Semantics, Metadata, Geographical Information and Users

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

[1]Department of Geography, University of Leicester, Leicester, UK. Tel +44 (0)116 252 3812 Fax +44 (0)116 252 3854 E-mail: ajc36@le.ac.uk, pff1@le.ac.uk
[2]CEH Lancaster, Bailrigg, Lancaster, LA1 4AP, UK. E-mail: rawad@ceh.ac.uk

* Corresponding author

1. Semantics and Geographical Information

Semantics is concerned with analysing the meaning encoded in language (Calvani, 2004). Within a technical description of data, semantic descriptions ought to be an important adjunct, filling out the labels and codings of classes and providing justification for measurements. Semantics is equally applicable whether applied to single word labels (Building, Tree etc.), short phrases (coniferous forest, upland moors, etc.), or to longer textual descriptions of a phenomenon. Data semantics also includes the general description of a dataset and its characteristics and limitations.

Spatial data and their semantics vary for a variety of reasons that are not to do with differences in the feature being measured. In the creation of any spatial data there are a series of choices about what to map and how to map it which will depend on a range of commissioning and institutional factors. Different choices result in different representations and variation between datasets. The variability between different, but equally valid, mappings of the same real world objects ultimately points to the social construction of spatial data (Harvey and Chrisman, 1998). Much valuable geographical information is therefore embedded in its semantics.

2. Metadata and semantics

ISO 19115 (ISO, 2003a) describe metadata as "*Data about data or a service. Metadata is the documentation of data. In human-readable form, it has primarily been used as information to enable the manager or user to understand, compare and interchange the content of the described data set*". It is clear that the semantics of a dataset are a legitimate area which might be considered by metadata and semantics are part of the metadata standards corpus, but they are treated very differently in different domains within and between standards agencies/groups. In the domain of spatial information semantics are poorly treated by metadata and data standards.

Metadata standards are primarily concerned with the 'discovery' of data, they therefore describe where it is and in what form, rather like a library catalogue tells you where a book is but not whether it is worth reading. Although metadata standards are often flexible enough to contain all sorts of descriptive elements, the proscriptive elements on 'content' are usually related to 'accuracy'. Typically, metadata for spatial data include descriptions of data quality in terms of the Positional Accuracy, Attribute Accuracy, Logical Consistency, Completeness, and Lineage. These were first suggested in the

Proposed Standard for Digital Cartographic Data (DCDSTF, 1988), and are included in many standards for spatial data quality and metadata reporting since (FGDC, 1998; ANZLIC, 2001; ISO, 2003a, 2003b). The specification of metadata standards for describing the components and character of information sources in general have been distilled into the Dublin Core (DCMI Usage Board, 2006). The Dublin Core Metadata Element Set contains 15 elements. No element relates to quality, information content (although this could be included in "description") or semantics. The availability of data for access over massively networked computer resources such as Spatial Data Infrastructures (SDIs) and the GRID has led to concern that metadata as currently specified may not provide enough information for informed data use (e.g. Comber et al, 2005a; Goodchild, 2006; Schuurman and Leszczynski, 2006).

The specification of standards for metadata is useful because in theory they provide a common framework, enabling parties to exchange data without misunderstandings. There two problems with current metadata as specified by standards. First, metadata specification is always a compromise and necessarily lags behind research activity and sometimes industrial practice. For example a recent book on spatial data standards took 10 years from inception to being published (Moellering, 2005). Second, they are grounded in data production rather than being focused on use or usability. There is no mechanism within current metadata to ensure that the specification of the conceptual model, including the semantics, is understood and shared. An example of this, which marks a retreat from the intention of metadata to describe fitness for use, is provided by the recent INSPIRE draft rules for metadata. INSPIRE is the EU SDI and the draft explicitly states: *'Attempts to objectively rate (and publish in metadata) the "usefulness" of a service, such as that it produces correct responses or behaviours, will almost certainly create problems among service vendors, and would likely do more harm than good to consumers. Most other markets rely on informal user feedback as the ultimate test as to whether or not a product or service is useful, a good value, etc. This feedback appears spontaneously in news and mail forums, in the popular press, and by word-of-mouth'* (INSPIRE, 2007, p. 17). The net result of these static standards is that users do not know how to relate data quality measures to their analyses and have trouble assessing the suitability of the data for their application (Hunter, 2001), or may not even be given the reports. In spite of the declared intention that metadata assist users in defining the fitness of a dataset for their application, the standards in general, and data quality and semantic descriptions in particular, are not easy to relate to use.

3. User focused extensions to metadata

As an alternative definition to metadata being "data about data", a user-focussed definition of metadata might be: *Information that helps the user assess the usefulness of a dataset relative to their problem*. In this definition metadata is not static information but is concerned with whether the data can address the task in hand. Many of the issues in data integration are concerned with how to relate one view of the world, as encapsulated by a particular dataset, to another. The GI community has looked to the ontological research community to provide standards for data catalogues (e.g. through OWL). But different standards support semantics in different ways and some standards for mapping

originate from other academic areas – i.e. users are developing *de facto* standards, and proposing *de jure* standards. There are a number of ways that metadata could be made more relevant to data users that were identified during a metadata workshop held at National Institute for Environmental eScience in the summer of 2005.

i). Descriptions of the socio-political context of data creation. Documents such as interim reports and minutes from steering group meetings describe the process of negotiating data specifications. Data commissioning includes a legitimising activity that involves the major data users, agencies and NGOs in ensuring that the product specification fulfils their policy requirements.

ii) Critiques of the data from academic and industrial papers. Academic or practical journal papers, magazine articles and technical reports which describe or critique uses of a dataset in particular contexts are a form of metadata. For users wishing to identify the suitability of any particular dataset for their problem, it would be useful to be directed to these papers as they provide an independent opinion of the data quality and fitness.

iii) Data producers' opinions of class separability. The data producer opinions on the separability of allows informed and dynamic assessments of data quality (i.e. fitness for the intended use) to be made.

iv) Expert opinions of relations to other datasets. Experts, familiar with the data can provide measures of how well the concepts or classes in one dataset relate to those of another. This generates measures of (external) data inconsistency which can be used as weights for applications.

v) Experiential metadata. Feedback from users about their experiences of using the data, either organised from an application or disciplinary perspective, would describe positive and negative experiences in using the data. The experience of other users would provide independent opinions of data quality and fitness for use.

vi) Free text mining of descriptions from producers. The existing and emerging metadata standards include elements for free text slots, for example the *Descriptions* in the Dublin Core specifications. If these are populated (they are not) with either producer or user community perspectives then they can be text mined.

4. Concluding comments

The proposals for the extension of metadata put forward in this editorial will not be novel to many in the GI community. Currently researchers use such information in a *de facto* way to overcome the semantic gap in current metadata specifications. Our argument is that as the number of users of spatial data increases e.g. through SDIs, there will be a need for semantic information about the data to be formally linked to it. We believe that what is considered to be metadata and even its specifications in standards should be expanded to accommodate the informal, *de facto* metadata currently that is being used. Of the six proposals above all relate to semantics and nearly all have been applied operationally in order to generate a better understanding of some dataset. Comber et al. (2003) analysed the socio-political context of data creation to better understand discordant mappings of land cover in the UK, their different socio-political contexts and their influence on data conceptualisations. Comber et al. (2004) and Fritz and See (2005; See and Fritz, 2006) have applied data producers descriptions of internal class separability as weights for assessing data quality and internal data inconsistency. Comber

et al. (2004) used expert opinions of how one dataset related to another to determine whether variations between different datasets were due to data inconsistencies (i.e. alternative 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). Wadsworth et al. (2006, forthcoming) have mined free text descriptions provided by data producers to identify overlaps between classes and datasets, providing information which is helpful to users who are unfamiliar with the data. In all of the cases above, some understanding of (and analysis of) the data semantics helped the user to better understand the relationships between the classes and other datasets

The need for semantics in to be included in metadata derives from the increasing distance between users and producers. Distributed computer architectures such as SDIs obviate the need for user and producers to communicate directly. For the user, the process of dialogue with the producer before obtaining a dataset is removed and the data producer can no longer prevent inappropriate use of their data. The survey memoir has been replaced short cryptic metadata statements that relate to production rather than understanding or meaning. Current metadata paradigms reflect the position articulated by Goodchild (2006: p690) that computers "replace the extended and often confused process by which we learn the meanings of terms and languages with precise, instantaneous translators". The typical data user is left in the paradoxical situation that on the one hand they have easier access to more data than ever before via SDIs, but on the other hand they know less about the meaning behind that data. This is analogous to the hoary joke "what is a lecture": *A lecture is the process whereby the notes of the lecturer are transferred to the notebooks of the students without going through the brain of either.* For these reasons Schuurman and Leszczynski (2006) and Comber *et al.* (2005b) have argued that metadata ought to include more than documentation of the technical aspects of data production. We hope that the proposal outlined in this editorial go some way to addressing this issue.

References

ANZLIC 2001 *Metadata Guidance: core metadata elements for geographic information in Australia and New Zealand*, Griffith ACT, Australia

Calvani, D 2004 Between interpreting and cultures: a community interpreters toolkit http://wwwaucegyptedu/academic/interpreters/documents/ManualforCommunityInterpreterspdf_000pdf [available 12 February 2007]

Comber, A, Fisher, P, and Wadsworth, R 2003 Actor Network Theory: a suitable framework to understand how land cover mapping projects develop? *Land Use Policy*, 20: 299–309

Comber, AJ, Fisher, P, and Wadsworth, R 2004 Integrating land cover data with different ontologies: identifying change from inconsistency. *International Journal of Geographical Information Science*, 18(7): 691-708

Comber, AJ, Fisher, PF, Wadsworth, RA 2005b 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

DCDSTF (Digital Cartographic Data Standards Task Force) 1988 The proposed standard for digital cartographic data. *American Cartographer* 15 (1): 9-140

DCMI Usage Board, 2006 DCMI Metadata Terms, http://dublincoreorg/documents/2006/12/18/dcmi-terms/

FGDC (Federal Geographic Data Committee), 1998 Content Standard for Digital Geospatial Metadata, FGDC-STD-001-1998, Reston, Virginia http://wwwfgdcgov/standards/projects/FGDC-standards-projects/metadata/base-metadata/v2_0698pdf , accessed 12 February 2007

Fritz, S, and See, L, 2005 Comparison of land cover maps using fuzzy agreement. *International Journal of Geographical Information Science* 19 (7): 787-807

Goodchild, MF 2006 GIScience Ten Years After Ground Truth. *Transactions in GIS*, 10(5): 687-692

Harvey, F, and Chrisman, N, 1998 Boundary objects and the social construction of GIS technology. *Environment and Planning A* 30: 1683-1694

Hunter, GJ (2001) Spatial Data Quality Revisited. In *Proceedings of GeoInfo 2001*, 04–05 October, Rio de Janeiro, Brazil, 1–7

INSPIRE (Infrastructure for Spatial Information in Europe), 2007 *Draft Implementing Rules Metadata* http://inspirejrcit/reports/ImplementingRules/draftINSPIREMetadataIRv2_20070202pdf, accessed 20 March 2007

ISO, 2003a ISO 19115 *Geographical Information – Metadata*, International Standards Organisation, Geneva

ISO, 2003b ISO 19114 *Geographical Information – Data Quality Principles*, International Standards Organisation, Geneva

Moellering, M (ed) 2005 *World Spatial Metadata Standards: Scientific and Technical Characteristics, and Full Descriptions with Crosstable*. International Cartographic Association / Pergamon.

Schuurman, N and Leszczynski, A 2006. Ontology-Based Metadata. *Transactions in GIS*, 10(5): 709-726.

See, L and Fritz, S 2006 Towards a global hybrid land cover map for the year 2000. *IEEE Transactions on Geosciences and Remote Sensing*, 44(7): 1740-1746

Wadsworth RA, Comber AJ, and Fisher PF, 2006 Expert knowledge and embedded knowledge: or why long rambling class descriptions are useful. In *Progress in Sptial Data Handling, Proceedings of SDH 2006*, (eds Andreas Riedl, Wolfgang Kainz, Gregory Elmes), Springer Berlin: 197 – 213

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