Proceeding of the 10th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences
Florianopolis-SC, Brazil, July 10-13, 2012.

Geospatial Data Quality Indicators

Victoria Lush[1], Lucy Bastin [1] and Jo Lumsden [1]

[1] Knowledge Engineering Group, Aston University, Birmingham, B4 7ET, UK
lushv@aston.ac.uk, j.lumsden@aston.ac.uk, l.bastin@aston.ac.uk

Abstract

Indicators which summarise the characteristics of spatiotemporal data coverages 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. Criteria which are commonly prioritised in evaluating spatial data quality and assessing a dataset's fitness for use include lineage, completeness, logical consistency, positional accuracy, temporal and attribute accuracy. However, user requirements may go far beyond these broadly-accepted spatial quality metrics, to incorporate specific and complex factors which are less easily measured. This paper discusses the results of a study of high level user requirements in geospatial data selection and data quality evaluation. It reports 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. We briefly describe the implications for tools and standards to support the communication and intercomparison of data quality, and the ways in which these can contribute to the generation of a GEO label.

Keywords: Geospatial data, geospatial data quality, geospatial data quality indicators, quality evaluation.

1. Introduction

Over recent years, the production and availability of geospatial data has significantly increased. Discovery and reuse of geospatial data has been particularly facilitated by the recent explosion of Web-based catalogues, portals, standards and services, and by initiatives such as INSPIRE and GEOSS. GIS professionals, decision makers and non-expert end-users are always interested in data of high quality (Wang and Huang, 2007) but may face problems in fully evaluating the resources available to them. Part of the problem comes from the inherent impossibility of a perfect representation of the real world with all its unlimited complexity and level of detail. However, another important challenge is the inconsistent and patchy nature of data quality information, which makes intercomparison very difficult (Boin and Hunter, 2006).

Generalised, abstracted and aggregated as it is, geospatial data can only provide an approximation of the real world and therefore almost always suffers from imperfect quality and limited accuracy (Goodchild, 1995; Couclelis, 2003). Unsurpris-

ingly, spatial data quality, accuracy and uncertainty is a longstanding area of active research within the Geographic Information (GI) community (Devillers, 2002, and many others) and there are many well-tested methods for reliably quantifying and representing data quality. Despite these detailed recommendations, and despite the existence of metadata standards such as ISO 19115 and FGDC, data quality information is often not communicated to users in a consistent, interoperable and standardised way (Boin and Hunter, 2006). In this paper, we investigate, through detailed interviews, the facets of data quality which users wish to see when making fitness-for-purpose judgements. The resulting requirements are being used to drive development of information models and APIs within the EU 'GeoViQua' project.

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; Boin, 2008). 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". 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). Commonly-accepted criteria for evaluating spatial data quality include lineage, completeness, logical consistency, positional, temporal and attribute accuracy (Caprioli *et al.*, 2003; Boin and Hunter, 2006; Devillers *et al.*, 2007). However, user requirements in terms of "fitness-for-use" may go far beyond these common spatial quality elements, to incorporate specific data features such as spatial and spectral resolution and complex factors, such as continuity of supply and reputation of the producer, which are less easily measured. The subjective and context-specific nature of these needs makes research into fitness-for-use more challenging than many of the more straightforwardly quantitative aspects of data quality assessment (Boin, 2008). Users must frequently assess fitness-for-use by mapping simplified quality indicators to their specific demands (Duckham, 2002). While no tangible user-defined quality indicators to specifically assist fitness-for-use evaluation have been identified, there are many existing forms of metadata 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.

The research we present in this paper represents a significant knowledge elicitation step in an ongoing agenda aimed at (1) identifying the key quality indicators of geospatial datasets upon which users in different application areas rely when selecting datasets for use on specific projects, and (2) developing and delivering novel means of representing and interrogating dataset quality indicators with a view to supporting efficient and effective geospatial dataset selection on the basis of quality and fitness for use.

2. Methodology

We carried out a series of semi-structured telephone and face-to-face interviews with geospatial data users and experts, to collect high-level user requirements relating to quality-aware data selection. The interviews were relatively informal, and were guided by a set of general questions; follow-up and clarification questions

were asked depending on specific interview circumstances. The general questions asked users to describe:

- a current area of their work in which they use external data sources;
- data they use in their work, and where it comes from;
- how they choose datasets, and the reasons for their decisions;
- whether they are aware of any data certificates or seals in selecting their data;
- whether the data they use come with sufficient supporting information to allow them to make an informed judgement;
- how much information they need.

A total of 18 interviewees were recruited, representing a variety of expert groups including end data users, researchers, data archivists, academics, and data producers. The range of expert groups allowed us to elicit an interesting variety of user stories and develop a wide-ranging picture of user needs. Information gathered from the interviews was used to derive user stories – very high-level informal statements of the requirements that capture what the users want to achieve. From these user stories we derived a set of user requirements and key geospatial data informational attributes that are used in selection and quality assessment.

3. Study Results

We should, at this point, stress that we are attributing no statistical significance to the findings reported here, given the small sample size. Our intention was to conduct an in-depth initial investigation to elicit geospatial data quality attributes that data users and experts consider when making a dataset selection decision. These initial observations will be more extensively researches through further surveys and prototype testing. A parallel survey with over 80 respondents has specifically addressed the community's requirements for a quality GEO label, but is outside the scope of the current discussion.

Our study helped us to identify common geospatial data informational attributes that are considered by geospatial data users and experts when selecting a dataset. These common attributes included metadata content, metadata visualisation, community advice, reputation of data provider, citation information, and 'soft knowledge'.

3.1. Metadata

Our analysis identified that geospatial data users are exceedingly interested in good quality metadata. Both users and producers stated that complete and well documented metadata records, which comply with ISO and Dublin Core standards, are essential in the assessment of geospatial data quality. Our survey revealed that, at present, users find metadata records are typically incomplete with a lot of essential data omitted. Our interviewees specifically highlighted provenance (i.e., lineage) and licensing as information that is typically missing from the data they come across. Users and experts listed the following provenance elements as desired to be provided with every dataset: original dataset provider; methodology adopted for dataset data collection; how a dataset was derived and on what it is based; the pur-

pose for which a dataset was originally collected; parties who have subsequently processed the dataset; parties who have used the dataset before; dataset harvesting pathway and processing log. Additionally, as mentioned above, a number of interviewed users and experts pointed out that the licensing information is nearly always missing. Despite the fact that standardisation bodies provide clear schemata for much of the above information, they are inconsistently used, as demonstrated by (Boin and Hunter, 2006). A related study by Maso *et al.* (2012) has demonstrated the patchy nature of metadata within the GEOSS clearinghouse, showing that 'the documentation of quality indicators and lineage is far from general in ... Earth observation data'.

3.2. Metadata Visualisation and Comparison

Our interviews revealed that geospatial data users and experts require more sophisticated tools for visualisation of metadata records, which, at present, are difficult to examine and assimilate. Non-expert users suffer the most from not being able to absorb and understand all of the information recorded in metadata. Effective visualisation methods for metadata records need to be developed to support users in data quality evaluation and decision making process. Another important aspect of metadata visualisation identified during our interviews is the ability to easily and systematically compare metadata records. Our users indicated that side-by-side visualisation of all metadata elements would allow them to systematically compare geospatial datasets more effectively, particularly where several similar datasets appear to fit the purpose, and differences are hard to distinguish. A prototype comparison stylesheet has been developed within the GeoViQua project, and tools/queries to enable such 'comparison shopping' are in development.

3.3. Community Advice

Users of geospatial data stated that they rely heavily on peer recommendations when selecting a dataset. They contact their peers to obtain valuable feedback on the context in which datasets were used, what these datasets were good for, problems with the datasets and other potentially useful information. A peer review functionality for geospatial data would facilitate improved data selection and quality evaluation by allowing users and experts to provide their comments on datasets they used, record publications which were generated, and flag any limitations or problems associated with the datasets. Geospatial data producers also stated that they are interested in having their datasets peer reviewed, as this would allow them to identify and resolve any issues within their datasets, and also to respond to users' comments. A fundamental component of any such feedback approach is the availability and management of unique dataset identifiers, to identify the target of any feedback item and to federate records from producer metadata and online databases. The proposed GEO label[1] could well be implemented as a dynamic summary of information aggregated in such a way.

3.4. Reputation of Data Provider

The reputation of data providers was identified as a key factor in dataset selection. Users typically rely on data from producers that they already know or those who have a very good established reputation in the community. This can mean that

[1] http://www.iiasa.ac.at/Research/FOR/downloads/ian/Egida/geo_label_concept_v01.pdf

unknown data producers experience much lower data demand, as in e-Commerce, where new, smaller vendors are much less trusted by customers than larger brands. Our interviewees also noted the tendency for well-organised and easily-accessible documentation to engender user trust in both data provider and the datasets they produce.

3.5. Citation Information

The majority of our interviewees base their dataset quality evaluation on dataset citation information. That is, when making dataset selections, users are largely interested in accessing the publications where data quality checks are reported for the dataset. As already mentioned in section 3.3, journal articles that describe dataset use and evaluation are considered to be very important in dataset quality assessment.

3.6. Soft Knowledge

Our interviewees highlighted that there are cases when data quality measures cannot be recorded in standard metadata records. Providers highlighted that they might be aware of problems with a satellite or a sensor but have no quantitative estimate to prove it. For example a sensor that has a particular range might work better in the middle of the range rather than at the edges; in such cases, they provide some soft knowledge (usually as free text) about data quality, including information which they think may be relevant to potential users. Users stressed the importance of data producers' comments and recommendations that are provided with the datasets they produce. They stated that having at least some soft knowledge about data uncertainty and error estimates would significantly help in more effective use of the data.

4. Conclusion

Our research thus far has identified potential quality indicators that geospatial data users and experts consider important when selecting geospatial data. Datasets with complete metadata records, good community reputation and a reputable data provider are more likely to be viewed as 'high-quality' by users. The results indicate that, when assessing data quality, users heavily rely on metadata records, community recommendations, reputation of dataset provider, citation information, and soft knowledge provided by the creator of the dataset. Visualisation of metadata records, with potential to compare two or more records side-by-side, would significantly simplify the assessment of datasets' relevance and quality. A standardised peer review of geospatial data would offer users invaluable information on usage and outcomes in specific application domains. Reputation of dataset provider presented as ratings and community comments would give users, especially novice ones, at least some indication of the quality of the datasets a producer provides. A complete list of citations, reporting a dataset's quality checks and usage, would provide users with a better understanding of previous data use. Finally, soft knowledge provided by the dataset producer can offer users any additional information that was not recorded in the metadata record. These informational attributes, if aggregated intelligently from distributed sources, may form the basis of the pro-

posed GEO label. When interrogated in more detail, they can potentially be integrated into decision support systems which allow a user to tune searches according to their specific needs.

Our research has so far, informed the development of producer and consumer quality information models for the generation of richer and better-linked metadata by users and producers. These models use and extend the existing ISO standards, GEOSS catalogs and OGC Web services. The original survey also represents some food for thought in terms of the current state of geospatial data and its quality assessment. We hope that our investigation will help to address issues with geospatial data quality and will lead to development of sophisticated tools to enable more effective data selection.

Acknowledgments

This project is funded by the EU Framework 7 Programme, contract no. 265178. We would like to thank all our participants without whom this research would not have been possible.

References

Boin, A., Hunter, G.J. (2006), "Do spatial data consumers really understand data quality information?". *In Proceedings of the 7th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*, Lisbon, Portugal, pp. 215–224.

Boin, A. (2008), *Exposing Uncertainty*. PhD thesis, The University of Melbourne, Australia.

Caprioli, M., Scognamiglio, A., Strisciuglio, G., Tarantino, E. (2003), "Rules and Standards for Spatial Data Quality in GIS Environments". *In Proceedings of the 21st International Cartographic Conference (ICC)*, Durban, South Africa, pp. 10-16.

Couclelis, H. (2003), "The certainty of uncertainty: GIS and the limits of geographic knowledge". *Transactions in GIS*, Vol. 7(2):165-175.

Devillers, R., Bédard, Y., Jeansoulin, R., 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*, Vol. 21(3): 261-282.

Devillers, R., Gervais, M., Bédard, Y., Jeansoulin, R. (2002), "Spatial Data Quality: From Metadata to Quality Indicators and Contextual End-User Manual". *Proceedings of OEEPE/ISPRS Joint Workshop on Spatial Data Quality Management*, Istanbul, Turkey.

Duckham, M. (2002), "A user-oriented perspective of error–sensitive GIS development". *Transactions in GIS*, Vol. 6(2): 179-194.

Goodchild, M. F. (1995), "Sharing Imperfect Data". *In:* Onsrud, H. J., Rushton, G. (eds.). *Sharing Geographic Information*, Rutgers University Press, New Brunswick, NJ, pp. 413-425.

Masó, J., Díaz, P., Ninyerola, M., Sevillano, E., Pons, X. (2012) "GEOSS clearinghouse quality metadata analysis". *Geophysical Research Abstracts*, 14, p. 8362.

Wang, F., Huang, Q. Y. (2007), "A methodology for definition and usage of spatial data quality rules". *Geoinformatics 2007: Geospatial Information Science*, Vol. 6753: D7531-D7531.