Developing a GEO Label: Providing the GIS Community with Quality Metadata Visualisation Tools

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

[1]Aston University, Aston Triangle, Birmingham, B4 7ET, UK.

Tel. (+44 (0)1212043000)

{lushv, l.bastin, j.lumsden}@aston.ac.uk

KEYWORDS: geospatial data quality, quality visualisation, quality labeling, fitness for use.

1. Introduction

Geospatial data quality and quality visualisation has always been an area of active research within the geographic information community. 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). Objective quality measures of geospatial data relate to the "*difference between the data and the real world that they represent*" (Goodchild, 2006, p. 13). Objective quality information (e.g., lineage, completeness, logical consistency, positional, temporal and attribute accuracy, uncertainty measures, etc.) is often found in formal metadata documents supplied by a dataset provider or in technical reports which describe quality checks. Subjective measures of 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 need (e.g., Chrisman, 1991). Subjective quality information 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.

With increased geospatial dataset production and increasing use of such datasets across ever more 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 clearinghouses (i.e., large data repositories that collect, store and make available geospatial data and metadata). 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. 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. However, search-by-quality is not currently available. This means that dataset users often have to manually inspect the data and metadata that is returned by search engines, in order to perform some form of quality assessment. To date, although the importance of fitness-for-use assessment is widely recognised in the geospatial community (e.g., Comber *et al.*, 2007), practical tools to support appropriate data quality interrogation and intercomparison based on the concept of fitness for use have been given relatively little research attention (Oort 2005). Via qualitative user studies, Boin and Hunter (2008) demonstrated that spatial data quality information is not communicated to data consumers effectively. Their findings showed that data consumers, when selecting a dataset to use,

heavily rely on subjective quality information (e.g., data source, provider's reputation, intercomparison with other data, cost, etc.). While these findings are very valuable, to date little practical work has been done to address the issue of conveying subjective quality information to data users. Consequently, geospatial data users still require accessible mechanisms for data quality discovery that recognise the subjective and use-case-dependent nature of geospatial data quality.

Our research therefore aims to:

1. identify the key informational aspects of geospatial datasets upon which users rely when selecting datasets for use; and

2. develop a GEO label that will support efficient and effective geospatial dataset quality representation and selection on the basis of quality and fitness for use.

The proposed GEO label will summarise and represent dataset quality information in a way which permits a user to easily assess the relevance of a dataset for their needs, and interrogate the specific aspects of metadata which are key to their application. This paper will discuss the process by which candidate GEO labels have been developed and evaluated, and will present the results of recent studies that elicited GIS community feedback on the GEO label.

2. Introduction to 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. 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, the GEO Science and Technology Committee (STC) propose 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*" (ST-09-02, 2010, p 2). The STC suggests 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). Furthermore, 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.

3. GEO Label Research to Date

Currently our project – GeoViQua (GeoViQua, 2011) – is undertaking active research to define, develop and evaluate a GEO label. The development and evaluation process is being carried out in three phases. In Phase I we conducted an online survey to identify initial user and producer views on a GEO label and its potential role. We analysed the results of the study and developed some GEO label examples based on responses gathered. Phase II consisted of a further study in which we presented the GEO label examples to potential end users and elicited feedback on their suitability and usability. In Phase III we are creating physical prototypes based on Phase II results for use in a human subject evaluation study.

Based the results of the Phase III study, the most successful prototypes will then be put forward as potential GEO label options for integration in the GEOSS. Each phase is discussed in more detail below.

The aim of our Phase I research was to investigate user and producer views on the role(s) that they think a GEO label should serve. Our intention was to elicit some initial insight into the functionality that a GEO label should comprise and to derive requirements for the development of GEO label prototypes. Our online GEO label questionnaire consisted of generic questions to identify: whether users and producers believe a GEO label is relevant to geospatial data; whether they want a single "one-for-all" label or separate labels that will serve a particular role; the function that would be most relevant for a GEO label to carry; and the functionality that users and producers would like to see based on common rating and review systems they already use. We collected 87 valid responses with overall results showing that users and producers of geospatial data appear to have generally very positive attitudes towards the development and introduction of a GEO label. The majority of respondents supported the notion of a GEO label providing an all-in-one drill-down interrogation facility that would combine:

- dataset producer information;
- producer comments on dataset quality;
- dataset's compliance with international standards;
- community advice;
- dataset ratings;
- links to dataset citations;
- expert value judgements; and
- quantitative quality information.

These study outcomes suggest that, in order to make an informed dataset selection decision, respondents require as much information as possible that is presented in one place in a format that allows for not only easy comparison but also deeper investigation of selected aspects. Consistent with earlier studies on fitness-for-use assessment (Comber *et al.*, 2007), users still feel that enhanced producer metadata (e.g., information on the documented uses of a dataset, and links to datasets against which its quality was evaluated) and ratings / comments on quality from their peers (e.g., user feedback) are lacking. Other work packages within the GeoViQua project are focussing on generating tools, services, data models and APIs which will support the gathering and aggregation of this supplementary information.

As proposed, the GEO labels will be a graphical representation generated individually for each dataset in the GEOSS (or other data portals and clearinghouses) based on the objective and subjective quality information that is available for that dataset. Based on the results of the Phase I study, we have developed three GEO label examples – graphic representations (i.e., static images) which could potentially be used to convey spatial dataset quality information. These GEO label representations were used in the Phase II study with an aim to identify the designs that convey quality information to users in the most efficient and comprehensible way. The online questionnaire was designed to first explore the level of understanding of the GEO label facets – the icons used within the labels to represent producer profile, user feedback, etc. – and the information they convey about the datasets they represent. Subsequently, collective visualisation of these facets within a single label was addressed using different scenarios and dataset ranking tasks. All three GEO label examples were presented separately in the context of different scenarios to assess their usability and effectiveness at conveying the availability of a dataset's quality information. The final

sections of the study were designed to elicit respondents' opinions on the informational aspects presented in the GEO label examples, effectiveness of using branding in the GEO label, and the GEO label examples in general. We will present the results of the GEO label studies conducted to date. The GEO label examples used in Phase II will be discussed and GIS community views on these will be presented. We hope that our investigation will help to address issues with geospatial data quality presentation and interrogation and will lead to development of an effective GEO label which will meet user and producer expectations, while supporting informed and efficient dataset selection and decision-making.

6. Acknowledgements

This project is funded by the EU Framework 7 Programme, contract no. 265178.

References

Boin A T, Hunter G J (2008) What communicates quality to the spatial data consumer? In: A Stein, W Shi, W Bijker, ed. *Quality Aspects in Spatial Data Mining*. New York: CRC Press, pp 285-296.

Chrisman N R (1991) The error component in spatial data. *Geographical Information Systems* **1** pp 165-174.

Comber A J, Fisher P F, Wadsworth R A (2007) User-focused metadata for spatial data, geographical information and data quality assessments. In: Proceedings of 10th AGILE International Conference on Geographic Information Science 2007, Denmark, Aalborg, 8 May 2007.

GeoViQua (2011) GeoViQua: Quality-aware visualisation for the Globasl Earth Observation System of Systems. [online] Available at <http://www.geoviqua.org/Docs/Poster.pdf > [Accessed on 14 November 2012]

Goodchild M F (1995) Sharing Imperfect Data. In: H J Onsrud and G Rushton, ed. *Sharing Geographic Information*. New Brunswick: Rutgers University Press.

Goodchild M F (2006) Foreword. In: R Devillers and R Jeansoulin, ed. *Fundamentals of Spatial Data Quality*. London: ISTE, pp 13-16.

Oort P (2005) Spatial data quality: from description to application. Delft: NCG.

ST-09-02 (2010) Draft GEO Label Concept. [online] Available at: < http://webarchive.iiasa.ac.at/Research/FOR/downloads/ian/Egida/geo_label_concept_v01.pdf> [Accessed on 14 November 2012].

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.

Biography

Miss Victoria Lush is a PhD student and a research assistant at Aston University. Her research is primarily in the field of geospatial data quality and human computer interaction with specific emphasis on data quality visualisation and trust.
Dr Lucy Bastin is a Senior Lecturer at Aston University whose current research interests include systematic conservation planning under real-world conditions, uncertainty management in modelling frameworks and in the Model Web, and spatial epidemiology.
Dr Jo Lumsden is a Senior Lecturer and manager of the Aston Interactive Media (AIM) Lab at Aston University. Her research is primarily in the field of human computer interaction with specific emphasis on mobile HCI, trust, and evaluation.