transcripts were validated by a third party to confirm accuracy and completeness. Because data collected was qualitative, the transcripts were subjected to thematic data analysis – “*a method for identifying, analysing, and reporting patterns (themes) within data*” – as described by Braun and Clarke (2006, p. 79) to identify patterns of meaning across the study data. Table 9.3 outlines the six phases of the thematic analysis process which were followed to analyse the study data.

Table 9.3: Phases of thematic analysis (Braun and Clarke, 2006, p. 87).

Aston University

Illustration removed for copyright restrictions

To become entirely familiar with the study data, the transcripts were first carefully read. Following this, a second reading was conducted to identify and summarise preliminary topics. Individual interesting data extracts were then systematically coded with a descriptive word or phrase summarising key points. Table 9.4 provides an example of how codes were applied to data extracts (see Appendix I for full data analysis). It should be noted that individual data extracts were at times associated with multiple codes, yet individual codes were reapplied to different data extracts only when the conveyed message of both extracts were almost identical. The potential themes were identified by sorting and combining relevant codes. Following this, the themes were reviewed to ensure their validity in relation to the coded extracts and to the entire data set. Finally, informative names for each theme were generated by analysing the key aspects each theme captured.

Table 9.4: Examples of data extracts with codes applied.

Data Extracts	Coded As
"*I've kind of picked that one first cause it's got most of the colours in. I don't know why* [laughing]*. But I figured that the fully-coloured one has more information about it* [dataset]*, so that's the theory.*"	1. Influenced by label colours 2. Associates colours with metadata completeness 3. Selected a fully-coloured label
"*I'd probably look at those three* [datasets]*, and the reason why is – after all that filtering it appears to be the same reviewers, based on the user comments and on the citations, they've used it for pretty much the same study. So if they've done it, I'll just copy what they've done.*"	1. Dataset selection decision 2. Trust into peers 3. Reliance on user feedback 4. Reliance on citations
"*I would just want a link to take me over to that* [producer] *site, to the same data that is viewed on here* [pointing at the tool]*. Would be handy. Cause otherwise, my next step would be to either contact the user* [pointing at the user feedback drilldown page] *or it would probably just be copy and paste the dataset name and try to find it somewhere on Google.*"	1. Acquiring data 2. Link to producer website 3. Link to the physical data 4. Obtaining information about physical data from peers
"*It would be good, if here I almost knew what extents these* [datasets] *were over. I don't know how hard it would be, almost to include a tiny map here* [pointing at the Dataset Details area] *with just the area that it* [the dataset] *is valid for highlighted, with the rest, kind of, faded out a bit.*"	1. Importance of having a small map showing dataset's spatial extents 2. Importance of knowing dataset's spatial extents
"*Once I got my head around with the colour scheme and things like that, it does work quite quickly for visually picking up what you need to. Or, I should say, not spotting before you filter, but once you filter knowing that you are on a right track without having to go and read quite so much. And then it does filter it down enough to read the actual comments and the actual, the longer text and that kind.*"	1. Learning effect 2. Learning colour scheme 3. Enables visual filtering 4. Minimises unnecessary reading 5. Keeps user on a right track 6. Filters results down to read longer text

The following sections discuss the themes that emerged from data analysis. As will be seen, some of the themes formed naturally around the GEO LINC tool functionality.

9.3.1 Dataset Discovery

The dataset discovery theme naturally emerged from the study participants' discussions about the GEO LINC initial search page and participants' interaction with the system. This theme encapsulated participants' comments and discussions about the search criteria and mechanisms that are essential in facilitating effective dataset discovery. As study results indicated, although the GEO LINC tool provided sufficient support to define a simple initial search query, participants questioned some search functionality and also proposed potential system improvements. While the dataset discovery theme resulted from the evaluation of a specific tool, the concepts discussed in this section directly relate to broader issues of dataset discovery and functionality that is currently provided by real-world geospatial data portals, catalogues and clearinghouses.

Boolean Queries

One of the most prominent topics that emerged from the study was the ability to define complex flexible search query strings when searching or filtering geospatial datasets. All six study participants either directly argued the need for more sophisticated SQL-like search queries or indirectly indicated this requirement by attempting to construct Boolean query expressions. When defining the search keywords or applying the free-text filters, all six participants were wondering whether they can construct queries using 'AND', 'OR' and 'NOT' clauses. For instance, while entering the search keywords, participant 3 speculated: "*can I AND it? Let's see. I am guessing that this is pushing* [the boundary] *but this is how I'd do it*". Participant 5 stated that she "*would like* [the keywords criteria] *to be more like an SQL query*" where she could use AND clauses, parenthesis, etc. When applying the dataset source filtering, 3 participants attempted to use Boolean clauses to construct an 'OR' query; participant 2 commented that "[he could not] *tell if* [he could] *do combined* [queries…] *like 'FAO or NASA', like filtering where* [one] *can do ANDs or anything*".

Participants argued that one of the issues with geospatial data portals is that "*there is not enough flexibility in the query construction*" which significantly complicates dataset discovery. Participant 3 explained that, when using bibliographic databases, he "*fairly routinely build*[s] *up quite complex query strings involving* [*…*] *a fairly standard practice of putting in quotes, using AND, OR and NOT as a combination, and bracketing for precedence*". It was also highlighted that, if functionality for defining Boolean expressions becomes available, additional mechanisms for saving complex search queries would be very useful.

Overall, participants' discussions, comments, and direct interaction with the system revealed that complex search queries are *absolutely* essential to support effective dataset discovery. Nevertheless, standard geospatial data portals and catalogues do not generally support 'sophisticated' Boolean query strings.

Autocomplete Suggestions

The study results also revealed that the autocomplete functionality is very valuable when searching for geospatial datasets. While completing the exercises participants were highly influenced by the autocomplete suggestions and agreed that "*the autocomplete thing is great* [… be]*cause it helps* [users] *to understand what the system actually knows about*" (participant 3). Great care must be taken when providing the autocomplete functionality to ensure that the suggestions are effective and do not mislead the users.

Spatial Extent

When defining the location (spatial extent) search constraint, three study participants were unsure how to use the interactive map. For instance, after hovering the mouse over the map, participant 1 stated: "*I would like to be able to just draw a selection box over the UK now*". Despite the fact that the location selection fields were accompanied with "*Click and Drag on the map holding the Shift key to select an area*" text, participants failed to notice this instruction. When selecting global spatial coverage, participants were unsure whether to draw a box over the map to cover the entire map area, leave the location options blank, or to select a 'global' option from the location dropdown menu. For instance, participant 3 assumed that "*if* [he] *do*[es]*n't select the location* […] *it's going to be global*". This clearly indicated the need for some guidelines on the effective selection of spatial extents. While modern geospatial data portals and catalogues typically provide interactive maps for defining the spatial extent, these maps might not be intuitive, particularly for novice users, unless accompanied by clear and prominent instructions.

Temporal Extent

Addressing more general usability aspects of the tool, study participants argued that the use of pop-up calendars for selecting the date range (temporal extent) was inefficient. Participants were actively trying to enter the dates manually rather than use the interactive pop-up calendars. While entering the start and end dates, participant 2 commented: "*I hate picking things. Yeah, just lazy*". When trying to select a start date, participant 3 commented that "[it is] *going to be annoying*", clearly expressing his dislike of the pop-up calendar functionality. Since some ill-formatted dates caused system errors, participants suggested adding clear instructions on what date formats are actually supported.

Study participants additionally highlighted the importance of the 'last updated' date, arguing that it is "*always a key bit*" (participant 2) and could potentially be included as part of the initial search criteria.

Access and Use Constraints

The analysis of the study transcripts revealed the great importance of access and use constraints criteria when selecting a dataset to use. While defining the use and access constraints criteria, participant 2 commented that these options "[are] *very useful to have on because the stuff* [he] *ever deal*[s] *with* […]*, particularly because it is regional Ordinance Survey stuff, usually got constraints*". Participant 3 also agreed with the importance of use and access constraints information stating that "*most of the time,* [he] *want*[s] *no access constraints or usage restrictions*". Participant 3 further explained that, at present, it is challenging to acquire datasets that are not subjected to any access or use restrictions. These study findings are in line with the results of the initial interviews, where interview respondents indicated the high importance of licensing and restrictions information (Section 4.2.11).

As could be observed throughout this section, the dataset discovery theme largely covered topics of interface usability and cosmetic issues which should presumably be addressed in modern geospatial dataset discovery applications, portals, catalogues, etc. Unfortunately, as the study results indicated, modern geospatial data portals and clearinghouses are often "*unfriendly to use*" and provide users with a bare minimum in terms of search query parameters and interactivity.

9.3.2 'At a Glance' Dataset Intercomparison

The analysis of the study transcripts identified that the proposed GEO LINC tool effectively facilitated 'at a glance' intercomparison of a large number of datasets. Participants were able to identify the most "*colourful*" or "*fully-coloured*" labels immediately after the search results appeared on screen.

When completing dataset selection exercises, most study participants were able to identify relevant datasets without applying any filtering options. With 130 search results being displayed on the screen, participant 4 elaborated: "*that one is looking quite good, and that one, and that one. Those three* [datasets] *got loads of information so it could be* [that] *all those three are the candidates*". When the first scenario search results appeared on screen, participant 1 immediately clicked on one of the most complete GEO labels, saying: "*let's start with the better looking ones*". After completing the exercise, the participant explained that he "*ha*[d]*n't initially clicked on* [the filtering options because] *it was relatively obvious from* [the search area] *which* [labels] *are the most 'colourful'*". The influential effect of the GEO label colours on the dataset selection was further confirmed by participant 2:

> "*I've kind of picked that* [label] *first* [be]*cause it's got most of the colours in. I don't know why* [laughing]*, but I figured that the fully-coloured* [label] *has more information about* [the dataset]*, so that's the theory.*"

The study results also demonstrated that participants were able to quickly learn the meaning of the GEO label facets. After using the system for only 8 minutes, participant 1 was able to identify missing information from just looking at the dataset labels: "*This one doesn't have the citations that I need. So does that one. That one doesn't have any expert reviews that I need*". Participant 2 was also able identify available information without applying any filtering, he remarked: "*Okay, so there is a lot of pink! Or red. Which is the producer stuff*". It was suggested that "*the same colour gives you an idea that it's the same area of concept*" (participant 4) which further facilitates facet recognition and recall. Overall, when completing the dataset selection exercises, none of the study participants expressed any difficulty with using the GEO labels and identifying the meaning of the facets. These results indicate that the clock-like positioning of the facets and colour grouping effectively facilitate ‘at a glance’ comprehension of information availability.

Regarding the starfield visualisation of the search results, the study results indicated that the fixed positioning of the labels supported visual recognition and recall. Participants stated that they were able to “*remember where* [the labels were] *on the screen*” and come back to the datasets that they have previously reviewed. Participant 1 stated that “*it is handy that* [the tool] *keeps* [the labels] *in the same place so that* [he] *know*[s] *that this* [label] *was the one that* [he] *clicked on in the first place*”.

While fixed positioning was deemed useful, it was also proposed to add functionality for regrouping the labels. Participants suggested grouping labels by created or updated dates; grouping labels by size when additional filtering is applied, for example, bringing larger labels to the top of the search area; or allow drag-and-drop functionality. To ensure that the benefits of fixed positioning are not lost, participant 4 proposed to have an ‘on’ and ‘off’ option for switching the regrouping off to return the labels to their original positions.

9.3.3 Side-by-side Metadata Comparison

Only 3 study participants took advantage of the dataset highlighting functionality. In the post-study discussion, participants admitted that, even after reading the user guide, they simply forgot that this functionality was available. In most cases, participants were able to identify the complete labels at a glance and remember their positioning for later reference, hence the highlighting functionality was not essential.

Those participants who utilised the highlighting functionality stated that it “*is really effective*” and useful for side-by-side metadata comparison. Participant 1 commented:

> “*So it’s now handy that I do have the ‘highlighted’ tab so that I can now go through each of the ones I flagged as being potentially useful dataset and just quickly compare them on the available metadata that they have*”.

Participant 2 also took advantage of the highlighting functionality “*to remember which* [datasets he] *looked at*”. After using this option for the first time, he happily remarked: “*Oh, look at that! I can compare. That’s what I was hoping.* [laughs]”. Participant 2 then admitted that the highlighting option “*does allow good comparison* [of datasets] *side by side*”. Although participant 3 did not use the highlighting functionality because he “*didn’t even know* [that he could] *do this*”, he stated that “*it could be useful if* [he was] *trying to build up a set of plausible datasets*”.

To enable more effective side-by-side dataset intercomparison, it was suggested to add functionality for reordering and customising the list of highlighted datasets. Participants argued that with too many datasets highlighted, it becomes quite challenging to manage the list. In such cases, drag-and-drop functionality would allow for bringing the most 'interesting' datasets to the top of the list or placing similar datasets next to each other. Participant 1 commented:

> "*So what I've just noticed, there are two datasets that seem almost identical, except one has user and expert comments associated with it. So if I was allowed to get this one and the one that I spotted at the bottom next to each other, so I could spot what exactly was the difference*".

9.3.4 Search Results Filtering

The search results filtering theme emerged from the study participants' discussions and interaction with the GEO LINC filtering functionality. Overall, the study results indicated that the filtering options allowed the participants to filter out irrelevant datasets "*without having to go and read quite so much*" text as they would have to otherwise. As participant 2 explained, the tool "*worked well to get* [him] *to the point where* [he was] *prepared to look at the* [longer] *text*" and comments. Participants also stated that the filtering functionality is not only effective for narrowing down the search results "*but it* [is] *also good for guidance*" and for "*knowing that you are on a right track*". Due to the fact that the tool provided participants with a large number of filtering options, participants argued that it should also support functionality for saving and reapplying the filtering queries. When performing a new dataset search with an alternative set of keywords, participant 6 commented that "*all* [her] *filters* [had] *gone* [and she has] *to put the same filters in* [again]". Participant 1 further elaborated:

> "*What I would really want to do is to reapply these* [filters] *again. I still want these constraints but I just wanted to change a keyword really. And I've now forgotten which ones I wanted*".

The study results also revealed that the additional filtering functionality was not particularly effective. While the additional filtering options (dataset source, average rating, etc.) were largely considered useful, the changes in size to convey relevance were very confusing. Most of the study participants believed that, similar to the facets filtering, additional filtering options remove the search results labels that do not match the filtering criteria. Consequently, when applying additional filtering options, participants were confused whether the filtering worked and whether they were "*doing something wrong*". Because the changes in size were not immediately obvious, some participants assumed that all the labels matched their filtering criteria. The study participants stated that the size variations were too subtle and not sufficiently noticeable. For instance, after applying the dataset source filtering, participant 3 was very puzzled until the evaluator has hinted the changes in the labels' size; he then commented:

> *"Oh, so they go ever so slightly smaller. So the smaller ones are the ones that* [the filter] *doesn't match. But that is too subtle, particularly when they all go smaller because nothing looks bigger* [laughs]. *I didn't notice that*".

The participants also argued that when several additional filters are applied, it is impossible to identify which filters the dataset actually matches unless detailed dataset description is manually inspected. Participant 6 commented:

> "*it would be nice to know how these* [dataset labels] *are responding to the filters that I'd set on here* [pointing at the additional filtering options]*, like the average rating and so on*".

To address this issue, the participants proposed to grey out the facets that do not match the additional filtering and provide a summary of all the information that relates to the additional filtering (dataset source, average rating, number of feedbacks, etc.) as part of the dataset details description. Participant 4 further suggested that the filtering area itself does not visually indicate whether the filters have been applied; she explained:

> "*I've got filtering applied here* [pointing at the additional filtering options] *and when I close* [the collapsible content panels] *there is no indication that I have filtering applied. So I could make a mistake of not removing a filter and therefore constraining my options*".

To provide visual feedback to users, it was suggested to colour the collapsible content headers in the same colour as the facets when the filters are applied.

When interacting with the tool, only participant 1 utilised the interactive GEO label to filter the search results. He intuitively clicked on the facets to set the filters and later suggested that the sliders are somewhat redundant. All the other study participants made use of the filtering sliders and in the post-study discussion admitted that they did not "*even think to click on* [the label] *because* [they] *thought it was a graphical representation of* [their] *filtering*", "*a summary… rather than an interactive thing in itself*" (participant 4). After learning the function of the interactive label, participants commented that the interactive label is "*really neat*" but they would still be inclined to use the sliders. These results indicate that various filtering mechanisms should be supported to suit different user needs.

Regarding the filtering options, participants argued that some of the additional filters were "*artificial*" and "[did not] *make a lot of sense to* [them] *in terms of the kind of data that* [they] *would find*". When discussing user feedback and expert review filters, participants suggested that having a distribution of ratings and number of feedbacks/reviews would be much more useful than filtering by average star rating and minimum number of feedbacks/reviews. It was argued that users do not typically think in terms fixed numbers when making a selection decision (e.g. "I want a dataset with at least 10 user feedbacks"). Participant 3 explained that he would "*want to look at what the distribution of the stars is and then* [he would] *make an informed decision* [*…*] *Obviously,* [he would] *look for the highest number of stars* [he] *can get*". Participant further commented that "*when it comes to TripAdvisor or any other equivalent rating systems,* [he] *certainly tend*[s] *to look at items which have significant number of feedback and take that feedback much more seriously*". Analogous comments applied to the 'minimum number of citations' filtering option, participants stated that "[it is a] *slightly strange way of filtering* […] *because datasets can be cited for a lot of reasons*" (participant 6). It was recommended to provide a filtering slider with the distribution of the number of citations and also allow filtering by citation date, author, and journal/conference name.

The analysis of the study transcripts further revealed that filtering by a number of processing steps is not useful since most of the geospatial datasets contain pre-processed data and it is highly unlikely that users "[would] *go anywhere near*" unprocessed Level 0 data. It was also suggested that *raw* data means different things to different geospatial data users. Participant 6 recommended that it would be more valuable to have information about the processing "*characteristics*" such as the number of datasets used in producing the product; the number of instruments used to collect the data; the types of processing applied, etc.

As already discussed in Section 9.3.1, the study revealed that the free-text filtering fields should support flexible Boolean query expressions.

9.3.5 Dataset Details Acquisition

The dataset details acquisition theme incorporated participants' feedback and comments about the mechanisms used to obtain detailed information about the dataset of interest and any additional dataset information that participants felt was missing. Overall, the study results indicated that the mechanisms provided for obtaining dataset details were effective, with study participants actively using the hover-over functionality to obtain facets' summaries, on-click functionality to display dataset details and drilldown to obtain detailed facet-related information. The results showed that none of the study participants were interested in raw metadata XML documents, with five out of six participants reacting very negatively when raw metadata XML appeared on the screen. For instance, participant 2 admitted that "[he] *kind of switched off when* [he] *saw* [a raw] *XML*" document.

The study results showed that the hover-over functionality was effective for obtaining quick summary information. The study participants were especially keen to use this function to obtain numerical summaries such as average user and expert ratings and citations count. Participants agreed that this function is useful for getting "*an idea*" without having to navigate to a new page. Participants however suggested that hover-over information should be reiterated as part of the dataset description.

The drilldown function was perceived as "*very useful*" at providing additional details about the datasets. Participant 3 stated that the drilldown function is "*the key bit…* [and] *is absolutely critical because* [the GEO label] *colours are not enough*". Participant 4 also commented:

> "[the drilldown functionality is] *handy, isn't it?* [Be]c*ause you've pulled out the metadata. Although it is not entirely obvious on this* [pointing at a GEO label in the dataset details area] *that you can click on one of these* [facets] *to open up a new window to give you that information. But that's one of those things where, once you've used it once or twice, you'll go "yes, of course I do that"*".

Participants recommended that user feedback and expert review drilldown pages should resemble common e-Commerce rating systems such as Amazon; provide quick summaries including average rating, ratings distribution, and number of feedbacks; and allow sorting the feedback/reviews lists. Regarding the citations information drilldown, it was recommended to adopt more academic citations style and provide citations filtering and sorting functionality.

Two study participants expressed a strong need for a "*visual cue of the spatial extent of the dataset*". When examining the datasets' descriptions, participant 1 struggled to identify the exact spatial extents of the datasets of interest, consequently, he proposed:

> "*It would be good, if here I almost knew what extents these* [datasets] *were over. I don't know how hard it would be, almost to include a tiny map here* [pointing at the detailed information area] *with just the area that* [the dataset] *is valid for highlighted with the rest kind of faded out a bit*".

Participant 3 also highlighted the importance of temporal coverage, including mean frequency, and recommended to provide "*a* [graphical] *timeline* [of temporal coverage] *showing* [a] *point at every time at which* [the data] *was available. Or, if it was continuous time, a shaded region along the timeline*". It was suggested that "*a* [spatial coverage] *map plus a timeline…* [would be] *much quicker for* [users] *to process than text*". It was also recommended to provide clear information about the creation, publication and updated dates because "*a lot of the time* [users] *need to understand the date range within the data*" and these are "*certainly one of the key criteria*" (participant 2).

9.3.6 Physical Data Acquisition

After identifying potential datasets of interest, participants stated that at that point they would want to look at the actual data to make the final dataset selection decision. For instance, participant 1 commented that "[he has] *identified a couple of datasets that sound promising but* [he] *need*[s] *to have that data and have a better look at them*". Participant explained that "[he] *would just want a link to take* [him] *over to that* [producer] *site, to the same data that is viewed* [in the tool]".

Participant 2 also explained that the data portals that he is accustomed to often provide the '*download in different formats*" buttons. He also noted that these long lists of download buttons are "*too busy*" and he would rather have a single button to navigate him to a separate page with a table of available download formats.

9.3.7 Alternative Search Results View

This theme emerged from participants' requirements for an alternative, more traditional visualisation of the dataset search results. As the study results indicated, two study participants found the starfield visualisation of the datasets insufficient to make a fully informed dataset selection decision.

After applying filtering options and inspecting a number of datasets, participant 1 struggled to identify a dataset that would fully satisfy the prescribed scenario 3 requirements. The participant cleared all the filtering and started to inspect each dataset one by one highlighting the datasets that seemed relevant. Participant explained that "[he] *could, because this is quite laborious going through* [the datasets] *one by one,* [he] *could filter* [the search results] *down by one of* [his] *criteria, but* [he was] *just worried* [he could] *miss something if* [he did]". He explained that the tool is brilliant for filtering the search results because "[it] *clearly showed* [him] *that it's only these three datasets that meet* [his] *criteria*", but if the filtered results do not satisfy the requirements, then clicking on 100s of dataset labels one by one becomes far too inefficient. Participant then suggested that a table view with the titles and abstracts would be more useful in this particular instance.

Participant 6 was quite doubtful of the benefits of a starfield display visualisation and explained that to her "*it seem*[ed] *to be strange to be selecting datasets off this sort of matrix*" and she "[did not] *really understand what this* [tool was] *showing* [her]". The participant argued that the tool is "*too impersonal*", "*not helpful*" and "*made* [dataset selection] *awkward by putting* [the datasets] *in different places* [on the screen]". She commented:

> "*I'm used to choosing my datasets from a list where you can see exactly what the name of the dataset is and* [… have] *a bit of information about each of them.*

> *So without seeing information on each* [… dataset] *it is hard to know if I've got my filtering right*".

The participant stressed that she personally "*would want to see a list*" of datasets that she could then rank and filter. She further explained that she usually searches for geospatial datasets in specific repositories and "*can do a lot of the filtering in* [her] *head from seeing the name of the dataset*". In the post-study discussion the participant admitted that a starfield visualisation tool could potentially be useful if she could apply very specific filters (e.g., snow depth, snow covered area, snow water equivalent, etc.).

These results are somewhat unsurprising since it cannot be expected that all geospatial data users will be inclined to use the starfield-based dataset discovery tools. Consequently, it is suggested that geospatial dataset discovery systems provide various options for visualising geospatial datasets to suit specific user needs. As already discussed in Section 9.3.1, more sophisticated dataset search queries and additional search options are *absolutely* essential to ensure effectiveness and acceptability of the starfield-based dataset discovery systems.

9.3.8 Informational Aspects

The informational aspects theme encompassed participants' comments about the informational facets included as part of the GEO label. The analysis of the study transcripts identified that geospatial data users highly value producer information and user feedback, with the compliance with standards being unimportant when making a dataset selection decision.

Producer Information

The study results highlighted that a significant importance is placed on producer profile information by geospatial data users when evaluating dataset quality and trustworthiness. Similar to the findings from the previous study (see Section 4.2.6), perceived producer credibility and availability of contact information were highly influential when making a dataset selection decision. After making a dataset selection decision, participant 1 commented:

> "*This dataset pretty much ticks all three of my boxes. It's good. It's just, again, it's missing the producer comments and the uncertainty measures. But, because I've got the producer information I could give them a ring and see what they have to say about the dataset that I needed*".

These findings indicate that geospatial data users are highly inclined to contact the dataset producer when supplied dataset information is not sufficient to make an informed dataset selection decision.

While inspecting producer profile information, participant 3 remarked that "*anyone called Freddy is not going to produce a decent dataset*". Participant 4 also seemed quite surprised to see "*Freddy!?*" in the producer profile information. It should be noted that this producer information was genuine and was not in any way modified for the purpose of the study. Participant 3 further supported the importance of perceived producer credibility:

> "*Right, I've got myself 6 datasets.* [selected the most complete label] *I'll have this one straight away, I know that, because it's from somebody from the JRC and I'm working with the JRC*".

These participants' reactions to the source of data confirm the previous findings (see Section 4.2.6) that producer reputation plays an important role in perceived quality and trustworthiness of geospatial datasets.

User Feedback

Similar to the results of the initial investigation (see section 4.2.3), one of the dominant themes to emerge from this study was the importance of user feedback, with five out of six study participants indicating its strong influence on the data selection process. Throughout the study, participants either directly argued the importance of user feedback or actively studied feedback comments to discover additional information on datasets' purpose and applicability to their needs. Participants stated that they "*would trust user feedback or an expert review more than* [they] *would just a standards stamp*" (participant 4). It was also suggested that "*the user feedback and the expert reviews stuff is probably the most important* […] *because that gives you the idea of what real people have used the data for and what their real experience was*". In cases when available datasets did not fully satisfy their requirements, participants consulted feedback of their peers. For instance, when none of the inspected datasets seemed to address the prescribed scenario requirements, participant 1 carefully inspected user and expert comments to identify any additional relevant information that could reassure his dataset selection decision. He further indicated his trust in peers by stating his willingness to "*contact*" other data users to acquire additional information about the datasets of interest.

Participant 5 went as far as to ignore the prescribed scenario requirements and focused instead on user feedback because this aspect was of personal value to her; she commented:

> "*So, although this exercise doesn't state that I should look at user commentaries, I personally find that very useful* […] *I would like to see users like me, not-expert users, their opinion, their experiences using the datasets*".

Participant 5 largely based her dataset selection decisions on user comments because she "*trust*[ed]" her peers. Only when none of the user comments seemed to answer respondent's dataset selection questions, she was referring to the expert reviews. She stressed the importance of non-expert user feedback by saying:

> "*For example this* [dataset] *looks very good because somebody like me, which is not an expert, has been studying what I need. Has been studying not exactly agricultural studies but he has done the study in the UK in the more or less same range of dates. And that is making me think that it is very likely that I will find all I need from this dataset.*"

Surprisingly, after inspecting a number of user feedbacks, participant 2 identified a connection between several users who provided similar feedback comments. At the end of exercise 3, the participant narrowed down his choice to 3 datasets and explained:

> "*I'd probably look at those three* [datasets]*, and the reason why is – after all that filtering, it appears to be the same reviewers, based on the user comments and on the citations, they've used it for pretty much the same study. So if they've done it, I'll just copy what they've done*".

This was a very interesting discovery since resemblance in feedback was not initially intended when generating the data, and resulted from reusing the fragments of fictitious feedback text to speed-up the sample data generation. Participant 2 further commented: "*I'm kind of a sheep, I follow everyone else really. If everyone else has done it before, I'll be like "yea, I'm on a right track!"*"". Emphasising the importance of user and expert feedback, participant 2 proposed adding a filtering option that would allow filtering or grouping datasets that contain feedback from a particular user. The participant also suggested grouping user feedback, expert reviews and citations information into a sole facet, stating that "[he] *would group users and academics* [be]*cause it's a feedback thing and, for* [him]*, there is always a potential with academic feedback to again go into gobbledygook that* [he] *do*[es]*n't understand, so* [he] *just look*[s] *into all feedback*".

Standards compliance

Contrary to the previous findings (see Section 4.2.1), the study results indicated low importance of geospatial datasets' compliance with standards. Participant 3 stated that, "*normally,* [he] *would never worry about the standards compliance*" when selecting a dataset to use. He explained that, while it is generally easier to work with standards complaint data, non-compliant geospatial data can be processed to a workable format. Perhaps, this indicates that standards are less important to expert users who are capable of processing raw data into a required product. Non-expert users, on the other hand, might not have the required knowledge to produce data products from non-compliant datasets.

Reliability of standards was also questioned, with participant 4 arguing that "*the fact that something's got an ISO standard just means that it adheres to some kind of processes and ticks some boxes… Having worked to standards,* [she] *know*[s] *they are a little bit dubious sometime*". Participant 4 also stated that she "*would trust a user feedback or an expert review more than* [she] *would just a standards stamp*".

9.3.9 Post-Study Session Feedback

In the post-study discussion, the participants agreed that the GEO label itself is intuitive, easy to use and is effective at "*very quickly* [providing] *a sense of what is available and what is not*" (participant 1). The participants stated that the colours clearly conveyed grouping of "*the same area of concept*[s]" and also supported facets' recognition and recall. All the eight GEO label informational aspects were perceived useful and it was suggested that eight facets is a sufficient number to effectively support dataset intercomparison. Only participant 1 suggested a modification, he argued that the expert review icon is not intuitive and proposed to use a user icon but with a '*professor hat*' to highlight the 'expert' aspect of the facet. Regarding the circular label design, participant 6 commented:

> "*I like the fact that it's in a circle, that sort of clock design, because it doesn't give any precedence to any one thing over any other thing, which is nice*".

As can be seen from Table 9.5, five out of six participants rated the GEO label as '*effective*' at conveying the availability of a dataset's quality information. None of the study participants rated the GEO label as ineffective to any degree. Overall, the participants agreed that ""*it is a good label. It's just getting over that first step of 'what does each one of these* [facets] *mean?', 'what does the half-shaded mean?', 'what do the colours mean?', 'do the colours represent something?'"*".

The post-study discussion revealed that, overall, participants found the GEO LINC tool "*very effective*" and "*fantastic at showing what's provided in terms of the metadata*" (participant 1). It was suggested that the interface is "*user friendly*", "*tremendously intuitive*" and "*it* [is] *clear what* [is] *happening straight away*" (participant 3). Participants appreciated the "*visual*" way of comparing metadata records and stated that the tool "*allow*[s] *good comparison side by side*" (participant 1) and effectively supports geospatial dataset selection. Participant 3 summarised his experience of using the GEO LINC tool as follows:

"*I really like that. I think, that works really, really well. I would be more likely to use a tool like this. I think what I like more than anything else is the interactivity. And that is quite important*".

Participant 5 additionally commented that the tool "*is very quick, it is very easy to apply the filters,* [and] *it looks very nice* [in terms of] *the design*". As can be seen from Table 9.5 and Table 9.6, four out of six study participants rated the GEO LINC tool as '*somewhat effective*' at supporting dataset intercomparison and selection and five out of six participants rated the tool as '*easy*' to use. None of the participants rated the tool as ineffective or difficult to use to any degree. To support their ratings decisions, participants commented that while the tool "*is effective*[, it] *could be more effective*" once their recommendations and comment are addressed.

Table 9.5: Participants' ratings of effectiveness of the GEO label and the GEO LINC tool.

Effectiveness	**GEO Label**	**GEO LINC**
Somewhat Effective	0	**4**
Effective	**5**	0
Very Effective	1	2

Table 9.6: Participants' ratings of ease of use of the GEO LINC tool.

Ease of Use	**GEO LINC**
Easy	**5**
Very Easy	1

9.4 Study Limitations

Since controlled studies are typically conducted in an artificial environment, the most widely cited limitation of the approach used here is that of external validity and generalisation to real-world settings (VanSchaik, 1990). To mitigate this issue, experienced geospatial data users and experts were engaged as participants in order to collect rich representative data, irrespective of environment of study. To ensure realism of the dataset selection exercises, the study scenarios were based on the user stories elicited from the initial investigation interviews. Finally, since most users would investigate datasets and make dataset selections in an office-like environment using the technology employed for this study, the actual study environment (set-up to be office like or actual offices) was, in actual fact, quite realistic.

The second major limitation of this study concerns the subjective nature of the coding and the derived themes. This is, however, inevitable when interpreting qualitative data that does not have well-defined constructs. To ensure validity of the study results, this research followed a structured thematic analysis process widely utilised for qualitative data analysis.

Another potential limitation of the study is use of fictitious geospatial datasets for the selection exercises. While it is recognised that fictitious user feedbacks, expert reviews, citations, etc. could potentially influence or even discourage study participants, at present there exist no real metadata records that include all GEO label informational aspects and so data had to be artificially created for use in the study. That said, great care was taken to ensure that all the sample metadata records were as realistic as possible. The core dataset information such as title, abstract, producer information, and dataset purpose were based on real-world geospatial dataset metadata records. A large variety of feedback comments, both positive and negative, were generated to ensure realism and consistency. In addition, all the citations were based on real journal publications and were carefully selected to match dataset description and purpose. Based on the fact that the study participants demonstrated active engagement in the dataset selection exercises, it is felt that use of fictitious data did not influence the study results.

9.5 Summary and Conclusions

This chapter presented the study that was conducted as part of the final phase of the GEO label research to evaluate the usability and effectiveness of the proposed GEO LINC tool. The findings of the human-subject evaluation study indicate that, with some modifications and improvements, the GEO LINC tool has real potential to effectively support geospatial dataset discovery and intercomparison. Overall, the results indicated that the tool intuitively and effectively facilitated 'at a glance' dataset intercomparison, dataset filtering, acquisition of detailed dataset information, and side-by-side metadata comparison. While information availability filtering was intuitive, additional filtering functionality (resizing of the search results GEO labels) was confusing with some filtering options perceived as "*artificial*" and not applicable to geospatial data. The analysis of the study transcripts revealed that the two most dominant themes were the ability to define complex SQL-like search query strings and the importance of user feedback when selecting a dataset to use.

Most importantly, the results indicated that, in order for a starfield display to be effective at facilitating geospatial dataset discovery, the initial search functionality must support sophisticated flexible queries and provide sufficient search parameters to return relevant to the user results. It is also suggested that, in additional to the starfield display, the GEO LINC tool should also support more traditional table views to better suit different user needs and to allow for an overview of detailed search results information to ensure the correctness of the defined search and filtering criteria.

Table 9.7 outlines recommended modifications and improvements to the GEO LINC tool based on the study results.

Table 9.7: Recommended modifications and improvements to the GEO LINC tool.

Feature	**Proposed Modifications/Improvements**
	Initial Search Criteria
Keywords	The tool *must* provide functionality to support complex Boolean expressions in the keywords search field.
Saving filtering queries	The tool should provide functionality for saving the search queries.
Instructions for map use	The tool should display a pop-up message with instructions when the interactive area selection map is clicked on.
Spatial coverage map	The tool *must* provide a small map that visualises dataset's spatial extent.
Start/end dates	The tool should provide clear instructions on supported date formats. The tool should support more flexible date formats.
Temporal coverage timeline	The tool *must* provide a visual timeline that shows dataset's temporal extent.
Access/use constraints	The access and use constraints fields should be modified. These options should be presented as check boxes to allow selection of several alternatives. The access and use constraints fields should be provided with clear guidelines on what each constraint means.
Last updated date	The tool should provide the 'last updated date' option in the search query.
	'At a Glance' Dataset Intercomparison
Drag-and-drop	The tool should allow drag-and-drop functionality for customising starfield visualisation of search results.
Option to return to original positions	The tool should provide functionality to return the search results to their initial position.
	Side-by-side Metadata Comparison
Drag-and-drop	The tool should allow drag-and-drop functionality for customising the list of highlighted datasets.
Filtering	The tool should support filtering of the highlighted datasets list.
	Side-by-side Metadata Comparison
Side-by-side comparison of metadata fields	The tool should allow side-by-side comparison of all metadata fields of two datasets.
	Filtering
Free-text filters	The tool *must* provide functionality to support complex Boolean expressions in the free-text filtering field.
Lineage filtering	The 'number of process steps' filter must be removed because it is "*artificial*" and does not add any value. The tool should provide filtering options to filter datasets on processing characteristics, e.g., instruments, processing types, types of data (e.g., Level 0), etc.
User/expert ratings	The 'user/expert ratings' filtering options should be modified to show star ratings distributions where average rating would be accompanied by a number of datasets that fall into this category.
User/expert feedbacks filters	The 'number of feedbacks and expert reviews' filtering options should be modified. The tool should provide a slider that would range from 0 to the maximum number of feedbacks/reviews available for a single dataset. The tool could provide a filtering option to allow filtering by user/expert name.

Feature	Proposed Modifications/Improvements
Citations filter	The 'number of citations' filtering option should be modified. The tool should provide a slider that would range from 0 to the maximum number of citations available for a single dataset. The tool should provide filtering options to allow filtering by citation type, date, and author.
Additional filtering functionality	The resizing functionality requires reconsidering because it is not intuitive. If label resizing is used, the changes must be more distinct.
Indication whether dataset matches filtering	The search results GEO labels should visually indicate whether the dataset matches additional filtering. Facets that do not match the filtering could be greyed-out.
Collapsible filtering area	The tool should provide a visual indication when the additional filtering options are applied. The collapsible content panel's header could be coloured in the same colour as the facet to indicate that filtering was applied.
Saving filtering queries	The tool should provide functionality for saving the filtering queries.
Dataset Details Acquisition	
Reiterate hover-over information	The tool should reiterate a summary of the hover-over information as part of detailed dataset description.
User feedback/expert reviews drilldown	User feedback/expert reviews drilldown pages should resemble common e-Commerce rating systems such as Amazon and should provide quick summaries including average rating, ratings distribution, and number of feedbacks; and should allow sorting the feedback/reviews lists.
Citations information	Citations information drilldown pages should adopt more academic citations style and provide citations filtering and sorting functionality.
Data Acquisition	
Link to producer website	The tool should provide links to the dataset producers' websites.
Data download page	Where applicable, the tool should provide a separate data download page with a list of all the available downloadable formats.
Alternative Search Results View	
Table view	The tool should additionally provide a table view of the datasets search results.

The modifications and improvements outlined in Table 9.7 not only applicable to the GEO LINC tool but also present more general user-defined guidelines on the functionality and features required to effectively facilitate evaluation of geospatial dataset quality and fitness for use. It is believed that with the proposed user-defined modifications, the GEO LINC tool will provide effective decision support and facilitate improved geospatial dataset discovery and intercomparison. The following chapter provides thesis conclusions and describes practical and scientific implications and contributions to knowledge.

Chapter 10 Conclusions and Future Research

10.1 Thesis Summary

This thesis described UCD research conducted to establish the concept of a GEO label and identify the role it should serve in the visualisation of geospatial data quality and trustworthiness. The thesis presented six phases of research and development conducted to: (a) identify the informational aspects upon which users rely when assessing geospatial dataset quality and trustworthiness; (2) elicit initial user views on the GEO label role in supporting dataset comparison and selection; (3) evaluate prototype label visualisations; (4) develop a Web service to support GEO label generation; (5) develop a prototype GEO label-based dataset discovery and intercomparison decision support tool; and (6) evaluate the prototype tool in a controlled human-subject study.

The results of the preparatory phase (initial investigation) revealed 11 GEO label-appropriate themes and, from these, 8 facets were further identified as potential candidates for inclusion in the GEO label, namely: dataset compliance with international standards; side-by-side metadata records comparison; community advice; dataset ratings; expert value judgments; the reputation of the data producer; producer comments on dataset quality; and links to dataset citations. The results also suggested that a GEO label would best serve a drill-down function whereby, at the top level, it visually represents the *availability* of specific informational elements for its associated dataset and, thereafter, permits users to click the label to drill down into and interrogate the detail for each informational element.

These findings then formed the basis of the Phase I study which investigated geospatial data producers' and users' views on the concept of a GEO label and the role it should serve. The results of the study revealed that the majority of users and producers strongly supported the notion of a GEO label providing an all-in-one drill-down interrogation facility that would combine 8 informational aspects identified in the initial investigation. On further reflection, however, since side-by-side metadata visualisation would require at least two datasets and does not represent an informational facet of a single dataset alone, it was decided not to include this function in the GEO label visualisation itself. Additionally, since the initial study interviewees and Phase I respondents indicated that objective (quantitative) quality information was important in the assessment of dataset fitness-for-use, it was decided to include an additional facet – 'dataset quality information' – in the GEO label. On the basis of these findings, three prototypic graphic representations (i.e., static images) of the GEO label were developed; these combined the 8 identified and confirmed informational aspects, namely: dataset producer information; producer comments on the dataset quality; dataset's compliance with standards; user feedback; user ratings of the dataset; expert reviews; dataset citations; and quantitative dataset quality information.

The phase II questionnaire-based study was then conducted to evaluate the effectiveness of, or potential issues with, the proposed GEO label designs. The findings of this study led to a conclusion that the final user-dictated graphical GEO label representation should either be a hybrid of two of the tested prototype designs (the circular and star-based designs) or should adopt a modified version of the rectangular design, comprising the 8 informational aspects but solely conveying information availability (i.e., changing the user rating meaning). Based on these findings and further feedback and recommendations from geospatial data experts, the graphical representations of the GEO label underwent modifications and improvements: a new lineage information facet was introduced to convey availability of provenance and lineage information; user feedback and user ratings facets were combined into a sole user feedback facet; the ISO symbol of the standards compliance facet was replaced with a more generic 'target with a tick' icon. The results of the geospatial community voting for the final GEO label design showed that the circular GEO label representation was the most favoured by the geospatial community because it was more effective at conveying information availability and would better support cognitive processing of a large number of dataset labels at once. Consequently, on the basis of the voting outcomes, it was concluded that the final GEO label visualisation should adopt the circular layout.

To support use of the graphical GEO label, a Web service was developed to generate GEO label representations for datasets by combining producer metadata (from standard catalogues or other published locations) with structured user feedback and, based on evaluated information availability, building a dynamic SVG (Scalable Vector Graphic).

Although the services primarily rely on the GeoViQua quality models, an external XPath configuration file which is used for determining whether information is available can be adapted to support any XML-based metadata models.

Following the service implementation, a prototype GEO LINC tool has been developed to provide an innovative approach to visualising geospatial dataset metadata records and selecting datasets that fit user needs. The tool effectively utilised the GEO label and transformed it into a powerful interrogation facility to facilitate dataset filtering based on the informational aspects' availability recorded in eight GEO label facets.

The final Phase III study was conducted to evaluate the usability and effectiveness of the GEO LINC tool. Overall, the results indicated that the tool intuitively and effectively facilitated 'at a glance' dataset intercomparison, dataset filtering, acquisition of detailed dataset information, and side-by-side metadata comparison. Most importantly, the results indicated that, in order for a starfield display to be effective at facilitating geospatial dataset discovery, the initial search functionality *must* support sophisticated flexible queries and provide sufficient search parameters to return relevant to the user results.

10.2 Practical Implications

Interoperability of the developed GEO label service facilitated GEO label integration into several real-world geospatial data portals and applications. In the server side, URLs for generating GEO labels are being integrated into the metadata records distributed by the GeoViQua Discovery and Access Broker (DAB-Q)[13] (see Figure 10.1). The GEO label is considered as a representation of the dataset and is being embedded in an ISO19115:2003 gmd:MD_BrowseGraphic element. Using this approach, the GEO label can be integrated into any ISO19115:2003 compliant metadata record.

```
[...]
<gmd:graphicOverview>
   <gmd:MD_BrowseGraphic>
      <gmd:fileName>
         <gmx:FileName src= "ADD_METADATA_URL">GeoViQua.GeoLabel</gmx:FileName>
      </gmd:fileName>
      <gmd:fileDescription>
         <gco:CharacterString>A GEO Label with drilldown</gco:CharacterString>
      </gmd:fileDescription>
      <gmd:fileType>
         <gmx:MimeFileType type="image/svg+xml">SVG</gmx:MimeFileType>
      </gmd:fileType>
   </gmd:MD_BrowseGraphic>
</gmd:graphicOverview>
[...]
```

Figure 10.1: An example of XML code for integrating the GEO label into metadata records.

[13] http://geoviqua.org/GeoViQuaBroker.htm

On the client side, the GEO label has been integrated into a prototype extension of the GEOSS portal where an individual GEO label representation is provided with each geospatial dataset (see Figure 10.2). As part of the 6th Phase of the GEOSS Architecture Implementation Pilot (AIP-6), the GEO label has been incorporated into the George Mason University (GMU) portal[14] (see Figure 10.3). The GEO label has also been integrated into a prototype development version of the Global Carbon Atlas platform (see Figure 10.4). The GeoViQua project has additionally integrated the GEO label into the popular metadata catalogue and editor application, GeoNetwork[15] (see Figure 10.5).

Figure 10.2: GEO label integration into a prototype extension of the GEOSS portal.

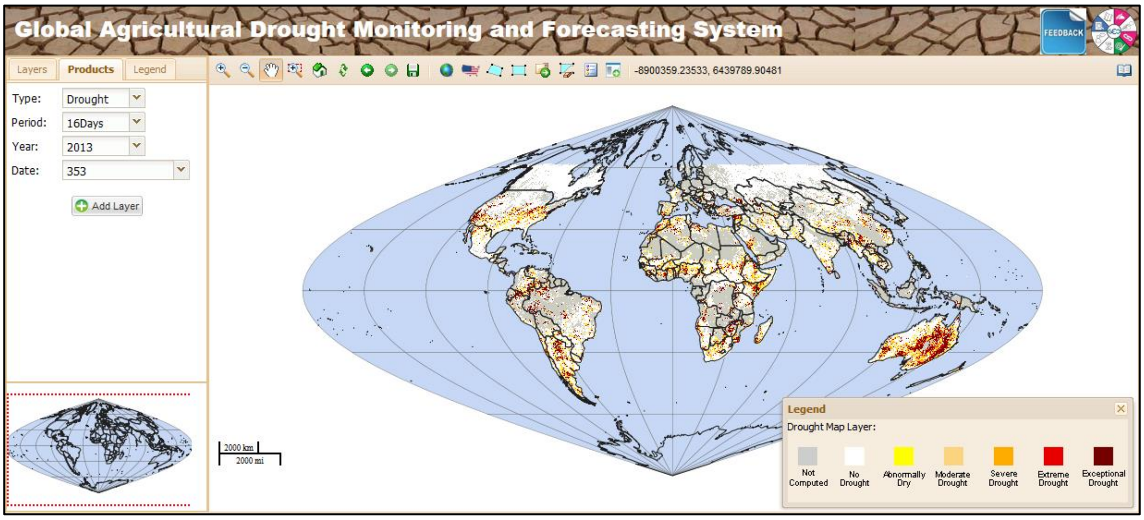


Figure 10.3: GEO label integration into the GMU portal.

[14] http://gis.csiss.gmu.edu/GADMFS/
[15] http://uncertdata.aston.ac.uk:8080/geonetwork/srv/eng/main.home

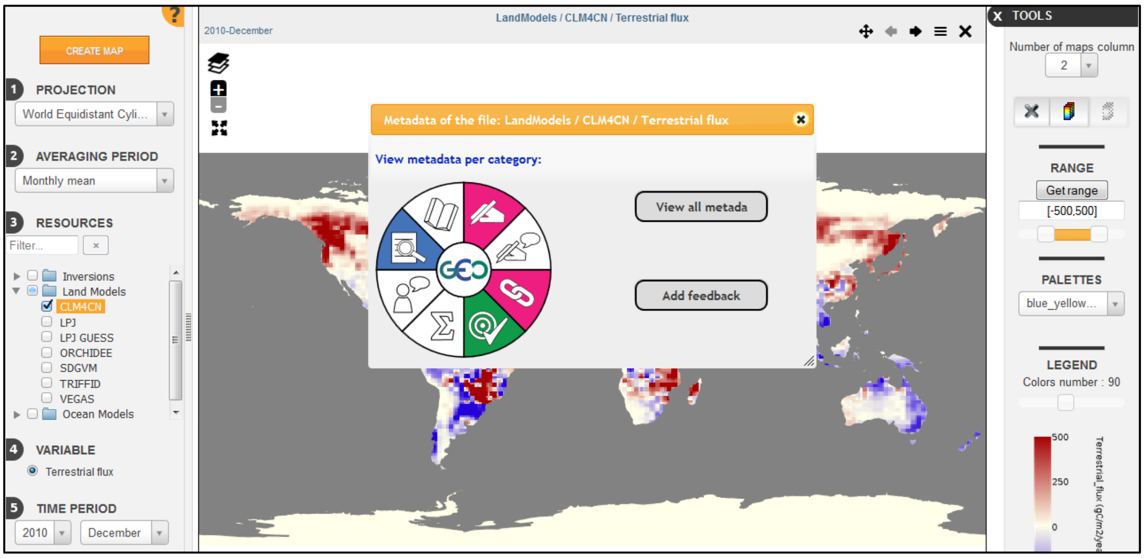


Figure 10.4: GEO label integration into a prototype of the Global Carbon Atlas platform.

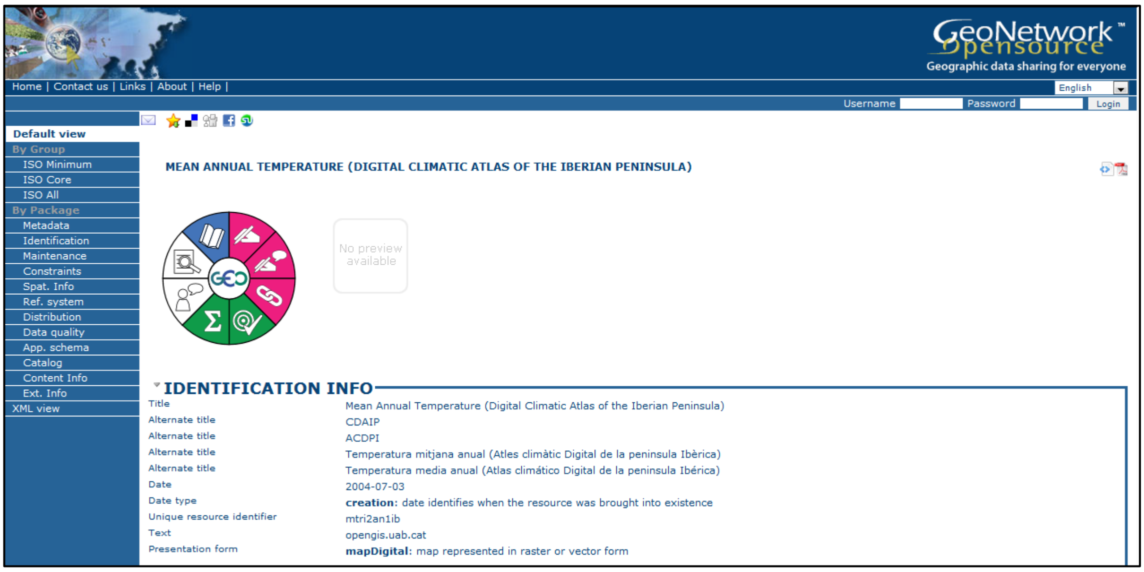


Figure 10.5: GEO label integration into the GeoNetwork catalog application.

In May 2014, the GEO Standards and Interoperability Forum (SIF) officially nominated the GEO label for the operational use within the GCI. As stated in the nomination letter, "*the SIF considers the GEO label an important contribution to enhance the user experience during data discovery and, therefore, recommends its adoption into the GEOSS Common Infrastructure*". The nomination letter has been forwarded to the GEO Infrastructure Implementation Board (IIB) and is now being reviewed by the board members. This official nomination demonstrates the GEO label acceptance by the GIS and GEO community and is a great success for this research project.

10.3 Scientific Implications

Section 1.2 presented four research questions that were the primarily focus of this scientific investigation. This section discusses how these research questions were answered.

1. **What are the key informational aspects of geospatial datasets upon which users rely when evaluating quality and trustworthiness of geospatial datasets and making a dataset selection decision?**

The answer to this question was provided via analysing and interpreting the initial investigation interview results (Chapter 4) and subsequently confirmed via the more structured Phase I study (Chapter 5). As the answer to this question, this research identified the following user-defined geospatial data quality and trust indicators:

- metadata completeness;
- compliance with international standards;
- producer comments (soft knowledge about the quality of the data);
- reputation of dataset provider/producer (producer profile information);
- user feedback and ratings of the dataset;
- expert reviews;
- citations information;
- provenance information;
- quantitative quality information; and
- licensing information (use and access constraints).

While these informational aspects resulted from the initial interviews and Phase I study, the subsequent studies confirmed the validity of these findings (see Chapter 6 and Chapter 9).

2. **What role should a GEO label serve to effectively support evaluation of geospatial dataset quality and trustworthiness?**

The answer to this major research question was obtained via the Phase I investigation (Chapter 5). It was established that the GEO label should provide an *all-in-one drill-down interrogation facility* whereby, at the top level, it should visually represent the availability of 8 informational aspects for its associated dataset and, thereafter, permit users to click the label to drill down into and interrogate the detail for each informational element. These findings were confirmed via the Phase III study where, following the direct interaction with the dynamic GEO label representations, geospatial data expert users agreed on the usefulness and effectiveness of the developed GEO label.

3. **How should a GEO label summarise and represent dataset quality and trustworthiness information in a way which permits a user to easily assess the relevance of a dataset for their needs, and interrogate the specific aspects which are key to their application?**

This research question was answered via the iterative GEO label representations' design and the Phase II studies (Chapter 6). The results of the studies indicated that the GEO label should adopt a segmented circular design, represent the informational aspects using icons, and convey information availability through the boldness of the facet's colour. The practical GEO label service implementation (Chapter 7) further demonstrated how the GEO label can provide drilldown and hover-over functions to facilitate the interrogation of specific informational aspects.

4. **How can a GEO label be applied to facilitate an innovative approach to decision support in geospatial dataset intercomparison and selection?**

The design, development and evaluation of the GEO LINC tool (see Chapter 8 and Chapter 9) provided answer to this research question. Building on the concepts of a starfield display and real-world geospatial data portals, GEO LINC offered a novel approach to the visualisation and intercomparison of a large numbers of datasets. The tool demonstrated how the GEO label can be transformed it into a powerful interrogation facility that allows dataset search- and interrogation-by-quality. The evaluation study results validated the effectiveness of the GEO label-based tool in facilitating an innovative approach to decision support in geospatial dataset intercomparison and selection.

10.4 Contributions to Knowledge

For over a decade, GIS researchers and scholars have been theorising about user-centric metadata and user-oriented quality indicators for geospatial datasets. While these theories provided useful insight into potential approaches to communicating geospatial data quality information and improving quality evaluation, there has been very little research involving real geospatial data users to confirm or reject these theories. This research therefore adopted a user-centred design approach and demonstrated how UCD methods can be applied in the GIS domain to develop solutions that are not just usable but useful to geospatial data users. Following the discovery and validation phases, this research established a set of *user-defined* geospatial data quality and trust indicators and confirmed the initial user-centric metadata proposals. The strongest theme revealed throughout all the phases of this research was the importance of user and expert feedback. It was discovered that peer recommendations are of greatest value to geospatial data users and that users would want to see e-Commerce review and rating functionality available in geospatial data portals, catalogues and clearinghouses. Practical tools to support feedback functionality are now being developed in the GIS domain. To ensure the effectiveness of these feedback tools, developers should draw on the strengths and weaknesses of e-Commerce websites where consumer feedback has been successfully used for many years to engender consumer trust in products and services. The second strongest theme revealed in this research was the reputation of dataset producers. The findings indicated that, to engender user trust, geospatial data producers need to supply complete metadata records, supporting documentation, and contact information with the datasets that they produce. Interestingly, views on standards compliance varied across the studies. While geospatial data users acknowledged the importance of international standards, it is also recognised that standards compliance does not necessarily guarantee that the data is of high quality. Perhaps, if the supplied metadata records were generally complete, users would place more confidence in compliance with international standards.

Via extensive literature review, discussions and reflection, this research also demonstrated how research on trust in other domains can be applied to geospatial data and GIS applications. It was revealed that geospatial data quality and trust indicators largely mirror e-Commerce trust triggers. Drawing on the extensive research and knowledge in the e-Commerce domain, it is suggested that the GIS domain should employ the trust promoting mechanisms to engender user trust in geospatial data and GIS applications.

The main contribution to the scientific knowledge is the establishment and development of the GEO label which visually summarises user-defined geospatial data quality and trust indicators and provides a novel approach to visualisation of metadata records. Via practical implementation, it was demonstrated that it is possible to develop an effective voluntary quality label without having to establish a new standard, standardisation body or a certification programme. The developed solution not only fulfilled the needs of the geospatial community, but also addressed the STC's initial vision of a GEO label that would comprise objective labelling (producer metadata) and subjective labelling (user-focused metadata). The success of this research is reflected in its acceptance by the GEO community and the official proposals to operationalise the GEO label.

The GEO label integration into geospatial data portals should raise community awareness of metadata incompleteness. It is much easier to conceal metadata incompleteness in complex XMLs files; even tabular views can give a false impression of information availability since some records provide long lists of keywords, responsible parties, and points of contact. The GEO label, on the other hand, provides an overall at-a-glance view of information availability. Consequently, for geospatial data producers, the GEO label can act as a graphical template of information that should be provided with every geospatial dataset and should encourage producers to supply rich metadata.

The development and evaluation of the prototype GEO LINC tool contributed a novel approach to decision support in geospatial dataset intercomparison and selection. The tool demonstrated a practical implementation of search- and interrogation-by-quality functionality which is not currently provided by real-world geospatial data portals and clearinghouses. This research further revealed that, to ensure effective dataset discovery, geospatial data portals and clearinghouses should support sophisticated flexible queries and provide sufficient search parameters.

The significance of this research is that it developed novel decision support mechanisms that enable a more efficient and informed evaluation of geospatial dataset quality and trustworthiness, and facilitate more effective dataset intercomparison and selection. While the GEO label has been developed to communicate the availability of geospatial data quality, these concepts can be applied in other domains where artefacts are evaluated and intercompared based on multiple attributes.

10.4.1 Future Research Directions

Since in the absence of this research it would not be possible to identify new research directions, further research is considered to be an important part of the contribution to knowledge.

The findings from this research opened a number of new research directions. First of all, the conducted studies only involved novice users who had no previous knowledge of the tools being evaluated. The studies elicited initial user and expert views on the GEO label and the GEO LINC tool. Future research could focus on long-term studies to evaluate the effect of learning on usability and effectiveness. Since the research presented here adopted a qualitative exploratory approach, further research could focus on applying quantitative methods to statistically validate the findings.

To conduct controlled evaluation studies, this research used fictitious data and fictitious GEO labels. While the results indicated effectiveness and intuitiveness of the GEO label, it would be beneficial to confirm these findings via evaluations in an operational environment. The GEO label is already integrated into several real-world GIS applications, if these tools become widely used by the GIS community, internal system logging could be used to collect statistical data such as frequency of access of the drilldown pages, time spent inspecting information, correlation between the GEO labels' completeness and the inspected datasets. Online feedback questionnaires could also be used to collect user views on the operational GEO label. Users could be prompted to complete short feedback surveys with a view of improving their dataset search experience.

The Phase III study revealed a number of modifications and improvements that should be applied to the GEO label tool to ensure its effectiveness. Additionally, the tool only provided a small local database of fictitious geospatial datasets. To further evaluate the GEO LINC tool, the proposed modifications should be implemented and the tool should be connected to a real geospatial data brokering system to provide the access to real geospatial datasets. The tool could then be made available online for public use to test its usability and effectiveness in the operational environment. Internal system logging and short feedback questionnaires could be applied to collect statistical data and subjective user feedback.

Another interesting research direction would be to investigate whether the GEO label could usefully fulfil a role of a global 'health' indicator for the geospatial data portals and clearinghouses, i.e., visualise the overall completeness of the metadata records. The hover-over or the drilldown function could then be used to provide global portal's statistics, for example, producer profile facet could provide the total number of datasets that contain producer contact information. When displaying the search query results, the 'health' indicator could visualise the overall information availability in the returned datasets. This global overview of the search results would allow the user to immediately identify whether any of the returned datasets contain the information that the user is interested in. For instance, the 'health' indicator could indicate that none of the search results contain any lineage information and only 2 out of 100 datasets contain user feedback.

References

5 Star Open Data, 2012. 5 Star Open Data [Online]. Available at: http://5stardata.info/ [Accessed on: 10 June 2014].

Abras, C., Maloney-Krichmar, D. and Preece, J., 2004. User-Centered Design. *In:* Bainbridge, W. (ed.) *Encyclopedia of Human-Computer Interaction*. Sage Publications: New York.

Ahlberg, C. and Shneidermnan, B., 1994a. Visual Information Seeking Using the FilmFinder. *CHI'94*, 24–28 April, Boston, Massachusetts, USA. pp. 433–434.

Ahlberg, C. and Shneidermnan, B., 1994b. Visual Information Seeking: Tight Coupling of Dynamic Query Filters with Starfield Displays. *CHI'94*, 24–28 April, Boston, Massachusetts, USA. pp. 313– 317.

Ahlberg, C., Williamson, C. and Shneiderman, B., 1992. Dynamic Queries for Information Exploration: An Implementation and Evaluation. *CHI '92*, 3–7 May, Monterey, California, USA. pp. 619–626.

Aiken, K. D. and Boush, D. M., 2006. Trustmarks, Objective-Source Ratings, and Implied Investments in Advertising: Investigating Online Trust and the Context-Specific Nature of Internet Signals. *Journal of the Academy of Marketing Science*, 34(3), pp. 308-323.

Alam, A., Subbiah, G., Khan, L. and Thuraisingham, B., 2007. DAGIS: A Geospatial Semantic Web Services Discovery and Selection Framework. *Lecture Notes in Computer Science*, 4853, pp. 268-277.

Alexander, M., 2012. Decision-Making using the Analytic Hierarchy Process (AHP) and SAS/IML. *Southeast SAS Users Group (SESUG) 2012*, 14–16 October, Durham, North Carolina. pp. 1–12.

Angehrn, A. A. and Lüthi, H.-J., 1990. Intelligent Decision Support Systems: A Visual Interactive Approach. *Interfaces*, 20(6), pp. 17–28.

Araujo, I. and Araujo, I., 2003. Developing Trust in Internet Commerce. *2003 Conference of the Centre for Advanced Studies on Collaborative Research*, 6 - 9 October 2003, Toronto, Ontario, Canada.

Arnott, D., 2008. Personal Decision Support Systems. *Handbook on Decision Support Systems 2*. Springer: Berlin Heidelberg. pp. 127–150.

Arnott, D. and Pervan, G., 2005. A Critical Analysis of Decision Support Systems Research. *Journal of Information Technology*, 20(2), pp. 67-87.

Arnott, D. and Pervan, G., 2008. Eight Key Issues for the Decision Support Systems Discipline. *Decision Support Systems*, 44(3), pp. 657–672.

Arnott, D. and Pervan, G., 2012. Design Science in Decision Support Systems Research: An Assessment Using the Hevner, March, Park, and Ram Guidelines. *Journal of the Association for Information Systems*, 13(11), pp. 923–949.

Baek, E.-O., Cagiltay, K., Boling, E. and Frick, T., 2008. User-Centered Design and Development. *In:* Spector, M. J. (ed.) *Handbook of Research on Educational Communications and Technology*. Lawrence Erlbaum Associates New York, US. pp. 660-668.

Balakrishnan, V. and Majd, E., 2013. A Comparative Analysis of Trust Models for Multi-Agent Systems. *The First International Conference on Advanced Data and Information Engineering (DaEng-2013)*, 16 - 18 December 2013, Kuala Lumpur, Malaysia.

Barbalet, J., 2009. A Characterization of Trust, and Its Consequences. *Theory and Society*, 38(4), pp. 367-382.

Bastin, L., Thum, S. and Masó, J., 2012. Deliverable D6.1 Data Quality Encoding as a Best Practice Paper [Online]. Available at: http://www.geoviqua.org/Docs/SubmittedDeliverables/D6_1_GeoViQua.pdf.

Beard, K., 1989. Use Error: The Neglected Error Component. *Auto-Carto*, 2 - 7 April 198, Baltimore, USA. pp. 808–817.

Bédard, Y., Devillers, R., Gervais, M. and Jeansoulin, R. 2004. Towards Multidimensional User Manuals for Geospatial Datasets: Legal Issues and their Considerations into the Design of a Technological Solution. *Third International Symposium on Spatial Data Quality (ISSDQ).* Bruck an der Leitha, Austria.

Belanger, F., Hiller, J. S. and Smith, W. J., 2002. Trustworthiness in Electronic Commerce: The Role of Privacy, Security, and Site Attributes. *The Journal of Strategic Information Systems*, 11(3-4), pp. 245-270.

Béquignon, J., Caughey, J., Cramer, W., Fellous, J.-L., Heip, C., Justice, C., Key, J. R., Koike, T., Lacaux, J.-P., Lafaye, M., Lafeuille, J., Mathieu, P.-P., Ranchin, T., Scholes, B. and Schroedter-Homscheidt, M., 2010. GEO and Science [Online]. Available at: http://www.earthobservations.org/documents/committees/stc/20100923_geo_and_science.pdf.

Berger, J., Sorensen, A. T. and Rasmussen, S. J., 2010. Positive Effects of Negative Publicity: When Negative Reviews Increase Sales. *Marketing Science*, 29(5), pp. 815-827.

Bertino, E., Thuraisingham, B., Gertz, M. and Damiani, M. L., 2008. Security and Privacy for Geospatial Data: Concepts and Research Directions. *International Workshop on Security and Privacy in GIS and LBS (SIGSPATIAL ACM GIS 2008)*, 4 November 2008, Irvine, California, USA. pp. 6-19.

Bevan, N., 2003. UsabilityNet Methods for User Centred Design. *HCI International 2003*, 22-27 June, Crete, Greece. pp. 434-438.

Beynon, M., Rasmequan, S. and Russ, S., 2002. A New Paradigm for Computer-based Decision Support. *Decision Support Systems*, 33(2), pp. 127–142.

Bharati, P. and Chaudhury, A., 2004. An Empirical Investigation of Decision-Making Satisfaction in Web-Based Decision Support Systems. *Decision Support Systems*, 37(2), pp. 187–197.

Bhargava, H. K., Power, D. J. and Sun, D., 2007. Progress in Web-based Decision Support Technologies. *Decision Support Systems*, 43(4), pp. 1083–1095.

Bishr, M. and Janowicz, K., 2010. Can we Trust Information? - The Case of Volunteered Geographic Information. *Towards Digital Earth: Search, Discover and Share Geospatial Data, Workshop at Future Internet Symposium*, 20 September 2010, Berlin, Germany.

Bishr, M. and Kuhn, W., 2007. Geospatial Information Bottom-up: A Matter of Trust and Semantics. *The European Information Society*, pp. 365-387.

Bishr, M. and Mantelas, L., 2008. A Trust and Reputation Model for Filtering and Classifying Knowledge about Urban Growth. *GeoJournal*, 72(3-4), pp. 229-237.

Boin, A. T., 2008. Exposing Uncertainty. Thesis. Melbourne, Australia.

Boin, A. T. and Hunter, G. J. 2006. Do Spatial Data Consumers Really Understand Data Quality Information? *In:* Caetano, M. and Painho, M. (eds.) *7th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences.* Lisboa, Portugal.

Boin, A. T. and Hunter, G. J. 2007. What Communicates Quality to the Spatial Data Consumer? *In:* Stein, A., Bijker, W. and Shi, W. (eds.) *5th International Symposium on Spatial Data Quality (ISSDQ 2007).* Enschede, The Netherlands.

Broek, M., Smeets, J., Thum, S. and Masó, J., 2013. Deliverable D3.2 User Feedback Elicitation Tool [Online]. Available at: http://twiki.geoviqua.org/twiki/pub/GeoViQuaIntranet/D3_2/D3.2_User_Feedback_Elicitation_Tool_Final.pdf.

Broutsou, A. and Fitsilis, P., 2012. Online Trust: The Influence of Perceived Company's Reputation on Consumers' Trust and the Effects of Trust on Intention for Online Transactions. *Journal of Service Science and Management*, 5(4), pp. 365-372.

Brown, M., Sharples, S., Harding, J., Parker, C., Bearman, N., Maguire, M., Forrest, D., Haklay, M. and Jackson, M., 2013. Usability of Geographic Information: Current Challenges and Future Directions. *Applied ergonomics*, 44(6), pp. 855–865.

Büyüközkan, G., Arsenyan, J. and Ertek, G., 2010. Evaluation of E-Learning Web Sites Using Fuzzy Axiomatic Design Based Approach. *International Journal of Computational Intelligence Systems*, 3(1), pp. 28-42.

Caprioli, M., Scognamiglio, A., Strisciuglio, G. and Tarantino, E. 2003. Rules and Standards for Spatial Data Quality in GIS Environments. *21st International Cartographic Conference (ICC).* Durban, South Africa.

CEN, 2014. Home Page [Online]. Available at: https://www.cen.eu/Pages/default.aspx [Accessed on: 10 June 2014].

CEN/TC287, 2010. CEN/TC 287 Business Plan [Online]. Available at: http://standards.cen.eu/BP/6268.pdf.

Chakraborty, R., Rao, H. R., Ramireddy, S. and Raghu, T. S., 2010. The Information Assurance Practices of Cloud Computing Vendors. *IT Professional*, 12(4), pp. 29-37.

Chang, H. H. and Chen, S. W., 2008. The Impact of Online Store Environment Cues on Purchase Intention: Trust and Perceived Risk as a Mediator. *Online Information Review*, 32(6), pp. 818-841.

Chang, M. K., Cheung, W. and Tang, M., 2013. Building Trust Online: Interactions Among Trust Building Mechanisms. *Information & Management*, 50(7), pp. 439–445.

Chen, S. C. and Dhillon, G. S., 2003. Interpreting Dimensions of Consumer Trust in E-Commerce. *Information Technology and Management*, 4(2-3), pp. 303-318.

Chen, Y. H., Hsu, I. C. and Lin, C. C., 2010. Website Attributes that Increase Consumer Purchase Intention: A Conjoint Analysis. *Journal of Business Research*, 63(9-10), pp. 1007-1014.

Chopra, K. and Wallace, W. A., 2003. Trust in Electronic Environments. *36th Annual Hawaii International Conference on System Sciences*, 6 - 9 January 2003, Hilton Waikoloa Village, Island of Hawaii. pp. 1-10.

Chrisman, N. R., 1991. The Error Component in Spatial Data. *In:* Maguire, D. A., Goodchild, M. F. and Rhind, D. W. (eds.) *Geographical Information Systems: Overview Principles and Applications*. Longman: White Plains, NY. pp. 165-174.

Chrisman, N. R., 2001. Defining a Geographic Information System. *Exploring Geographic Information Systems*. John Wiley & Sons: New York.

Christian, H., 2010. Crowdsourcing Geospatial Data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 65(6), pp. 550–557.

Cockcroft, S., 1997. A Taxonomy of Spatial Data Integrity Constraints. *GeoInformatica*, 1(4), pp. 327-343.

Collins, F. C. and Smith, J. L., 1994. Taxonomy for Error in GIS. *International Symposium on the Spatial Accuracy of Natural Resource Data Bases*, 16 - 20 May 1994, Williamsburg, USA. pp. 1-7.

Colquitt, J. A., Scott, B. A. and LePine, J. A., 2007. Trust, Trustworthiness, and Trust Propensity: A Meta-Analytic Test of Their Unique Relationships with Risk Taking and Job Performance. *Journal of Applied Psychology*, 92(4), pp. 909-927.

Comber, A. J., Fisher, P. F. and Wadsworth, R. A., 2007a. Approaches for Providing User Relevant Metadata and Data Quality Assessments. *Geographical Information Science Research UK Conference (GISRUK)*, 11 - 13 April 2007, National Centre For Geocomputation, National University Of Ireland, Maynooth, Ireland. pp. 79–82.

Comber, A. J., Fisher, P. F. and Wadsworth, R. A., 2007b. User-Focused Metadata for Spatial Data, Geographical Information and Data Quality Assessments. *10th AGILE International Conference on Geographic Information Science 2007*, 8 - 11 May 2007, Aalborg University, Denmark. pp. 1-13.

Corbitt, B. J., Thanasankit, T. and Yi, H., 2003. Trust and e-Commerce: A Study of Consumer Perceptions. *Electronic Commerce Research and Applications*, 2(3), pp. 203-215.

Costigan, R. D., Ilter, S. S. and Berman, J. J., 1998. A Multi-Dimensional Study of Trust in Organizations. *Journal of Managerial Issues*, 10(3), pp. 303-317.

Couclelis, H., 2003. The Certainty of Uncertainty: GIS and the Limits of Geographic Knowledge. *Transactions in GIS*, 7(2), pp. 165-175.

Crompvoets, J., Bregt, A., Rajabifard, A. and Williamson, I., 2004. Assessing the Worldwide Developments of National Spatial Data Clearinghouses. *International Journal of Geographical Information Science*, 18(7), pp. 665-689.

Cvetkovich, G., 2013. Social Trust and the Management of Risk. Routledge: New Uork, US.

Dahwa, M. P., Al-Hakim, L. and Ng, E., 2013. The Importance of Trust in Procurement Practices and Its Impact on Business Performance: An Empirical Investigation From the Perspective of the Buyer–Supplier Dyad. *Journal of Relationship Marketing*, 12(4), pp. 280-300.

Daignault, M., Shepherd, M., Marche, S. and Watters, C., 2002. Enabling Trust Online. *Third International Symposium on Electronic Commerce*, 19 October 2002 Research Triangle Park, NC, USA. pp. 3-12.

DCMI, 2014a. About Us [Online]. Available at: http://dublincore.org/about-us [Accessed on: 01 May 2014].

DCMI, 2014b. Dublin Core Metadata Element Set, Version 1.1 [Online]. Available at: http://dublincore.org/documents/dces/ [Accessed on: 10 June 2014].

DCMI, 2014c. Dublin Core Metadata Initiative Home Page [Online]. Available at: http://dublincore.org/ [Accessed on: 10 June 2014].

Densham, P. J., 1991. Spatial Decision Support Systems. *In:* Maguire, D. J., Goodschild, M. F. and Rhind, D. W. (eds.) *Geographical Information Systems: Principals and Applications*. Longman Scientific & Technical: Harlow, Essex, UK. pp. 403–412.

Deutsch, M., 1958. Trust and Suspicion. *Journal of Conflict Resolution*, 2(4), pp. 265-279.

Devillers, R., Bédard, Y. and Jeansoulin, R., 2005. Multidimensional Management of Geospatial Data Quality Information for its Dynamic Use Within GIS. *Photogrammetric Engineering Remote Sensing*, 71(2), pp. 205-215.

Devillers, R., Bédard, Y., Jeansoulin, R. and Moulin, B., 2007. Towards Spatial Data Quality Information Analysis Tools for Experts Assessing the Fitness for Use of Spatial Data. *International Journal of Geographical Information Science (IJGIS)*, 21(3), pp. 261-282.

Devillers, R., Gervais, M., Bédard, Y. and Jeansoulin, R., 2002. Spatial Data Quality: From Metadata to Quality Indicators and Contextual End-User Manual. *OEEPE/ISPRS Joint Workshop on Spatial Data Quality Management*, 21 - 22 March 2002, Istanbul, Turkey.

Devillers, R. and Jeansoulin, R., 2006. Spatial Data Quality: Concepts. *In:* Devillers, R. and Jeansoulin, R. (eds.) *Fundamentals of Spatial Data Quality*. ISTE Ltd: London, UK. pp. 31-42.

Devillers, R., Stein, A., Bédard, Y., Chrisman, N. R., Fisher, P. and Shi, W., 2010. Thirty Years of Research on Spatial Data Quality Achievements, Failures, and Opportunities. *Transactions in GIS*, 14(4), pp. 387–400.

Doney, P. M. and Cannon, J. P., 1997. An Examination of the Nature of Trust in Buyer-Seller Relationships. *Journal of Marketing*, 6(2), pp. 35-51.

Duckham, M., 2000. Error-Sensitive GIS Development: Technology and Research Themes. *4th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*, 12–14 July 2000, Amsterdam, The Netherlands. pp. 183-190.

Dunn, J., 2000. Trust and Political Agency. *In:* Gambetta, D. (ed.) *Trust: Making and Breaking Cooperative Relations*. University of Oxford: Oxford, UK. pp. 73-93.

Egger, F. N., 2001. Affective Design of E-Commerce User Interfaces : How to Maximise Perceived Trustworthiness. *Conference on Affective Human Factors Design (CAHD2001)*, 27-29 June 2001, Singapore. pp. 317-324.

EGIDA, 2011. D.3.2 Proposal for a GEO Label [Online]. Available at: http://www.egida-project.eu/images/documents/proposalforageolabel.pdf.

Ellul, C., Foord, J. and Mooney, J., 2013. Making Metadata Usable in a multi-National Research Setting. *Applied Ergonomics*, 44(6), pp. 909-918.

Eom, S. and Kim, E., 2006. A Survey of Decision Support System Applications (1995-2001). *Journal of the Operational Research Society*, 57(11), pp. 1264-1278.

Evangelos, T., 2000. Multi-criteria Decision Making Methods: A Comparative Study. Kluwer Academic Publishers: Dordrecht.

Fan, S. and Shen, Q., 2011. The Effect of Using Group Decision Support Systems in Value Management Studies: An Experimental Study in Hong Kong. *International Journal of Project Management*, 29(1), pp. 13–25.

FGDC, 2006. The Federal Geographic Data Committee: Historical Reflections – Future Directions [Online]. Available at: http://www.fgdc.gov/library/whitepapers-reports/white-papers/fgdc-history [Accessed on: 10 June 2014].

FGDC, 2014a. Home Page [Online]. Available at: http://www.fgdc.gov/ [Accessed on: 10 June 2014].

FGDC, 2014b. National Spatial Data Infrastructure [Online]. Available at: http://www.fgdc.gov/nsdi/nsdi.html [Accessed on: 10 June 2014].

Flanagin, A. J. and Metzger, M. J., 2008. The Credibility of Volunteered Geographic Information. *GeoJournal*, 72(3-4), pp. 137-148.

Fogg, B. J., Marshall, J., Laraki, O., Osipovich, A., Varma, C., Fang, N., Paul, J., Rangnekar, A., Shon, J., Swani, P. and Treinen, M., 2001. What Makes Web Sites Credible? A Report on a Large Quantitative Study. *CHI 2001 Conference on Human Factors in Computing Systems*, 31 March - 5 April 2001, Seattle, Washington, US.

Fogg, B. J., Soohoo, C., Danielson, D., Marable, L., Stanford, J. and Tauber, E. R., 2002. How Do People Evaluate a Web Site's Credibility? Results from a Large Study [Online].

Francis, J. J., Johnston, M., Robertson, C., Glidewell, L., Entwistle, V., Eccles, M. P. and Grimshaw, J. M., 2010. What is an Adequate Sample Size? Operationalising Data Saturation for Theory-Based Interview Studies. *Psychology & Health*, 25(10), pp. 1229-1245.

Fulmer, C. A. and Gelfand, M. J. 2011. How Do I Trust Thee? Dynamic Trust Profiles and Their Individual and Social Contextual Determinants. *24rd Annual International Association of Conflict Management Conference.* Istanbul, Turkey.

Gahegan, M., 2005. The Grid. Bringing Data Producers and Consumers Closer? *NIEeS Workshop on Activating Metadata*, 6 - 7 July 2005, Cambridge, UK.

Gambetta, D., 2000. Can we Trust Trust? *In:* Gambetta, D. (ed.) *Trust: Making and Breaking Cooperative Relations*. University of Oxford,: Oxford, UK. pp. 213-237.

Ganesan, S. and Hess, R., 1997. Dimensions and Levels of Trust: Implications for Commitment to a Relationship. *Marketing Letters*, 8(4), pp. 439-448.

Gefen, D., 2000. E-Commerce: The Role of Familiarity and Trust. *Omega*, 28, pp. 725-737.

Gefen, D., 2002. Reflections on the Dimensions of Trust and Trustworthiness Among Online Consumers. *ACM SIGMIS Database*, 33(3), pp. 38-53.

Gefen, D., Karahanna, E. and Straub, D. W., 2003. Trust and TAM in Online Shopping: An Integrated Model. *MIS Quarterly*, 27(1), pp. 51-90.

GEO, 2011a. GEO Task ST-09-02: Promoting Awareness and Benefits of GEO [Online]. Available at: http://www.geo-tasks.org/st0902/ [Accessed on: 10 June 2014].

GEO, 2011b. ST-09-02: Promoting Awareness and Benefits of GEO in the Science and Technology Community [Online]. Available at: http://www.earthobservations.org/ts.php?id=91 [Accessed on: 10 June 2014].

GEO, 2014a. EuroGEOSS Broker [Online]. Available at: http://www.eurogeoss-broker.eu/ [Accessed on: 01 September 2014].

GEO, 2014b. GEO Task ID-03: Science and Technology in GEOSS [Online]. Available at: http://www.geo-tasks.org/id03 [Accessed on: 10 June 2014].

GEO, 2014c. What is GEOSS?: The Global Earth Observation System of Systems [Online]. Available at: https://www.earthobservations.org/geoss.shtml [Accessed on: 10 June 2014].

Gervais, M., 2006. On the Importance of External Data Quality in Civil Law. *In:* Devillers, R. and Jeansoulin, R. (eds.) *Fundamentals of Spatial Data Quality*. ISTE Ltd: London, UK. pp. 283-300.

Geyskens, I., Steenkamp, J.-b. E. M., Scheer, L. K. and Kumar, N., 1996. The Effects of Trust and Interdependence on Relationship Commitment: A Trans-Atlantic Study. *International Journal of Research in Marketing*, 13(4), pp. 303-317.

Goodchild, M. F., 1995. Sharing Imperfect Data. *In:* Onsrud, H. J. and Rushton, G. (eds.) *Sharing Geographic Information*. Rutgers University Press: New Brunswick, Canada. pp. 413–425.

Goodchild, M. F., 2006. Foreword. *In:* Devillers, R. and Jeansoulin, R. (eds.) *Fundamentals of Spatial Data Quality*. ISTE Ltd: London, UK. pp. 13-16.

Goodchild, M. F., 2009. Putting Research into Practice. *In:* Stein, A., Shi, W. and Bijker, W. (eds.) *Quality Aspects of Spatial Data Mining*. CRC Press: Boca Raton. pp. 345–356.

Goodchild, M. F., 2012. The Future of Digital Earth. *Annals of GIS*, 18(2), pp. 93-98.

Goodchild, M. F., Yuan, May and Cova, T. J., 2007. Towards a General Theory of Geographic Representation in GIS. *International Journal of Geographical Information Science*, 21(3), pp. 239–260.

Google Drive, 2014. Home Page [Online]. Available at: https://drive.google.com/ [Accessed on: 10 June 2014].

Grandison, T. and Sloman, M., 2000. A Survey of Trust in Internet Applications. *Communications Surveys & Tutorials, IEEE*, 3(4), pp. 2-16.

Guerlain, S., Brown, D. E. and Mastrangelo , C., 2000. Intelligent Decision Support Systems. *2000 IEEE International Conference on Systems, Man, and Cybernetics*, 8–11 October, Nashville, Tennessee, US. pp. 1934–1938.

Guest, G., Bunce, A. and Johnson, L., 2006. How Many Interviews Are Enough?: An Experiment with Data Saturation and Variability. *Field Methods*, 18(1), pp. 59-82.

Guo, S., Wang, M. and Leskovec, J., 2011. The Role of Social Networks in Online Shopping: Information Passing, Price of Trust, and Consumer Choice. *ACM Conference on Electronic Commerce 2011*, 5 - 9 June 2011, San Jose, CA, USA. pp. 157-166

Ha, H. Y., 2004. Factors Influencing Consumer Perceptions of Brand Trust Online. *Journal of Product & Brand Management*, 13(5), pp. 329-342.

Harding, J., 2006. Vector Data Quality: A Data Provider's Perspective. *In:* Devillers, R. and Jeansoulin, R. (eds.) *Fundamentals of Spatial Data Quality*. ISTE Ltd: London, UK. pp. 141-159.

Hartono, E., Holsapple, C. W., Kim, K.-Y., Na, K.-S. and Simpson, J. T., 2014. Measuring Perceived Security in B2C Electronic Commerce Website Usage: A Respecification and Validation. *Decision Support Systems*.

Harvey, F., 2003. Developing Geographic Information Infrastructures for Local Government: The Role of Trust. *The Canadian Geographer*, 47(1), pp. 28–36.

Hasan, L. and Abuelrub, E., 2011. Assessing the Quality of Web Sites. *Applied Computing and Informatics*, 9(1), pp. 11-29.

Hassanein, K. S. and Head, M. M., 2004. Building Online Trust through Socially Rich Web Interfaces. *2nd Annual Conference on Privacy, Security and Trust, Fredericton*, 13 - 15 October 2004, New Brunswick, Canada. pp. 15-22.

Head, M. M. and Hassanein, K., 2002. Trust in e-Commerce: Evaluating the Impact of Third-Party Seals. *Quarterly Journal of Electronic Commerce*, 3(3), pp. 307-326.

Heikkila, E. J., 2007. GIS is Dead; Long Live GIS! *Journal of the American Planning Association*, 64(3), pp. 350-360.

Higgins, E. T., 2000. Making a Good Decision Value from Fit. *American Psychologist*, 55(11), pp. 1217–1230.

Hoffman, D. L., Novak, T. P. and Peralta, M., 1999. Building Consumer Trust Online. *Communications of the ACM*, 42(4), pp. 80-85.

Hong-ling, M. and Guang-xing, S., 2011. An Overview of Trust Mechanism and Applications of E-Commerce *International Conference on Management and Service Science (MASS)*, 12 - 14 August 2011 Wuhan, China.

Hong, I. B. and Cha, H. S., 2013. The Mediating Role of Consumer Trust in an Online Merchant in Predicting Purchase Intention. *International Journal of Information Management*, 33(6), pp. 927–939.

Hosmer, L. T., 1995. Trust: The Connecting Link Between Organizational Theory and Philosophical Ethics. *Academy of Management Review*, 20(2), pp. 379-403.

Hu, X., Lin, Z. and Zhang, H., 2001. Myth or Reality : Effect of Trust-Promoting Seals in Electronic Markets. *Eleventh Annual Workshop on Information Technologies and Systems*, 15 - 16 December 2001, New Orleans, Louisiana, USA. pp. 143-150.

Hu, X., Wu, G., Wu, Y. and Zhang, H., 2010. The Effects of Web Assurance Seals on Consumers' Initial Trust in an Online Vendor: A Functional Perspective. *Decision Support Systems*, 48(2), pp. 407-418.

INSPIRE, 2008. INSPIRE Metadata Implementing Rules: Technical Guidelines based on EN ISO 19115 and EN ISO 19119 [Online]. Available at: http://inspire.ec.europa.eu/index.cfm/pageid/101/list/2 [Accessed on: 10 June 2014].

INSPIRE, 2014a. About INSPIRE [Online]. Available at: http://inspire.ec.europa.eu/index.cfm/pageid/48 [Accessed on: 10 June 2014].

INSPIRE, 2014b. Data and Service Sharing [Online]. Available at: http://inspire.ec.europa.eu/index.cfm/pageid/62 [Accessed on: 10 June 2014].

INSPIRE, 2014c. Data Specifications [Online]. Available at: http://inspire.ec.europa.eu/index.cfm/pageid/2/list/2 [Accessed on: 10 June 2014].

INSPIRE, 2014d. Home Page [Online]. Available at: http://inspire.jrc.ec.europa.eu/index.cfm [Accessed on: 01 May 2014].

INSPIRE, 2014e. Monitoring and Reporting [Online]. Available at: http://inspire.ec.europa.eu/index.cfm/pageid/182 [Accessed on: 10 June 2014].

INSPIRE, 2014f. Network Services [Online]. Available at: http://inspire.ec.europa.eu/index.cfm/pageid/5 [Accessed on: 10 June 2014].

ISO, 2005. ISO 19119:2005 Geographic information - Services [Online]. Available at: http://www.iso.org/iso/catalogue_detail.htm?csnumber=39890 [Accessed on: 10 June 2014].

ISO, 2014a. About ISO [Online]. Available at: http://www.iso.org/iso/home/about.htm [Accessed on: 10 June 2014].

ISO, 2014b. ISO 19115-1:2014 Geographic information – Metadata – Part 1: Fundamentals [Online]. Available at: http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?csnumber=53798.

ISO, 2014c. Standards [Online]. Available at: http://www.iso.org/iso/home/standards.htm [Accessed on: 10 June 2014].

ISO and SEN, 2001. Agreement on Technical Co-Operation between ISO and CEN (Vienna Agreement) [Online]. Available at: http://boss.cen.eu/ref/Vienna_Agreement.pdf.

ISO/TC211, 2003. ISO 19115:2003 Geographic information – Metadata [Online]. Available at: http://www.iso.org/iso/catalogue_detail.htm?csnumber=26020 [Accessed on: 10 June 2014].

ISO/TC211, 2009. ISO 19115-2:2009 Geographic information – Metadata – Part 2: Extensions for imagery and gridded data [Online]. Available at: http://www.iso.org/iso/catalogue_detail.htm?csnumber=39229 [Accessed on: 10 June 2014].

ISO/TC211, 2013. ISO 19157:2013 Geographic information – Data quality [Online]. Available at: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=32575 [Accessed on: 10 June 2014].

ISO/TC211, 2014a. Home Page [Online]. Available at: http://www.isotc211.org [Accessed on: 01 May 2014].

ISO/TC211, 2014b. ISO/TC 211 Geographic information/Geomatics - Scope [Online]. Available at: http://www.isotc211.org/ [Accessed on: 10 June 2014].

Ivánová, I., Morales, J., de By, R. A., Beshe, T. S. and Gebresilassie, M. A., 2013. Searching for Spatial Data Resources by Fitness for Use. *Journal of Spatial Science*, 58(1), pp. 15-28.

Jankowski, P., 1995. Integrating Geographical Information Systems and Multiple Criteria Decision-making Methods. *International Journal of Geographical Information Systems (IJGIS)*, 9(3), pp. 251–273.

Jarvenpaa, S. L., Noam, T. and Vitale, M., 2000. Consumer Trust in an Internet Store. *Information Technology and Management*, 1(2), pp. 45-71.

Jarvenpaa, S. L., Tractinsky, N. and Saarinen, L., 1999. Consumer Trust in an Internet Store: A Cross-Cultural Validation. *Journal of Computer-Mediated Communication*, 5(2), pp. 0-0.

Jelenc, D., Hermoso, R., Sabater-Mir, J. and Trček, D., 2013. Decision Making Matters: A Better Way to Evaluate Trust Models. *Knowledge-Based Systems*, 52, pp. 147-164.

Jiang, P., Jones, D. B. and Javie, S., 2008. How Third-Party Certification Programs Relate to Consumer Trust in Online Transactions: An Exploratory Study. *Psychology and Marketing*, 25(9), pp. 839-858.

Jones, K., 1996. Trust as an Affective Attitude. *Ethics*, 107(1), pp. 4-25.

Kaihong, X. and Mingxia, W., 2007. Economic Function of Trust Seal in E-Commerce: An Experiment Study Based on Chinese Subjects. *International Conference on Service Systems and Service Management*, 9-11 June 2007, Chengdu, China. IEEE, pp. 1-5.

Keßler, C. and de Groot , R. T. A., 2013. Trust as a Proxy Measure for the Quality of Volunteered Geographic Information in the Case of OpenStreetMap. *Geographic Information Science at the Heart of Europe*, pp. 21-37.

Kim, C., Tao, W., Shin, N. and Kim, K.-S., 2010. An Empirical Study of Customers' Perceptions of Security and Trust in e-Payment Systems. *Electronic Commerce Research and Applications*, 9(1), pp. 84–95.

Kim, D., Ferrin, D. and Rao, R., 2008a. A Trust-Based Consumer Decision-Making Model in Electronic Commerce: The Role of Trust, Perceived Risk, and Their Antecedents. *Decision Support Systems*, 44(2), pp. 544-564.

Kim, D., Steinfield, C. and Lai, Y., 2008b. Revisiting the Role of Web Assurance Seals in Business-to-Consumer Electronic Commerce. *Decision Support Systems*, 44(4), pp. 1000-1015.

Kim, D. J., Song, Y. I., Braynov, S. B. and Rao, H. R., 2005. A Multidimensional Trust Formation Model in B-to-C e-Commerce: A Conceptual Framework and Content

Analyses of Academia/Practitioner Perspectives. *Decision Support Systems*, 40(2), pp. 143-165.

Kim, J. B., 2012. An Empirical Study on Consumer First Purchase Intention in Online Shopping: Integrating Initial Trust and TAM. *Electronic Commerce Research*, 12, pp. 125-150.

Kim, K. and Kim, J., 2011. Third-party Privacy Certification as an Online Advertising Strategy: An Investigation of the Factors Affecting the Relationship between Third-party Certification and Initial Trust. *Journal of Interactive Marketing*, 25(3), pp. 145–158.

Kim, K. and Prabhakar, B. 2000. Initial Trust, Perceived Risk, and the Adoption of Internet Banking. *Twenty-First International Conference on International Conference on Information Systems (ICIS 2000).* Brisbane, Australia.

Kini, A. and Choobineh, J., 1998. Trust in Electronic Commerce: Definition and Theoretical Considerations *Thirty-First Hawaii International Conference on System Sciences*, 6 - 9 January 1998 Kohala Coast, Hawaii. pp. 51-61 vol.4.

Koufaris, M. and Hampton-Sosa, W., 2004. The Development of Initial Trust in an Online Company by New Customers. *Information & Management*, 41(3), pp. 377-397.

Lee, A. J. and Yu, T., 2009. Towards a Dynamic and Composite Model of Trust. *14th ACM Symposium on Access Control Models and Technologies* 03 - 05 June 2009 Stresa, Italy.

Li, D., Zhang, J. and Wu, H., 2012. Spatial Data Quality and Beyond. *International Journal of Geographical Information Science*, 26(12), pp. 2277-2290.

Litwin, L. and Rossa, M., 2011. Standards and Interoperability. *Geoinformation Metadata in INSPIRE and SDI: Understanding. Editing. Publishing*. Springer Verlag Berlin. pp. 39-74.

Liu, Y., Li, H. and Hu, F., 2013. Website Attributes in Urging Online Impulse Purchase: An Empirical investigation on Consumer Perceptions. *Decision Support Systems*, 55(3), pp. 829-837.

Livingston, J. A., 2005. How Valuable is a Good Reputation? A Sample Selection Model of Internet Auctions. *Review of Economics and Statistics*, 87(3), pp. 453-465.

Lohse, G. L. and Spiller, P., 1998. Electronic Shopping. *Communications of the ACM*, 41(7), pp. 81-87.

Longhorn, R. A., 2005. Geospatial Standards, Interoperability, Metadata Semantics and Spatial Data Infrastructure. *NIEeS Workshop on Activating Metadata*, 6 – 7 July 2005, Cambridge, UK.

Lowry, P. B., Vance, A., Beckman, B., Moody, G. and Read, A., 2008. Explaining and Predicting the Impact of Branding Alliances and Web Site Quality on Initial Consumer Trust of E-Commerce Web Sites. *Journal of Management Information Systems*, 24(4), pp. 199-224.

Luhmann, N., 1979. Trust and Power. Wiley: Chichester, UK.

Lumsden, J., 2008. A Method for Systematic Artifact Selection Decision Making. *In:* Frederic, A. and Humphreys, P. (eds.) *Encyclopedia of Decision Making and Decision Support Technologies*. IGI global: Hershey, US.

Lumsden, J., 2009. Triggering Trust: To What Extent Does the Question Influence the Answer When Evaluating the Perceived Importance of Trust Triggers? *23rd British HCI Group Annual Conference on People and Computers: Celebrating People and Technology*, 1 - 5 September 2009, Cambridge, UK. British Computer Society, pp. 214-223.

Lumsden, J. and MacKay, L., 2006. How Does Personality Affect Trust in B2C e-Commerce? *8th International Conference on Electronic Commerce: The New e-Commerce: Innovations for Conquering Current Barriers, Obstacles and Limitations to Conducting Successful Business on the Internet*, 13 - 16 August 2006, Fredericton, New Brunswick, Canada.

Lyons, B. G. and Stuth, J. W., 1992. Decision Support Systems for the Management of Grazing Lands: Emerging Issues. Parthenon Publishing Group Ltd: Carnforth.

Maguire, D. J. and Longley, P. A., 2005. The Emergence of Geoportals and Their Role in Spatial Data Infrastructures. *Computers, Environment and Urban Systems*, 29(1), pp. 3-14.

Malhotra, N. K., 1984. Reflections on the Information Overload Paradigm in Consumer Decision Making. *Journal of Consumer Research*, 10(4), pp. 436–440.

Manchala, D. W., 2000. E-Commerce Trust Metrics and Models. *IEEE Internet Computing*, 4(2), pp. 36–44.

Marsh, S., 1994a. Optimism and Pessimism in Trust. *Ibero-American Conference on Artificial Intelligence (IBERAMIA'94)*, 25 - 28th October 1994, Caracas, Venezuela. McGraw-Hill Publishing, pp. 1-12.

Marsh, S. and Meech, J., 2000. Trust in Design. *CHI Conference on Human Factors in Computing Systems (CHI'00)*, 1 - 6 April 2000, The Hague, The Netherlands. ACM Press, pp. 45.

Marsh, S. P., 1994b. Formalising Trust as a Computational Concept. Thesis. Stirling, UK.

Martinez, L., Ruan, D. and Herrera, F., 2010. Computing with Words in Decision support Systems: An overview on Models and Applications. *International Journal of Computational Intelligence Systems*, 3(4), pp. 382–395.

Mascha, M. F., Miller, C. L. and Janvrin, D. J., 2011. The Effect of Encryption on Internet Purchase Intent in Multiple Vendor and Product Risk Settings. *Electronic Commerce Research*, 11(4), pp. 401-419.

Mayer, R. C., Davis, J. H. and Schoorman, F. D., 1995. An Integrative Model of Organizational Trust. *Academy of Management Review*, 20(2), pp. 709-734.

Mccord, M. and Ratnasingam, P., 2004. The Impact of Trust on the Technology Acceptance Model in Business to Consumer E-Commerce. *International Conference of the Information Resources Management Association: Innovations Through Information Technology*, 23 – 26 May 2004, New Orleans, USA. pp. 921-925.

Mcknight, D. H., 2000. Trust in e-Commerce Vendors: A Two-Stage Model. *Twenty First International Conference on Information Systems (ICIS 2000)*, 10-13 December 2000, Brisbane, Australia.

Mcknight, D. H. and Chervany, N. L. 1996. The Meanings of Trust. Minneapolis, US: University of Minnesota.

McKnight, D. H. and Chervany, N. L., 2001. Trust and Distrust Definitions: One Bite at a Time. *Lecture Notes in Computer Science*, 2246, pp. 27-54.

McKnight, D. H. and Chervany, N. L. 2006. Reflections on an Initial Trust-Building Model. *In:* Bachmann, R. and Zaheer, A. (eds.) *Handbook of Trust Research.* Cheltenham, UK: Edward Elgar Publishing.

McKnight, D. H., Choudhury, V. and Kacmar, C., 2002a. Developing and Validating Trust Measures for e-Commerce: An Integrative Typology. *Information Systems Research*, 13(3), pp. 334-359.

McKnight, D. H., Cummings, L. L. and Chervany, N. L., 1998. Initial Trust Formation in New Organizational Relationships. *The Academy of Management Review*, 23(3), pp. 473-490.

Mcknight, D. H., Kacmar, C. J. and Choudhury, V., 2004. Shifting Factors and the Ineffectiveness of Third Party Assurance Seals: A Two-Stage Model of Initial Trust in a Web Business. *Electronic Markets*, 14(3), pp. 252-266.

McKnight, H. D., Choudhury, V. and Kacmar, C., 2002b. The Impact of Initial Consumer Trust on Intentions to Transact with a Web Site: A Trust Building Model. *Journal of Strategic Information Systems*, 11(3), pp. 297-323.

Meziane, F. and Nefti, S., 2007. Evaluating e-Commerce Trust Using Fuzzy Logic. *International Journal of Intelligent Information Technologies*, 3(4), pp. 25–39.

Mohammadi, H., Rajabifard, A. and Williamson, I. P., 2009. Enabling Spatial Data Sharing through Multi-source Spatial Data Integration. *GDSI 11*, 12 August 2011, Rotterdam, The Netherlands. pp. 1-19.

Moores, T., 2005. Do Consumers Understand the Role of Privacy Seals in e-Commerce? *Communications of the ACM*, 48(3), pp. 86-91.

Moores, T. T. and Dhillon, G., 2003. Do Privacy Seals in e-Commerce Really Work? *Communications of the ACM*, 46(12), pp. 265-271.

Moorman, C., Deshpande, R. and Zaltman, G., 1993. Factors Affecting Trust in Market Research Relationships. *Journal of Marketing*, 57(1), pp. 81-101.

Moyano, F., Fernandez-Gago, C. and Lopez, J., 2012. A Conceptual Framework for Trust Models. *Trust, Privacy and Security in Digital Business*, 7449, pp. 93-104.

Mudambi, S. M. and Schuff, D., 2010. What Makes a Helpful Online Review? A Study of Customer Reviews on Amazon.com. *MIS Quarterly*, 34(1), pp. 185-200.

Mudambi, S. M., Schuff, D. and Zhang, Z., 2014. Why Aren't the Stars Aligned? An Analysis of Online Review Content and Star Ratings. *47th Hawaii International Conference on System Sciences (HICSS)*, 6 - 9 January 2014 Waikoloa, Hawaii. pp. 3139-3147.

Nemati, H. R., Steiger, D. M., Iyer, L. S. and Herschel, R. T., 2002. Knowledge Warehouse: An Architectural Integration of Knowledge Management, Decision Support, Artificial Intelligence and Data Warehousing. *Decision Support Systems*, 33(2), pp. 143–161.

Noteberg, A., Christiaanse, E. and Wallage, P., 2003. Consumer Trust in Electronic Channels: The Impact of Electronic Commerce Assurance on Consumers' Purchasing Likelihood and Risk Perceptions. *E-Service Journal*, 2(2), pp. 46-67.

OGC, 2014a. About OGC [Online]. Available at: http://www.opengeospatial.org/ogc [Accessed on: 10 June 2014].

OGC, 2014b. Home Page [Online]. Available at: http://www.opengeospatial.org/ [Accessed on: 01 May 2014].

Parasuraman, A., Zeithaml, V. A. and Malhotra, A., 2005. E-S-QUAL A Multiple-Item Scale for Assessing Electronic Service Quality. *Journal of Service Research*, 7(3), pp. 213-233.

Pavlou, P. A., 2003. Consumer Acceptance of Electronic Commerce: Integrating Trust and Risk with the Technology Acceptance Model. *International Journal of Electronic Commerce*, 7(3), pp. 101-134.

Payne, J. W., 1976. Task Complexity and Contingent Processing in Decision Making: An Information Search and Protocol Analysis. *Organizational Behavior and Human Performance*, 16(2), pp. 366–387.

Pennanen, K., Paakki, M.-K. and Kaapu, T., 2008. Consumers Views on Trust, Risk, Privacy and Security in e-Commerce A Qualitative Analysis. *In:* Kautonen, T. and Karjaluoto, H. (eds.) *Trust and New Technologies Marketing and Management on the Internet and Mobile Media*. Edward Elgar Publishing: Cheltenham, UK. pp. 108-123.

Plag, H.-P. 2012. Labeling Geo - Referenced Information in Support of Data Sharing and the Facilitating of Societal Benefits of Earth Observations. *16th International Conference on Heavy Metals in the Environment.* Rome, Italy: E3S Web of Conferences.

Poortinga, W. and Pidgeon, N. F., 2003. Exploring the Dimensionality of Trust in Risk Regulation. *Risk Analysis*, 23(5), pp. 961-972.

Pundt, H., 2002. Field Data Collection with Mobile GIS : Dependencies Between Semantics and Data Quality. *GeoInformatica*, 6(4), pp. 363-380.

QA4EO, 2014. Home Page [Online]. Available at: http://qa4eo.org/index.html [Accessed on: 01 May 2014].

QuestionPro, 2014. Home Page [Online]. Available at: http://www.questionpro.com/ [Accessed on: 10 June 2014].

Ratnasingham, P., 1998. The Importance of Trust in Electronic Commerce. *Internet Research*, 8(4), pp. 313-321.

Resnick, P., Zeckhauser, R., Swanson, J. and Lockwood, K., 2006. The Value of Reputation on eBay: A Controlled Experiment. *Experimental Economics*, 9(2), pp. 79-101.

Riegelsberger, J. and Sasse, M. A., 2001. Trustbuilders and Trustbusters: The Role of Trust Cues in Interfaces to e-Commerce Applications. *1st IFIP Conference on e-Commerce, e-Business, e-Government (i3e'2001)*, pp. 17-30.

Rifon, N. J., LaRose, R. and Choi, S. M., 2005. Your Privacy Is Sealed: Effects of Web Privacy Seals on Trust and Personal Disclosures. *The Journal of Consumer Affairs*, 39(2), pp. 339-362.

Rinner, C., 2003. Web-based Spatial Decision Support: Status and Research Directions. *Journal of Geographic Information and Decision Analysis*, 7(1), pp. 14–31.

Rotter, J. B., 1967. A New Scale for the Measurement of Interpersonal Trust. *Journal of Personality*, 35(4), pp. 651-665.

Rotter, J. B., 1980. Interpersonal Trust, Trustworthiness, and Gullibility. *American Psychologist*, 35(1), pp. 1-7.

Rousseau, D. M., Sitkin, S. B., Burt, R. S. and Camerer, C., 1998. Not So Different After All: A Cross-Discipline View Of Trust. *Academy of Management Review*, 23(3), pp. 393-404.

Sadana, R. and Stasko, J., 2014. Designing and Implementing an Interactive Scatterplot Visualization for a Tablet Computer. *2014 International Working Conference on Advanced Visual Interfaces*, 27–29 May, Como, Italy. pp. 265–272.

Schoorman, F. D., Mayer, R. C. and Davis, J. H., 2007. An Integrative Model of Organizational Trust: Past, Present, and Future. *The Academy of Management Review*, 32(2), pp. 344-354.

Sensio Labs, 2011. Home Page [Online]. Available at: http://silex.sensiolabs.org/.

Servigne, S., Lesage, N. and Libourel, T., 2006. Quality Components, Standards, and Metadata. *In:* Devillers, R. and Jeansoulin, R. (eds.) *Fundamentals of Spatial Data Quality*. ISTE Ltd: London, UK. pp. 179-210.

Sharma, D., 2009. Decision Making Style: An Introduction. *Decision Making Style: Social and Creative Dimensions*. Global India Publications: Delhi, India.

Shim, J. P., Warkentin, M., Courtney, J. F., Power, D. J., Sharda, R. and Carlsson, C., 2002. Past, Present, and Future of Decision Support Technology. *Decision Support Systems*, 33(2), pp. 111–126.

Shirgaonkar, S., Rathi, S. and Rajkumar, T., 2010. Overview of Real Time Decision Support System. *International Conference and Workshop on Emerging Trends in Technology (ICWET 2010)*, 26–27 February, Mumbai, India. pp. 179–181.

Shneiderman, B., 1996. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. *IEEE Symposium on Visual Languages*, 3–6 September, Boulder, Colorado, USA. pp. 336–343.

Silva, N. N., Sánchez, J. A., Proal, C. and Rebollar, C., 2003. Visual Exploration of Large Collections in Digital Libraries. *Latin American Conference on Human-Computer Interaction*, 17–20 November, Rio de Janeiro, Brazil. pp. 147–157.

Sivaji, A., Downe, A. G., Mazlan, M. F., Soo, S.-T. and Abdullah, A., 2011. Importance of Incorporating Fundamental Usability with Social & Trust Elements for E-Commerce Website. *2011 International Conference on Business, Engineering and Industrial Applications (ICBEIA)*, 5 - 7 June 2011 Kuala Lumpur, Malaysia. pp. 221-226

Skarlatidou, A., Cheng, T. and Haklay, M., 2013. Guidelines for Trust Interface Design for Public Engagement Web GIS. *International Journal of Geographical Information Science*, 27(8), pp. 1668-1687.

Skarlatidou, A., Haklay, M. and Cheng, T., 2010a. Preliminiary Investigation of Web GIS Trust: The Example of the “WIYBY” Website. *Joint International Conference on*

Theory, Data Handling and Modelling in GeoSpatial Information Science, 26 - 28 May 2010, Hong Kong.

Skarlatidou, A., Haklay, M. and Cheng, T., 2011a. Trust in Web GIS : The Role of the Trustee Attributes in the Design of Trustworthy Web GIS Applications. *International Journal of Geographical Information Science*, 25(12), pp. 1913-1930.

Skarlatidou, A., Haklay, M., Cheng, T. and Francis, N., 2010b. Trust in Web GIS: A Preliminary Investigation of the Environment Agency's WIYBY Website with non-Expert Users. *In:* Haklay, M., Morley, J. and Rahemtulla, H., eds. *GIS Research UK 18th Annual Conference*, 14-16 April 2010, London, UK. pp. 439-446.

Skarlatidou, A., Wardlaw, J., Haklay, M. and Cheng , T., 2011b. Understanding the Influence of Specific Web GIS Attributes in the Formation of Non-Experts' Trust Perceptions. *Advances in Cartography and GIScience*, 1, pp. 219-238

Slovic, P. and Lichtenstein, S., 1971. Comparison of Bayesian and Regression Approaches to the Study of Information Processing in Judgment. *Organizational Behavior and Human Performance*, 6, pp. 649-744.

Smith, D., Menon, S. and Sivakumar, K., 2005. Online Peer and Editorial Recommendations, Trust, and Choice in Virtual Markets. *Journal of Interactive Marketing*, 19(3), pp. 15-37.

Spillinger, A. and Parush, A., 2012. The Impact of Testimonials on Purchase Intentions in a Mock E-commerce Web Site. *Journal of Theoretical and Applied Electronic Commerce Research*, 7(1), pp. 51-63.

Sprague, R. H., 1980. A Framework for the Development of Decision Support Systems. *MIS Quarterly*, 4(4), pp. 1–26.

ST-09-02, 2010. A GEO Label: Informing Users About the Quality, Relevance and Acceptance of Services, Data Sets and Products Provided by GEOSS [Online].

Stein, A. and van Oort, P., 2006. The Impact of Positional Accuracy on the Computation of Cost Functions. *In:* Devillers, R. and Jeansoulin, R. (eds.) *Fundamentals of Spatial Data Quality*. ISTE Ltd: London, UK. pp. 107-122.

Steiniger, S. and Hunter, A. J. S., 2012. Free and Open Source GIS Software for Building a Spatial Data Infrastructure. *Geospatial Free and Open Source Software in the 21st Century*, pp. 247-261

Sui, D., 2014. Opportunities and Impediments for Open GIS. *Transactions in GIS*, 18(1), pp. 1-24.

Tan, F. B. and Sutherland, P., 2004. Online Consumer Trust: A Multi-Dimensional Model. *Journal of Electronic Commerce in Organizations*, 2(3), pp. 40-58.

Teisman, G. R., 2000. Models for Research into Decision-Making Processes: On Phases, Streams and Decision-Making Rounds. *Public Administration*, 78(4), pp. 937–956.

Teo, T. S. H. and Liu, J., 2007. Consumer Trust in e-Commerce in the United States, Singapore and China. *Omega*, 35(1), pp. 22–38.

The British Assessment Bureau, 2014. ISO 9000 - Quality management [Online]. Available at: http://www.iso.org/iso/home/standards/management-standards/iso_9000.htm [Accessed on: 10 June 2014].

Triglav, J., Petrovič, D. and Stopar, B., 2011. Spatio-Temporal Evaluation Matrices for Geospatial Data. *International Journal of Applied Earth Observation and Geoinformation*, 13(1), pp. 100-109.

Tucci, M. and Giordano, A., 2011. Positional Accuracy, Positional Uncertainty, and Feature Change Detection in Historical Maps: Results of an Experiment. *Computers Environment and Urban Systems*, 35(6), pp. 452-463.

Tumba, A. G. and Ahmad, A., 2014. Geographic Information System and Spatial Data Infrastructure: A Developing Societies' Perception. *Universal Journal of Geoscience*, 2(3), pp. 85-92.

Umuhoza, D., Agbinya, J. I., Moodley, D. and Vahed, A., 2008. A Reputation based Trust Model for Geospatial Web Services. *1st WSEAS International Conference on Environmental and Geological Science and Engineering (EG'08)*, 11 - 13 September 2008, Malta. pp. 220-225.

Uran, O. and Janssen, R., 2003. Why are Spatial Decision Support Systems not Used? Some Experiences from the Netherlands, Computers. *Environment and Urban Systems*, 27(5), pp. 511–526.

USGCRP, 1999. Global Change Science Requirements for Long-Term Archiving. *Workshop on Global Change Science Requirements For Long-Term Archiving*, 28 - 30 October 1999, Washington, DC, US. USGCRP Program Office.

Uslaner, E. M., 1999. Trust and Consequences.

Uslaner, E. M., 2002. The Moral Foundations of Trust. Cambridge University Press: Cambridge, UK.

Utz, S., Kerkhof, P. and Van den Bos, J., 2012. Consumers Rule: How Consumer Reviews Influence Perceived Trustworthiness of Online Stores. *Electronic Commerce Research and Applications*, 11(1), pp. 49–58.

van Dijk, G., Minocha, S. and Laing, A., 2007. Consumers, Channels and Communication: Online and Offline Communication in Service Consumption. *Interacting with Computers*, 19(1), pp. 7–19.

Van Schaik, F. D. J., 1988. Effectiveness of Decision Support Systems. Koninklijke Bibliotheek: The Hague.

Veregin, H., 1999. Data Quality Parameters. *In:* Longley, P. A. G., M F; Maguire, D J; Rhind, D W (ed.) *Geographical Information Systems*. Wiley: New York, US. pp. 177 - 189.

Verhagen, T., Meents, S. and Tan, Y.-H., 2006. Perceived Risk and Trust Associated with Purchasing at Electronic Marketplaces. *European Journal of Information Systems*, 15(6), pp. 542-555.

Vroom, V. H. and Yetton, P. W., 1973. Leadership and Decision-Making. University of Pittsburgh Press: Pittsburgh, Pennsylvania, US.

Waloszek, G., 2013. FilmFinder [Online]. Available at: http://www.sapdesignguild.org/goodies/controls/FilmFinder.htm [Accessed on: 01 September 2014].

Wang, F. and Huang, Q. Y., 2007. A Methodology for Definition and Usage of Spatial Data Quality Rules. *SPIE 6753 Geoinformatics 2007: Geospatial Information Science*, 26 July 2007, Nanjing, China. SPIE, pp. D7531-D7531.

Wang, Y. D. and Emurian, H. H., 2005. An Overview of Online Trust: Concepts, Elements, and Implications. *Computers in Human Behavior*, 21(1), pp. 105–125.

Worboys, M., 1998. Imprecision in Finite Resolution Spatial Data. *GeoInformatica*, 2(3), pp. 257-280.

Yamamoto, Y., 1990. A Morality Based on Trust: Some Reflections on Japanese Morality. *Philosophy East and West*, 40(4), pp. 451-469.

Yang, Y., Hu, Y. and Chen, J., 2005. A Web Trust-Inducing Model for e-Commerce and Empirical Research. *7th International Conference on Electronic Commerce (ICEC 05)*, 15 - 17 August 2005, Xi'an, China. ACM Press.

Ye, Q., Law, R., Gu, B. and Chen, W., 2011. The Influence of User-Generated Content on Traveler Behavior: An Empirical Investigation on the Effects of e-Word-of-Mouth to Hotel Online Bookings. *Computers in Human Behavior*, 27(2), pp. 634–639.

Zeleny, M., 1982. The Decision Process and Its Stages. *In:* Cochrane, J. L. (ed.) *Multiple Criteria Decision Making*. McGraw-Hill: New York.

Zhang, H., 2005. Trust Promoting Seals in Electronic Markets: Impact on Online Shopping Decisions. *Journal of Information Technology Theory and Application JITTA*, 6(4), pp. 29-40.

Zhang, P. and von Dran, G. M., 2000. Satisfiers and Dissatisfiers: A Two-Factor Model for Website Design and Evaluation. *Journal of the American Society for Information Science and Technology*, 51(14), pp. 1253-1268.

Zhang, Z., Ye, Q., Law, R. and Li, Y., 2010. The Impact of e-Word-of-Mouth on the Online Popularity of Restaurants: A Comparison of Consumer Reviews and Editor Reviews. *International Journal of Hospitality Management*, 29(4), pp. 694–700.

Zhu, D. S., Lee, Z. C. and O'Neal, G. S., 2011. Mr. Risk! Please Trust Me: Trust Antecedents that Increase Online Consumer Purchase Intention. *Journal of Internet Banking and Commerce*, 16(3), pp. 1-23.

Appendix A. Initial Investigation Materials

A.1. Initial User Interviews Template

This form is used to capture “user stories” from informal interviews with potential users.

Date:

Place:

Interviewer(s) and their institution:

Interviewee(s) and their institution:

Brief summary:

Notes from each question:

1. **Please describe a current area of your work in which you use external data sources**

2. **What data do you use in your work, and where do they come from?**

3. **How do you choose which datasets to use in your work? What are the reasons for your decisions?**

4. **Are you aware of any data certificates or seals in selecting your data? Do you look for specific certificates or meta-information in a data set you use? How do you know whether to trust the data?**

5. **Do the data you use come with sufficient supporting information to allow you to make an informed judgement about which one(s) to choose? How much information do you need?**

A.2. Example Interview Transcript

Initial user interviews – Interview 1

Date: 31/05/11

Place: Aston University via Telephone

Interviewer(s) and their institution: Jo Lumsden, Victoria Lush (Aston University)

Interviewee(s) and their institution: Interviewee 1

Brief summary:

Interviewee 1 is working at the [Organisation] – a science network of people, organisations, and, most importantly, observation platforms, that perform **Long-Term Ecological Research (LTER)** in [location] and its surrounding oceans. Interviewee 1 works on the development of the [Organisation] platform as a system architect. He works on establishing the data portal and archiving facilities. The [Organisation's] main objective is to look for and gather environmental observation data in the country and filling in the gaps that may exist in their own research. The gathered datasets are to be archived and properly described in some multiple metadata standards.

The data that is used from datasets:
- in this case data is not used but archived; and
- [Organisation] is trying to obtain data for [location], archive/preserve it, and describe it.

Sources used:
- single researchers – smaller datasets that come from single research projects;
- national spatial infrastructure;
- major government departments; and
- major national or international funded effort to obtain data in a specific domain

What influences selection of the datasets:
- they do not make any value judgement but some [Organisation] nodes do;
- should at least have metadata for all data that is found;
- optimisable and interoperable data formats;
- coverages;
- peer review:
    - likes or dislikes;
    - frequency of use; and
    - citation;
- have some kind of value judgement by experts.

Missing data/challenges:
- citation type information;
- coverages;
- licensing;
- methodology; and
- harvesting route.

Standards:
- support ISO family of spatial metadata standards;
- ISO 19115;
- Dublin Core;
- Darwin Core; and
- FGDC.

Quality assessment:
- syntactic – data source conforms to some standardised interoperability specification;
- schematic – comply with a certain conceptual model, or schematic model; and

- semantic – people try and apply the data and then provide feedback. Most complicated as cannot be done automatically. Require peer review system.

[Organisation] has its own set of minimum metadata implementation standards. These are listed in the [Name] document in Annexure [X, Y, Z]. Annexure [X] describes the minimum implementation of EML (Ecological Mark-up Language). High-level elements include:

- DataSet
- System Meta-Data
- Data Set Owner(s)
- Associated Party
- Abstract
- Keywords
- License and Usage Rights
- Coverages
- Geographic Coverage
- Temporal Coverage
- Taxonomic Coverage
- Access Control
- Access Rules
- Contact
- Methods Info
- Data Tables, Images, and Other
- Physical Structure Description
- Attribute(s) Info
- Online Distribution Info

Annexure [Y] describes the minimum ISO implementation:

Citation (what it is about and who created it or owns it) • Title • Publication Date • Author/Responsible Party • Topic/Category/Keywords • Abstract	Technical Data Set Parameters • Language • Character Set • Scale • Format • Format Version • Data Representation • Spatial Reference System	Coverage (main dimensions of the data) • Spatial Coverage/Bounding Coordinates • Temporal Coverage/Start and End Dates for Data Meta-Data Elements • Online Resource • Meta-Data Standard • Custodian's Reference • Meta-Data Creation Date • Custodian

And the Annexure [Z] describes the minimum Dublin Core implementation:

Citation (what it is about and who created it or owns it) • Title • Subject • Publication Date • Author/Responsible Party • Associated Party or Parties (co-authors, etc.) • Topic/Category/Keywords • Abstract	Usage, Restrictions, Caveats • Licensing and Usage Rights Technical Data Set Parameters • Language • Format Coverage • Spatial Coverage • Temporal Coverage/Start and End Dates for Data Meta-Data Elements • Online Resource Custodian's Reference

Notes from each question:

[the following represents a verbatim transcript of the interview[16]]

[16] "?" inserted to indicate where speech in audio recording of interview was no clear and so an estimate has had to be made.

1. **Please describe a current area of your work in which you use external data sources.**

The [Organisation] was created by the government to find and preserve environmental observation data for the long term. The [Organisation] constitutes a national government response to the [event name] and is a component of the [organisation name]. The [Organisation] was in existence for maybe 7 or 8 years now, and initially they really concentrated on establishing important monitoring sites throughout the country, which are now fixed.

And I became involved with [Organisation] semi-permanently about 18 months ago to establish a data portal and archiving facilities for them.

The idea being is that [Organisation] would look for and gather environmental observation data in the country, as well as filling some other gaps that may exist with their own research, and that all of these datasets would be archived for a very long term future use, then properly described in some multiple metadata standards depending on the user community. So that technical work is more or less complete and we are now focusing on acquiring datasets and describing them and getting the metadata records sorted out and quality controlled, and so on.

In parallel to that there have been other developments. First of all, the [department] became the major contributor and a founding member of [Organisation 2]. And in response to that they started the project called [project name] and they approached [Organisation] because they saw that the physical infrastructure and the software that they needed to participate in [Organisation 2] is probably 80% the same as the platform that [Organisation] had already established. So they are now co-funding [Organisation] and [Organisation] acts as implementation agent to also provide the [project name] platform.

And then, if it was not complicated enough, about 6 months later they decided to establish something called the [name] and again [Organisation] was contracted to provide the software and hardware infrastructure to support the applet. And that process now is seemingly ongoing because it looks like [Organisation] platform will also be used for one or two the other science councils – [Council 1] already using it; maybe extended to the [Council 2] and the [Council 3] shortly; and also some other [Organisation 2]-related initiatives such as [name 2]. So it is becoming a fairly comprehensive and wide ranging initiative to find, describe, and reference data sources whether they are locally archived or not. And specifically outside the [Organisation] environment there is an obvious need to be able to find and use data that may be anywhere in the world for that matter. And, increasingly, if those data sources are standardised and automated then one can obviously do a bit more with it. I am a system architect and expeditor of these platform initiatives so it gives me a bit wider view of how we work with scientific data. In [location], research experience is mostly in the open environmental observation domain but that is probably now going to extend to other domains as well.

2. **What data do you use in your work, and where does it come from?**

Unlike some of the typical [Organisation 2] sector providers we face what we call the “long tail” because many of the datasets come from single researchers that are doing some kind of postgraduate degree and move on. So [Organisation] has identified many of these to date and these are smaller datasets that are not necessarily part of some international or even large national effort but are the result of individual research projects. So that is the one kind of dataset that is maybe fairly unique at this point in time, where, let’s say for argument’s sake, an American postgraduate student would visit one of our national parks for 2 years and generate some environmental observation data, and then they move on. But [Organisation] is fairly well placed to try and obtain that data, and archive it and describe it. So we have this one source that is going to grow extensively, I think, in the future because [Organisation] has started to directly engage now with academic institutions to preserve and describe their datasets and their underlying theses and journal articles and so on. So that's one source.

Then we have a programming conjunction with our [name] effort to engage with all the major line departments of the national government, that will typically be things like water affairs and the departments of the environment and a few others, to describe and make available their major datasets whether it's deemed to be part of the national spatial data infrastructure or not, because a lot of the data that they own will not necessarily be seen as part of that infrastructure but it is still of use to, or maybe of use to, someone in the future. So that is another major source of datasets.

That process has only started now and I think one of the things that we can face from our experience, and it is probably is going to be true in many other parts of the world where the legislation is not the major driver, is that there is a lot of institutional, I would not call it resistance, but inertia in the process because, the way I figured it for myself is that, typically these government departments' employees are quite willing to collaborate at first. But they are not being measured in terms of what they provide to us, by way of datasets and metadata, it does not form part of their job descriptions or the way in which their merit assessments are going to be determined. Until that changes, I think, where there is some kind of incentive for government in general to participate, we're going to have this inertia. They are basically doing us favours most of the time and that's not sustainable.

So the major government departments are lined up and then increasingly we'll start focusing also on provincial departments, because that maybe a little different in some places in [location 1] but I think in [location 2] it is very much the same situation where there is large scale duplication of effort at the state level. And the same happens in [location], in terms of provincial government, where they have their own data management environment, especially for spatial data. So I would say that's the second major source.

And then the third major source of datasets is an ad hoc process that revolves around major national or international funded effort to obtain data in a specific domain. So the good examples would be something like the [name] that is hopefully going to kick off in a few months, where [location] will be contributing to the [Organisation 2] initiative to create the bio-energy atlas for [location]. And it is probably going to last for three years or so, with data gathering and modelling and the dissemination of the data. But we need to preserve that data for the future in some way either in [Organisation's] archives or in [name]-funded archive. So that's I would say is the third major source of datasets.

3. **How do you choose which datasets to use in your work? What are the reasons for your decisions?**

At the moment I am not selective. I have a process in place that says we should at least have metadata for all data that we can find, so that we know what exists. Whether it is useful to someone is much more difficult to determine because it may be of some use 20 years in the future for addressing some kind of question or problem that we are unaware of right now. So that is one of the major problems, especially with the [Organisation] mandate, because the other aspect to it is that the data that we need to keep is not always only the latest data. For argument's sake, something like a land cover map of the country: it's of huge interest to the long term environmental monitoring domain to have prior versions of that land cover map available, going back as far as possible, so that they can track the changes and the trends.

So we do not make any value judgement about the data that we keep. We try to make the quality assurance all technical, in the sense that we try and insist on metadata records in the specific standards that we support and also data formats that are easily automatable and interoperable with other similar services in the rest of the world.

We have defined, let's say, a framework for quality assessment that relies on peer review of some kind, almost "crowd sourcing" the assessment of the data by way of likes or dislikes, or frequency of use, or citation but we haven't implemented that yet. So I don't see that we will ever personally make a value judgement on the framework or platform side about the data, except to provide the tools whereby it can automatically be managed if need be.

Having said that, the [Organisation's] nodes do make value judgements for the data that they gather. So because they are domain experts, they will say that it's not worth keeping datasets x, y or z, in which case we just don't publish it. But I've been around for a while so I keep the metadata available, even though is not published, because I am sure that the circumstances might change in the future.

And then, the initiative like the [name] maybe has an interesting perspective, because I agree that not all datasets are useful or have equal value. And I try always to explain that to myself, even by way of thinking about the difference between something like a Google search, which finds everything and anything irrespective of the usefulness to the person researching and maybe Wikipedia which offers a lot of discussion and maybe only a few sources. And I think our [name 1] and [name] type implementation tried to bridge that gap by saying that in the case of the [name 1] they have seven, or ten, or eleven themes, if I remember correctly. Those are similar almost to the societal benefits areas that they use in [Organisation 2] and they are funded theme conveners and scientists that work on those themes describe the datasets that are available and most useful and so on. So maybe that's the way that we are going to keep on doing it, by having some kind of value judgement by experts, on top of, let's say, more common requests for data supporting a given theme.

- **You mentioned that your quality assurances are, as you said, are all technical based on metadata and standards that you support. Can you elaborate on those standards that you support?**

At the moment we support the ISO family of spatial metadata standards. Our national standard is a profile of that called [name], that is probably unimportant at the moment to go outside our borders, it is essentially ISO 19115 profile. We support Dublin Core metadata because not all of our datasets are spatial.

I think that is a bit different between the platform we are building and many of the others like Inspire or even the GEOSS Portal, because overwhelmingly the datasets there are spatial data. We increasingly get datasets that don't have much of a spatial reference. That maybe be something like the relationship between rainfall and bio-mass in one specific location for a very long period of time, for argument's sake, and there is not really any spatial reference except a point where the data is collected. So we support Dublin Core for non-spatial datasets.

[Organisation] especially does a lot of work with [organisation name] network. They store all of their metadata, and sometimes their data, in a product called MetaCat. And the metadata description language that they are using is called **Ecological Mark-up Language (EML)**. It is a bit troublesome for us, because it is a very well engineered environment but it is not mainstream and we continuously need to translate that to maybe other metadata standards that are more commonly used.

We also support FGDC (http://www.fgdc.gov/metadata/geospatial-metadata-standards) but I think it will probably disappear in the next few years.

And we have projects underway to support the Darwin Core metadata. That is slightly different, because it is essentially the data and the metadata of some kind of species observation that's recorded in Darwin Core and that is the basis of all the data that goes to [organisation name], as I am sure you'll know. We don't want to generate millions of records in Darwin Core so we are looking at a way of almost something like a Meta-Darwin Core record that is aggregated to some kind of higher level of spatial and temporal resolution. So that people can search maybe for occurrences of a genus in a quarter degree square for the decade, something along those lines.

4. **Are you aware of any data certificates or seals in selecting your data? Do you look for specific certificates or meta-information in a data set you use? How do you know whether to trust the data?**

No, I haven't. But I think it is probably an idea for which the time has come because it was recently discussed at some length at the [Organisation 2] work plan symposium. And it was also discussed...I am a member of the [name] scientific network for the world data system(?). And there that is even more, let's say, pertinent topic because the world data system places much stronger emphasis on data quality and usability. So we haven't done anything like it but we've done some conceptual work in this regard and we have came up with the three peer statements:

One is essentially **syntactic**, and the other one is **schematic**, and the other one is **semantic**...if I have to substitute the, let's say the commonly used terms for it. But in essence what it is saying that if a data source conforms to some standardised interoperability specification, for argument's sake, it is a WebMap service or NetCDA data source or whatever, then we can certify its syntax and say that yes it sort of pacifies the condition for being interoperable. And if someone uses the correct syntax, they can access the data, maybe even query it in some cases.

But then there is also a schematic requirement. For argument's sake, you cannot say that the data is a network of roads if it doesn't comply with a certain conceptual model, or schematic model of what data like that should look like, that it must be composed of nodes and links, or something along those lines. So, many of the datasets can be verified automatically in respect of the schema or, let's say, their compliance with some kind of widely accepted schema for data such as that. It's a bit more difficult but, I guess, there are conventions for it that one could work on.

The most difficult one is the usability or the semantic interoperability. Where really the only way to judge it is for people to try and apply the data and come back and say "no it is not useful because it's inaccurate, or methodology wasn't appropriate", or whatever the case may be. And that I think is probably the challenge, because the others can be evaluated automatically with a scan. So that's why we are looking for some kind of a peer review system for people to pick the usability of the data, in respect of its accuracy and its usefulness to them.

5. **Does the data you use come with sufficient supporting information to allow you to make an informed judgement about which one(s) to choose? How much information do you need?**

No, I don't think that is the case. At the moment our strategy, as I said, is first to try and find metadata references so that we know what exists. And then slowly, but surely, trying to work down the pipeline to improve the quality of the data and understand what it might be useful for and so on. But that is obviously is a huge task. So I don't think that metadata as it's provided at present contains enough information about the data to make it useful. And it's gets worse because you will find the datasets that, for instance, if we look at EML, that is very well engineered, and it is well capable of describing a dataset in terms of let's say spatial coverage, and temporal coverage, and taxonomic coverage, which are the three main ones that it deals with. But very often people have the species or taxonomical references inside the dataset and there is no mention of the species that are covered in the dataset at all in the metadata. Now this is very difficult to discover because you have to manually inspect every single dataset before you start understanding whether the metadata adequately describes the data to start with. So I guess it's a huge issue, and maybe one that, if people start applying their minds to this and probably a mixture of better tools and guidelines, then one can maybe hope to improve on it. But I think historical data, the best option almost would be to say that, if someone really uses that dataset one could almost try and rely on them to improve the metadata because they going to use the dataset anyway.

- **What information would you ideally like to see in the metadata associated with a dataset?**

We've derived a minimum or, let's say, a mandatory metadata elements set. But generally speaking, the things that I think are often missing are, apart now from the obvious descriptions like citation type information and the coverages, the other things that are generally really always missing are things like the licence under which it can be distributed...that's something very often one has to find out afterwards by contacting the providers of the data again. And that is critical, I think, for the kinds of systems that we want to build to understand what licence applies to the dataset.

Then, I think, the methodology is probably also important in the metadata because if we are going to preserve data for the very long term, in 50 years from now the people who generated that dataset are not going to be around anymore. If it's a year from now you can probably contact the provider and find out more about the approach that they followed to generate the data and so on. But in the very long term that's not going to be possible. So, I think, that's another aspect that one needs to look at improving substantially.

And then the other thing which is maybe not the fault of the researchers, but it's an issue I think in the standards and how people work with metadata...because of the growth in a number of portals that rely on metadata and on harvesting metadata from one another, the path that the metadata follows to arrive at your doorstep is lost. You only ever, usually, have a portal that you harvested it from, and may be the starting point but you don't have the intermediaries. So one other thing we are looking at is to improve the way in which we store maybe a pathway where each portal that touches a metadata record has an element to describe who they are and where they are located and who to contact at the end of the metadata record, so that one can trace the path. We are thinking along those lines, so that we have an opportunity to maybe optimise the harvesting in future, because we think it's going to become an onerous path to keep these portals up-to-date and synchronised and one that we want to harvest unnecessary...if you can cut out a very long harvesting route by going directly to another source then maybe one should be able to do so.

A.3. User Stories

User Type Abbreviations

User Type Abbreviation	Description
ACAD	Academic
ARCH	Archivist
CF	Climate Forecaster
DU	Data User
PROD	Producer
RES	Researcher
S-ARC	System Architect

User Stories Derived from Aston University Interviews

User Story ID	User Story (Derived from User Interviews)	User Type
US-1	As a data archivist/system architect, I want to be able to accurately describe and archive/preserve data to make it available longterm for selection and use.	Archivist, System Architect
US-1.1	As a data archivist/system architect, I want to look for and gather environmental observation data so that I can archive/preserve that data for a very long term future use.	
US-1.2	As a data archivist/system architect, I want to be able to access all the information provided with the dataset.	
US-1.3	As a data archivist/system architect, I want to find and use data that may be anywhere in the world so that I can describe and reference data sources.	
US-1.4	As a data archivist/system architect, I want to have quality assessment that relies on peer review, almost “crowd sourcing”, the assessment of the data by way of likes or dislikes, or frequency of use, or citation.	
US-2	As a climate forecaster/data user I want to be able to ascertain key metadata information as well as to investigate the uncertainty of information in a dataset so that I can accurately determine its quality, decide on pre-processing that needs to be done, and make a decision as to whether or not to trust the source.	Climate Forecaster, Data User
US-2.1	As a climate forecaster/data user, I want to be able to get a community advice on the datasets.	
US-2.2	As a climate forecaster/data user, I want dataset providers to record some soft knowledge about their datasets so that I can see the purpose for which the datasets were created, what are the boundaries of the datasets and when should I trust or not to trust the datasets.	

US-3	As a researcher/data user, I want to be able to ascertain key metadata information as well as to investigate the quality of information at specific points in the dataset so that I can accurately determine its quality, usefulness for my purpose, and the extent of callibration required before being able to use the dataset.	Researcher, Data User
US-3.1	As a researcher/data user, I want data providers to have certain quality standards and provide the documentation on the standards that they support.	
US-3.2	As a researcher/data user, I want to have more complete metadata records provided with the datasets so that I can do proper calibration if data is not calibrated.	
US-4	As a land use researcher, I want to be able to ascertain key metadata information as well as to contribute in a peer review sense to metadata for existing datasets and to compare dataset metadata across a range of datasets in order to make an informed dataset selection.	Researcher, Data User
US-4.1	As a land use researcher/data user, I want to know who to contact (someone who actually replies) about the data so that I can get additional information.	
US-4.2	As a land use researcher/data user, I want to access dataset citations so that I know exactly why and how the data was collected, and what the errors are in the dataset	
US-5	As an academic/researcher, I want to be able to trace the processing steps that have been applied to the data as well as to access the raw data so that I can go back to the root of the data processing, if the processing applied is not useful for me, and process the data myself.	Academic, Researcher
US-5.1	As an academic/researcher, I am interested in the data itself rather than metadata records so I only want a light number of elements from the metadata record as well as the log of all the processing steps applied to the data.	
US-5.2	As an academic/researcher, I want to acquire the datasets that are free or at a very low cost.	
US-6	As a researcher/data producer, I want to be able to populate metadata records in multiple, ad-hoc metadata standards and also flag (warn in general terms) the potential lack of quality in the data I produce.	Researcher, Data Producer
US-6.1	As a researcher/data producer, I want to be able to find out about the data from other sources (word of mouth) and accept recommendations from others as well as obtain feedback on the data that I produce.	
US-6.2	As a scientist/data producer I want to have access to data provider's documentation on the datasets they provide.	

User Stories Derived from Reading University Interviews

User Story ID	User Story (Derived from User Interviews)	User Type
US-READ-1	As an environmental researcher/data user, I want to be able to ascertain key metadata information, particularly methodology and data citations, as well as to accept community advice and recommendations in order to make an informed dataset selection.	RES, DU
US-READ-1.1	As a scientist, for a reanalysis product, I want to know: what version of the code that was used in generating the reanalysis, what observation data used to generate reanalysis, what methods used in generating reanalysis, what assimilation system used in generating the reanalysis, and documentation of the reanalysis of the product, etc.	RES, DU
US-READ-1.2	As a generic user or scientist, I want to know the up-to-date version/ beta version of this dataset, build-up history of the dataset (derived or measured dataset (e.g. reanalysis product, in-situ observation)), methodology consistency, retrieval algorithm (if used), whether known issues fixed, etc. so that I can latter derive more quality information on this dataset.	RES, DU
US-READ-1.3	As a generic user or a scientist, I want to know the methodology that is used to derive or measure the dataset (e.g. reanalysis product, in-situ observation), so that I can have more quality information about this dataset.	RES, DU
US-READ-1.4	As a scientist or a generic user, I want to know what communities or projects are using or have used this dataset; I also want to know the associated information about the communities, so that it can help me to make a judgement about how much I can trust the dataset.	RES, DU
US-READ-2	As a researcher/data user, I want to be able to ascertain key metadata information, particularly information on data errors, as well as to obtain community advice and recommendations on what datasets can be trusted so that I can make an informed dataset selection.	RES, DU
US-READ-2.1	As a scientist or a generic user, I want to know where and when this dataset has been published, so that it can help me to make a judgement about how much I can trust the dataset.	RES, DU
US-READ-2.2	As a scientist or a generic user, I want to know the error estimates for this dataset, so that it can help me to make a judgement about how much I can trust the dataset.	RES, DU
US-READ-2.3	As a scientist or a generic user, if the data has been published, I want to know how many citations in scientific literature, so that it can help me to make judgement about how much I trust the dataset (i.e. as a trust trigger).	RES, DU
US-READ-2.4	As a generic user or scientist, I want to know some key 'profile' information of the dataset: for example, spatial resolution, temporal resolution, time period coverage, geolocation coverage, etc., these key profile information can help me to better understand the quality of the dataset.	RES, DU

US-READ-2.5	As a scientist or generic user, I want to know whether the dataset has gone through certain 'validation campaign', what the 'validation campaign ' is, and what the validation results are, etc. so that I make a judgement about much I can trust the dataset.	RES, DU
US-READ-2.6	As a generic user or scientist, I want the system to tell me any inter-comparison result(s) with other dataset, so that it can help me to make a judgement about how much I can trust this dataset. The item to be compared can include: geolocation resolution, temporal resolution, time period coverage, geospatial coverage, time of the data produced, citations, etc. (More items can be identified at a later stage)	
US-READ-2.7	As a generic user, I want the system to provide a mechanism that can encourage the data users to make comments on the data they have used.	DU
US-READ-2.8	As a scientist or a generic user, I want to know what communities or projects are using or have used this dataset; I also want to know the associated information about the communities, so that it can help me to make a judgement about how much I can trust the dataset.	RES, DU
US-READ-3	As an environmental researcher/data user, I want to be able to ascertain key metadata information, particularly information on data provenance; I want to use the datasets that come from reputable source as well as to accept community recommendations when selecting a dataset to use.	RES
US-READ-3.1	As a scientist, for a reanalysis product, I want to know: what version of the code that was used in generating the reanalysis, what observation data used to generate reanalysis, what methods used in generating reanalysis, what assimilation system used in generating the reanalysis, and documentation of the reanalysis of the product, etc.	RES
US-READ-3.2	As a scientist, I want to know whether the dataset has gone through certain 'validation campaign', what the 'validation campaign ' is, and what the validation results are, etc. so that I make a judgement about much I can trust the dataset.	RES
US-READ-4	As a meteorology researcher/data user, I want to be able to compare a new dataset with a “trusted” one as well as accept community recommendations in order to make an informed dataset selection.	RES, DU
US-READ-4.1	As a generic user or a scientist, if a quality label is provided, I want to know the method(s) used to create the label, who creates the label, so that it can help me to make a judgment whether I should trust the label	RES, DU
US-READ-4.2	As a generic user or scientist, I want to know the up-to-date version / beta version of this dataset, build-up history of the dataset (derived or measured dataset (e.g. reanalysis product, in-situ observation)), methodology consistency, retrieval algorithm (if used), whether known issues fixed, etc. so that I can latter derive more quality information on this dataset.	RES, DU

US-READ-4.3	As a generic user or scientist, I want to know which data centre or organisation is hosting this dataset, what is access methods (e.g. FTP), any standard(s) (e.g. IPCC standard) are imposed on the archival of dataset, any access restrictions, etc., so that it can help me to make a judgement whether I will use this dataset.	RES, DU
US-READ-4.4	As a scientist, if the quantified error estimates of a dataset are provided, I want to know what methodology is used to determine the error estimates, and the peer-reviewed comments or publications for the methodology, so that I can understand better about how much uncertainties on this error estimates.	RES, DU
US-READ-4.5	As a scientist, I want the system to provide a kind of mechanism for me to compare the dataset with a 'trusted' reference dataset, so that I can make a judgement about much I can trust the dataset.	RES, DU
US-READ-4.6	As a scientist or a generic user, I want to know what communities or projects are using or have used this dataset; I also want to know the associated information about the communities, so that it can help me to make a judgement about how much I can trust the dataset.	RES, DU
US-READ-4.7	As a scientist or generic user, I want to know the problems identified and highlighted about the dataset, so that it can help me to make a judgement about how much I can trust the dataset.	RES, DU
US-READ-5	As a meteorology researcher/data user, I want to be able to ascertain key metadata information, particularly price, resolution and availability of the data, as well as to obtain community advice and recommendations on datasets and data providers' reputation so that I can make an informed dataset selection.	RES, DU
US-READ-5.1	As a data provider, I want the system to provide me a kind of mechanism (e.g. template) that can guide me to provide required quality information for the dataset.	RES, DU (PROD)
US-READ-5.2	As a scientist or a generic user, I want to know more information (e.g. website, email) about the data provider, or more information (e.g. email, background, publications) about the data author, so that I can make a judgement about how much I can trust the dataset, or I can ask them specific questions straightaway.	RES, DU
US-READ-5.3	As a generic user or scientist, I want to know some key 'profile' information of the dataset: for example, spatial resolution, temporal resolution, time period coverage, geolocation coverage, etc., these key profile information can help me to better understand the quality of the dataset.	RES, DU
US-READ-5.4	As a scientist or a generic user, I want to know what communities or projects are using or have used this dataset; I also want to know the associated information about the communities, so that it can help me to make a judgement about how much I can trust the dataset.	RES, DU

US-READ-6	As a meteorology researcher/data user, I want to be able to ascertain key metadata information, particularly error estimates, correlations between error estimates and provenance information, as well as to obtain users' comments on the datasets so that I can make an informed dataset selection.	RES, DU
US-READ-6.1	Sometimes choosing which dataset to use depends on specific research question, model used, etc. As a scientist, I want the system to have some kind of intelligence: after entering specific information (e.g. research question, model information), the system can recommend me the appropriate dataset.	RES
US-READ-6.2	As a generic user or a scientist, I want to know which model(s) (if any) are suggested to use with this reanalysis product.	RES, DU
US-READ-6.3	As a scientist or generic user, I want to know whether the dataset has gone through certain 'validation campaign', what the 'validation campaign ' is, and what the validation results are, etc. so that I make a judgement about much I can trust the dataset.	RES, DU
US-READ-6.4	As a scientist, if some error estimates of a dataset are provided, I want to know the correlations between the error estimates, so that I can understand whether there is any error over-lapping.	RES
US-READ-6.5	As a generic user, I want to see more user comments and any recommendations on this dataset, etc. so that I make a judgement about much I can trust the dataset.	DU

A.4. User Requirements

A.4.1 Metadata Querying and Standards

Req. ID	Requirement	User Story	User Type
R-3.4	The system should support the querying/search of a wide range of meta data values by individuals selecting datasets to use.	US-1, US-3.2, US-6	ARCH, S-ARC, RES, DU, PROD
R-3.4.2	The system should support the querying of the meta-data standard attribute from the dataset metadata record.	US-1	ARCH, S-ARC

A.4.2 Peer Review and Rating Functionality

Req. ID	Requirement	User Story	User Type
R-10.	The system should support peer review ("crowd sourcing") of datasets (and subsequent querying of such reviews) such that dataset users can recommend or otherwise datasets to other users in the community.	US-1.4, US-2.1, US-4, US-6.2, US-READ-6.5	ARCH, S-ARC, CF, DU, RES, PROD
R-10.1	The system should support the reviewing of the datasets by providing the users with a standardised form to record their dataset evaluation.	US-4	RES, DU
R-10.3	The system should support the querying of all the submitted peer reviews for the selected dataset.	US-1.4, US-2.1, US-4, US-6.2	ARCH, S-ARC, CF, DU, RES, PROD
R-10.4	The system should support the querying of all the comments on metadata records submitted by the users.	US-4	RES, DU
R-11	The system should support the ability to rate (potentially by peer-review) the reputation of a dataset and to allow this rating to be subsequently edited, queried, and searched and to allow the dataset provider to respond to the rating.	US-4	RES, DU
R-11.1	The system should support the ability to rate the reputation of a dataset.	US-4	RES, DU
R-11.2	The system should support the ability to edit the rating of the reputation of a dataset.	US-4	RES, DU
R-11.3	The system should support the ability to query the ratings of the reputation of a dataset.	US-4	RES, DU
R-11.4	The system should support the ability to search the ratings of the reputation of a dataset.	US-4	RES, DU
R-11.5	The system should allow the dataset provider to respond to the rating of the dataset reputation.	US-4	RES, DU
R-12	The system should support the ability to rate (potentially by peer-review) the reputation of a dataset provider and to allow this rating to be subsequently edited, queried, and searched and to allow the dataset provider to respond to the rating.	US-2.1, US-3, US-4	CF, DU, RES
R-12.1	The system should support the ability to rate the reputation of a dataset provider.	US-2.1, US-3, US-4	CF, DU, RES
R-12.2	The system should support the ability to edit the rating of the reputation of a dataset provider.	US-2.1, US-3, US-4	CF, DU, RES

R-12.3	The system should support the ability to query the ratings of the reputation of a dataset provider.	US-2.1, US-3, US-4	CF, DU, RES
R-12.4	The system should support the ability to search the ratings of the reputation of a dataset provider.	US-2.1, US-3, US-4	CF, DU, RES
R-12.5	The system should allow the dataset provider to respond to the rating applied to them.	US-2.1, US-3, US-4	CF, DU, RES

A.4.3 Citations Information

Req. ID	Requirement	User Story	User Type
R-13	The system should support the recording, editing and querying of the dataset citations information.	US-3.1, US-4.2, US-6.2, US-READ-2.1, US-READ-2.3	RES, DU, PROD
R-13.1	The system should support the recording of links to journal articles (or other publications) in which dataset quality checks are reported.	US-3.1, US-4.2, US-6.2	RES, DU, PROD
R-13.1.1	The system should support the recording of links to journal articles in which dataset quality checks are reported.	US-3.1, US-4.2	RES, DU
R-13.1.2	The system should support the recording of links to dataset provider documentation in which any additional information on the dataset is provided.	US-3.1, US-6.2	RES, DU, PROD
R-13.2	The system should support the editing of links to journal articles (or other publications) in which dataset quality checks are reported.	US-3.1, US-4.2, US-6.2	RES, DU, PROD
R-13.2.1	The system should support the editing of links to journal articles in which dataset quality checks are reported.	US-3.1, US-4.2	RES, DU
R-13.2.2	The system should support the editing of links to dataset provider documentation in which any additional information on the dataset is provided.	US-3.1, US-6.2	RES, DU, PROD
R-13.3	The system should support the querying of links to journal articles (or other publications) in which dataset quality checks are reported.	US-3.1, US-4.2, US-6.2	RES, DU, PROD
R-13.3.1	The system should support the querying of links to journal articles in which dataset quality checks are reported.	US-3.1, US-4.2	RES, DU
R-13.3.2	The system should support the querying of links to dataset provider documentation in which any additional information on the dataset is provided.	US-3.1, US-6.2	RES, DU, PROD

A.4.4 Dataset Provider Information

Req. ID	Requirement	User Story	User Type
R-14	The system should support the recording, editing, reviewing, and querying of the information on the dataset provider.	US-1.2, US-3.1, US-4.1, US-5, US-READ-5.2	ARCH, S-ARC, RES, DU, ACAD
R-14.1	The system should support the recording, editing, reviewing, and querying of the quality standards that the dataset provider supports.	US-3.1, US-READ-4.3	RES, DU
R-14.1.1	The system should support the recording of the quality standards that the dataset provider supports.	US-3.1	RES, DU
R-14.1.2	The system should support the editing of the quality standards that the dataset provider supports.	US-3.1	RES, DU

R-14.1.3	The system should support the reviewing of the quality standards that the dataset provider supports.	US-3.1	RES, DU
R-14.1.4	The system should support the querying of the quality standards that the dataset provider supports.	US-3.1	RES, DU
R-14.2	The system should support the recording, editing, reviewing, and querying of the links to the textual documents in which dataset providers' quality standards are reported.	US-3.1	RES, DU
R-14.2.1	The system should support the recording of the links to the textual documents in which dataset providers' quality standards are reported.	US-3.1	RES, DU
R-14.2.2	The system should support the editing of the links to the textual documents in which dataset providers' quality standards are reported.	US-3.1	RES, DU
R-14.2.3	The system should support the reviewing of the links to the textual documents in which dataset providers' quality standards are reported.	US-3.1	RES, DU
R-14.2.4	The system should support the querying of the links to the textual documents in which dataset providers' quality standards are reported.	US-3.1	RES, DU
R-14.3	The system should support the recording, updating, referencing and querying of contact details for the person(s) responsible for the creation of a given dataset.	US-1.2, US-4.1	ARCH, S-ARC, RES, DU
R-14.3.1	The system should support the recording of contact details for the person(s) responsible for the creation of a given dataset.	US-1.2, US-4.1	ARCH, S-ARC, RES, DU
R-14.3.2	The system should support the updating of contact details for the person(s) responsible for the creation of a given dataset.	US-1.2, US-4.1	ARCH, S-ARC, RES, DU
R-14.3.3	The system should support the referencing of contact details for the person(s) responsible for the creation of a given dataset.	US-1.2, US-4.1	ARCH, S-ARC, RES, DU
R-14.3.4	The system should support the querying of contact details for the person(s) responsible for the creation of a given dataset.	US-1.2, US-4.1	ARCH, S-ARC, RES, DU
R-14.4	The system should support the recording, editing, reviewing, and querying of the location of the dataset producer.	US-1.2	ARCH, S-ARC
R-14.4.1	The system should support the recording of the location of the dataset producer.	US-1.2	ARCH, S-ARC
R-14.4.2	The system should support the editing of the location of the dataset producer.	US-1.2	ARCH, S-ARC
R-14.4.3	The system should support the reviewing of the location of the dataset producer.	US-1.2	ARCH, S-ARC
R-14.4.4	The system should support the querying of the location of the dataset producer.	US-1.2	ARCH, S-ARC

A.4.5 Soft Knowledge

Req. ID	Requirement	User Story	User Type
R-19	The system should support the recording, editing, reviewing, and querying of soft knowledge about a dataset as provided by the creator of the dataset and to include: why it was originally derived - for what purpose; anticipated limitations of the dataset; anticipated boundaries of use for the dataset; conditions under which the data should be trusted or not trusted, etc.	US-2.2, US-4	CF, DU, RES
R-19.1	The system should support the recording of soft knowledge about a dataset as provided by the creator of the dataset.	US-2.2, US-4	CF, DU, RES
R-19.1.1	The system should support the recording of soft knowledge about why/for what purpose the dataset was originally derived as provided by the creator of the dataset.	US-2.2, US-4	CF, DU, RES
R-19.1.2	The system should support the recording of soft knowledge about the anticipated limitations of the dataset as provided by the creator of the dataset.	US-2.2, US-4, US-READ-4.7	CF, DU, RES
R-19.1.3	The system should support the recording of soft knowledge about the anticipated boundaries of use for the dataset as provided by the creator of the dataset.	US-2.2, US-4	CF, DU, RES
R-19.1.4	The system should support the recording of soft knowledge about the conditions under which the data should be trusted or not trusted as provided by the creator of the dataset.	US-2.2, US-4, US-READ-4.7	CF, DU, RES
R-19.1.5	The system should support the recording of at least soft knowledge about the dataset uncertainty as provided by the creator of the dataset.	US-2.2	CF, DU
R-19.1.6	The system should support the recording of soft knowledge about the model(s) that are suggested to use with the selected dataset.	US-READ-6.2	RES, DU
R-19.2	The system should support the editing of soft knowledge about a dataset as provided by the creator of the dataset.	US-2.2, US-4	CF, DU, RES
R-19.2.1	The system should support the editing of soft knowledge about why/for what purpose the dataset was originally derived as provided by the creator of the dataset.	US-2.2, US-4	CF, DU, RES
R-19.2.2	The system should support the editing of soft knowledge about the anticipated limitations of the dataset as provided by the creator of the dataset.	US-2.2, US-4, US-READ-4.7	CF, DU, RES
R-19.2.3	The system should support the editing of soft knowledge about the anticipated boundaries of use for the dataset as provided by the creator of the dataset.	US-2.2, US-4	CF, DU, RES
R-19.2.4	The system should support the editing of soft knowledge about the conditions under which the data should be trusted or not trusted as provided by the creator of the dataset.	US-2.2, US-4	CF, DU, RES
R-19.2.5	The system should support the editing of at least soft knowledge about the dataset uncertainty as provided by the creator of the dataset.	US-2.2	CF, DU

R-19.2.6	The system should support the editing of soft knowledge about the model(s) that are suggested to use with the selected dataset.	US-READ-6.2	RES, DU
R-19.3	The system should support the querying of soft knowledge about a dataset as provided by the creator of the dataset.	US-2.2, US-4	CF, DU, RES
R-19.3.1	The system should support the querying of soft knowledge about why/for what purpose the dataset was originally derived as provided by the creator of the dataset.	US-2.2, US-4	CF, DU, RES
R-19.3.2	The system should support the querying of soft knowledge about the anticipated limitations of the dataset as provided by the creator of the dataset.	US-2.2, US-4, US-READ-4.7	CF, DU, RES
R-19.3.3	The system should support the querying of soft knowledge about the anticipated boundaries of use for the dataset as provided by the creator of the dataset.	US-2.2, US-4	CF, DU, RES
R-19.3.4	The system should support the querying of soft knowledge about the conditions under which the data should be trusted or not trusted as provided by the creator of the dataset.	US-2.2, US-4	CF, DU, RES
R-19.3.5	The system should support the querying of at least soft knowledge about the dataset uncertainty as provided by the creator of the dataset.	US-2.2	CF, DU
R-19.3.6	The system should support the querying of soft knowledge about the model(s) that are suggested to use with the selected dataset.	US-READ-6.2	RES, DU

A.4.6 Inter-comparison

Req. ID	Requirement	User Story	User Type
R-15	The system should support side-by-side visual comparison of metadata attributes across a range of datasets.	US-4	RES, DU
R-32	The system should provide the results of the dataset inter-comparison with other dataset(s).	US-READ-2.6	RES, DU
R-32.1	The system should provide the results of inter-comparison of the datasets' geolocation.	US-READ-2.6	RES, DU
R-32.2	The system should provide the results of the inter-comparison of the datasets' resolution.	US-READ-2.6	RES, DU
R-32.3	The system should provide the results of the inter-comparison of the datasets' temporal resolution.	US-READ-2.6	RES, DU
R-32.4	The system should provide the results of the inter-comparison of the datasets' time period coverage.	US-READ-2.6	RES, DU
R-32.5	The system should provide the results of the inter-comparison of the datasets' geospatial coverage.	US-READ-2.6	RES, DU
R-32.6	The system should provide the results of the inter-comparison of the datasets' citation information.	US-READ-2.6	RES, DU
R-33	The system should provide the results of inter-comparison of the selected dataset with a 'trusted' dataset.	US-READ-4.5	RES, DU

A.4.7 GEO Label Provenance Information

Req. ID	Requirement	User Story	User Type
R-27	The system should provide the GEO Label provenance information.	US-READ-4.1	RES, DU
R-27.1	The system should provide the users with information on the methods used to generate the label.	US-READ-4.1	RES, DU
R-27.2	The system should provide the users with information on the label provider.	US-READ-4.1	RES, DU
R-27.3	The system should provide the users with information on underlying quality assessment mechanisms if the GEO Label is used to convey quality assessment of the datasets.	US-READ-4.1	RES, DU

Appendix B. Phase I Study GEO Label Questionnaire

GEO Label Questionnaire

Introduction

GeoViQua is an EU project that is working on providing the GEO data user community with innovative quality-aware visualisation and advanced geo-search capabilities. One of the objectives of our project is to contribute to defining a concept of a GEO Label. This questionnaire aims to collect initial user views on a GEO Label and its potential role.

We would be very grateful if you could spare some time to complete this questionnaire. It should take you no more than approximately 30 minutes to complete. Your responses will be completely anonymous and confidential, and will only be used in our study. Your responses will provide important information that will help us in defining what role a GEO Label should fulfil.

Before beginning this questionnaire, please identify yourself as one of the following:
(Tick one that applies)

- ☐ Data user ➔ Please go to **Section A – Data User**
- ☐ Data producer ➔ Please go to **Section A – Data Producer**

Section A – Data User

Section A – Background Information

This section consists of questions to gather background information about you, the data sources you use in your work and your awareness of any certificates or seals that apply to geospatial data.

A1. Please select the description(s) of user that most accurately describe you: (Tick all that apply)

☐ Group on Earth Observations committee member
☐ Private sector data user
☐ Governmental data user
☐ Researcher data user
☐ Academic data user
☐ Other: please specify ____

A2. In your typical work do you use data from external data providers?

☐ Yes
☐ No

If you answered **No** skip to question **A14**, otherwise continue to question **A3**

A3. Please list the external data providers from whom you source your data. If possible, describe the type(s) of data that you use.

A4. Is there more than one data provider supplying the types of datasets you need?

☐ Yes
☐ No

If you answered **YES** skip to question **A6**, otherwise continue to question **A5**

A5. From the sole data provider, do you have a choice of datasets to meet your needs?

☐ Yes
☐ No

If you answered **NO** skip to question **A14**, otherwise continue to question **A6**

A6. Do you consider metadata records or any other supporting information when selecting datasets for use?

☐ Yes
☐ No

If **Yes,** what metadata or supporting information do you consider when working with geospatial datasets?

A7. When selecting a dataset to use, how important to you is an expert's judgement of the dataset and its quality?
(Tick one that applies)

☐	☐	☐	☐	☐	☐	☐
Very Unimportant	Unimportant	Somewhat Unimportant	Neutral	Somewhat Important	Important	Very Important

A8. When selecting a dataset to use, how important to you is a dataset's compliance with international standards? E.g., compliance with Dublin Core, ISO19115 standards, etc.
(Tick one that applies)

☐	☐	☐	☐	☐	☐	☐
Very Unimportant	Unimportant	Somewhat Unimportant	Neutral	Somewhat Important	Important	Very Important

A9. When selecting a dataset to use, how important to you are community advice and recommendations on what datasets are best to use for your application?
(Tick one that applies)

☐	☐	☐	☐	☐	☐	☐
Very Unimportant	Unimportant	Somewhat Unimportant	Neutral	Somewhat Important	Important	Very Important

A10. When selecting a dataset to use, how important to you is the reputation of the dataset provider?
(Tick one that applies)

☐	☐	☐	☐	☐	☐	☐
Very Unimportant	Unimportant	Somewhat Unimportant	Neutral	Somewhat Important	Important	Very Important

A11. When selecting a dataset to use, how important to you are dataset citations (e.g., a list of journal articles or other publications where the dataset has been used and quality checks have been reported)?
(Tick one that applies)

☐	☐	☐	☐	☐	☐	☐
Very Unimportant	Unimportant	Somewhat Unimportant	Neutral	Somewhat Important	Important	Very Important

A12. When selecting a dataset to use, how important to you is 'soft knowledge' (subjective and informal statements) about the dataset quality that is provided by the creator or provider of the dataset?
(Tick one that applies)

☐	☐	☐	☐	☐	☐	☐
Very Unimportant	Unimportant	Somewhat Unimportant	Neutral	Somewhat Important	Important	Very Important

A13. When selecting a dataset to use, how useful would it be for you to be able to visualise metadata records side-by-side when comparing two or more datasets?
(Tick one that applies)

☐	☐	☐	☐	☐	☐	☐
Very Useless	Useless	Somewhat Useless	Neutral	Somewhat Useful	Useful	Very Useful

A14. Are you aware of any certificates or seals that certify spatial datasets or metadata records?

☐ Yes
☐ No

If you answered **No** skip to **Section B**, otherwise continue to question **A15**

A15. Please provide the names of the certificates or seals that you have encountered:

A16. In your opinion, are these certificates or seals useful to the user community?

☐ Yes
☐ No

If **Yes**, please elaborate on why these certificates or seals are **USEFUL** to the user community:

If **No**, please elaborate on why these certificates or seals are **NOT USEFUL** to the user community:

Now proceed to **Section B**

Section A – Data Producer

Section A – Background Information
This section consists of questions to gather background information about you, the data you produce in your work and your awareness of any certificates or seals that apply to geospatial data.

A1. Please select the description(s) of producer that most accurately describe you:

(Tick all that apply)

- ☐ Group on Earth Observations (GEO) committee member
- ☐ Data producer with some dataset(s) in the Global Earth Observation System of Systems (GEOSS) Common Infrastructure
- ☐ Data producer not part of Global Earth Observation System of Systems (GEOSS)
- ☐ Private sector data producer
- ☐ Governmental data producer
- ☐ Research/academic data producer
- ☐ Other: please specify ____

A2. Please list and describe the type(s) of data that you produce.

A3. Do you provide metadata records or any other supporting information with the datasets that you produce?

- ☐ Yes
- ☐ No

If **Yes,** what metadata or supporting information do you provide with your datasets?

If **No**, please explain why not:

A4. Do you support any data and/or metadata standards, such as ISO, Dublin Core, Darwin Core, etc?

☐ Yes
☐ No

If **Yes,** list the data and/or metadata standards that you support:

If **No**, please explain why not:

A5. Are you aware of any certificates or seals that certify spatial datasets or metadata records?

☐ Yes
☐ No

If you answered **No** skip to **Section B**, otherwise continue to question **A6**

A6. Please provide the names of the certificates or seals that you have encountered:

A7. In your opinion, are these certificates or seals useful to the user community?

☐ Yes
☐ No

If **Yes**, please elaborate on why these certificates or seals are **USEFUL** to the user community:

If **No**, please elaborate on why these certificates or seals are **NOT USEFUL** to the user community:

Now proceed to **Section B**

Section B – Initial User Perspectives on a GEO Label

For the purpose of this questionnaire we consider a GEO Label to be some form of graphical representation that accompanies a dataset.
The following set of questions is designed to gather information about your initial view on the role that a GEO Label should serve.

B1. Do you think geospatial data or metadata records would benefit from certification programme(s) being applied to them?

☐ Yes
☐ No

If **Yes**, please provide a brief explanation as to why you think certification programme(s) **WOULD** be beneficial:

If **No**, please provide a brief explanation as to why you think certification programme(s) **WOULD NOT** be beneficial:

B2. If a GEO Label is provided, what role would you want it to serve? Please provide a brief explanation as to what would you want a GEO Label to represent:

B3. If a GEO Label is provided with datasets to certify their quality/trustworthiness, do you think the presence of such a label would influence your dataset selection decision?

☐ Yes
☐ Depends on whether I have previously used the data
☐ No
☐ Don't know

Please provide a brief explanation to accompany your answer:

Now proceed to **Section C**

Section C – Your Awareness of Everyday Rating Systems and Certification Programmes

To inform the definition and role of a GEO Label, it is useful to first consider common certificates, labels, rating systems, and feedback facilities that are used in everyday life to help in product or service selection. We are interested in your opinion on such systems so that we can draw on the strengths of these when developing the GEO Label.

C1. Which of the following review/rating systems are you aware of?
(Tick all that apply)

- ☐ eBay
- ☐ Amazon
- ☐ TripAdvisor
- ☐ I am not aware of any rating/review systems
- ☐ Other ____

If you answered "**I am not aware**…" skip to question **C3**, otherwise continue to question **C2**

C2. Please indicate how often do you use each of the following review/rating systems?

eBay:	Amazon:	TripAdvisor:	Other:
☐ Daily	☐ Daily	☐ Daily	☐ Daily
☐ 1 – 2 times per week	☐ 1 – 2 times per week	☐ 1 – 2 times per week	☐ 1 – 2 times per week
☐ Weekly	☐ Weekly	☐ Weekly	☐ Weekly
☐ Monthly	☐ Monthly	☐ Monthly	☐ Monthly
☐ Less than monthly	☐ Less than monthly	☐ Less than monthly	☐ Less than monthly

Below is a screenshot taken from the eBay products listing page. Some of the listed products carry a ***Top-rated Seller*** label.

Aston University

Illustration removed for copyright restrictions

For each of the following statements, **indicate your level of agreement** using the scale provided.

C3. The Top-*rated seller* label encourages me to trust the vendor more than I would otherwise.
(Tick one that applies)

☐	☐	☐	☐	☐
Strongly Disagree	Disagree	Neither Agree Nor Disagree	Agree	Strongly Agree

C4. The *Top-rated seller* label would have a strong negative effect on my intention to purchase.
(Tick one that applies)

☐	☐	☐	☐	☐
Strongly Disagree	Disagree	Neither Agree Nor Disagree	Agree	Strongly Agree

C5. I would be more likely to purchase a product from a vendor that carries the *Top-rated seller* label.
(Tick one that applies)

☐	☐	☐	☐	☐
Strongly Disagree	Disagree	Neither Agree Nor Disagree	Agree	Strongly Agree

C6. A vendor that carries the *Top-rated seller* label provides products of high quality.
(Tick one that applies)

☐	☐	☐	☐	☐
Strongly Disagree	Disagree	Neither Agree Nor Disagree	Agree	Strongly Agree

C7. A vendor that carries the *Top-rated seller* label has a good reputation.
(Tick one that applies)

☐	☐	☐	☐	☐
Strongly Disagree	Disagree	Neither Agree Nor Disagree	Agree	Strongly Agree

C8. A vendor that carries the *Top-rated seller* label is more reliable than vendors that do not carry such a label.
(Tick one that applies)

☐	☐	☐	☐	☐
Strongly Disagree	Disagree	Neither Agree Nor Disagree	Agree	Strongly Agree

C9. A vendor that carries the *Top-rated seller* label has my best interests at heart.
(Tick one that applies)

☐	☐	☐	☐	☐
Strongly Disagree	Disagree	Neither Agree Nor Disagree	Agree	Strongly Agree

Below is a screenshot taken from the TripAdvisor hotel listing page. The listed hotels have user ratings and reviews.

Aston University

Illustration removed for copyright restrictions

For each of the following statements, **indicate your level of agreement** using the scale provided.

C10. A positive review encourages me to trust a hotel more than I would otherwise.
(Tick one that applies)

☐	☐	☐	☐	☐
Strongly Disagree	Disagree	Neither Agree Nor Disagree	Agree	Strongly Agree

C11. A positive review would have a strong negative effect on my intention to book a hotel.
(Tick one that applies)

☐ Strongly Disagree ☐ Disagree ☐ Neither Agree Nor Disagree ☐ Agree ☐ Strongly Agree

C12. I would be more likely to book a hotel that has positive reviews.
(Tick one that applies)

☐ Strongly Disagree ☐ Disagree ☐ Neither Agree Nor Disagree ☐ Agree ☐ Strongly Agree

C13. The hotels that have positive reviews are of high quality.
(Tick one that applies)

☐ Strongly Disagree ☐ Disagree ☐ Neither Agree Nor Disagree ☐ Agree ☐ Strongly Agree

C14. A hotel that has good reviews has a good reputation.
(Tick one that applies)

☐ Strongly Disagree ☐ Disagree ☐ Neither Agree Nor Disagree ☐ Agree ☐ Strongly Agree

C15. A hotel that has good reviews is more reliable than hotels that have bad reviews.

☐ Strongly Disagree ☐ Disagree ☐ Neither Agree Nor Disagree ☐ Agree ☐ Strongly Agree

C16. Good reviews do not promote trust towards a hotel.
(Tick one that applies)

☐ Strongly Disagree ☐ Disagree ☐ Neither Agree Nor Disagree ☐ Agree ☐ Strongly Agree

C17. I would be less likely to book a hotel that has positive reviews.
(Tick one that applies)

☐	☐	☐	☐	☐
Strongly Disagree	Disagree	Neither Agree Nor Disagree	Agree	Strongly Agree

C18. Considering your previous experience with rating/review systems or in light of the examples provided, should a GEO Label support a similar role?

☐ Yes
☐ No
☐ Don't know

Please provide a brief explanation of your answer:

Now proceed to **Section D**

Section D – Drill-Down Functionality

We believe that a GEO Label could potentially be clickable and fulfil a drill-down function such that if the GEO Label was clicked, the user would be navigated to a GEO Label certification programme home page. The home page would provide information on the certification programme, describe the underlying criteria used when assigning the label to a dataset and provide more detailed information about the dataset itself.

E-Commerce privacy, security and business integrity seals of approval use click-to-verify functionality to confirm that the seal displayed by the vendor is genuine. Below (Figure D1) are example images of the most commonly used seals that use click-to-verify functionality.

Aston University

Illustration removed for copyright restrictions

Figure D1: Example seals with click-to-verify functionality.

D1. Have you ever come across any e-Commerce seals (including those in Figure D1) that use click-to-verify functionality?

☐ Yes
☐ No
☐ I was not aware of click-to-verify functionality

If you answered **No** or **I was not aware** skip to question **D4**, otherwise continue to

D2. Have you ever clicked on any of the approval seals to verify if they are genuine?

- ☐ Yes
- ☐ No

Please provide a brief explanation as to why you have/have not clicked on approval seals for verification:

D3. Considering your previous experience with approval seals, do you think a GEO Label should fulfil a drill-down function and have click-to-verify functionality?

- ☐ Yes
- ☐ No
- ☐ Don't know

Please provide a brief explanation to accompany your answer:

D4. In your opinion, how appropriate would it be to include an expert's judgement of the dataset and its quality as part of the GEO Label drill-down function?
(Tick one that applies)

☐	☐	☐	☐	☐	☐	☐
Extremely Inappropriate	Inappropriate	Somewhat Inappropriate	Neutral	Somewhat Appropriate	Appropriate	Extremely Appropriate

D5. In your opinion, how appropriate would it be to include information about a dataset's compliance with international standards (e.g., Dublin Core, ISO19115, etc) as part of the GEO Label drill-down function?
(Tick one that applies)

☐	☐	☐	☐	☐	☐	☐
Extremely Inappropriate	Inappropriate	Somewhat Inappropriate	Neutral	Somewhat Appropriate	Appropriate	Extremely Appropriate

D6. **In your opinion, how appropriate would it be to include community (i.e., not necessarily expert) review functionality as part of the GEO Label drill-down function?**
(Tick one that applies)

☐	☐	☐	☐	☐	☐	☐
Extremely Inappropriate	Inappropriate	Somewhat Inappropriate	Neutral	Somewhat Appropriate	Appropriate	Extremely Appropriate

D7. **In your opinion, how appropriate would it be to establish a mechanism for rating datasets and including the rating as part of the GEO Label drill-down function?**
(Tick one that applies)

☐	☐	☐	☐	☐	☐	☐
Extremely Inappropriate	Inappropriate	Somewhat Inappropriate	Neutral	Somewhat Appropriate	Appropriate	Extremely Appropriate

D8. **In your opinion, how appropriate would it be to establish a mechanism for rating dataset's <u>providers</u> and including this rating as part of the GEO Label drill-down function?**
(Tick one that applies)

☐	☐	☐	☐	☐	☐	☐
Extremely Inappropriate	Inappropriate	Somewhat Inappropriate	Neutral	Somewhat Appropriate	Appropriate	Extremely Appropriate

D9. **In your opinion, how relevant would it be to include dataset citations (e.g., a list of journal articles or other publications where the dataset has been used and quality checks have been reported) as part of the GEO Label drill-down function?**
(Tick one that applies)

☐	☐	☐	☐	☐	☐	☐
Extremely Irrelevant	Irrelevant	Somewhat Irrelevant	Neutral	Somewhat Relevant	Relevant	Extremely Relevant

D10. **In your opinion, how useful would it be to include side-by-side metadata records visualisation for comparing two or more datasets as part of the GEO Label drill-down function?**
(Tick one that applies)

☐	☐	☐	☐	☐	☐	☐
Extremely Useless	Useless	Somewhat Useless	Neutral	Somewhat Useful	Useful	Extremely Useful

D11. In your opinion, how relevant would it be to include 'soft knowledge' (subjective and informal statements) about the dataset quality that is provided by the creator or provider of the dataset as part of the GEO Label drill-down function?
(Tick one that applies)

☐	☐	☐	☐	☐	☐	☐
Extremely Irrelevant	Irrelevant	Somewhat Irrelevant	Neutral	Somewhat Relevant	Relevant	Extremely Relevant

D12. What other functionality (if any) do you think would be appropriate to include as part of the GEO Label drill-down function?

Now proceed to **Section E**

Section E – Closing Summary

E1. Please indicate your preference(s) for the role of a GEO Label
(Tick all that apply)

☐ Certification seal
☐ Drill down interrogation facility
☐ Other ______________________________
☐ Don't know

Please provide a brief explanation for your stated preference(s):

E2. In your opinion, should a GEO Label combine multiple functions (data ratings, reviews, quality assurance, etc.) and represent an all-in-one quality/trustworthiness indicator?

☐ Yes
☐ No

If **Yes**, please provide a brief explanation as to why do you think a GEO Label **SHOULD** be an all-in-one seal:

If **No**, please provide a brief description of what single function a GEO Label should adopt and explain why:

E3. Please feel free to leave any other comments or suggestions on a GEO Label or its potential role:

Thank you for taking the time to complete this questionnaire!
Please leave your contact details if you would like to be informed about the outcome of this survey or would be happy to participate in further studies associated with the GeoViQua research project.

Name: ____

E-mail: ____

☐ Yes, I would like to be informed about the results of this survey.

☐ Yes, I would be happy to participate in further studies.

Appendix C. GEO Label Leaflet

GEOVIQUA OBJECTIVES

1. Provision of innovative tools to enhance the current infrastructure capability. GeoViQua's major technical innovation is search and visualization tools for the community which communicate and exploit data quality information from GEOSS catalogues.

2. Development of the GEOLabel. GEOLabel requirements are identified (Work Package WP2), integrated with components (WP6), validated and applied into pilot cases (WP7), and disseminated to the community (WP8). It will be completed in collaboration with the GEO task ST-09-02.

3. Harmonization, exploitation and dissemination of project outputs. A careful validation process is conducted in collaboration with a number of communities of practice and standards committees to ensure that the project contributes effectively to the GCI architecture. Collaboration to AIP will be continuous.

Pilot Cases (PC)

CONTACT

www.geoviqua.org
contact@geoviqua.org

Edifici C
Campus de Bellaterra (UAB)
08193 Cerdanyola del Vallès
Barcelona, Catalonia. Spain

Phone: +34 93 5811312
Fax: +34 93 5814151

For more information about the GEO Label Questionnaire please contact:

Victoria Lush
School of Engineering and Applied Science
Aston University, Aston Triangle,
Birmingham, B4 7ET, UK

E-mail: lushv@aston.ac.uk

GEO Label Questionnaire

SEVENTH FRAMEWORK PROGRAMME (n° 265178)
THEME [ENV.2010.4.1.2-2]
Integrating new data visualisation approaches of earth Systems into GEOSS development

GeoViQua

QUAlity aware VIsualisation for the Global Earth Observation system of systems

01/02/2011 – 31/01/2014

INTRODUCTION

What is GEO Label:
At present the Science and Technology Committee (S&T) is actively discussing the topic of establishing a GEO Label to assist the user in assessing the quality of geospatial data. Thus far, however, this initiative has scarce research on which to base its agenda: hence, it remains open to proposals for establishing the GEO Label.

GeoViQua's Role:
GeoViQua is expected to highly contribute to define the concept of the GEO Label so to allow for an increasing trustworthiness over GEOSS data and services delivery. GeoViQua will conduct online surveys, focus groups and controlled structured studies to define the role that a GEO Label should serve. We will apply the defined GEO Label protocol to the pilot case studies and identify major issues and possibility for improvements.

GEO Label Aim:
• Encourage scientists, researchers, and others to contribute their data and systems to GEOSS by offering an accepted voluntary label that provides recognition that their contribution is valued by the GEO community.
• Differentiate components, data and products delivered through GEOSS and provide a "trusted brand" to GEOSS users; member governments may base their decisions on data/products of such contributions.
(Extracted from: ST-09-02 Draft GEO Label Concept document)

GEO LABEL DEVELOPMENT PHASES

Phase I :
We will conduct an online survey (GEO Label Questionnaire) that will be looking at :
• background information about the survey participants, the data sources they use or produce in their work and their awareness of any certificates or seals that apply to geospatial data;
• initial user and producer view on the role that a GEO Label should serve;
• some examples of common review and rating systems to identify the participants' opinion on such systems so that we can draw on the strengths of these when developing the GEO Label;
• some examples of commonly used seals that use click-to-verify functionality and the relevance of such functionality to the GEO Label;
• further comments and suggestions on a GEO Label or its potential role.

Phase II :
When Phase I results are analysed, we will conduct a further study presenting some GEO Label examples that will be based on our first study results. We will elicit feedback on these examples under controlled conditions and in a well-managed and structured way.

Phase III :
We will create physical prototypes which will be used in a human subject study. The most successful prototypes will then be used to define the GEO Label concept and the role that a GEO Label will serve.

GEO LABEL QUESTIONNAIRE

As part of Phase I of the GEO Label development process, we produced an online questionnaire to collect the initial GEO Label requirements and to identify the role that a GEO Label should serve. The questionnaire consists of more generic questions not to bias the users in any direction. Initial questions attempt to identify whether users believe a GEO Label is relevant to geospatial data; whether users want a single "one-for-all" label or separate labels that will serve particular roles; the function that would be most relevant for a GEO Label to carry; the functionality that users would want to transfer to a GEO Label from common rating and review systems they use.

PARTICIPANTS NEEDED!

We would be very grateful if you could spare some time to complete our GEO Label questionnaire. Your responses will provide important information that will help us in defining what role a GEO Label should fulfil.

For more information on how to complete our online questionnaire, please visit:

http://www.geoviqua.org/Eng/GeoLabel.htm

Appendix D. Phase II Study GEO Label Questionnaire

Aston University
Birmingham

GeoViQua

Designing a GEO Label:
Evaluating Prototype Label Designs

Introduction

GeoViQua is an EU research project that is working to provide the GEO data user community with innovative quality-aware visualisation and advanced geo-search capabilities. One of the objectives of our research is to contribute to defining the concept of a GEO label – that is, a label to assist users in quality assessment of geospatial datasets. This questionnaire represents the second phase of our GEO label research: having already solicited initial opinions from the user community as to the role the GEO label should serve and the information it should convey, we have consequently developed some prototype GEO label visualisations and it is on these we would like to elicit your opinion via this survey.

We would be very grateful if you could spare some time to complete this questionnaire. It should take you no more than approximately 40 minutes to complete. Your responses will be completely anonymous and confidential, and will only be used in our research. Your responses will provide important information that will help us to fully define and establish a GEO label that meets the needs of the geodata user community.

Section A – Background Information

Section A – Background Information
This section consists of a small number of questions to gather background information about you and your requirements when selecting geospatial data to use.

A1. Please identify yourself as one of the following:
(Tick one that applies)

- ☐ Primarily a data user
- ☐ Primarily a data producer
- ☐ Equally a data user and data producer

A2. Please select the description(s) that most accurately describe you:
(Tick all that apply)

- ☐ Group on Earth Observations committee member
- ☐ Private sector data user OR data producer
- ☐ Governmental data user OR data producer
- ☐ Research data user OR data producer
- ☐ Academic data user OR data producer
- ☐ Other: please specify ________________________________

A3. What type of organization do you work for?
(Tick one that apply)

- ☐ Private company
- ☐ Public corporation
- ☐ Federal ministry or governmental organization
- ☐ Provincial ministry or governmental organization
- ☐ Municipal governmental organization
- ☐ Academic institution
- ☐ Non-for-profit organization
- ☐ Other: please specify ________________________________

A4. In the context of your current and previous position(s), for how long have you been working with geospatial data or maps?
(Tick one that apply)

- ☐ Less than 2 years
- ☐ 2 to 9 years
- ☐ 10 to 19 years
- ☐ More than 20 years

A5. In the context of your current position, what approximate percentage (between 0 and 100) of your time is spent working directly with geospatial data?
(Please enter a number between 0 and 100)

☐ %

A6. In your typical work is there a choice of datasets you can use?

- ☐ Yes
- ☐ No
- ☐ Not applicable

If you answered **No** or **Not applicable** skip to **Section B**, otherwise continue to question **A7**

A7. Do you use any data portals or clearinghouses[1] for selecting datasets to use in your work?

[1] By ‘data portals and clearinghouses’ we mean large data repositories that collect, store and make available geospatial data and metadata, e.g. GEOSS clearinghouse, Geo.Data.gov, etc.

☐ Yes
☐ No
☐ Not sure

A8. If you answered Yes to A7, please list the data portals and clearinghouses that you use.

A9. In general, do you find selecting datasets that fit your needs a challenging task?

☐ Yes
☐ No
☐ Not sure

Please provide a brief explanation to your answer:

Now proceed to **Section B**

Section B – Understanding GEO Label Facets

Section B – Understanding GEO Label Facets

Geospatial data quality and the GEO label:
When selecting and using geospatial data, users typically need information on its quality. *Objective* quality information is often found in formal metadata documents supplied by the dataset provider or in technical reports which describe quality checks. *Subjective* quality information is also available and can include informal reports from other users describing how they used a dataset, users' ratings of data or assessment of data relevance, recommendations for appropriate/inappropriate uses of the data, or supplementary advice from dataset providers, such as warnings about problems in specific areas. At present, most standards-compliant geospatial data has a metadata record, but other (subjective) quality information can be scattered or unavailable. For this reason, the GEO label is proposed as a representation and interrogation facility to combine objective and subjective quality information in one place.
The GEO label itself will be a graphic representation (i.e., a static image) which will be generated individually for each dataset in the GEOSS (or other data portals and clearinghouses[1]) based on the quality information that is available for that dataset. The role of the GEO label in the data selection process can be best explained using a small scenario:

Consider yourself searching for a dataset in the GEOSS (or any other data portal or clearinghouse). You enter a search query and the search engine returns a number of datasets that match your criteria. Each of these datasets is accompanied by a GEO label which visually summarises the availability of quality information for that dataset. This allows you to make a quick assessment as to whether or not the information available suits your needs in terms of making an informed dataset selection. If you decide to investigate a dataset further, you can click on its GEO label and access the actual quality information for the dataset.

[1]*By 'data portals and clearinghouses' we mean large data repositories that collect, store and make available geospatial data and metadata, e.g. GEOSS clearinghouse, Geo.Data.gov, etc.*

Aim of this section:
From the results of a previous study, we established a set of eight user information needs that dataset users and producers felt a GEO label should address. In this section we present eight GEO label facets (that is, components of the overall GEO label) designed to represent those information needs. We want to explore **your level of understanding of the GEO label facets and the information they convey** about the datasets they represent before considering their collective visualisation within a single label. We are also keen to solicit your feedback and opinion on the examples presented.

B1. **Please carefully inspect each of the icons presented below. For each icon, briefly describe what geospatial dataset informational aspects you believe the icon represents.**

____

____

____

____

____

____

____

____

In the previous question we asked you to give your initial impression of the meaning of each icon. Now we will explain the *intended* meaning of each icon and ask you to consider how intuitive YOU think they are.

Producer Profile – information about the producer of the dataset, e.g., organisation or individual who produced the dataset, their contact information, etc.

B2. **Please indicate how intuitive YOU think the Producer Profile icon is in terms of its intended meaning?**
(Tick one that applies)

☐	☐	☐	☐	☐	☐	☐
Very Unintuitive	Unintuitive	Somewhat Unintuitive	Neutral	Somewhat Intuitive	Intuitive	Very Intuitive

B3. **Please provide any comments or suggestions on the <u>Producer Profile</u> icon. For example, describe how easy/difficult it was to interpret its meaning, what you like/dislike about the icon, and provide any suggestions on how the icon can be improved, etc.**

Producer Comments – any informal comments about the dataset quality as provided by the dataset producer, e.g., any identified problems, suggested use, etc.

B4. **Please indicate how intuitive YOU think the <u>Producer Comments</u> icon is in terms of its intended meaning?**
(Tick one that applies)

☐	☐	☐	☐	☐	☐	☐
Very Unintuitive	Unintuitive	Somewhat Unintuitive	Neutral	Somewhat Intuitive	Intuitive	Very Intuitive

B5. **Please provide any comments or suggestions on the <u>Producer Comments</u> icon. For example, describe how easy/difficult it was to interpret its meaning, what you like/dislike about the icon, and provide any suggestions on how the icon can be improved, etc.**

Compliance with Standards – information about dataset's compliance with international standards, e.g., compliance with ISO 19115, Dublin Core, etc.

B6. **Please indicate how intuitive YOU think the <u>Compliance with Standards</u> icon is in terms of its intended meaning?**
(Tick one that applies)

☐	☐	☐	☐	☐	☐	☐
Very Unintuitive	Unintuitive	Somewhat Unintuitive	Neutral	Somewhat Intuitive	Intuitive	Very Intuitive

B7. **Please provide any comments or suggestions on the <u>Compliance with Standards</u> icon. For example, describe how easy/difficult it was to interpret its meaning, what you like/dislike about the icon, and provide any suggestions on how the icon can be improved, etc.**

User Feedback – feedback and comments provided by the users of the dataset, e.g., general comments on dataset quality, identified problems, suggested use for the dataset, etc.

B8. **Please indicate how intuitive YOU think the <u>User Feedback</u> icon is in terms of its intended meaning?**
(Tick one that applies)

☐	☐	☐	☐	☐	☐	☐
Very Unintuitive	Unintuitive	Somewhat Unintuitive	Neutral	Somewhat Intuitive	Intuitive	Very Intuitive

B9. **Please provide any comments or suggestions on the <u>User Feedback</u> icon. For example, describe how easy/difficult it was to interpret its meaning, what you like/dislike about the icon, and provide any suggestions on how the icon can be improved, etc.**

User Ratings – ratings of the dataset quality as provided by the users of the dataset (based on a 5-star rating system).

B10. **Please indicate how intuitive YOU think the <u>User Ratings</u> icon is in terms of its intended meaning?**
(Tick one that applies)

☐	☐	☐	☐	☐	☐	☐
Very Unintuitive	Unintuitive	Somewhat Unintuitive	Neutral	Somewhat Intuitive	Intuitive	Very Intuitive

B11. **Please provide any comments or suggestions on the <u>User Ratings</u> icon. For example, describe how easy/difficult it was to interpret its meaning, what you like/dislike about the icon, and provide any suggestions on how the icon can be improved, etc.**

Expert Review – domain experts' comments on dataset quality, e.g., results of formal quality checks, expert suggestions on the dataset applications, etc.

B12. **Please indicate how intuitive YOU think the <u>Expert Review</u> icon is in terms of its intended meaning?**
(Tick one that applies)

☐	☐	☐	☐	☐	☐	☐
Very Unintuitive	Unintuitive	Somewhat Unintuitive	Neutral	Somewhat Intuitive	Intuitive	Very Intuitive

B13. **Please provide any comments or suggestions on the <u>Expert Review</u> icon. For example, describe how easy/difficult it was to interpret its meaning, what you like/dislike about the icon, and provide any suggestions on how the icon can be improved, etc.**

Citation Information – list of citations where the dataset was used and cited, e.g., formal reports on dataset quality checks, journal articles, etc.

B14. **Please indicate how intuitive YOU think the <u>Citation Information</u> icon is in terms of its intended meaning?**
(Tick one that applies)

☐	☐	☐	☐	☐	☐	☐
Very Unintuitive	Unintuitive	Somewhat Unintuitive	Neutral	Somewhat Intuitive	Intuitive	Very Intuitive

B15. Please provide any comments or suggestions on the <u>Citation Information</u> icon. For example, describe how easy/difficult it was to interpret its meaning, what you like/dislike about the icon, and provide any suggestions on how the icon can be improved, etc.

Quality Information – formal quality measures of the dataset, e.g., uncertainty measures recorded in UncertML, errors, accuracy information, etc.

B16. Please indicate how intuitive YOU think the <u>Quality Information</u> icon is in terms of its intended meaning?
(Tick one that applies)

☐	☐	☐	☐	☐	☐	☐
Very Unintuitive	Unintuitive	Somewhat Unintuitive	Neutral	Somewhat Intuitive	Intuitive	Very Intuitive

B17. Please provide any comments or suggestions on the <u>Quality Information</u> icon. For example, describe how easy/difficult it was to interpret its meaning, what you like/dislike about the icon, and provide any suggestions on how the icon can be improved, etc.

Visualisation of information availability:
As the GEO label is intended to convey the availability of quality information for a given dataset, each informational facet can represent one of three availability states: '***available***', '***not available***', and '***available only at a higher level***'[2]. These three information availability states will be expressed through varying the appearance of the facet icons. The following set of questions is designed to explore **the intuitiveness of these icon variations and how effective they are at conveying information availability.**

[2] 'Available only at a higher level' indicates that information is not immediately available for the dataset, but is available for a parent dataset. For example, quality information may be available for the Landsat 7 data product, but not for a specific tile.

B18. Please carefully inspect the icons presented below. For each icon, briefly describe the information availability state you believe the icon represents.

____

____

____

In the previous question we asked you to give your initial impression of the meaning of the icon variations. Now we will explain the *intended* meaning of each icon variation and ask you to consider how intuitive YOU think they are.

Blue background + white icon – information is **available** for this dataset.

White background + icon outline – information is **not available** for this dataset.

White background + blue icon – information is **available only at a higher level** for this dataset.

B19. Please indicate how intuitive YOU think the icon variations are in terms of their intended meaning?
(Tick one that applies)

☐	☐	☐	☐	☐	☐	☐
Very Unintuitive	Unintuitive	Somewhat Unintuitive	Neutral	Somewhat Intuitive	Intuitive	Very Intuitive

B20. Please comment on the design of the availability icons. For example, describe how easy/difficult was it to interpret their meaning, what do you like/dislike about the icons, and any suggestions on how the icons can be improved, etc.

Now proceed to **Section C**

Section C – GEO Label Examples

Section C – GEO Label Examples
The following section is designed to explore the effectiveness of the proposed GEO label prototypes at conveying the availability of a dataset's quality information. Here we will present you with three different designs for the GEO label, two of which use the icons you have previously considered, to solicit your feedback and opinion on these designs.

GEO label – Example 1:

Below is a prototype example of the GEO label. As described in the previous section, the GEO label is intended as a representation and interrogation facility that combines several informational aspects: **dataset producer information**, **producer comments on the dataset quality, dataset's compliance with standards, user feedback**, **user ratings of the dataset**, **expert reviews**, **dataset citations**, and **dataset quality information**. As you may remember from Section B, each informational facet shows whether the information it represents is 'available', 'not available' or 'only available at a higher level' for the dataset with which it is associated.

Please carefully inspect the label and proceed to the questions below.

C1. Based on the GEO label provided, please select the statements about the dataset that you believe are TRUE.
(Tick all that apply)

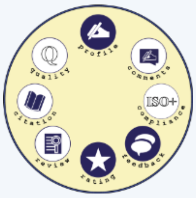
- ☐ Expert review is **available**
- ☐ User rating is **only available at a higher level**
- ☐ Information about the dataset's compliance with international standards is **available**
- ☐ Quality information is **not available**
- ☐ User feedback is **not available**
- ☐ Producer comments are **only available at a higher level**
- ☐ Citation information is **available**
- ☐ Producer profile information is **not available**
- ☐ User feedback is **available**

Please read this carefully before proceeding to the questions:
Consider a scenario where you are searching an established data portal or a clearinghouse for a dataset to use in your work. You enter your search criteria and get a list of datasets that match your query. The datasets returned are accompanied by a GEO label.

Scenario 1:
For this scenario we would like you to assume that the following information is of high importance to you and will heavily influence your dataset selection:
a) Any feedback on the dataset's previous use, such as discovered issues or suggested applications;
b) Journal publications or quality reports which refer to the dataset; and
c) Contact information of the dataset provider in case you require additional information about the dataset.

Please carefully inspect the example labels presented and proceed to the questions.




Dataset 1 **Dataset 2** **Dataset 3** **Dataset 4** **Dataset 5**

C2. **In light of the scenario presented above and based on the GEO labels provided, please rank the datasets in order in terms of their relative ability to provide the quality information you seek.**
(Enter a number from 1 to 5 for each dataset with 1 being the first dataset to inspect.)

☐ Dataset 1
☐ Dataset 2
☐ Dataset 3
☐ Dataset 4
☐ Dataset 5

C3. **Please provide a brief explanation to your ranking decision, for example explain how the informational aspects represented within the label influenced your decision.**

Scenario 2:
As in the previous scenario, you are selecting a dataset for further inspection. Here we would like you to assume that the following information is of high importance to you and will heavily influence your dataset selection:

a) Formal quality information such as uncertainty measures and dataset accuracy;
b) The dataset's compliance with international standards;
c) Experts' opinions on the dataset quality; and
d) Reports on quality checks which refer to the dataset.

Please carefully inspect the example labels presented and proceed to the questions.

Dataset 1 **Dataset 2** **Dataset 3** **Dataset 4** **Dataset 5**

C4. **In light of the scenario presented above and based on the GEO labels provided, please rank the datasets in order in terms of their relative ability to provide the quality information you seek.**
(Enter a number from 1 to 5 for each dataset with 1 being the first dataset to inspect.)

☐ Dataset 1
☐ Dataset 2
☐ Dataset 3
☐ Dataset 4
☐ Dataset 5

C5. **Please provide a brief explanation to your ranking decision, for example explain how the informational aspects represented within the label influenced your decision.**

C6. **Please indicate how difficult it was to rank the datasets in these scenarios based on the example GEO labels provided.**
(Tick one that applies)

☐	☐	☐	☐	☐	☐	☐
Very Difficult	Difficult	Somewhat Difficult	Neutral	Somewhat Easy	Easy	Very Easy

C7. In your opinion, how effective is this proposed GEO label design at conveying the availability of a dataset's quality information?
(Tick one that applies)

☐	☐	☐	☐	☐	☐	☐
Very Ineffective	Ineffective	Somewhat Ineffective	Neutral	Somewhat Effective	Effective	Very Effective

C8. Please describe which aspects of this proposed GEO label design you find most effective/ineffective in conveying the availability of a dataset's quality information.

Currently we are considering a number of different versions of the information 'available only at a higher level' icon designs. After reviewing the first proposed GEO label design, we would like your opinion on these alternatives so that we can select the most effective and intuitive version for the GEO label.

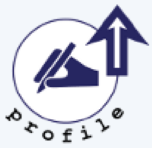
Version 1

Version 2

Version 3

Version 4

Version 5



Version 6

C9. Of the GEO label icon versions you have just viewed, please rank the versions in the order of YOUR preference from 1 to 6.
(Enter a number from 1 to 6 for each version with 1 being the most preferable.)

☐ Version 1
☐ Version 2
☐ Version 3
☐ Version 4
☐ Version 5
☐ Version 6

C10. Please provide any comments about or suggestions related to the icon versions you have just viewed.

GEO label – Example 2:

Below is a second prototype example for the GEO label. As you can see, this example is very similar to the prototype shown in example 1. It also shows availability of **dataset producer information**, **producer comments on the dataset quality, user feedback**, **expert reviews**, **dataset citations**, and **dataset quality information**. **User ratings of the dataset** in this example are presented as an average 5-star rating. The label also provides information on the **specific international standard(s)** with which the dataset complies.

Please carefully inspect the label and proceed to the questions below.

Please read this carefully before proceeding to the questions:
As in the previous example, consider a scenario where you are searching an established data portal or a clearinghouse for a dataset to use in your work. You enter your search criteria and get a list of datasets that match your query. The datasets returned are accompanied by a GEO label.
Scenario 1:
For this scenario we would like you to assume that the following information is of high importance to you and will heavily influence your dataset selection:

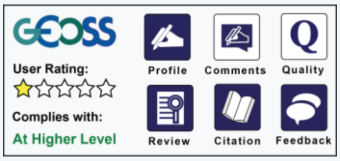
a) The dataset's average rating which should be of 3 or above;
b) Experts' opinions on the dataset quality;
c) The dataset's compliance with international standards; and
d) Journal publications or quality reports which refer to the dataset.

Please carefully inspect the example labels presented and proceed to the questions.

Dataset 1

Dataset 2

Dataset 3



Dataset 4

Dataset 5

C12. In light of the scenario presented above and based on the GEO labels provided, please rank the datasets in order in terms of their relative ability to provide the quality information you seek.
(Enter a number from 1 to 5 for each dataset with 1 being the first dataset to inspect.)

☐ Dataset 1
☐ Dataset 2
☐ Dataset 3
☐ Dataset 4
☐ Dataset 5

C13. Please provide a brief explanation to your ranking decision, for example explain how the informational aspects represented within the label influenced your decision.

C14. Please indicate how difficult it was to rank the datasets in these scenarios based on the example GEO labels provided.
(Tick one that applies)

☐	☐	☐	☐	☐	☐	☐
Very Difficult	Difficult	Somewhat Difficult	Neutral	Somewhat Easy	Easy	Very Easy

C15. In your opinion, how effective is this proposed GEO label design at conveying the availability of a dataset's quality information?
(Tick one that applies)

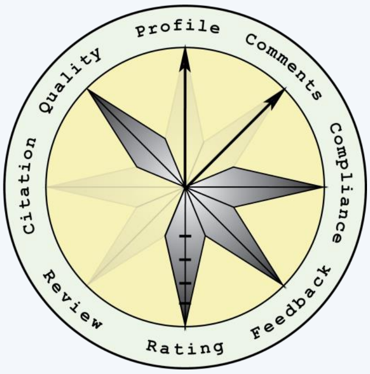
☐	☐	☐	☐	☐	☐	☐
Very Ineffective	Ineffective	Somewhat Ineffective	Neutral	Somewhat Effective	Effective	Very Effective

C16. Please describe which aspects of this proposed GEO label design you find most effective/ineffective in conveying the availability of a dataset's quality information.

GEO label – Example 3:

Below is the third prototype example of the GEO label. Unlike the first two examples which used icons to represent informational aspects, this GEO label uses eight star points to convey whether the information is 'available', 'not available' or 'only available at a higher level' for the dataset with which it is associated. As in the previous examples, the label combines eight informational aspects: **dataset producer information**, **producer comments on the dataset quality, dataset's compliance with standards, user feedback**, **user ratings of the dataset**, **expert reviews**, **dataset citations**, and **dataset quality information**.

Please carefully inspect the label and proceed to the questions below.

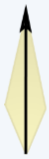


Visualisation of information availability:
As you may remember from the previous section, each informational facet can represent one of three availability states: '***available***', '***not available***', and '***available only at a higher level***'[2]. In this example design, these three information availability states are represented by varying the appearance of the arm of the star. The following set of questions is designed to explore **the intuitiveness of these star arm variations and how effective they are at conveying information availability.**

[2] 'Available only at a higher level' indicates that information is not immediately available for the dataset, but is available for a parent dataset. For example, quality information may be available for the Landsat 7 data product, but not for a specific tile.

C17. Please carefully inspect the icons presented below. For each icon, briefly describe what information availability state you believe the icon represents.

____



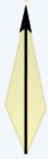
____

____

In the previous question we asked you to give your initial impression of meaning of the variations in the star arms. Now we will explain the *intended* meaning of each star arm variation and ask you to consider how intuitive YOU think they are.

Gray star arm with black outline – information is **available** for this dataset.

Transparent star arm – information is **not available** for this dataset.



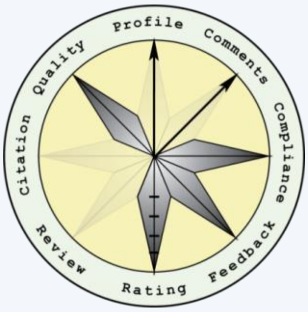
Transparent star arm with black arrow – information is **available only at a higher level** for this dataset.

C18. Please identify how intuitive do YOU think the variations in the star arms are in terms of their intended meaning?
(Tick one that applies)

☐	☐	☐	☐	☐	☐	☐
Very Unintuitive	Unintuitive	Somewhat Unintuitive	Neutral	Somewhat Intuitive	Intuitive	Very Intuitive

C19. **Please comment on the design of the availability icons. For example, describe how easy/difficult was it to interpret their meaning, what do you like/dislike about the icons, and any suggestions on how the icons can be improved, etc.**

Please again carefully inspect the label and proceed to the questions below.



C20. **Based on the GEO label provided, please select the statements about the dataset that you believe are TRUE.**
(Tick all that apply)

- ☐ Citation information is **available**
- ☐ Information about the dataset's compliance with international standards is **not available**
- ☐ Expert review is **not available**
- ☐ Producer comments are **only available at a higher level**
- ☐ User feedback is **available**
- ☐ Citation information is **only available at a higher level**
- ☐ Quality information is **only available at a higher level**
- ☐ User rating is **not available**
- ☐ Producer profile information is **only available at a higher level**
- ☐ Expert review is **available**

Please read this carefully before proceeding to the questions:
As in the previous example, consider a scenario where you are searching an established data portal or a clearinghouse for a dataset to use in your work. You enter your search criteria and get a list of datasets that match your query. The datasets returned are accompanied by a GEO label.

Scenario 1:
For this scenario we would like you to assume that the following information is of high importance to you and will heavily influence your dataset selection:

a) The dataset's average rating which should be of 4 or above;
b) Formal quality information such as uncertainty measures and dataset accuracy;
c) Information about the dataset's producer, such as who produced the data and how to contact them; and
d) Producer comments on the dataset quality.

Please carefully inspect the example labels presented and proceed to the questions.

Dataset 1 **Dataset 2** **Dataset 3** **Dataset 4** **Dataset 5**

C21. In light of the scenario presented above and based on the GEO labels provided, please rank the datasets in order in terms of their relative ability to provide the quality information you seek.
(Enter a number from 1 to 5 for each dataset with 1 being the first dataset to inspect.)

☐ Dataset 1
☐ Dataset 2
☐ Dataset 3
☐ Dataset 4
☐ Dataset 5

C22. Please provide a brief explanation to your ranking decision, for example explain how the informational aspects represented within the label influenced your decision.

C23. Please indicate how difficult it was to rank the datasets in these scenarios based on the example GEO labels provided.
(Tick one that applies)

☐	☐	☐	☐	☐	☐	☐
Very Difficult	Difficult	Somewhat Difficult	Neutral	Somewhat Easy	Easy	Very Easy

C24. In your opinion, how effective is this proposed GEO label design at conveying the availability of a dataset's quality information?
(Tick one that applies)

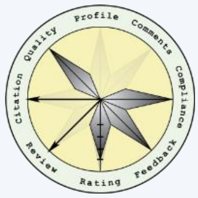
☐	☐	☐	☐	☐	☐	☐
Very Ineffective	Ineffective	Somewhat Ineffective	Neutral	Somewhat Effective	Effective	Very Effective

C25. Please describe which aspects of this proposed GEO label design you find most effective/ineffective in conveying the availability of a dataset's quality information.

Please review the three GEO label examples presented in this study.

Example 1

Example 2



Example 3

C26. After reviewing the examples presented here, please rank the GEO label examples in order of YOUR preference (with 1 being your most preferred version).

☐ GEO label – Example 1
☐ GEO label – Example 2
☐ GEO label – Example 3

Please provide a brief explanation for your answer:

Now proceed to **Section D**

Section D – GEO Label Facets

Section D – GEO Label Facets
This section consists of a small number of questions to gather your opinion on the informational aspects presented in the GEO label examples.

D1. Of the informational aspects included in the GEO labels you have viewed, please select the informational aspects that YOU think are useful in dataset selection and that YOU think should be included in the GEO label.
(Tick all that apply)

- ☐ Dataset producer information
- ☐ Producer comments
- ☐ Compliance with international standards
- ☐ User feedback
- ☐ User rating
- ☐ Expert Review
- ☐ Citation information
- ☐ Dataset quality information
- ☐ None of these

Please provide a brief explanation for your answer:

D2. Of the informational aspects included in the GEO labels you have viewed, please rank the aspects in the order of importance to YOU from 1 to 8.
(Enter a number from 1 to 8 for each informational aspect with 1 being the most important)

- ☐ Dataset producer information
- ☐ Producer comments
- ☐ Standards Compliance
- ☐ User feedback
- ☐ User rating
- ☐ Expert Review
- ☐ Citation information
- ☐ Dataset quality information

Please provide a brief explanation for your ranking:

D3. Please select any informational aspect(s) that you believe are <u>redundant</u> and should NOT be present in the GEO label.
(Tick all that apply)

- ☐ Dataset producer information
- ☐ Producer comments
- ☐ Standards Compliance
- ☐ User feedback
- ☐ User rating
- ☐ Expert Review
- ☐ Citation information
- ☐ Dataset quality information
- ☐ None of these

Please provide a brief explanation for your answer:

D4. Please provide any comments or suggestions about the informational aspects presented in the GEO label examples. List any informational aspects that you believe are missing and should be included in the GEO label.

Now proceed to **Section E**

Section E – GEO Label Branding

Section E – GEO Label Branding
This section consists of a small number of questions to gather your opinion on the use of branding in the GEO label.

Use of branding in the GEO label:
In this section we would like to explore your opinion on the use of GEO branding in the GEO label.

Please carefully inspect the labels and proceed to the questions below.

For each of the following statements, **indicate your level of agreement** using the scale provided.

E1. Presence of the GEO branding encourages me to trust the GEO label more than I would otherwise.
(Tick one that applies)

☐ Strongly Disagree ☐ Disagree ☐ Neither Agree Nor Disagree ☐ Agree ☐ Strongly Agree

E2. A GEO label that carries the GEO branding is more reliable than a label that has no branding.
(Tick one that applies)

☐ Strongly Disagree ☐ Disagree ☐ Neither Agree Nor Disagree ☐ Agree ☐ Strongly Agree

E3. I would be less likely to trust a label that has the GEO branding.
(Tick one that applies)

☐ Strongly Disagree ☐ Disagree ☐ Neither Agree Nor Disagree ☐ Agree ☐ Strongly Agree

E4. Of the GEO label examples provided above, please indicate which version you prefer:

☐ The GEO label **WITH the GEO branding**
☐ The GEO label **WITHOUT the GEO branding**
☐ Not sure

Please provide a brief explanation for your answer:

Now proceed to **Section F**

Section F – Closing Summary

Section F – Closing Summary

F1. Please feel free to leave any other comments or suggestions on a GEO label and the examples presented here:

Thank you for taking the time to complete this questionnaire!
Please leave your contact details if you would like to be informed about the outcome of this survey or would be happy to participate in further studies associated with the GeoViQua research project.
Name: ____
E-mail: ____
☐ Yes, I would like to be informed about the results of this survey.
☐ Yes, I would be happy to participate in further studies.

Appendix E. Phase II Study Respondents' Profiles

ID	Respondent's Profile
9896673	**Respondent 1** is a research data producer who works for academic institution. Respondent 1 has been working with geospatial data for more than 20 years and, in the context of the current position, he/she spends approximately 80% of his/her time working directly with geospatial data. Respondent 1 has a choice of datasets to use but he/she does not use any data portals or clearinghouses for selecting datasets. Respondent 1 finds selecting datasets that fit his/her needs a challenging task.
9921227	**Respondent 2** is a governmental data user and data producer who works for a federal ministry or a governmental organization. Respondent 2 has been working with geospatial data for more than 20 years and, in the context of the current position, he/she spends approximately 40% of his/her time working directly with geospatial data. Respondent 2 has a choice of datasets to use and he/she uses WMO WIS, WMO GTS, GEOSS and Exeter GISC data portals or clearinghouses for selecting datasets. Respondent 2 **does not** find selecting datasets that fit his/her needs a challenging task because he/she has a lot of experience and "*routinely use*[s] *(undocumented) reputation of many data producers*".
9944350	**Respondent 3** is a research data user and data producer who works for an academic institution. Respondent 3 has been working with geospatial data for about 2 to 9 years and, in the context of the current position, he/she spends approximately 50% of his/her time working directly with geospatial data. Respondent 3 has a choice of datasets to use and he/she uses NASA, ESA; GEOSS, Spanish SDI data portals or clearinghouses for selecting datasets. Respondent 3 finds selecting datasets that fit his/her needs a challenging task because of "*lack of clear metadata information, that easily describe*[s] *the data*".
9956809	**Respondent 4** is an academic data user and data producer who works for an academic institution. Respondent 4 has been working with geospatial data for more than 20 years and, in the context of the current position, he/she spends 100% of his/her time working directly with geospatial data. Selection of geospatial data is not applicable to the respondent 4.
9959670	**Respondent 5** is an academic data user who works for an academic institution. Respondent 5 has been working with geospatial data for about 2 to 9 years and, in the context of the current position, he/she spends only about 5% of his/her time working directly with geospatial data. Respondent 5 has no choice of datasets to use.
9973692	**Respondent 6** is a research data user who works for an academic institution. Respondent 6 has been working with geospatial data for about 2 to 9 years and, in the context of the current position, he/she spends approximately 20% of his/her time working directly with geospatial data. Respondent 6 has a choice of datasets to use but he/she does not use data portals or clearinghouses for selecting datasets. Respondent 6 finds selecting datasets that fit his/her needs a challenging task because he/she does not know how to use data portals and there is no clear explanation on how data portals work. Also "*at local scale data is not available*" and "*thematic areas (e.g. soil data) have specialized portals* [users] *are not aware of*".

9980194	**Respondent 7** is a research data user and data producer who works for a non-for-profit organization. Respondent 7 has been working with geospatial data for about 10 to 19 years and, in the context of the current position, he/she spends approximately 25% of his/her time working directly with geospatial data. Respondent 7 has a choice of datasets to use and he/she uses EDGAR and GENESI-DEC data portals or clearinghouses for selecting datasets. Respondent 7 finds selecting datasets that fit his/her needs a challenging task because "*it is often hard to find a suitable data set which is freely available for direct download*".
10748369	**Respondent 8** is a research data user who works for a federal ministry or a governmental organization. Respondent 8 has been working with geospatial data for about 2 to 9 years and, in the context of the current position, he/she spends approximately 30% of his/her time working directly with geospatial data. Respondent 8 has a choice of datasets to use and he/she uses INSPIRE Geoportal, RNDT, and SILVenezia data portals or clearinghouses for selecting datasets. Respondent 8 finds selecting datasets that fit his/her needs a challenging task because of "*bad metadata, not optimal discovery tools, lack of licences associated with data*".
10765670	**Respondent 9** is a private company service provider and data user who works for a private company. Respondent 9 has been working with geospatial data for about 10 to 19 years but, in the context of the current position, he/she spends only about 10% of his/her time working directly with geospatial data. Respondent 9 has a choice of datasets to use and he/she uses GEO portal, ESA, NASA, USGS, and a number of thematic portals for selecting datasets. Respondent 9 finds selecting datasets that fit his/her needs a challenging task because of "*too many choices, not good description of meta data, what the data can be used for*" and because data selection "*require*[s] *too much specialist knowledge*".
10789209	**Respondent 10** is a research and academic data user who works for an academic institution. Respondent 10 has been working with geospatial data for about 2 to 9 years and, in the context of the current position, he/she spends only about 10% of his/her time working directly with geospatial data. Selection of geospatial data is not applicable to the respondent 10.
10810835	**Respondent 11** is a research data user who works for a non-for-profit organization. Respondent 11 has been working with geospatial data for about 10 to 19 years and, in the context of the current position, he/she spends approximately 75% of his/her time working directly with geospatial data. Respondent 11 has a choice of datasets to use but he/she does not use any data portals or clearinghouses for selecting datasets. Respondent 11 **does not** find selecting datasets that fit his/her needs a challenging task because in general his/her organisation knows which datasets it needs. "*However* [his/hers organisation is] *always looking for updated versions, and of course new datasets which could be used*".

10812895	**Respondent 12** is an academic data user and data producer who works for an academic institution. Respondent 12 has been working with geospatial data for more than 20 years and, in the context of the current position, he/she spends approximately 50% of his/her time working directly with geospatial data. Respondent 12 has a choice of datasets to use and he/she uses ESPON, OSI.IE, CSO.IE, and various EU for selecting datasets. Respondent 12 finds selecting datasets that fit his/her needs a challenging task however he/she argues that this depends on a situation. "*Sometimes the problems arises from the emergence of a new dataset (it permits some questions that we've had rumbling around to be answered). Other times, the spatial resolution of a dataset (this is a problem with Irish censuses) is a challenge. You conclusions are necessarily conditioned on the data (this is a statement of the blindingly obvious, but it's not always recalled).*"
10820938	**Respondent 13** is a governmental data user and data producer who works for a federal ministry or a governmental organization. Respondent 13 has been working with geospatial data for about 10 to 19 years and, in the context of the current position, he/she spends approximately 50% of his/her time working directly with geospatial data. Respondent 13 has a choice of datasets to use but he/she does not use any data portals or clearinghouses for selecting datasets. Respondent 13 finds selecting datasets that fit his/her needs a challenging task. The "*most problematic is to find a data set that is of good quality, has a decent resolution and is consistent, i.e. there are no sudden jumps in time series*".
10829438	**Respondent 14** is a governmental, private sector, research and academic data producer who works for a non-for-profit organization. Respondent 14 has been working with geospatial data for about 2 to 9 years and, in the context of the current position, he/she spends approximately 50% of his/her time working directly with geospatial data. Respondent 14 has a choice of datasets to use and he/she uses NASA OBPG, ESA portals, GEOSS registry, and NERC portal for selecting datasets. Respondent 14 finds selecting datasets that fit his/her needs a challenging task. "*When a data source is known (e.g. by a colleague), then it's relatively easy. Finding a new data source for a new problem is often trickier. Google is normally the best way to find something!*"
10854825	**Respondent 15** is a research data producer who works for a provincial ministry or a governmental organization. Respondent 15 has been working with geospatial data for More than 20 years and, in the context of the current position, he/she spends approximately 90% of his/her time working directly with geospatial data. Respondent 15 has a choice of datasets to use and he/she uses I.Stat, FAO and EEA data portals and clearinghouses for selecting datasets. Respondent 15 finds selecting datasets that fit his/her needs a challenging task but does not provide more detail on the actual reasons.
10867181	**Respondent 16** is an academic data user who works for an academic institution. Respondent 16 has been working with geospatial data for more than 20 years and, in the context of the current position, he/she spends approximately 25% of his/her time working directly with geospatial data. Respondent 16 has no choice of datasets to use.

10892939	**Respondent 17** is an academic data user and data producer who works for an academic institution. Respondent 17 has been working with geospatial data for about 2 to 9 years and, in the context of the current position, he/she spends approximately 80% of his/her time working directly with geospatial data. Respondent 17 has a choice of datasets to use but he/she does not use any data portals or clearinghouses for selecting datasets. Respondent 17 finds selecting datasets that fit his/her needs a challenging task but does not provide more detail on the actual reasons.
10896211	**Respondent 18** is a governmental data user and data producer who works for a federal ministry or a governmental organization. Respondent 18 has been working with geospatial data for about 2 to 9 years and, in the context of the current position, he/she spends 100% of his/her time working directly with geospatial data. Respondent 18 has a choice of datasets to use but he/she does not use any data portals or clearinghouses for selecting datasets. Respondent 18 finds selecting datasets that fit his/her needs a challenging task but does not provide more detail on the actual reasons.
10901021	**Respondent 19** is a private sector and research data user and data producer who works for a non-for-profit organization. Respondent 19 has been working with geospatial data for about 10 to 19 years but, in the context of the current position, he/she spends only about 5% of his/her time working directly with geospatial data. Respondent 19 has a choice of datasets to use and he/she uses SeaZone product set, Geo.Data.gov, ArcGIS supplied datasets, and Natural England for selecting datasets. Respondent 19 is not sure whether selecting datasets that fit his/her needs is a challenging task because he/she is "…*well aware of datasets available directly in* [his/her] *area of expertise, but less sure when* [he/she] *need*[s] *something else*."
10922275	**Respondent 20** is an academic data user and data producer who works for a public corporation. Respondent 20 has been working with geospatial data for about 10 to 19 years and, in the context of the current position, he/she spends 100% of his/her time working directly with geospatial data. Respondent 20 has a choice of datasets to use but he/she does not use any data portals or clearinghouses for selecting datasets. Respondent 20 finds selecting datasets that fit his/her needs a challenging task because "[he/she] *work*[s] *with soil profiles and the major challenge is to determine whether 2 soil profiles are in reality different or identical but with different reference sources or semantics*".
10923696	**Respondent 21** is an academic data user and data producer who works for an academic institution. Respondent 21 has been working with geospatial data for more than 20 years and, in the context of the current position, he/she spends approximately 50% of his/her time working directly with geospatial data. Respondent 21 has a choice of datasets to use and he/she uses www.nationaalregister.nl, www.geocommons.com, and pdok.pleio.nl for selecting datasets to use. Respondent 21 finds selecting datasets that fit his/her needs a challenging task because, "*provided metadata is mostly insufficient to decide on fitness for use, therefore the pick involves risk of dataset's misfit for* [his/her] *task*".

10952298	**Respondent 22** is an academic data user who works for an academic institution. Respondent 22 has been working with geospatial data for about 10 to 19 years but, in the context of the current position, he/she only spends approximately 10% of his/her time working directly with geospatial data. Respondent 22 has a choice of datasets to use and he/she uses BADC archive, Streetmap, and Google for selecting datasets to use. Respondent 22 finds selecting datasets that fit his/her needs a challenging task. "*Finding the right data - fragmented nature of data on the web, hence google being the main tool! Main challenge is to compare datasets ... this is very tricky since no two datasets are that similar*".
11032412	**Respondent 23** is an academic data user and data producer who works for an academic institution. Respondent 23 has been working with geospatial data for more than 20 years but, in the context of the current position, he/she only spends approximately 10% of his/her time working directly with geospatial data. Respondent 23 has a choice of datasets to use but he/she does not use any data portals or clearinghouses for selecting datasets. Respondent 23 finds selecting datasets that fit his/her needs a challenging task. "*I develop methodologies and need data sets to test and illustrate them. I don't collect data myself so I rely on others. often it is difficult to get hold of data that are useful.*".
11138974	**Respondent 24** is an academic data user who works for an academic institution. Respondent 24 has been working with geospatial data for about 2 to 9 years and, in the context of the current position, he/she only spends approximately 20% of his/her time working directly with geospatial data. Respondent 24 has a choice of datasets to use but he/she does not use any data portals or clearinghouses for selecting datasets. Respondent 24 is not sure whether selecting datasets that fit his/her needs is a challenging task because he/she is "*mostly a technology provider so do*[es] *not use geospatial data to answer questions of* [his/her] *own*". Respondent 24 is therefore does not have "*specific scientific 'needs' about data*".
11163053	**Respondent 25** is an academic data user and data producer who works for a non-for-profit organization. Respondent 25 has been working with geospatial data for more than 20 years but, in the context of the current position, he/she only spends approximately 15% of his/her time working directly with geospatial data. Respondent 25 has a choice of datasets to use but he/she does not use any data portals or clearinghouses for selecting datasets. Respondent 25 finds selecting datasets that fit his/her needs a challenging task because, as he/she says, "*you never really know there isn't a better one!*".
11289301	**Respondent 26** is an academic data user who works for an academic institution. Respondent 26 has been working with geospatial data for about 2 to 9 years and, in the context of the current position, he/she only spends approximately 20% of his/her time working directly with geospatial data. Respondent 26 has a choice of datasets to use and he/she uses National Snow and Ice Data Centre (NSIDC) and MODIS LP DAAC for selecting datasets to use. Respondent 26 **does not** find selecting datasets that fit his/her needs a challenging task because his/her "*work is around the comparison of datasets to assess their consistency, or to compare to models*". The database that Respondent 26 uses – CEOS database – is "*an excellent resource for finding out which missions have flown at which times, delivering which parameters*". Although, "*the difficulty can come in accessing the data and finding out enough ancillary information about it, rather than selection per se*".

Appendix F. GEO Label Voting Leaflet

INTRODUCTION

What is GeoViQua:
GeoViQua is an FP7 project focused on adding rigorous quality specifications to the Global Earth Observation System of Systems (GEOSS) spatial data in order to improve reliability in scientific studies and policy decision making. One of the aims of the project is to define the concept of a **GEO label** so to allow for an increasing trustworthiness over GEOSS data and services delivery.

What is GEO Label:
The Global Earth Observation System of Systems (GEOSS) is estimated to contain more than 28 million dataset records and is constantly growing. To tackle the problem of data quality assessment and dataset selection decision making, our project – **GeoViQua** – is undertaking active research to define, develop and evaluate a GEO label. **The GEO label will visually summarise and allow interrogation of key informational aspects of geospatial dataset records** upon which users rely when selecting datasets for use.

GEO Label Studies:
To date, we have conducted 3 **user studies** to **(1)** identify the informational aspects of geospatial datasets upon which users rely when assessing dataset quality and trustworthiness, **(2)** elicit initial user views on a GEO label and its potential role and **(3)**, evaluate prototype label visualisations.

GEOVIQUA OBJECTIVES

1. **Provision of innovative tools to enhance the current infrastructure capability.** GeoViQua's major technical innovation is search and visualization tools for the community which communicate and exploit data quality information from GEOSS catalogues.

2. **Development of the GEO label.** GEO label requirements are identified (Work Package WP2), integrated with components (WP6), validated and applied into pilot cases (WP7), and disseminated to the community (WP8). It will be completed in collaboration with the GEO task ST-09-02.

3. **Harmonization, exploitation and dissemination of project outputs.** A careful validation process is conducted in collaboration with a number of communities of practice and standards committees to ensure that the project contributes effectively to the GCI architecture. Collaboration to AIP will be continuous.

CONTACT

www.geoviqua.org
contact@geoviqua.org

Edifici C
Campus de Bellaterra (UAB)
08193 Cerdanyola del Vallès
Barcelona, Catalonia. Spain
Phone: +34 93 5811312
Fax: +34 93 5814151

For more information about the GEO label, please contact:

Victoria Lush
School of Engineering and Applied Science
Aston University, Aston Triangle,
Birmingham, B4 7ET, UK
E-mail: lushv@aston.ac.uk

GEOVIQUA

Please use this leaflet to VOTE for the GEO label representation.

OR

QUAlity aware VIsualisation for the Global Earth Observation system of systems
01/02/2011 – 31/01/2014

GEO LABEL FACETS

Producer Profile facet conveys availability of information about the producer of the dataset.

Producer Comments facet conveys availability of informal comments about the dataset as provided by dataset producer.

Quality Information facet conveys availability of formal quality measures of the dataset, e.g., uncertainty measures.

Compliance with Standards facet conveys availability of information about dataset's compliance with international standards.

User Feedback facet conveys availability of feedback and comments provided by the users of the dataset.

User Ratings facet conveys availability of ratings of the dataset quality as provided by the users of the dataset.

Expert Reviews facet conveys availability of domain experts' comments on dataset quality, e.g., results of formal quality checks.

Citations Information facet conveys availability of citations where the dataset was used and cited.

Each informational facet can visually represent one of three availability states: **'available'**, **'not available'**, and **'available only at a higher level'** (information is available for a parent dataset):

Available | Not Available | Available at a Higher Level

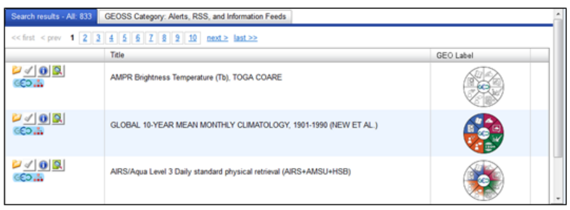
PLEASE VOTE FOR THE GEO LABEL REPRESENTATION

Our studies results indicated that the GEO label representation should resemble one of the following:

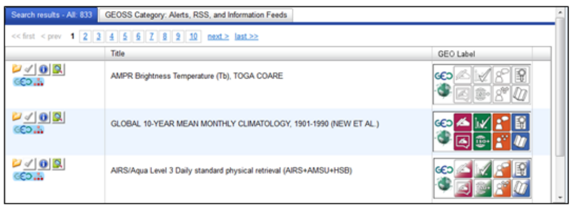
PLEASE VOTE

Below are mock-up GEO portal pages that incorporate two different GEO label designs. To arrive at a final GEO label representation, **we are asking you to vote for the representation, that you think is most effective at conveying information availability.**

Tick ONE that applies:

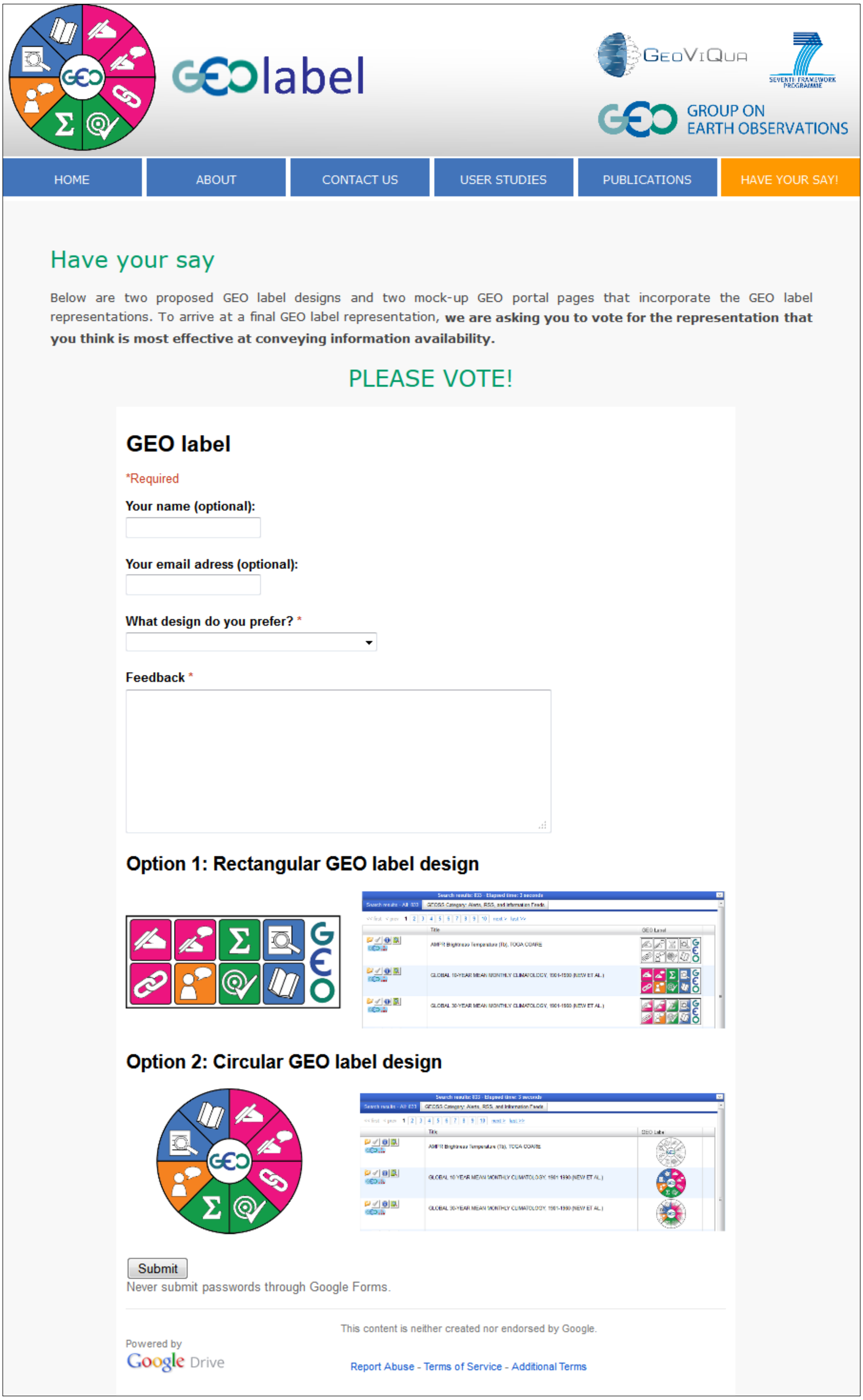


☐ **Circular Design**



☐ **Rectangular Design**

Appendix G. GEO Label Online Voting Form



GEOlabel

GeoViQua

SEVENTH FRAMEWORK PROGRAMME

GROUP ON EARTH OBSERVATIONS

HOME | ABOUT | CONTACT US | USER STUDIES | PUBLICATIONS | HAVE YOUR SAY!

Have your say

Below are two proposed GEO label designs and two mock-up GEO portal pages that incorporate the GEO label representations. To arrive at a final GEO label representation, **we are asking you to vote for the representation that you think is most effective at conveying information availability.**

PLEASE VOTE!

GEO label

*Required

Your name (optional):

Your email adress (optional):

What design do you prefer? *

Feedback *

Option 1: Rectangular GEO label design

Option 2: Circular GEO label design

Submit

Never submit passwords through Google Forms.

This content is neither created nor endorsed by Google.

Powered by Google Drive

Report Abuse - Terms of Service - Additional Terms

Appendix H. Phase III Study Materials

H.1. Participant Information Sheet

Investigating effectiveness of the GEO label in supporting geospatial datasets intercomparison and selection.

You are being invited to take part in a research study that will be conducted at Aston University or, if more convenient for you, directly at your work place. Before you decide whether or not to participate, it is important for you to understand why the research is being done and what your participation will involve. Please take time to read the following information carefully and discuss it with others if you wish. Ask us if there is anything that is not clear or if you would like more information. Take time to decide whether or not you wish to take part.

What is the purpose of the project?
With increased geospatial dataset production and increasing use of such datasets intercomparison of dataset quality and the evaluation of a dataset's fitness for use can present a major challenge for geospatial data users. Our European Union research project – GeoViQua – is investigating innovative quality-aware visualisation and dataset discovery tools. As part of our research on visualisation of quality information about geospatial datasets, we have developed a GEO label – a label that visually summarises the availability of and allows interrogation of key informational aspects of geospatial datasets upon which users rely when selecting datasets for use. We have also developed a prototype dataset selection/decision support system which utilises the GEO label and allows dataset filtering based on informational aspects' availability recorded in eight GEO label facets to support the selection process. Consequently, this research study aims to gather user feedback on both the GEO label and the prototype comparison tool that incorporates it.

Why have I been chosen?
You have been asked to take part in this research because you use geospatial datasets in your work and have knowledge of using geospatial data discovery tools.

Do I have to take part?
You are under no obligation to take part in this study. Your participation is entirely voluntary. If you do decide to take part in the study you can still stop and withdraw at any point without having to provide an explanation and without penalty.

What will I have to do if I decide to take part?
If you decide to take part in the study, you will first be asked to complete a very short questionnaire to determine your current use of geospatial datasets and your processes for selection of such datasets. You will then be asked to use our prototype dataset intercomparison decision support system to complete a series of prescribed tasks and then to provide feedback on its usability and effectiveness. You will be given a set of three dataset selection exercises and will be required to use the system to select a dataset that fits the specified criteria according to each scenario. Your interaction with the system will be audio-video recorded and you will be asked to think-aloud as you use the system. Recording your interaction with the system in this way supports detailed post-session analysis of the system's usability. The investigator will be able to answer any questions that you may have while carrying out the exercises. On completing the set of exercises, you will be asked to complete a short questionnaire about your experience of using the system.

The study will approximately take 60 – 90 minutes to complete.

What are the possible disadvantages and risks of taking part?

There are no risks associated with the study procedures and it is not expected that you will experience any disadvantages as a result of participating in the study.

What are the possible benefits of taking part?

Your participation in this study will provide important data that will help us to evaluate and validate the proposed GEO label. The study results will also be used to inform future development of better geospatial dataset intercomparison tools that meet the needs of the geodata user community.

Will my taking part in this project be kept confidential?

Your anonymity and any personal information will be protected. All data collected in the study will be stored in a secure area and on a password-protected computer. The collected data will only be accessible to the investigators.

What will happen to the results of the research project and how will participant anonymity be protected?

The audio-video recording will be analysed and transcribed into an anonymous record of your interaction with the system. Aggregated and anonymised findings will be published as part of a GeoViQua project deliverable, a Ph.D. thesis and related academic publications. You will not be identifiable from any use of the data in such publications. You will be given the opportunity to give the researcher your contact details if you would like to be sent a copy of the study findings.

Who is organising and funding the research?

Miss Victoria Lush, a Ph.D. student from the School of Engineering and Applied Sciences at Aston University is organising and conducting the research. The research is supervised by Drs Jo Lumsden and Lucy Bastin from the School of Engineering and Applied Sciences at Aston University. The research is funded by the GeoViQua research project (see www.geoviqua.org).

Who has reviewed the project?

The study has been reviewed and approved by Aston University's School of Engineering and Applied Sciences Ethics Committee.

Who do I contact if something goes wrong or I need further information?

If you have any questions, concerns or would like further information about the study please feel free to contact the researcher, Victoria Lush, at lushv@aston.ac.uk. You may also contact Victoria's supervisor (Dr Jo Lumsden at j.lumsden@aston.ac.uk).

Who do I contact if I wish to make a complaint about the way in which the research is conducted?

If you have any concerns about the way in which the study has been conducted you should contact the Secretary of Aston University Research Ethics Committee, John Walter, on j.g.walter@aston.ac.uk or telephone 0121 204 4665.

H.2. Participant Consent Form

RESEARCH PARTICIPANT CONSENT FORM

NAME OF PARTICIPANT ________________________ **DATE OF BIRTH:** DD / MM / YY

Study Number:

Title of Project: **Investigating effectiveness of the GEO label in supporting geospatial datasets intercomparison and selection.**

Project investigators: Victoria Lush

		Tick Box
1.	I confirm that I have read and understood the information sheet for the above study. I have had the opportunity to consider the information, ask questions and have had any questions answered satisfactorily.	
2.	I understand that my participation is voluntary and that I am free to withdraw at any time without giving any reason, without my legal rights being affected.	
3.	I agree to my interaction with the software being digitally audio-video recorded and later transcribed. I understand that such transcriptions will be anonymised such that I cannot be identified from the record.	
4.	I understand that data collected from me during this study will be available only to the study team, including the investigator and her supervisors.	
5.	I understand that data collected from me may be published in aggregated and/or anonymised form but that any publication will not contain any personal information that could identify me.	
6	I agree to take part in the above study.	

Name of Participant	Date	Signature
Name of Person taking consent (If different from researcher)	Date	Signature
Investigator	Date	Signature

STUDY OUTCOMES

Please leave your contact details if you would like to be informed about the outcome of this study or would be happy to participate in further studies associated with the GeoViQua research project.

Name: ____

E-mail: ____

☐ Yes, I would like to be informed about the results of this survey.
☐ Yes, I would be happy to participate in further studies.

H.3. Participant Study Materials

Investigating effectiveness of the GEO label in supporting geospatial datasets intercomparison and selection

Introduction

GeoViQua is an EU research project that is working to provide the GEO data user community with innovative quality-aware visualisation and advanced geo-search capabilities. One of the objectives of our research is to contribute to defining the concept of a GEO label – that is, a label to assist users in quality assessment of geospatial datasets. This study represents the third phase of our GEO label research. Having already solicited opinions from the user community as to the role the GEO label should serve and the information it should convey, we have consequently developed a prototype GEO label-based dataset discovery system for datasets intercomparison and selection. This study is intended to elicit your opinion on the usability and effectiveness of the GEO label and the prototype dataset discovery tool that incorporates the label to support fit-for-purpose dataset selection.

As part of this study, you will first be asked to complete a very short questionnaire to determine your current use of geospatial datasets and your processes for selection of such datasets. You will then be asked to use our prototype dataset intercomparison decision support system to complete a series of prescribed tasks and then to provide feedback on its usability and effectiveness. Your interaction with the system will be audio-video recorded and you will be asked to think-aloud as you use the system. The video camera will be focused on the screen of the computer, not on you. Recording your interaction with the system in this way supports detailed post-session analysis of the system's usability. The investigator will be able to answer any questions that you may have while carrying out the exercises. On completing the set of exercises, you will be asked to complete a short questionnaire about your experience of using the system.
The study should take you no more than approximately 90 minutes to complete. Your responses will remain anonymous and confidential, and will only be used in our research. Your responses will provide important information that will help us to fully define and establish a GEO label that meets the needs of the geodata user community.

Section A – Background Information

Section A – Background Information

This section consists of a small number of questions to gather background information about you and your requirements when selecting geospatial data to use.

A1. Please identify yourself as one of the following:
(Tick one that applies)

☐ Primarily a data user
☐ Primarily a data producer
☐ Equally a data user and data producer

A2. Please select the description(s) that most accurately describe you:
(Tick all that apply)

☐ Group on Earth Observations committee member
☐ Private sector data user OR data producer
☐ Governmental data user OR data producer
☐ Research data user OR data producer
☐ Academic data user OR data producer
☐ Other: please specify ____

A3. What type of organization do you work for?
(Tick one that apply)

☐ Private company
☐ Public corporation
☐ Federal ministry or governmental organization
☐ Provincial ministry or governmental organization
☐ Municipal governmental organization
☐ Academic institution
☐ Non-for-profit organization
☐ Other: please specify ____

A4. In the context of your current and previous position(s), for how long have you been working with geospatial data or maps?
(Tick one that apply)

☐ Less than 2 years
☐ 2 to 9 years
☐ 10 to 19 years
☐ More than 20 years

A5. In the context of your current position, what approximate percentage (between 0 and 100) of your time is spent working directly with geospatial data?
(Please enter a number between 0 and 100)

☐ %

A6. In your typical work is there a choice of datasets you can use?

☐ Yes
☐ No
☐ Not applicable

If you answered **No** or **Not applicable** to A6 skip to **Section B**, otherwise continue

A7. Do you use any data portals or clearinghouses[1] for selecting datasets to use in your work?

[1] By 'data portals and clearinghouses' we mean large data repositories that collect, store and make available geospatial data and metadata, e.g. GEOSS clearinghouse, Geo.Data.gov, etc.

☐ Yes
☐ No
☐ Not sure

A8. If you answered Yes to A7, please list the data portals and clearinghouses that you use.

A9. In general, do you find selecting datasets that fit your needs a challenging task?

☐ Yes
☐ No
☐ Not sure

Please provide a brief explanation to your answer:

Now proceed to **Section B**

Section B – Introduction to the GEO Label

Section B – Introduction to the GEO label

Aim of this section:
This section will provide a brief introduction to the GEO label and its role.

Introduction to the GEO label:
A GEO label is a graphic representation which visually summarises the availability of quality information for the dataset it represents. Producer and feedback metadata documents are being used to dynamically assess information availability and generate a label. The GEO label representation comprises 8 informational facets:

‘*Producer profile*’: this facet conveys availability of information about the producer of the dataset – e.g., organisation or individual who produced the dataset, their contact information, etc.

‘*Producer comments*’: this facet conveys availability of any informal comments about the dataset quality as provided by the dataset producer – e.g., any identified problems, suggested use, etc.

‘*Lineage information*’: this facet conveys availability of lineage/provenance information – e.g., processing applied to data and number of process steps.

‘*Compliance with standards*’: this facet conveys availability of information about a dataset’s compliance with international standards – e.g., compliance with ISO 19115, Dublin Core, etc.

‘*Quality information*’: this facet conveys availability of formal quality measures of the dataset – e.g., uncertainty measures recorded in UncertML, errors, accuracy information, etc.

‘*User feedback*’: this facet conveys availability of feedback, comments and ratings provided by the users of the dataset – e.g., general comments on dataset quality, identified problems, suggested use for the dataset, etc.

‘*Expert reviews*’: this facet conveys availability of domain experts’ comments on dataset quality – e.g., results of formal quality checks, expert suggestions on the dataset applications, etc.

‘*Citations information*’: this facet conveys availability of citations where the dataset was used and cited – e.g., formal reports on dataset quality checks, journal articles, etc.

Introduction to the GEO label continued:

As the GEO label is intended to convey the availability of quality information for a given dataset, each informational facet can represent one of three availability states: '*available*'; '*not available*'; and '*available only at a higher level*' (indicating that information is not immediately available for the dataset, but is available for a parent dataset). These three information availability states are expressed through varying the appearance of the facet icons.

Fully filled-in background + white icon – indicates that **information is available** for this dataset.

White background + icon outline – indicates that **information is not available** for this dataset.

Partially filled-in background + icon outline – indicates that **information is available only at a higher level**[1] for this dataset.

[1] *'Available only at a higher level' indicates that information is not immediately available for the dataset, but is available for a parent dataset. For example, quality information may be available for the Landsat 7 data product, but not for a specific tile.*

Based our previous user studies, we developed user-informed proposal for the GEO label representation.

User-informed circular GEO label design (information is available).

User-informed circular GEO label design (information is available at a higher level).

User-informed circular GEO label design (information is not available).

Now proceed to **Section C**

Section C – Introduction to the Dataset Discovery Tool

Section C – Introduction to the Dataset Discovery Tool

Aim of this section:
This section will provide a brief introduction to the GEO label-based dataset discovery tool.

Introduction to the GEO label-based dataset discovery tool:
The GEO label-based dataset discovery tool is a prototype online system which is designed to support geospatial dataset selection. The system utilises the GEO label and allows dataset filtering based on the informational aspects' availability recorded in eight GEO label facets.
The initial datasets discovery page (see Figure 1) provides an interface for discovering geospatial datasets through searching the metadata records available in the system catalogue. The **Query Constraints** tab is used to define the search criteria, it allows user to:

a) specify query keywords (e.g., cloud cover, precipitation, sea surface temperature, etc.);
b) specify required spatial coverage by either selecting a predefined location option or selecting a custom area using an interactive map;
c) specify required temporal coverage by selecting the start and end dates; and
d) select dataset access and use constraints.

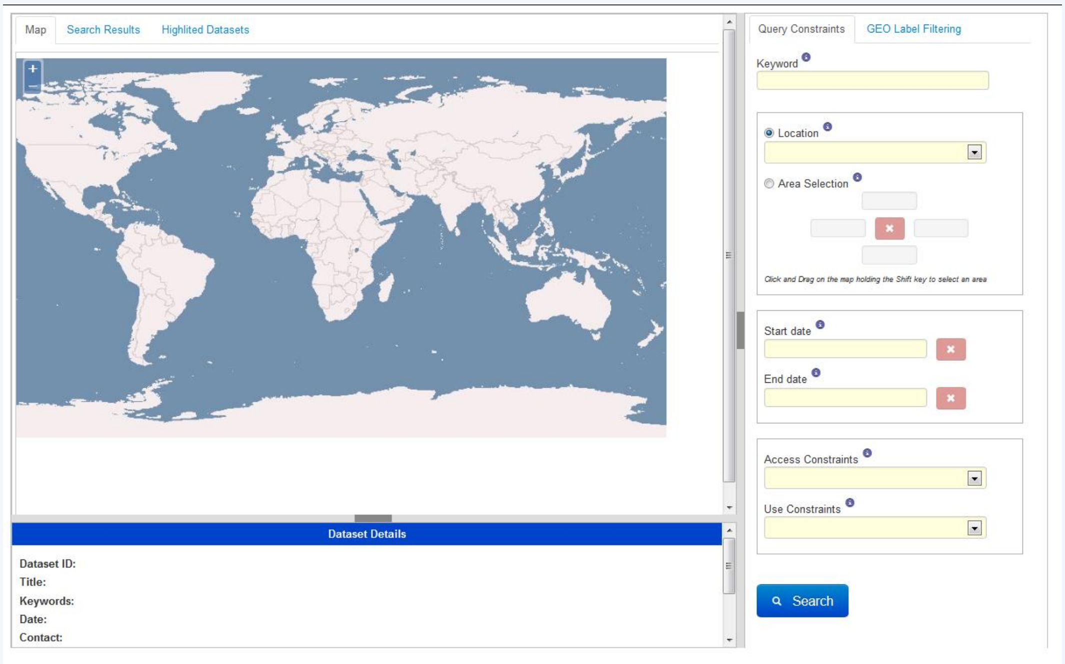


Figure 1: The GEO label-based dataset discovery tool (initial dataset discovery page).

Filtering search results:

After the search query has been submitted by clicking the **Search** button, **the system returns the GEO label representations of all the datasets that match the search criteria** (see Figure 2).

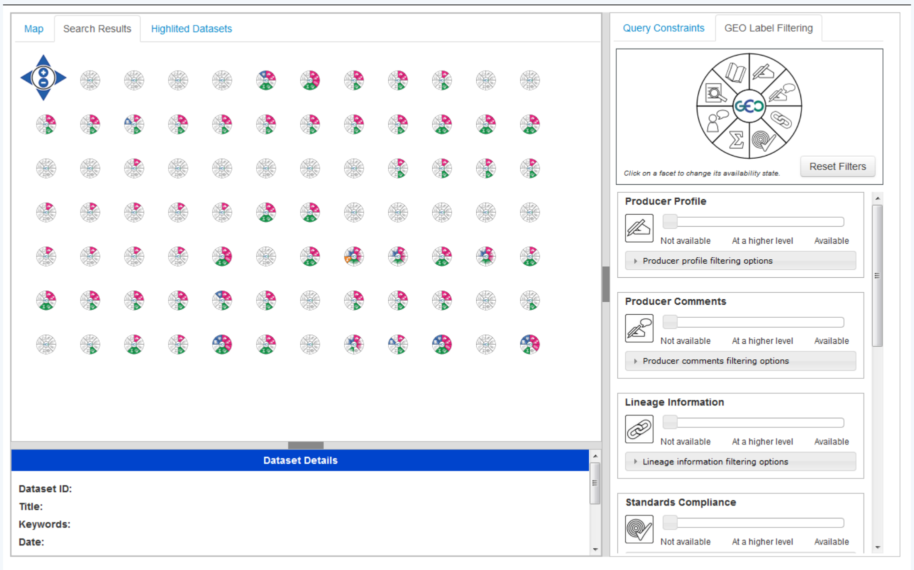


Figure 2: The GEO label-based dataset discovery tool (search results page).

With the search results being displayed, the **GEO Label Filtering** tab allows to apply information availability filtering based on the informational aspects' availability recorded in eight GEO label facets. Using an interactive clickable label (located at the top of the **GEO Label Filtering** tab) or eight facet sliders, user can select one of three availability states ('*available*', '*not available*' or '*available only at a higher level*') for each GEO label facet. For instance, a user might only be interested in datasets that have producer information immediately available (i.e., available for the dataset itself and not its parent dataset), therefore he/she would set producer profile availability to '*available'* to filter out all the datasets that do not contain this information. Figure 3 presents an example of producer profile filtering being set to '*available'* state. As can be noted from the example, when facet filtering is applied, the GEO labels that do not match the specified availability state are removed from the search results.

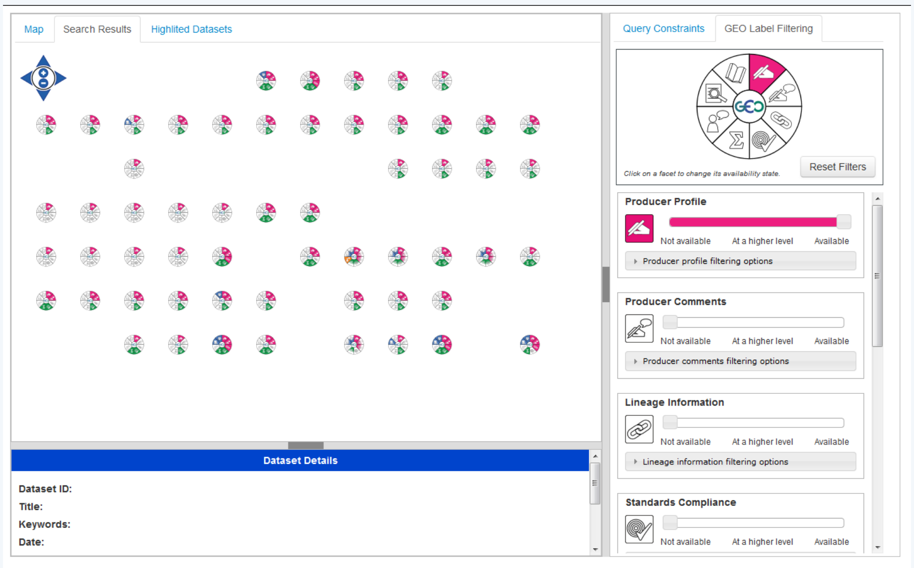


Figure 3: The GEO label-based dataset discovery tool (producer profile filtering applied).

Filtering search results continued:

The dataset discovery tool also offers additional filtering options for every GEO label informational aspect. These additional filtering options allow user to specify:

a) dataset source, i.e., name of the dataset producer;
b) producer comments' type (supplemental information, known problems, or both supplemental information and known problems);
c) maximum number of process steps that have been applied to the data;
d) name of metadata standard to which the dataset complies;
e) quality information scope (dataset or pixel level);
f) average user rating and minimum number of user feedbacks;
g) average expert rating and minimum number of expert reviews; and
h) minimum number of citations which refer to the dataset.

Unlike the information availability filtering, the additional filtering does not remove the GEO labels that do not match the specified criteria. These filtering options alter the GEO labels' size to indicate most relevant datasets. Figure 4 presents an example of additional filtering being applied. As can be noted from the example, the datasets' label representations that match the specified filtering criteria are larger in size.

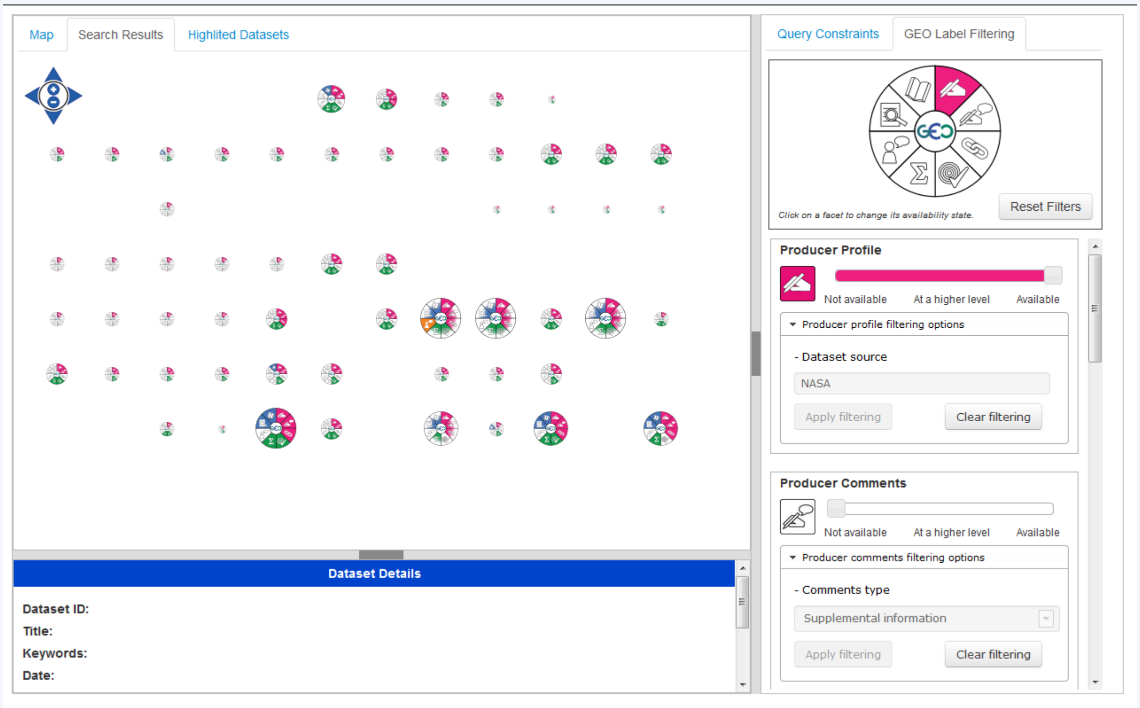


Figure 4: The GEO label-based dataset discovery tool (dataset source filtering applied).

Obtaining detailed information about a dataset:

The dataset's ID can be inspected by hovering over its GEO label representation. The detailed information about a dataset can be obtained by clicking on its GEO label representation. The selected GEO label will be highlighted and the dataset's title, abstract, producer details, link to a full metadata record, etc. will be displayed in the **Dataset Details** section of the discovery tool (see Figure 5).

Obtaining detailed information about a dataset continued:

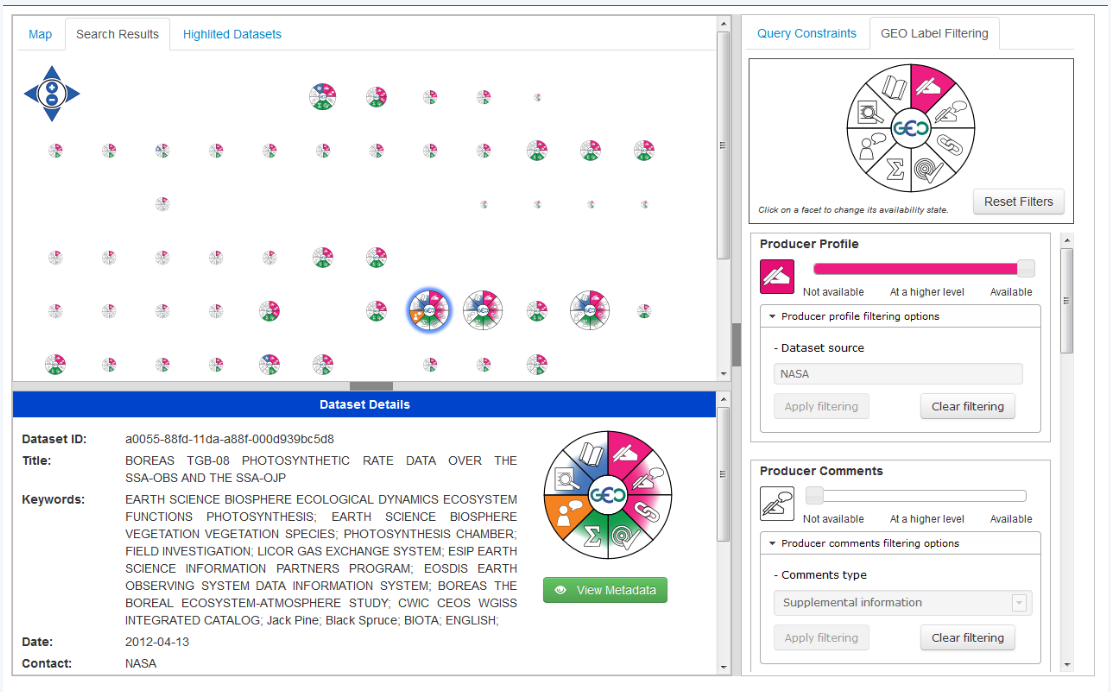


Figure 5: The GEO label-based dataset discovery tool (dataset details are displayed).

The **Dataset Details** section will also provide an enlarged GEO label representation of the dataset which can be used to obtain further details about the dataset it represents. Hovering over each facet of the enlarged dataset label will display a facet summary, e.g., name of the dataset producer, producer comments, number of process steps applied to the data, etc. The enlarged dataset label also offers drill-down functionality, i.e., when a facet is clicked, styled structured information extracted from the dataset's metadata record will be displayed in a new browser window. Figure 6 shows a user feedback summary displayed after '*user feedback*' facet was clicked.

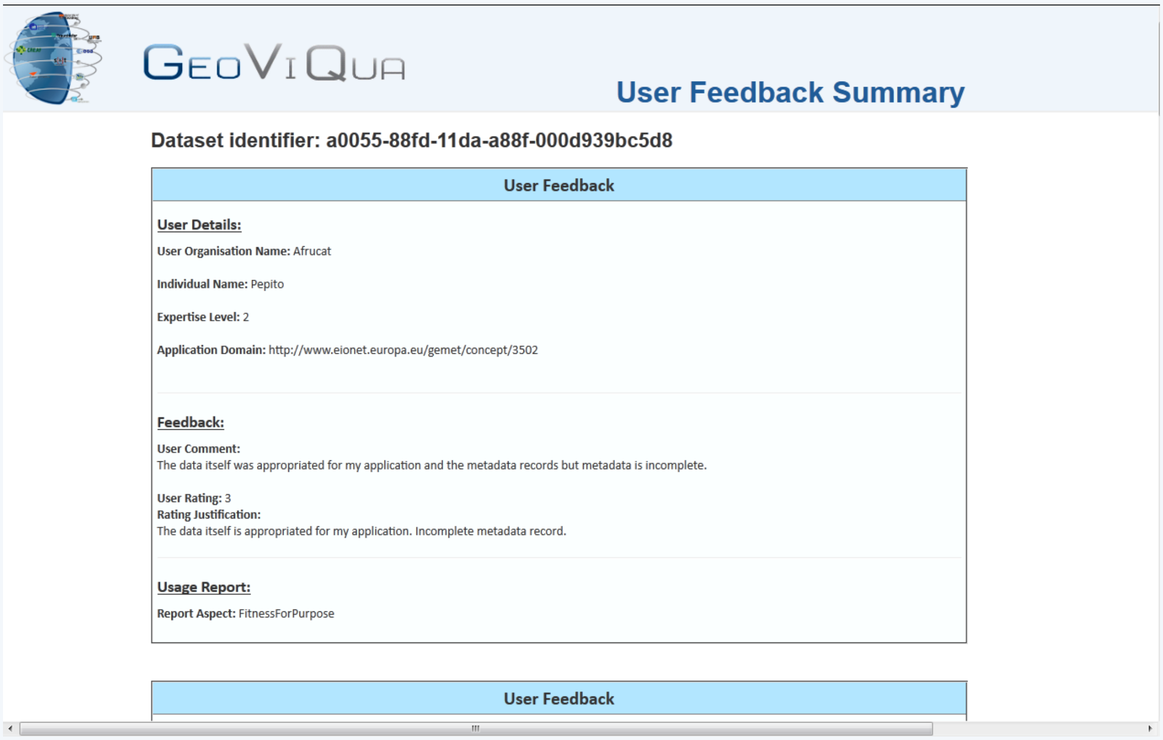


Figure 6: GEO label drill-down functionality (user feedback summary).

Highlighting the datasets of interest:

The datasets of interest can be highlighted for later reference by right-clicking on their GEO label representation and selecting the ‘Highlight’ option (see Figure 7). As can be noted from Figure 7, an outer glow is applied to the highlighted GEO labels for a visual distinction.

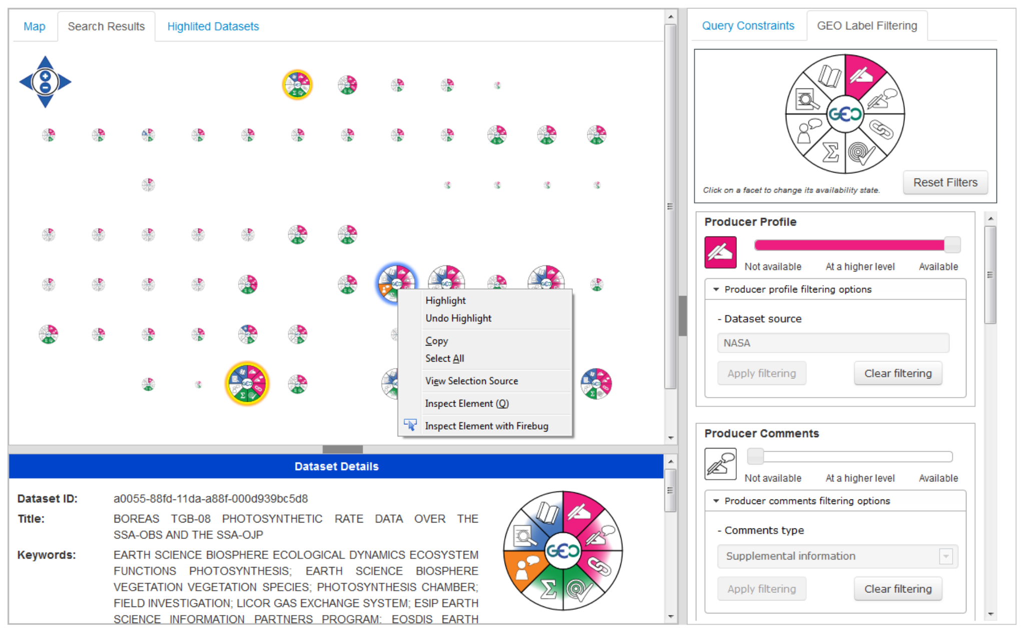


Figure 7: The GEO label-based dataset discovery tool (highlighting a dataset of interest).

When a dataset is highlighted, its detailed information is displayed in the **Highlighted Datasets** tab (see Figure 8). The highlighted datasets can be removed from the highlighted list by either clicking on a **Remove from List** button in the **Highlighted Datasets** tab or by right-clicking the dataset’s label representation in the **Search Results** tab and selecting **Undo Highlight**.

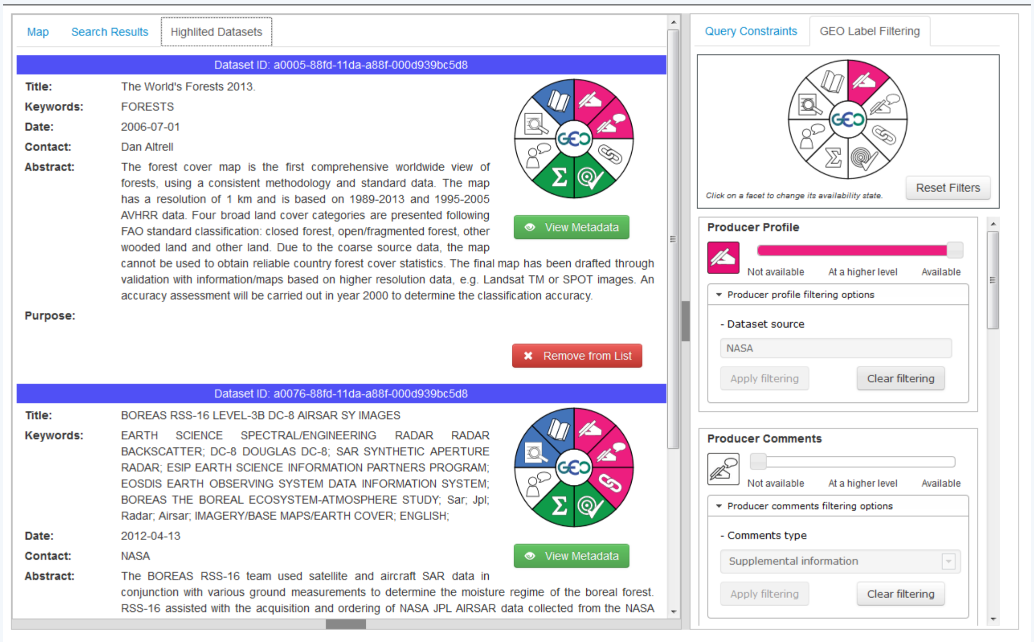


Figure 8: The GEO label-based dataset discovery tool (highlighted datasets).

Now proceed to **Section D**

Section D – Dataset Selection Exercises

Section D – Dataset Selection Exercises

In this part of the study you will be asked to use the above described prototype dataset intercomparison decision support system to complete a series of prescribed tasks. Your interaction with the system will be audio-video recorded and you will be asked to think-aloud as you use the system. The video camera will be focused on the screen of the computer, not on you. The investigator will be able to answer any questions that you may have while carrying out the exercises.

Dataset Selection – Exercise 1

Please read this carefully before proceeding to the dataset discovery tool:
Consider a scenario where you are searching our proposed dataset discovery tool for a dataset to use in your work. Your current task involves monitoring changes in global forest cover in the last two decades to identify where the changes occurred and to what degree (for instance, where were trees cut or newly planted). For this task, you require a dataset or a combination of datasets with global spatial coverage, temporal coverage between 1993 and 2013, and no access or use constraints.

We would also like you to assume that the following information is of high importance to you and will heavily influence your dataset selection:

a) Availability of contact information of the dataset provider in case you require additional information about the dataset. You are particularly interested in the datasets that are provided by the Joint Research Centre (JRC) because you have used its data in the past.
b) Availability of formal quality information such as uncertainty measures and dataset accuracy. For your current task quality information at a dataset level is acceptable.
c) Availability of information on dataset's compliance with international standards. You are specifically looking for the datasets that comply with ISO 19115 or at least FGDC standards.
d) Availability of experts' opinions on the dataset quality. If available, you would prefer the datasets that have been reviewed by at least 10 experts and received an average rating of 4 stars or higher.
e) Availability of reports on quality checks or journal publications which refer to the dataset. You would be particularly interested in the datasets that have been cited in 5 or more publications.

Now proceed to the dataset discovery tool and attempt to locate the dataset(s) that fit the above described scenario.
After completing the dataset selection exercise 1, please proceed to question **D1**.

D1. Please write down the ID(s) of the dataset(s) that you selected in the exercise 1.

D2. Please provide a brief explanation to your dataset(s) selection decision.

Please let the investigator know that you have now completed the task so that he/she can reset the system ready for the next exercise.
After the system was reset, please proceed to **Dataset Selection – Exercise 2.**

Dataset Selection – Exercise 2

Please read this carefully before proceeding to the dataset discovery tool:
As in the previous example, consider a scenario where you are searching our proposed dataset discovery tool for a dataset to use in your work. Your current task involves investigating agricultural land use in the UK to monitor any changes in fields' geometry, rotation of crops and farmers' land use through the year. For this task, you require a dataset or a combination of datasets with the UK spatial coverage and temporal coverage between May 2012 and July 2013. The datasets(s) should preferably not have any access or use restrictions.
We would also like you to assume that the following information is of high importance to you and will heavily influence your dataset selection:

a) Availability of contact information of the dataset provider in case you require additional information about the dataset. You are particularly interested in the datasets that are provided by the Food and Agriculture Organization (FAO) or NASA.
b) Availability of producer comments on dataset quality. In particular you are interested in information on known problems so that you can decide whether the dataset is fit for purpose.
c) Availability of user feedback on the dataset's previous use, such as discovered issues or suggested dataset applications. If available, you would prefer the datasets that have at least 10 user feedbacks and received an average rating of 3 stars or more.
d) Availability of lineage information to identify what processing has been applied to the data. For your current task you require pre-processed datasets, therefore you will be looking for the datasets that have at least one processing step applied.
e) Availability of any journal publications or quality reports which refer to the dataset. You would be particularly interested in the datasets that have been cited in at least 2 publications.

Now proceed to the dataset discovery tool and attempt to locate the dataset(s) that fit the above described scenario.
After completing the dataset selection exercise 2, please proceed to question **D3**.

D3. Please write down the ID(s) of the dataset(s) that you selected in the exercise 2.

D4. Please provide a brief explanation to your dataset(s) selection decision.

Please let the investigator know that you have now completed the task so that he/she can reset the system ready for the next exercise.
After the system was reset, please proceed to **Dataset Selection – Exercise 3.**

Dataset Selection – Exercise 3

Please read this carefully before proceeding to the dataset discovery tool:
As in the previous examples, consider a scenario where you are searching our proposed dataset discovery tool for a dataset to use in your work. Your current task involves identifying the effect of climate change on protected areas in South Africa in the past 5 years. For this task, you require climate datasets that contain temperature, precipitation and aridity data, have spatial coverage of South Africa region and temporal coverage between September 2008 and September 2013. For this task you would prefer the datasets that do not have any access or use restrictions.
We would also like you to assume that the following information is of high importance to you and will heavily influence your dataset selection:

a) Availability of producer comments on dataset quality. In particular you are interested in any supplemental information on suggested dataset application and use.
b) Availability of information on dataset's compliance with international standards. You are specifically looking for the datasets that comply with the Inspire or ISO 19115 standards.
c) Availability of any feedback/reviews on dataset's quality. You are equally interested in feedback from general users and reviews from the domain experts. If such feedback/reviews are available, you would be interested in the datasets that received an average user or expert rating of at least 4.5 stars.
d) Availability of lineage information to identify what processing has been applied to the data. For your current task you require raw data, therefore you will be looking for the datasets that have zero processing steps applied.
e) Availability of formal quality information such as uncertainty measures and dataset accuracy. For your current task you require quality information at a pixel level.

Now proceed to the dataset discovery tool and attempt to locate the dataset(s) that fit the above described scenario.
After completing the dataset selection exercise 3, please proceed to question **D5**.

D5. **Please write down the ID(s) of the dataset(s) that you selected in the exercise 3.**

D6. **Please provide a brief explanation to your dataset(s) selection decision.**

Now proceed to **Section E.**

Section E – GEO Label Evaluation

Section E – GEO Label Evaluation
The following section is designed to solicit your feedback and opinion on the GEO label visualisation presented to you in this study.

E1. **In your opinion, how effective is this proposed GEO label design at conveying the availability of a dataset's quality information?**
(Tick one that applies)

☐	☐	☐	☐	☐	☐	☐
Very Ineffective	Ineffective	Somewhat Ineffective	Neutral	Somewhat Effective	Effective	Very Effective

E2. **Please describe which aspects of the proposed GEO label design you find most effective/ineffective in conveying the availability of a dataset's quality information.**

E3. **Please describe any modifications or improvements that you would apply to the proposed GEO label. For instance, identify any informational aspect(s) that you believe are redundant or list any informational aspects that you believe are missing and should be included in the GEO label.**

Section F – Dataset Discovery Tool Evaluation

Section F – Dataset Discovery Tool Evaluation
The following section is designed to solicit your feedback and opinion on the dataset discovery tool presented to you in this study.

F1. Please indicate how difficult it was to complete the dataset selection exercises using the proposed GEO label-based dataset discovery tool.
(Tick one that applies)

☐	☐	☐	☐	☐	☐	☐
Very Difficult	Difficult	Somewhat Difficult	Neutral	Somewhat Easy	Easy	Very Easy

F2. What aspects of the system did you find the most challenging when comparing and selecting geospatial datasets, if any?

F3. In your opinion, how effective is the proposed GEO label-based dataset discovery tool at supporting dataset intercomparison and selection?
(Tick one that applies)

☐	☐	☐	☐	☐	☐	☐
Very Ineffective	Ineffective	Somewhat Ineffective	Neutral	Somewhat Effective	Effective	Very Effective

F4. Please describe which aspects of the proposed GEO label-based dataset discovery tool you find most effective/ineffective in supporting dataset intercomparison and selection.

F5. Please describe any modifications or improvements that you would apply to the proposed GEO label based dataset discovery tool.

Section G – Closing Summary

Section G – Closing Summary

G1. Please feel free to leave any other comments or suggestions on the GEO label and the prototype dataset discovery tool:

Thank you for taking the time to participate in our study!

Appendix I. Phase III Thematic Analysis

I.1. Dataset Discovery

Data Extracts	**Coded As**
"*The ones where you get the whole list of downloads is too busy really. I like the cleanness of this* [pointing at the Dataset Details area], *I like the limited data you've got, which is perfect*". (participant 2).	5. Likes the cleanliness of the interface 6. Prefers limited amount of information 7. Standard portals provide too many download buttons
"*So for me, these [pointing at the filtering options] give additional ways of filtering the information once you've discovered the datasets that you actually potentially could use*" (participant 3).	1. Filtering supports dataset discovery 2. Additional way of filtering once discovered the datasets 3. Filtering allows to find potentially useful datasets
"*And I really like the visual way of doing that* [filtering the results]". (participant 3).	1. Likes the visual way of filtering the results
"*Wow!* [looking at the search results] *I never find a hundred datasets. I'm lucky if I find one*" (participant 3).	1. Surprised to get hundred search results 2. Limited number of available datasets 3. Availability of datasets is an issue
"*I'd almost have to go through every single one of these to see if there is anything now at the UK resolution. I don't think you can find the datasets to answer that question with the system as it is at the minute. I don't know if there is one in here or not. But I suspect that you need a specific UK dataset for that. So I'd want to know whether it was created in the UK*" (participant 3).	1. Needs more filtering functionality 2. Needs more search functionality 3. Has to go through every label to acquire information 4. Cannot find suitable dataset with the current state of the system 5. Needs a dataset with UK resolution
"*I would expect that* [keywords search through all text]*, and I don't know whether the keyword there searches through the whole thing, I sort of assume that it searches through the whole thing. But it may only look at the keywords itself. I don't know actually. I'm assuming it's within any of the text and sort of title and the abstract But I could be wrong*" (participant 3).	1. Unsure whether the keywords are being searched in all text 2. Assumes that the keywords are being searched in whole of the text
"*I know that there are better datasets out there. So I would use those*" (participant 3).	1. Knows better datasets to use 2. Knows were to search for datasets

"*The other portal that I use a lot is the BADC one. And it's fine because it actually has lists of the datasets as well. And that BADC portal is just so unfriendly to use. It just spews back a massive list of things and it's like "oh, my god, how do I even start with this?*" (participant 3).	1. Likes lists of datasets 2. Data ports is not user friendly 3. Data portal returns too many results 4. Difficult to find suitable dataset
"*Ok, I am interested if data comes from FAO or NASA, but if it was collected by DEFRA or whatever and it did what I wanted, I'd probably be, in practice, quite happy to use it*" (participant 3).	1. Happy to use any data that fits needs 2. Happy to use data from different providers 3. Loosens up search constraints

Boolean Queries

Data Extracts	**Coded As**
"*I can't tell if I can do combined ones* [queries] *on this, so like 'FAO or NASA'. Like filtering where you can do ANDs or anything along those lines*" (participant 2).	5. Sophisticated filtering 6. Cannot tell if can define combined queries
"*It's OR! Oh, it's OR in there. Cramps! Can I AND it? Let's see. I am guessing that this is pushing* [the boundary] *but this is how I'd do it. Oh, OK, I see* [participant saw same results as for the previous query]" (participant 3).	1. Trying to apply a Boolean query 2. Dislikes OR clause 3. Wants to use AND clause 4. Would normally use Boolean queries
"*Oh, I think* [I would add] *just the ability to have these Boolean constructs to do more refined, to develop more refined queries*" (participant 3).	1. Add the ability to construct Boolean search queries 2. Wants the ability to construct refined queries 3. GEO LINC modification suggestion
"Because when I'm looking for something, let's not use Google but let's say Web of Science or bibliographic database, then I'll fairly routinely build up quite complex query strings involving, I think, a fairly standard practice of putting in quotes, phrases you want to be, you know, "land" followed by "use" not just "land AND use" or any combination. And then using AND, OR and NOT as a combination, and bracketing for precedence" (participant 3).	1. Building up a complex query strings 2. Putting in quotes to construct a complex query 3. Using AND, OR and NOT as a combination to construct a query 4. Using query bracketing for precedence
"And I fairly routinely have quite complex saved queries built to find papers in areas that I'll be specifically interested in. Because if I do a general query on "gassing process" it'll flow up ten thousand papers or something and that's no use" (participant 3).	1. Constructing a complex query string 2. Saving complex query strings 3. A general query returns thousands of results
"*And I really like the visual way of doing that* [filtering the results]. *But what I don't like is that there is not enough flexibility in the query construction. And this is one of the issues within portals but it is also a generic issue within finding data*" (participant 3).	1. Likes visual way of filtering 2. Dislikes lack of flexibility in constructing queries 3. Portals do not support flexible querying 4. Flexibility of query construction is a generic issue

"*I'd want this free text. I want this really powerful query language, well not really powerful but just standard query language that I can use*" (participant 3).	1. Wants powerful query language in the initial search
"*Oh, I wonder if I could put in, that would be nice, can you put in 'FAO or NASA'. Am I allowed logical thingies in there?* [looking at the results] *I'm thinking it doesn't like logical ORs*" (participant 4).	1. Wondering if Boolean constructs are supported 2. Trying to construct a query with an OR clause 3. Disappointed that logical ORs are not supported
"*Is there a way to say OR here? Or AND?* [pointing at the metadata standard name]" (participant 5).	1. Wondering if can use an OR clause
"*Maybe here, that's where I would like to have something like OR or AND or something, in the selection of keywords*" (participant 5).	1. Would like to be able to use logical operators 2. Wants to use OR or AND in the keywords
"*For example here, I am not clear if I separate things with commas will I get anything – a dataset that contains all of those, or one of those or several of those. How do I know?*" (participant 5).	1. Not clear about keyword functionality 2. Not sure how to define keywords query
"*I would like it to be more like an SQL query or something like that, like in programming. So you can say "this OR that OR that", or maybe parenthesis, but maybe that's too much*" (participant 5).	1. Would like to have SQL-like queries 2. Would like to use logical operators

Autocomplete Suggestions

Data Extracts	Coded As
"*And I think what you've done with keywords, the autocomplete thing, is great cause that kind of dropdown is good cause it helps you to understand what the system actually knows about. But you need a bit more flexibility*" (participant 3).	1. Likes autocomplete options 2. Dropdown menus and autocomplete help to understand what the system 'knows' about 3. Need to ensure flexibility in the autocomplete suggestions

Spatial Extent

Data Extracts	Coded As
When selecting a location, participant 1: "*I would like to be able to just draw a selection box over the UK now*". Hinted the instructions to the respondents.	4. Selecting spatial extent 5. Did not notice the instructions 6. Wants to draw a selection box on the map 7. Needs clear instructions
"*And the location I need global. Well, if I don't select the location, I assume, it's going to be global*" (participant 3).	7. Selecting global spatial extent 8. Associates empty location fields with global coverage 9. Leaves location options empty to find datasets with global coverage
"*Oh, hold the Shift key! That's a bit small for my glasses*" (participant 4).	1. Instructions text is too small 2. Did not see the instructions

Temporal Extent

Data Extracts	**Coded As**
"*I hate picking things. Yea, just lazy*" (participant 2).	1. Needs flexible date selection 2. Wants to enter the dates manually 3. Dislikes selecting options
"*I am guessing this is 'last updated' date* [pointing at the date field]*, that's always a key bit*" (participant 2).	1. Importance of last updated date 2. Last updated date is a key bit
Clicked on date field: "*Oh, that's going to be annoying*" (participant 3).	1. Annoyed with a pop-up calendar
"*I hope it does any format.* [after format did not work] *Maybe it should be more tolerant of date formats*" (participant 4).	1. Hopes that the systems supports various date formats 2. The system should be more tolerant to date formats 3. GEO LINC modification suggestion

Access and Use Constraints

Data Extracts	**Coded As**
"*No use or access constraints* [while selecting dropdown options]. *Which is very useful to have on because the stuff I ever deal with is always, particularly because it is regional Ordinance Survey stuff, usually got constraints*" (participant 2).	3. Usefulness of use constraints and access constraints query options 4. Works with data that usually has constraints 5. Regional Ordinance Survey data has constraints
"*Yeah,* [use and access constraints option are] *bloody useful! Most of the time, I want none. I want no access constraints or usage restrictions. So I would often turn these on. Actually, thinking about it, it is unlikely I'd select these to 'none' when I'm searching for a dataset. I think what I'd do is… because my prior assumption is that there isn't a dataset. So what I actually want to do is – I want to just find if there a dataset first and, I guess, if there were 20 datasets, I might come back and say "just show me those that don't have access restrictions or use restrictions"*" (participant 3).	1. Use and access constraints options are very useful 2. Would use access and use constraints options if can find sufficient number of datasets

I.2. 'At a Glance' Dataset Intercomparison

Data Extracts	Coded As
Without even looking at the scenario requirements: "*Let's start with the better looking ones*" (participant 1).	1. Perceived 'goodness' of data 2. Identifying 'better looking' labels 3. At a glance intercomparison 4. Full labels are associated with better datasets
"*I haven't initially clicked on these* [pointing at the filtering options] *just cause it was relatively obvious from here* [pointing at the search area] *which ones are the most 'colourful'*" (participant 1).	1. At a glance intercomparison 2. Most '*colourful*' labels are obvious 3. Did not apply any filtering 4. Did not need filtering to identify complete labels
From just clicking through the search area labels and looking at the enlarged labels: "*This one doesn't have the citations that I need. So does that one. That one doesn't have any expert reviews that I need*" (participant 1).	1. Identified missing information at a glance 2. Missing citations 3. Missing expert review 4. At a glance information gathering 5. Facet recognition
"*It's fantastic at showing what's provided in terms of the metadata. It's so easy to, sort of, quickly look at here even when it's loads returned*" (participant 1).	1. Tool is fantastic at showing information availability 2. Easy to identify metadata completeness even with many search results
"*Once I got my head around with the colour scheme and things like that, it does work quite quickly for visually picking up what you need to. Or, I should say, not spotting before you filter, but once you filter knowing that you are on a right track without having to go and read quite so much. And then it does filter it down enough to read the actual comments and the actual, the longer text and that kind*" (participant 2).	1. Learning effect 2. Learning colour scheme 3. Enables visual filtering 4. Minimises unnecessary reading 5. Keeps user on a right track 6. Filters results down to read longer text
"*I've kind of picked that one first cause it's got most of the colours in. I don't know why* [laughing]. *But I figured that the fully-coloured one has more information about it [dataset], so that's the theory*" (participant 2).	1. Influenced by label colours 2. Associates colours with metadata completeness 3. Selected a fully-coloured label
"*OK, so there is a lot of pink! Or red. Which is the producer stuff*" (participant 2).	1. Facet recognition and recall 2. Associates pink/red colours with producer information 3. Recognised producer-related facets

"*Right, I've got myself 6 datasets.* [selected the most complete label] *I'll have this one straight away, I know that, because it's from somebody from the JRC and I'm working with the JRC*" (participant 3).	1. Selected the most complete label 2. Trusts familiar data producer 3. Trusts data producer because working with them
"*Well, it looks like it should be that one [dataset] is coming up the biggest, so that's what I'm going to put in [select]*" (participant 4).	1. Selected the dataset represented by the biggest label 2. Associates label size with fitness for purpose
"*This is pretty snazzy!* [laughing]" (participant 4).	1. Snazzy display 2. Feeling happy about the tool
As 130 results appeared on the screen, the participant was able to identify the most complete GEO labels: "*Given it's got quite a few. Oh, that one is looking quite good, and that one, and that one* [pointing at the labels]*. Those three got loads of information so it could be all those three are the candidates*" (participant 4).	1. Associates label completeness with metadata completeness 2. Interested in complete labels 3. Complete labels are perceived as potential candidates
"*So let's have a look at each of these. That one is quite small, so let's choose the bigger ones first*" (participant 4).	1. Influenced by label size 2. Interested in larger labels 3. Wants to inspect larger labels first
"*OK, none of them are complete circle of information but let's filter it down a bit*" (participant 4).	1. Associates label completeness with information completeness 2. Sees label incompleteness at a glance 3. Wants to apply filtering
"*It appears bigger so it seems more acceptable for my, for the filters I've applied*" (participant 4).	1. Associates bigger labels with fitness for purpose 2. Bigger labels are more acceptable
"*Yellow ones are the ones that I highlighted*" (participant 5).	1. Recognised highlighted labels 2. Associates yellow colour with highlighting
"*I want this one because it also has expert reviews*" (participant 5).	1. Facet recognition 2. Interested in expert reviews
"*And the explanation is, after applying filtering in the exercise I went to the results with the biggest label and the second biggest label, and I highlighted all those, and compared user commentaries, and I just picked the ones with the similar users to me that were focusing on the studies that I needed and commented that it was good. So I trust them*" (participant 5).	1. Selected biggest labels 2. Used highlighting functionality 3. Compared user comments 4. Selected datasets with comments from similar users 5. Influenced by user comments

I.3. Starfield Display

Data Extracts	Coded As
"*It is handy that it* [the tool] *keeps them* [the labels] *in the same place. So that I know that this* [label] *was the one that I clicked on in the first place*" (participant 1).	1. Likes fixed location of the labels 2. Fixed location of the labels is useful 3. Knows previously visited labels 4. Can recall labels by their location
"*So the one I clicked on first ticks all the boxes*" (participant 1).	1. Recalled previously inspected labels 2. Visual recognition 3. Recalling label's position
"*Perhaps, in terms of the visualisation side of things, grouping by date or something like that. So you've got all your circles all over the place, but straight away if they were box-grouped into date, whether that's date created or date updated, that might be the first thing as well. That's just a personal preference*" (participant 2).	1. Labels need grouping 2. Grouping labels by date
"*Visually I kind of remember where they are on the screen*" (participant 2).	1. Visually remembers labels' position 2. Recalling labels' position
"*If there was another option kind of saying "show me all the ones that this person rated or used", a way of grouping by users or studies, because I've spotted a pattern amongst these, same couple of people doing it for their, particularly this German, study. So you could have it as another way of looking at things. And that would definitely affirm or confirm that, what I was looking after, that was the right stuff I needed to and that I haven't missed out, or everyone missed out. One of the two, you never know!*" (participant 2).	1. Needs an option to group datasets by users who have previously used the data 2. Needs an option to filter datasets by users who have previously used the data 3. User feedback affirms dataset selection 4. Identified a pattern in data 5. Identified a group of users conducting same study
"*I think that graphical display here is brilliant. I really, really like it. And there is a way of navigating datasets. It's much, much nicer that sort of page-based. I really like being able to split the page up and just move through and view each one in turn and this hover thing*" (participant 3).	1. Likes graphical display 2. Likes ability to navigate the datasets 3. Starfield display is nicer than page-based display 4. Likes dataset details functionality 5. Likes hover-over functionality
"*One thing I was wondering if that would be possible to regroup them*" (participant 3).	1. Wants functionality to regroup dataset labels

I.4. Side-by-side Metadata Comparison

Data Extracts	Coded As
"*So what I've just noticed, there are two datasets that seem almost identical, except one has user and expert comments associated with it. So if I was allowed to get this one and the one that I spotted at the bottom next to each other, so I could spot what exactly was the difference*" (Participant 1).	1. Identified almost identical datasets 2. Wants to sort highlighted list 3. Wants to compare dataset descriptions next to each other 4. Wants to bring similar datasets together
"*I have now selected too many options here* [laughing]*, so it would be useful if I could shuffle these up and change the position almost just so that I could see the two that I really want side by side*" (participant 1).	1. Selected too many datasets 2. Wants to shuffle the datasets in the highlighted list 3. Wants to change datasets' position in the list 4. Wants to compare datasets side by side 5. Customise highlighted list
Looking at the Highlighted Datasets tab: "*Oh, look at that! I can compare. That's what I was hoping.* [laughs]" (participant 2).	1. Comparing datasets in the highlighted list 2. Happy to be able to compare datasets
"It does allow good comparison side by side" (participant 2).	1. Tool allows good side by side dataset comparison
"*Oh, I didn't even know you can do this, to be honest. I like it. I think it could be useful for if I'm trying to build up a set of plausible datasets and just add them to my dataset set*" (participant 3).	1. Did not know about the highlighting functionality 2. Would use highlighting functionality to build a custom list
"*So it's now handy that I do have the 'highlighted' tab so that I can now go through each of the ones I flagged as being potentially useful dataset and just quickly compare them on the available metadata that they have*" (participant 1).	1. Compare datasets in the 'highlighted' tab 2. Use highlighting functionality to flag potentially useful datasets 3. Highlighted list allows quick comparison
After applying all the filters and looking at two datasets: "*So this is where I could do with this highlighting, I suppose, to remember which ones I looked at*" (participant 2).	1. Use highlighting to remember visited datasets
"*The highlighting with colours, I think, is really effective, it shows up very well*" (participant 6).	1. Highlighting with colours is very effective 2. Highlighting is visible

I.5. Search Results Filtering

Data Extracts	Coded As
"*Is there a way of, let's say we are looking for NASA or Food and Agricultural Organisation, can I add NASA as a search term in a keyword search?*" (participant 1).	1. Wants to apply producer filtering 2. Does not know how to apply filtering 3. Needs further instructions 4. Misusing dataset search functionality
"*What I would really want to do is to reapply these again* [pointing at the filtering options]. *I still want these constraints but I just wanted to change a keyword really. And I've now forgotten which ones I wanted*" (participant 1).	1. Wants to save filtering options 2. Wants to reapply filtering 3. Forgot filtering criteria
"*I like the sliders and the additional filtering. The cons and pros of the filtering, I am sure you thought through with the filtering on the whole is – it can become a bit restrictive. But it's also good for guidance*" (participant 2).	1. Filtering options as guidance 2. Worried that filtering can be restrictive 3. Likes sliders 4. Likes additional filtering
"*I can't tell if I can do combined ones* [queries] *on this, so like 'FAO or NASA'. Like filtering where you can do ANDs or anything along those lines*" (participant 1).	1. Wants sophisticated filtering 2. Wants to use logical operators in query 3. Uncertain about filtering functionality
After applying the user feedback and expert review filters, only 7 datasets were left on the screen: "*So how that narrows it straight away*" (participant 2).	1. Filtering narrows down search results
"*I don't think there were problems with the system, it was more to do with text. The system worked well to get me to the point where I'm prepared to look at the text. That was the main thing*" (participant 2).	1. Dislikes reading long text 2. Tool helps to minimise reading
"*Once I got my head around with the colour scheme and things like that, it does work quite quickly for visually picking up what you need to. Or, I should say, not spotting before you filter, but once you filter knowing that you are on a right track without having to go and read quite so much. And then it does filter it down enough to read the actual comments and the actual, the longer text and that kind*" (participant 2).	1. Learning effect 2. Learning colour scheme 3. Enables visual filtering 4. Minimises unnecessary reading 5. Keeps user on a right track 6. Filters results down to read longer text
"*I mean the other thing would be a sort of like a Google search on a field or field level. I suspect I can't do that here*" (participant 3)	1. Wants Google-like search to filter on a particular metadata field
Participant 3 could not find a relevant dataset and reset all the filters: "*And now I want is a sort of keyword search or something where I can really look into all of them. Because what I want to do now is to narrow it down not just on this* [pointing at the filtering area]*…*" (participant 3).	1. Wants keyword search to filter results 2. Needs to narrow results down by additional criteria

"*I'd almost have to go through every single one of these to see if there is anything now at the UK resolution. I don't think you can find the datasets to answer that question with the system as it is at the minute. I don't know if there is one in here or not. But I suspect that you need a specific UK dataset for that. So I'd want to know whether it was created in the UK*" (participant 3).	1. Needs more filtering functionality 2. Needs more search functionality 3. Has to go through every label to acquire information 4. Cannot find suitable dataset with the current state of the system 5. Needs a dataset with UK resolution
"*So for me, these* [pointing at the filtering options] *give additional ways of filtering the information once you've discovered the datasets that you actually potentially could use*" (participant 3).	1. Filtering supports dataset discovery 2. Additional way of filtering once discovered the datasets 3. Filtering allows to find potentially useful datasets
"*I think this 'smaller' thing is too subtle there*" (participant 3).	1. Changes in size are too subtle
"*So it's experts reviews that is giving me pain*" (participant 3).	1. Identified that expert reviews filter out a lot of datasets
"*I really appreciated in a lot of these things [pointing at additional filtering options] that I had dropdowns rather than free text. Because I really wouldn't have known what to type*" (participant 3).	1. Likes dropdown lists
"*I like the sliders, I must admit. It is neat [clicking on the filtering label] but I don't think I'd use that*" (participant 3).	1. Prefers sliders for filtering 2. Would not use filtering label
"*I didn't even think to click on it because I thought it was a graphical representation of my filtering. It does say "Click on a facet…" [laughs] What it does show that people don't read!*" (participant 3).	1. Did not notice the instructions 2. Did not know how to use the dynamic filtering label
"*Oh, that's cool! I like how it jumps to them [facet sections]. That's really neat*" (participant 3).	1. Likes that appropriate filtering area becomes visible when filter is applied
"*The one issue for me – I don't know which filters do the resizing for me*" (participant 3).	1. Issues with resizing 2. Unsure which filters resize the labels 3. Confused with additional filtering
Participant filtered by 'FGDC': - "What do you think it is doing?" - "*Well, I am assuming it is removing the ones that do not comply with that. But these things don't seem to have FGDC… Oh, so they go ever so slightly smaller. So the smaller ones are the ones that it* [filtering] *doesn't match. But that is too subtle, particularly when they all go smaller because nothing looks bigger* [laughs]. *I didn't notice that*" (participant 3).	1. Did not notice the changes in size 2. The changes in size are too subtle 3. When all labels become smaller, nothing looks bigger 4. Assumed that additional filtering removes the labels that do not match the filter

"*So for me, these* [pointing at the filtering options] *give additional ways of filtering the information once you've discovered the datasets that you actually potentially could use*" (participant 3).	1. Filtering supports dataset discovery 2. Additional way of filtering once discovered the datasets 3. Filtering allows to find potentially useful datasets
Applied quality filter: "*Oh, that's good!*" (participant 4).	1. Likes dynamic filtering
When all labels changed in size, participant stopped and looked at the screen wondering what just happened. Then applied citations filter: "*So they've gone a bit smaller probably because they have fewer than 5* [citations]" (participant 4).	1. Associates changes in size with matching the filters 2. Uncertain about additional filtering effect
"*Another thing that would be handy. I've for filtering applied here* [pointing at additional filtering options] *and when I close it there is no indication that I have filtering applied. So I could make a mistake of not removing a filter and therefore constraining my options. Because I might not realised that I'd set a filter in an area. So a feedback to the user could be that, maybe, one of these bars [pointing at the filtering options area], that this bar, if you've got filters set, is coloured in the same colour or something like that. So there is feedback. So that you know that you've done it*" (participant 3).	1. Filtering options should provide visual feedback when applied 2. GEO LINC improvement suggestion
"*If I say "4.5 stars" am I removing all the results that have no user feedback?*" (participant 5).	1. Unsure how the additional filtering affects the search results 2. Assumes that the additional filtering removes the labels
"*I think I might be removing results that are not complaint with these* [pointing at the additional filtering options]" (participant 5).	1. Assumes that the additional filtering removes the labels
"*But it's not like that because it makes it a little bit small. I mean the label, smaller or bigger*" (participant 5).	1. Uncertain about the effect of the additional filtering
"*I have this feeling, because sometimes when I apply a filter some of them disappear, and I automatically make that… I know how it works but I've been assuming that* [additional filtering also removes the labels], *for some reason, because sometimes they disappear*" (participant 5).	1. Confused about filtering functionality 2. Understands additional filtering functionality but sometimes confuses it with facet filtering
"*Now, that I have all these* [filtered labels] *I would like to make a quick search to see if I have some climate change focused datasets*" (participant 5).	1. Needs additional keyword filtering
"*I think in this case, I wouldn't worry much about the number of processing steps. Also because, if you have raw data you don't really have temperature in your data, temperature data must be processed. So I think I would skip the bit of how many processing steps have occurred and I would focus more on what other people said*" (participant 5).	1. Number of processing steps is not important 2. Temperature data is always processed 3. More interested in user feedback

"*I am wondering how many am I left with. I have a total of 101 but I would quite like to know how many passed my test so far*" (participant 6).	1. Wants to know the number of datasets available after filtering
"*Did that work?* [when labels changed the size] *I'm not sure what that did. I might clear it. I am doing something obviously wrong. Let's apply these again. Ah, okay, some of them had gone smaller. Some of them are slightly bigger. It is not immediately obvious to me which ones have gone bigger and which haven't*" (participant 6).	1. Unsure if the additional filtering worked 2. Changes in size are not immediately obvious
"*See, it would be nice to know how these things* [dataset labels] *are responding to the filters that I'd set on here* [pointing at additional filtering options], *like the average rating and so on*" (participant 6).	1. Unsure which datasets match the additional filters 2. Needs more information
"[hovering over facets] *So that tells me how those fulfil those criteria or not. But then I'd have to click on another one* [dataset label] *to see that one and I'd probably forgotten what the last one was. So it would be nice to see those next to each other*" (participant 6).	1. Wants to see information that relates to additional filtering 2. Need to add hover-over text as part of dataset description
"*Apparently, I want to look for the datasets that have at last one processing step applied. That seems a strange way of filtering datasets but I will follow the instructions. Yeah, I have to say, this doesn't make a lot of sense to me in terms of the kind of data that I would find*" (participant 6).	1. Filtering by number of process steps does not make sense 2. Strange way of filtering
"*Anything that's geophysical has had a lot of the processing steps applied to it already because the only unprocessed dataset you can get is Level 0 of an instrument which is voltages. And no one who's interested in any kind of application is going to go anywhere near voltages. That's really specific thing*" (participant 6).	1. Most of the data is processed 2. Users are not interested in raw data
"*So I would expect any dataset that I am actually using in my work to have gone through maybe 4 or 5 processing steps already. It doesn't matter to me how many there've been, I don't care, I just want to know what level of dataset this is, how many datasets it depends on, you know, whether it's a single instrument or whether this is actually a blend of lots of instruments. So it's more of characteristics of that processing than the fact that there've been any. It's not really helpful thing to me to say*" (participant 6).	1. Does not care about number of process steps 2. Needs to know processing level 3. Needs to know dataset dependency on other datasets 4. Needs to know how many instruments collected data 5. Needs to know processing characteristics
"*And now all my filters have gone. I have to put the same filters in*" (participant 6).	1. Disappointed that filters were not saved 2. Has to reapply filters again
"*And I'm wondering if the things in these boxes are ORs or ANDs, so if I put more than one in, would it find things that did all of them or will it do one or the other.* [applied filter] *Impossible to tell*" (participant 6).	1. Uncertain about free text filtering options 2. Wonders what logical query constructs are supported
"*It might be nice to do that with a slider or something to see were the breakpoints are in the things that I'm picking*" (participant 6).	1. Wants to see the range of available data (average ratings, number of citations, etc.)

I.6. Dataset Details Acquisition

Data Extracts	**Coded As**
"When I hover-over every one of these, it is good that it shows "that many reviews". Maybe if it said what the symbol mean" (participant 1).	1. Likes hover-over functionality 2. Did not recognise the facet names in the hover-over text
"[opened metadata XML] *Ugh, blimey!* [closed metadata XML page]. *So, let's see what quality information is there.* [clicked on quality information facet drilldown] *That's better!*" (respondent 3).	1. Dislikes XML 2. Looking at quality information 3. Likes drilldown pages
"[looking at the lineage drill-down] *This is very useful!* [Looking at quality information drill-down] *I like the additional details*" (participant 2).	1. Drilldown is useful 2. Likes additional drilldown information
"*Definitely handy to have this box* [datasets details section] *to, kind of, work out what the data actually is*" (participant 1).	1. Dataset details are useful 2. Dataset details help to understand the data
"*What also occurred to me, it is not obvious, unless I have a good read through it, over what domain theses datasets are for*" (participant 1).	1. Lacking information about dataset domain
"*How would you get detailed information about the dataset?"* "*Yeah… That was my thought with that button first of all – the 'View Metadata' – but I kind of switched off when I saw XML*" (participant 2).	1. Wanted to see detailed information 2. Wondering how to get additional information about the dataset 3. Not interested in XML
Looking at the 'date' field the participant 2 was wondering what date this actually is: "*A am pretty sure that within the ISO, well, Dublin Core or ISO1900… or whatever it is, I always forget the number, there is an element for all the different dates, but that kind of goes overboard on the date sense, it goes mental. And a lot of the time we, perhaps, need to understand the date range within the data, which can be automatically generated. And then creation date stroke publication date is almost the same thing, in my opinion. A creation date vs. a publication date, it doesn't seem to be a big enough difference. Most of the times you just want to know when it was released to general use and what the actual date range within it*" (participant 2).	1. Importance of date range within the data 2. Importance of release date 3. Wants to know when the data was released for general use
"*I found myself always looking at date. That's certainly one of the key criteria that I pick up on because you do want to know the currency or you either doing a study over a certain period of time*" (participant 2).	1. Importance of date field 2. Wants to know currency of data 3. Creation and update dates are one of the key criteria
"*I think the key bit is this drilldown. I think it is absolutely critical because these* [GEO label] *colours are not enough. And you always need to read what they actually mean*" (respondent 3).	1. Drilldown is the key bit 2. Drilldown is absolutely critical 3. GEO label colours are not enough 4. Need to know what colours mean

"*I do like the hover because I don't always want to go to a new page. I just want to get an idea. If anything, I'd like hover* [text] *to come up slightly more quickly than that*" (participant 3).	1. Likes hover-over functionality 2. Hover-over is quicker than drilldown
"*I like filtering*" (participant 3).	1. Likes filtering
"*Getting an idea of the temporal, so a temporal coverage and a spatial coverage, I think, are often very key things when I am looking for data. And when I say temporal coverage, I also mean frequency*" (participant 3).	1. Importance of knowing dataset's temporal coverage 2. Importance of knowing the frequency of temporal coverage
"*It would be good, if here I almost knew what extents these* [datasets] *were over. I don't know how hard it would be, almost, to include a tiny map here* [pointing at the detailed information area] *with just the area that it* [the dataset] *is valid for highlighted with the rest, kind of, faded out a bit*" (participant 1).	1. Importance of having a small map showing dataset's spatial extents 2. Importance of knowing dataset's spatial extents
"*Actually, a visual cue of the spatial extent of the dataset. When I click on the dataset, here [pointing at the dataset details area] there is a little insert map which has a rectangle of the area it applies to immediately gives a cue, a visual cue that I can very quickly process*" (respondent 3).	1. Importance of knowing dataset's spatial extents 2. Small map as a visual cue of spatial extent.
"*So kind of a map plus a timeline showing sort of point at every time at which it was available. Or, if it was really continuous time, a shaded region along the timeline. Again, much quicker for me to process than text*" (respondent 3).	1. Importance of having a visual map 2. Importance of having a visual timeline of temporal coverage
"[discussing drilldown] *That's handy, isn't it? Cause you've pulled out the metadata. Although it is not entirely obvious on this [pointing at a GEO label in the search area] that you can click on one of these things to open up a new window to give you that information. But that's one of those things where, once you've used it once or twice, you'll go "yes, of course I do that". So that's just a learning thing. So that's not a big deal at all*" (respondent 4).	1. Usefulness of drilldown 2. Drilldown pulls out metadata 3. Drilldown is not entirely obvious 4. Learning about drilldown
"*I think that graphical display here is brilliant. I really, really like it. And there is a way of navigating datasets. It's much, much nicer that sort of page-based. I really like being able to split the page up and just move through and view each one in turn and this hover thing*" (participant 3).	1. Likes graphical display 2. Likes ability to navigate the datasets 3. Starfield display is nicer than page-based display 4. Likes dataset details functionality 5. Likes hover-over functionality

I.7. Physical Data Acquisition

Data Extracts	Coded As
"*I would just want a link to take me over to that* [producer] *site, to the same data that is viewed on here* [pointing at the tool]. *Would be handy. Cause otherwise, my next step would be to either contact the user* [pointing at the user feedback drilldown page] *or it would probably just be copy and paste the dataset name and try to find it somewhere on Google*" (participant 1).	1. Acquiring data 2. Wants a link to producer page 3. Link to producer website 4. Link to the physical data 5. Obtaining information about physical data from peers 6. Search for data on Google
"*So how do I get this data? Where do I get the data from?*" (participant 1).	1. Requires a link to the actual data 2. Wondering how to get the data.
"*So I've identified a couple of datasets that sound promising. But I need to have that data and have a better look at them*" (participant 1).	1. Identified datasets of interest 2. Wants to inspect physical data 3. Narrow down selection options
"*The next bit is actually getting it, the data. I can't immediately spot and it might be conditioning from the portals that I am used to, alongside the 'View Metadata' you tend to get the 'Download in different formats' buttons*" (participant 2).	1. Wants to download data 2. Looking for download buttons 3. Searching how to get the data
"*The ones where you get the whole list of downloads is too busy really. I like the cleanness of this* [pointing at the Dataset Details area], *I like the limited data you've got, which is perfect. I am guessing this is 'last updated' date* [pointing at the date field], *that's always a key bit*" (participant 2).	1. Likes the cleanliness of the interface 2. Prefers limited amount of information 3. Standard portals provide too many download buttons 4. Importance of the last updated date

I.8. Alternative Search Results View

Data Extracts	Coded As
Because none of the datasets seemed to perfectly fit the specified requirements, the participant started to go through each dataset (clicking on the labels) and look at the titles and descriptions: "*I could, because this is quite laborious going through this one by one, I could filter it down by one of my criteria, but I am just worried I might miss something if I do*" (participant 1).	1. Afraid to miss important dataset 2. Manual dataset inspection 3. Laborious to inspect datasets manually
"Do you think a table view would be useful as well with this?" "*Yea, that could be useful. In this particular instance, if I could just flag up just the titles of each one, that would be quite useful*" (participant 1).	1. Alternative table view 2. Wants to see datasets' titles

"*Just keeping out the ones that sound like what I want because I ended up filtering them out by making my criteria too strict. Because the GEO label is quite simple, you could have the title with the GEO label next to each one. So I could quickly find it*" (participant 1).	1. Filtered out too many datasets 2. Wants to see datasets' titles 3. Display the titles next to the labels 4. Needs alternative view
"*Because this is brilliant if you have to have these criteria. This clearly showed me that it's only these three datasets that meet your* [my] *criteria. But, if I didn't have the options to be picky, then this* [going through each label one by one] *does become a bit more laborious and then just be looking through the titles would be quite useful*" (participant 1).	1. Usefulness of seeing the titles 2. Laborious to go through each label 3. Tool is brilliant tool for filtering by specific criteria
"*I mean, it seems to be strange to be selecting datasets of this sort of matrix. I'm used to choosing my datasets from a list where you can see exactly what the name of the dataset is and what all of the… a bit of information about each of them*" (participant 6).	1. Strange to select datasets from a starfield display 2. Accustomed to choosing datasets from a list 3. Needs information about each dataset
"*So without seeing information on each while they are all there* [pointing at the dataset search results window] *it is hard to know if I've got my filtering right. Because this one* [selected a dataset] *arbitrarily it might be the one I am really looking for but you can't tell at a glance how it might be or might not be matching up with the criteria that you've set. How sensitive these choices to exactly what we've done here*" (participant 6).	1. Hard to know if the filtering is right 2. Hard to say which datasets fits requirements best 3. Can't say how the datasets match the filtering
"*I don't really understand what this is showing me. So this is sort of redundant information at the moment. I'm just seeing this as a list of datasets and it's made it awkward by putting them in different places*" (participant 6).	1. Does not understand the starfield view 2. Starfield shows redundant information
"*It's too impersonal, I guess*" (participant 6).	1. Tool is too impersonal
"*I can do a lot of the filtering in my head from seeing the name of the dataset, I suppose, as to whether it's the kind of thing I'm looking for. Whereas, this is deliberately hiding what the datasets actually are which to me, as someone who's used to knowing the sort of thing I am looking for, it's not helpful*" (participant 6).	1. Can do filtering from dataset names 2. Tools is deliberately hiding important information
"*Because I can't filter on the things that I would want to filter on, like if I'm looking for my snow data, for instance, from the snow data archive that I use*" (participant 6).	1. Not enough filtering options 2. Cannot filter on important to user criteria
"*I find this a bit weird* [pointing at the star display], *this way of displaying the data that's been produced. I would want to see a list and I would want to see how, see more kind of quantitatively, how things are fitting my criteria or not fitting my criteria. Maybe rank them in the list in the order that they fulfil my criteria or not so that I'm continuously look at the ones that make a difference. And then as I change the filters you can see exactly which ones are changing in status, I suppose*" (participant 6).	1. Starfield display is weird 2. Wants to see a list 3. Wants to rank the list of datasets 4. Wants to see how the filtering affects search results

I.9. Informational Aspects

Producer Information

Data Extracts	Coded As
"*This dataset pretty much ticks all three of my boxes. It's good. It's just, again, it's missing the producer comments and the uncertainty measures. But, because I've got the producer information I could give them a ring and see what they have to say about the dataset that I needed*" (participant 1).	1. Wants to contact the producer 2. Wants to obtain additional information from producer 3. Importance of producer contact information
"*Right, I've got myself 6 datasets.* [selected the most complete label] *I'll have this one straight away, I know that, because it's from somebody from the JRC and I'm working with the JRC*" (participant 3).	1. Selected the most complete label 2. Trusts familiar data producer 3. Trusts data producer because working with them
"*I would choose this one because I know where it's come from*" (participant 3).	1. Selecting datasets from a known producer 2. Trusts producer
Looking at producer profile: "*Anyone called Freddy is not going to produce a decent dataset*" (participant 3).	1. Does not trust producer

User Feedback

Data Extracts	Coded As
"*I would just want a link to take me over to that* [the producer] *site, to the same data that is viewed on here* [pointing at the tool]. *Would be handy. Cause otherwise, my next step would be to either contact the user* [pointing at the user feedback drilldown page] *or it would probably just be copy and paste the dataset name and try to find it somewhere on Google*" (respondent 1).	1. Acquiring data 2. Wants a link to producer page 3. Link to producer website 4. Link to the physical data 5. Obtaining information about physical data from peers 6. Search for data on Google
"*If there was another option kind of saying "show me all the ones that this person rated or used", a way of grouping by users or studies, because I've spotted a pattern amongst these, same couple of people doing it for their, particularly this German, study. So you could have it as another way of looking at things. And that would definitely affirm or confirm that, what I was looking after, that was the right stuff I needed to and that I haven't missed out, or everyone missed out. One of the two, you never know!*" (participant 2).	1. Needs an option to group datasets by users who have previously used the data 2. Needs an option to filter datasets by users who have previously used the data 3. User feedback affirms dataset selection 4. Identified a pattern in data 5. Identified a group of users conducting same study

After setting all the filters and inspecting dataset titles, abstracts and drill-down information, participant 2 narrowed his choices to 3 datasets: "*I'd probably look at those three* [datasets]*, and the reason why is – after all that filtering it appears to be the same reviewers, based on the user comments and on the citations, they've used it for pretty much the same study. So if they've done it, I'll just copy what they've done.*"	1. Dataset selection decision 2. Trust in peers 3. Reliance on user feedback 4. Reliance on citations 5. Other users used data for similar study 6. Trusts work done by other peers
"*I'm kind of a sheep, I follow everyone else really. If everyone else has done it before, I'll be like "yea, I'm on a right track*"" (participant 2).	1. Influenced by peers 2. Follow the crowd
In post scenario discussion, respondent 2: "*The feedback is very useful. I would group users and academics* [pointing at user feedback, expert review and citations information facets] *cause it's a feedback thing and, for me, there is always a potential with academic feedback to again go into gobbledygook that I don't understand, so I just look into all feedback"* (respondent 2).	1. Combine feedback, review and citations 2. Does not always understand academic feedback
"*… I do not put numbers like this in… because this is looking a bit artificial because I know I've got to look at 3 stars or more, I wouldn't think that. I would think that I'd want to look at what the distribution of the stars is and then I'll make an informed decision…*" (participant 3).	1. Artificial feedback scenario 2. Looks for star-ratings distribution
"*…obviously, I'm gonna look for the highest number of stars I can get*" (participant 3).	1. Trusts higher ratings 2. Looks for the highest number of stars
"Certainly, when it comes to TripAdvisor or any other equivalent rating systems, I do certainly tend to look at items which have significant number of feedback and take that feedback much more seriously. So if they've got 500 reviews and 4.7 out of 5 I'd think that place was pretty good" (participant 3).	1. Looking for significant number of feedbacks 2. Trusts ratings supported by a large number of reviews
"*Let's look at the feedbacks.* [scrolling through feedbacks] *Everyone quite likes it. So the feedbacks are quite good*" (participant 4).	1. Found positive user feedbacks
"*The user feedback and the expert reviews stuff is probably the most important thing, in a way, because that gives you the idea of what real people have used the data for and what their real experience was. The fact that something's got an ISO standard just means that it adheres to some kind of processes and ticks some boxes. And having worked to standards, I know they are a little bit dubious sometime anyway. It doesn't necessarily mean they are brilliant. So the fact that somebody've used it is probably of more value to somebody who's looking at this dataset*" (participant 4).	1. Importance of user feedback and expert reviews 2. User review is more important that compliance with standards 3. Standards are dubious 4. Compliance with standards does not indicate quality
"*I would trust a user feedback or an expert review more than I would just a standards stamp*" (participant 4).	1. Trusts user more than standards
"*So, although this exercise doesn't state that I should look at user commentaries, I personally find that very useful*" (participant 5).	1. Wants to look into comments

"*I would like to see users like me, not-expert users, their opinion, their experiences using the datasets*" (participant 5).	1. Interested in non-expert feedback
"*For example this one* [dataset] *looks very good because somebody like me, which is not an expert, has been studying what I need. Has been studying not exactly agricultural studies but he has done the study in the UK in the more or less same range of dates. And that is making me think that it is very likely that I will find all I need from this dataset*" (participant 5).	1. Trusts dataset that was used for a similar study 2. Trusts user feedback 3. Influenced by user feedback
"*And the explanation is, after applying filtering in the exercise I went to the results with the biggest label and the second biggest label, and I highlighted all those, and compared user commentaries, and I just picked the ones with the similar users to me that were focusing on the studies that I needed and commented that it was good. So I trust them*" (participant 5).	1. Selected biggest labels 2. Used highlighting functionality 3. Compared user comments 4. Selected datasets with comments from similar users 5. Influenced by user comments
"*So probably because this one has got more commentaries, it has been more used*" (participant 5).	1. Associates user feedback with frequency of dataset's use
"*So this C001 has this very useful expert saying that he used this to get climate change data. Well, he was studying climate change and he was studying South Africa. So this one* [dataset] *is a very good one*" (participant 5).	1. Found a very useful expert review 2. Expert used data for a similar study 3. Influenced by expert review
"*So one thing that I might do is look at the name of whoever made the commenting that I think was studying the same thing I was studying. I would probably look in other datasets for this person making comments. Because that's very likely that we are studying the same kind of thing*" (participant 5).	1. Wants to find feedback from same user 2. Found a user who has been studying the same kind of thing
"*So I think I would skip the bit of how many processing steps have occurred and I would focus more on what other people said*" (participant 5).	1. Number of processing steps is not important 2. Skip number of process steps functionality 3. More interested in user feedback
"*The way I find out about datasets is generally sort of word of mouth. You talk to a colleague saying "I need some precipitation information, what do you think the best products are?" and they'll tell you about what they use and why, and so then you might go and use that. And you'd never necessarily, you wouldn't always, be in a situation where you go out and try to choose between things*" (participant 6).	1. Finds datasets through word of mouth 2. Talks to colleagues 3. Use suggested datasets 4. Does not always have to search for datasets
"*I'm interested in datasets that've been cited in at least 2 publications. And again, that seems a slightly strange way of filtering it because datasets can be cited for a lot of reasons. They can be cited because they are very poor, or they can be cited because they are good, or because they've been used, or they've been compared to something. So just filtering on number is a little bit artificial but I get what we are trying to do here*" (participant 6).	1. Number of citations is a strange way of filtering 2. Datasets can be cited for various reasons 3. Artificial filtering option

"*I don't know how I would interpret star ratings on the datasets*" (participant 6).	1. Not sure how to interpret star ratings of the datasets
"*User feedback* [is useful]*, yes, absolutely. I just don't know if there is a good way of having user ratings for a dataset*" (participant 6).	1. User feedback is useful 2. Not sure how feedback should be presented
"[looking at the user feedback] *So it's this guy again, he popped up before*" (participant 6).	3. Recognised feedback from same user

Compliance with Standards

Data Extracts	**Coded As**
"Normally, I would never worry about the standards compliance" (participant 3).	1. Unimportance of standards compliance 2. Never worried about standards
"*I would trust a user feedback or an expert review more than I would just a standards stamp*" (participant 4).	1. Trusts user more than standards
"*The fact that something's got an ISO standard just means that it adheres to some kind of processes and ticks some boxes. And having worked to standards, I know they are a little bit dubious sometime anyway. It doesn't necessarily mean they are brilliant*" (participant 4).	1. Previously worked with standards 2. Standards are dubious 3. Compliance with standards does not indicate quality