Approaches for providing user relevant metadata and data quality assessments

A.J. Comber[1], P.F. Fisher[2], R.A. Wadsworth[3],

[1]Department of Geography,
University of Leicester,
Leicester, UK.
E-mail: ajc36@le.ac.uk

[2]Department of Information Science,
City University,
London, UK
e-mail: pff1@city.ac.uk

[3]CEH Monks Wood,
Abbots Ripton,
Cambridgeshire, UK
E-mail: rawad@ceh.ac.uk

1. Data discordance

Spatial data, especially natural resource inventories, vary for a variety of reasons that are not to do with differences in the feature being measured. Often these differences in data well known amongst geographers: the real world is infinitely complex and all representations (such as are contained in a map) involve the processes of abstraction, aggregation, simplification etc. In the creation of any spatial data there are series of choices about what to map and how to map it. These choices over representation will depend on:

- The commissioning context specifically legislation and policy (often related to who "paid" for it?);
- Observer variation such as the classic geography field trip (what do you see?);
- Institutional variation in classes and definitions (why do you see it?);
- Representational variation over map scale, minimum unit, (how do you record it?).

A second set of factors that contribute to data discord and variation originate in the demand for 'better' science. New technologies, improved techniques and changes in the understanding of the phenomenon offer greater insight into the process under investigation. Such changes in representation and understanding have a profound effect on the end data product and the meaning of the data in its widest sense. They change the data collection context in terms of data ontologies (specifications), data epistemologies (measurement) and data semantics (conceptualisations).

2. Metadata

Prior to its inclusion under the wider umbrella of information sciences, the GI community developed metadata standards for reporting data quality. Metadata for spatial data focussed on the need to document information about data for data quality assessments. The FGDC Content Standards for Digital Geospatial Metadata places an emphasis on using metadata elements in a discovery and query environment to provide “fitness for use” information to prospective users of digital geospatial data. In these standards metadata typically describes data quality in terms of the Positional Accuracy, Attribute Accuracy, Logical Consistency, Completeness, and Lineage. Consequently standards for data quality and metadata reporting have been based on these measures.

More recently the GIS community and spatial data standards have been included within the wider informatics and computing science community. There are a number of organisations concerned with the specification of metadata standards for describing the components and character of spatial data which are converging to differing degrees (e.g. OGC, Dublin Core and ISO). Despite the stated objectives of enabling users to understand data, typically these standards comprise a number of elements that principally specify how to document information relating to the cataloguing, finding and retrieval of data. Metadata standards are useful because they provide a common language, enabling parties to exchange data without misunderstandings. However their specification (content) is always a compromise and consequently they do not represent the depth of knowledge held within scientific community. They are:

- Focussed on aspects relating to data production and data mediation rather than the use of the data;
  Passive, rather than active descriptions relating to potential applications;
- Recording the easily measurable aspects of data rather than the most pertinent aspects of the data;
- Providing overall or global measures of data quality, not ones that relate to individual map objects;
- They are difficult for users to interpret in relation to a specific application

The data quality parameters reported in metadata do not communicate the producer’s wider knowledge of the data and relate to use, rather they reflect data production interests, reporting the easily measurable and showing that the data producer can follow a recipe (Comber et al., 2005).

We propose that the focus of metadata be towards the user. As an alternative definition to metadata being “data about data”, a user-focussed definition of metadata is:

> *Information that helps the user assess the usefulness of a dataset relative to their problem*.

Any measure of dataset quality can only be relative to its intended use. However it is impossible to predict every possible future use.

3. Recommendations for User-focussed Metadata

1. Socio-political context of data creation: actors and their influence

By examining the negotiation and discussion within the project documentation it is possible to identify the major actors and the nature of the influence they exert over the project. Comber et al. (2003) applied this approach to provide insights and to reveal fundamental differences between different land cover mappings in the UK in terms of the different socio-political context of the data creation.

2. Critiques of the data: academic papers

Academic papers could either be in the form of a critique of the data or describe their application to a specific problem. They would provide an independent opinion of the quality and fitness for use.

3. Data producers opinions: class separability

The opinions of the data producers on how separable classes are allow informed assessments of data quality to be made. Comber et al (2004a, 2004b) and Fritz and See (2005; See and Fritz, 2006) have applied such descriptions of class separability as weights for assessing data quality for assessing internal data inconsistency.

4. Expert opinions: relations to other datasets

Experts, familiar with the data, through experience of applying it in their analysis, can provide measures of how well the concepts or classes in one dataset relate to those of another. Comber et al. (2004a, 2004b) applied this approach to determine whether differences between different datasets were due to data inconsistencies (i.e. different specifications) or due to actual changes in the features being recorded. Expert opinions of how datasets relate have also been used to identify relative data inconsistencies for global land cover data (Fritz and See, 2005; See and Fritz, 2006) and for international soil classifications (Zhu et al., 2001).

5. Experiential metadata

Users could provide feedback about their experience of using the data. This could be from an application or disciplinary perspective in order to describe positive and negative experiences in using the data. Possible solutions are a metadata wiki and a system for use case logging where the data use was monitored via a web portal. User experience would provide independent opinions of data quality and fitness, would allow different user communities to be differentiated and provide a framework within which new potential data users could learn form the experience of others.

6. Free text descriptions from producers

The existing and emerging metadata standards include elements for free text slots – "Descriptions" in the Dublin Core and "Generic" and "Extra" in the NERC DataGrid specifications. Currently these are not extensively used. Wadsworth et al (2005, 2006, in submission) have concluded that free-form *descriptions* of classes longer than about 100 words provides sufficient information to be processed and used by someone unfamiliar with the epistemology, ontology and semantics of the data.

7. Tools for mining free text metadata slots

In order to identify suitable data, of a phenomenon that may not be familiar to the user, tools are needed to assist them make sensible and appropriate selections over their data choices. If free text slots are populated then novel approaches to metadata mining and analysis are needed. Wadsworth et al. (2005, 2006, submitted) and Comber et al. (submitted) have shown how simple text mining analyses can be used to generate measures of semantic and conceptual overlap between different datasets and different classes. The inclusion of free text descriptions of the data, coupled with text mining tools would allow users to identify consistencies and inconsistencies between the user and the data concepts.

References

COMBER, A., FISHER, P., WADSWORTH, R., (2003) Actor Network Theory: a suitable framework to understand how land cover mapping projects develop? *Land Use Policy*, 20: 299–309.

COMBER, A.J., FISHER, P.F., WADSWORTH, R.A., (2004a). Assessment of a Semantic Statistical Approach to Detecting Land Cover Change Using Inconsistent Data Sets. *Photogrammetric Engineering and Remote Sensing*, 70(8): 931-938.

COMBER, A., FISHER, P., WADSWORTH, R., (2004b). Integrating land cover data with different ontologies: identifying change from inconsistency. *International Journal of Geographical Information Science*, 18(7): 691-708.

COMBER, A.J., FISHER, P.F., WADSWORTH, R.A., (2005). You know what land cover is but does anyone else?…an investigation into semantic and ontological confusion. *International Journal of Remote Sensing*, 26 (1): 223-228

COMBER. A.J., FISHER, P.F., WADSWORTH, R.A (submitted). Using semantics to clarify the conceptual confusion between *land cover* and *land use*: the example of 'forest'. Paper submitted to *International Journal of Land Use Science*

FGDC (Federal Geographic Data Committee). 1998. *Content Standard for Digital Geospatial Metadata*, FGDC-STD-001-1998, National Technical Information Service, Computer Products Office, Springfield, Virginia, USA.

FRITZ, S., AND SEE, L., 2005. Comparison of land cover maps using fuzzy agreement. *International Journal of Geographical Information Science* 19 (7), 787-807

WADSWORTH R.A., FISHER P.F., COMBER A., GEORGE C., GERARD F. & BALTZER H. 2005. Use of Quantified Conceptual Overlaps to Reconcile Inconsistent Data Sets. Session 13 Conceptual and cognitive representation. *Proceedings of GIS Planet 2005*, Estoril Portugal 30th May - 2nd June 2005. ISBN 972-97367-5-8. 13pp

WADSWORTH R.A, COMBER A.J., & FISHER P.F., (2006). Expert knowledge and embedded knowledge: or why long rambling class descriptions are useful. pp 197 – 213 in *Progress in Sptial Data Handling, Proceedings of SDH 2006*, (eds. Andreas Riedl, Wolfgang Kainz, Gregory Elmes), Springer Berlin.

WADSWORTH R.A, COMBER A.J., & FISHER P.F., (SUBMITTED). The Application of Simple Text Mining Techniques to Physical Geography. Paper submitted to the *Journal of Environmental Management*.

Zhu, A. X., HUDSON, B., BURT, J., LUBICH, K. AND SIMONSON, D. (2001.) Soil Mapping Using GIS, Expert Knowledge, and Fuzzy Logic. *Soil Science Society of America Journal* 65:1463-1472

Biography

Lex Comber gained his PhD from the Macaulay Institute and the University of Aberdeen in 2001. Up to 2003 he worked as an RA on the EU REVIGIS project developing methods for integrating semantically discordant data. After a year in GIS consultancy with ADAS, Lex took up a lectureship at the University of Leicester where he now directs the MSc in GIS.