User-focused metadata for spatial data, geographical information 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. Tel +44 (0)116 252 3812 Fax +44 (0)116 252 3854 E-mail: ajc36@le.ac.uk
[2.] The giCentre, Department of Information Science, City University, London, UK e-mail: pff1@city.ac.uk
[3.] CEH Monks Wood, Abbots Ripton, Cambridgeshire PE28 2LS, UK. E-mail: rawad@ceh.ac.uk

ABSTRACT

This paper suggest a number of ways that metadata for spatial data could be expanded and activated to make it more user relevant. We review current metadata standards and specifications and show that they are grounded in data production. They do not provide information to help users assess the relative uncertainties associated with using data for their application. As a result of the recent EU REVIGIS project and a metadata workshop at the National Institute for Environmental eScience, we suggest ways in which metadata could be re-focussed towards the operational use of data. These include:

- Socio-political context of data creation, actors and their influence;
- Critiques of the data such as academic papers;
- Data producers opinions of class separability;
- Expert opinions of relations to other datasets;
- Experiential metadata;
- Free text descriptions from producers;
- Tools for mining free text metadata.

The objective of activating and expanding metadata in this way two-fold. First, to close the gap between users and producers that are emerging as a result of spatial data infrastructure initiatives like the INSPIRE which eliminate the need for dialogue between the user and data producer. Second, to provide metadata and tools for users so that they are able to assess the spatial data quality relative to their intended application.

1. INTRODUCTION

This paper pulls together a number of research strands that emerged and were developed in the course addressing a data revision and integration problem during recent EU research project[1]. In the course of exploring approaches specific to the integration of discordant land cover data, we realised that many of our findings and the issues we were addressing were generic to spatial data. More importantly many of the conclusions we reached and the future research directions we indicated (some of which we have since pursued) applied to spatial data generally and the reporting of metadata. Many of these issues are increasingly relevant to the efficient development of emerging spatial data infrastructures and to the increased distance they place between data users and data producers.

We start by introducing the generic problem of data discordance: each time we go out and measure the world we do so in different ways. This results in data that report the same features (e.g. land cover, soils or

[1] http://www.lsis.org/REVIGIS/

geology) but which are difficult to combine due to differences in recording, measurement and conceptualisation (Section 2). In Section 3 the current specifications of metadata for spatial data are reviewed. Current metadata standards can be characterised as being static and grounded in the production of spatial data. We describe the decline of user focussed metadata in the form of the survey memoir and its implications for users unfamiliar with the data. Because of this we propose that metadata be re-focussed towards the operational use of data, specifically to allow users determine the 'usefulness' of the data for their task in hand. We note that measures of data quality in are relative rather than absolute (Section 4). The core of this paper (Section 5) proposes a series of directions for the expansion and refocusing of metadata. These have arisen from the REVIGIS work of the authors, including a subsequent workshop on metadata for spatial data at the National Institute for Environmental eScience[2]. We discuss the need for metadata to focus on users and use within a context of increased user access to spatial data under a range of policy and infrastructural initiatives (Section 6) before presenting some conclusions.

2. GENERIC PROBLEM: 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. This is in the Descartian tradition where greater certainty and understanding are sought through the application of scientific reason in an objective manner in order to better explain and predict. New technologies, improved techniques and changes in the understanding of the phenomenon offer greater insight into the process under investigation.

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). Differences over representation choices results in different types and directions of information loss and uncertainty. The result is variability between different, but equally valid, mappings of the same real world objects and the social construction of spatial data due to technology institutions and policy and explains why discordance in spatial data and geographical information are endemic.

In previous work we have given many examples of this data discord, much of it relating to land cover in the UK (see Comber et al., 2003; 2004a; 2004b; 2005a; 2005b). However the problem is by no means unique to land cover. The US Soil Conservation Service most of the 1938 classification was revised through a series of approximations culminating in the 1975 Soil Taxonomy which divided the world's soils into 10 soil orders. This itself was revised in 1999 into 12 soil orders.

[2] http://archive.niees.ac.uk/cgi-bin/search.cgi?event_id=metadata2

As well as complete changes in the way that a particular phenomenon is conceived (such as land cover) specific features are also subject to the same process. For example, the 1990 and 2000 land cover maps of the UK produced by the same organisation, show a profound shift in the conceptualisation of the class of "Bog". In 1990 the land cover of bog was defined by the presence of "permanent waterlogging, resulting in depositions of acidic peat, mostly herbaceous communities of wetlands with permanent or temporary standing water"[3]. In 2000 bog was defined as "ericaceous, herbaceous and mossy vegetation in areas with peat >0.5 m deep"[4]. Comber et al (2005c) have shown the potential implications of this situation of 2 datasets mapping the class of "Bog" for a 100km x 100km area: in 1990 there were 12 pixels (<1 ha) of bog and in 2000 some 120,728 pixels (75 km^2). To the informed (and initiated) users this represents an unreasonable change due to the ecological inertia of bog plant communities, however as will be described later, there are many more users of spatial data and access to the data is no longer mediated through a gatekeeper ensuring appropriate use.

Over time scientific knowledge advances, policy objectives change and technology develops. This contributes to a common problem in the environmental sciences where most surveys create a new "base line" rather than being part of a sequence. They describe the features of interest in different ways even if the phenomenon has not changed. For some of these such as solid geology differences between "maps" (generally) represent changes in understanding, technology or objectives. For other phenomena, like land cover, such differences may also be because the phenomenon has changed in significant and interesting ways.

3. METADATA

3.1 Metadata, standards and data quality reporting

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. It aims are[5]:

- to help organize and maintain an organization's internal investment in spatial data;
- to provide information about an organization's data holdings to data catalogues, clearinghouses, and brokerages;
- to provide information to process and interpret data received through a transfer from an external source.

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 (e.g. STDS, 1998; FGDC, 1998) and the interested reader is directed to Guptill and Morrison (1995) and Veregin (1998).

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

[3] http://science.ceh.ac.uk/data/lcm/classM.htm
[4] http://www.ceh.ac.uk/sections/seo/dataspec.html
[5] http://www.fgdc.gov/dataandservices/fgdcmeta/

converging to differing degrees (e.g. OGC, Dublin Core and ISO). All of them adopt the ISO 19115 definition of 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. In the Web Services context, XML-encoded (machine-readable and human-readable) metadata stored in catalogs and registries enables services to use those catalogs and registries to find data and services*"[6]

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. For instance the Dublin Core Metadata Element Set[7] contains 15 elements (Contributor, Coverage, Creator, Date, Description, Format, Identifier, Language, Publisher, Relation, Rights, Source, Subject, Title, Type) each to be specified with a minimum set of attributes (Name, Uniform Resource Identifier, Label, Definition, Type of Term, Status, Date Issued).

By way of an example consider the metadata specification for the NERC DataGrid in the UK which specifies 5 types of metadata rather than one schema:

- A: Usage metadata to provide information about the data needed by the processing and visualisation services;
- B: Generic complete metadata to provide high-level statement of entities and the relations between them and to summarise the data model in a user friendly form;
- C: Metadata generated to describe documentations and annotations;
- D: Discovery metadata to be created by data producers, conforming to ISO19115;
- E: Extra discipline specific information

The argument is that communities of users will identify portions of the metadata that are appropriate to their domain, will populate slots where necessary (using the catchall E if necessary). This view of data and user communities is similar to that put forward by the OGC (OGC refers to them as 'information communities' see Visser et al., 2000). There are two general problems with approaches based on interest or use-communities. First, they are not only diverse (witness the range of users and uses of land cover data – see Comber et al., 2005a, 2005b), but they are difficult to define and identify. Visser et al (2000) note that communication between different information communities can be very difficult because they do not agree on common conceptualisations. If you enter "land cover" in the NERC metadata gateway[8] then you get only a fraction of the data that NERC funded (LCMGB but not LCM2000, CS1990 but not CS1984 or CS2000 etc.). The land cover community within NERC may be rather reticent, but, it raises the question of who is part of that community and how do you determine degree of membership to a community. On the other hand looking at the thread of e-mails from the "CF community"[9] it is obvious that the community is advanced, but comparing threads for "omega" and the "shape of the earth" some threads are trying to get technical agreement on a physical measurement whilst others are concerned with what to include or exclude. Second, if one accepts that it is possible to identify various user communities, any taxonomy of metadata (as A, B, C, etc) may have to be populated *N* times for the *N* different communities who may want to use the data, however this is not straightforward as within one community (e.g. land cover in the

[6] http://www.opengeospatial.org/resource/glossary/m
[7] http://dublincore.org/documents/dcmi-terms/
[8] http://www.nmp.rl.ac.uk/
[9] http://www.cgd.ucar.edu/mailman/listinfo/cf-metadata

point above) there may be issues of meta-semantics (what community do I think I belong. This is notwithstanding that it is only possible to interpret the metadata in B (descriptions of the "generic complete metadata, semantic including syntactic, not including discipline specific") if the specification of the conceptual model is understood. Implicit in this view of metadata, standards and communities is that identifying the 'universe of discourse' is unproblematic. The theory is that from an agreed universe of discourse encoding protocols, application schemas, conceptual modelling and feature catalogues – each with their own ISO standard – can be developed as a process for within community governance of semantics. However defining a universe consistently is problematic for many within the GI community, which recognises the deeply socially constructed nature of much geographic information.

3.2 Metadata and data quality measures are grounded in data production

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. More pertinently they have the following characteristics:

- They are focussed on aspects relating to data production and data mediation rather than the use of the data;
  They are passive, rather than active descriptions relating to potential applications;
- They record the easily measurable aspects of data rather than the most pertinent aspects of the data;
- They provide overall or global measures of data quality, not ones that relate to individual map objects;
- They do not communicate the producer's wider knowledge of the data;
- They are difficult for users to interpret in relation to a specific application

The data quality parameters reported in metadata do not 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., 2005b) rather than more fully communicating the producer's knowledge of the data. The net result is that users do not know how to apply data quality measures in their analyses and cannot assess the suitability of the data for their application (Hunter, 2001). Metadata should allow the user to determine data quality and fitness for their analysis or data integration activity. Necessarily this involves relating one view of the world, as captured within the specification of a particular dataset, to another such as the objective of the intended analysis.

Current metadata reporting does not give any information to the user about how to best exploit the data. Rather they reflect data production and a cartographic legacy. Comber et al., 2005b) have commented that data quality measures originate from the data production process, and reflect the stages involved in data creation: methods and data sources define *lineage* and assessment of results defines *accuracy*, *consistency* and *completeness*. Most data quality measures derive from the historical cartographic legacy of geospatial data production, where the need is to report the extent to which database objects generalise complex information for the purposes of display (Fisher, 1998). Data quality measures can also be seen in the context of "economic game theory" as the user has imperfect knowledge of how much effort the producer puts in to data quality reporting. The standard measures of data quality are a signal of an accepted paradigm that the producer has put in the required data quality reporting effort. The net result is two-fold. First, data quality information communicates overall measures that do not relate to individual database objects. They implicitly assume that data quality is evenly distributed across the entire dataset and local variations in quality are unreported. Secondly, measures of data quality are difficult for users to interpret in relation to a specific application. In order to assess data quality (a relative concept) users need to relate

their intended use of the data with the semantic and ontological concepts of the data – the concept of "semantic accuracy" as introduced by Salgé (1995). This concept of quality has had little impact upon the standards process and data quality specifications (Fisher, 2003). This is because the specification of data quality standards, previously dominated by the national mapping agencies and software companies, has now become the preserve of dedicated standards and industry specifications organisations such as ISO and the Open GIS Consortium.

4. UNDERSTANDING DATA UNCERTAINTY AND DATA QUALITY

4.1 User-focused metadata

There has been a decline in metadata reporting of spatial data or mapped information. For example contrast the current situation as documented above with the land use surveys of Stamp in the 1930s and Coleman in the 1960s: each of Stamp's maps was accompanied with a book that described what it was that being mapped, the precise descriptions of the concepts and objects that were identified. The implications the lack of current metadata relating to use are profound: users will assuming they understand what it is that is being mapped in the fullest sense whilst in reality they may not understand the meaning of the information they incorporate into their analysis. For instance, we all have a notion of the prototypic 'forest' and we may assume that what is mapped as forest matches our conceptualisations (Comber et al., 2005a). The more familiar the class names and labels the less we think about them. Fisher (2003) suggests that this problem is getting worse and that before the widespread introduction of computers geographic phenomenon like land cover or soils were described by experts (often in state sponsored organisations) in extensive memoirs and maps were used to *support* the written description. For scientists, the actual map was almost incidental to the recording methods and class descriptions (ontologies and conceptualisations) documented in the books. This metadata embedded in paper reports allowed users (as well as extensive dialogue with the data producer) to understand the wider meaning of the data. Now it is just the map that is wanted, often to be immediately downloaded and incorporated inside an application

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.*

In this definition metadata is not information relating to the "usability" of the data as anything is useable but, whether the data is "useful" for the task in hand. Many of the issues of data integration are concerned with how to relate one view of the world, as encapsulated by a particular dataset, to another. In the context of the above statements, I would like metadata should include information that helps users to:

- Assess the suitability of data for their use;
- Integrate the data (spatially and thematically);
- Understand the limitations and uncertainties of any integrating activities.

4.2 Uncertainty and data quality are relative

Most of the organisations concerned with specifying metadata standards argue that metadata is already user focussed. However their perceptions of and frameworks for metadata do not support the heavily socially constructed nature of much geographical information, natural resource inventories and spatial data. Users have to be pragmatic and use the data that is available (there is often a monopolistic supply) despite the fact that the data will usually have been collected for a different purpose (and those purposes will change). At the same time data producers don't communicate all they know about the data how it links to their especially their "model" of the world. Users are encouraged by this scenario to assume that

the data classes match their prototypic ones and thus to treat heavily constructed information (an interpretation) as if it were data (an objective measurement).

The uncertainties in geographic information originate from three general different sources (Comber et al., 2006):

- Cartographic uncertainty resulting in positional and attribute errors;
- Uncertainty due to the inherent nature of geography: different interpretations can be equally valid;
- Conceptual uncertainty as a result of differences in 'what it is that is being mapped'.

Conventional approaches to metadata reporting to facilitate data quality assessments including uncertainty modelling have addressed the first of these. However as the distance between users and producers grows there is a need to consider the impact of uncertainties due to different conceptualisations and interpretations. The metadata standards embodied by the FGDC and by more recent developments such as the Dublin Core are focussed on static metadata elements.

Any measure of dataset quality can only be relative to its intended use. For any given application variations in the way that bog is defined as described above, may or may not be important: important if the data was to be used to monitor changes in upland habitat character but unimportant if the data was to be used to create and upland / lowland mask.

This relative and non-static view of data quality reporting underpins the need for metadata data to support assessments of data usefulness of rather than usability. However it is impossible to predict every possible future use. For these reasons Chrisman, in 1988, advised on the need for user experience of the data be included in the specification of data quality (DCDSTF, 1988) but these recommendations were omitted from final Spatial Data Transfer Standards specification (FGDC, 1998), an omission which has been propagated through most standards specifications.

The provision of user-friendly, user orientated, dynamic metadata for the creation of relative measures data quality leads to a number of unresolved questions:

- How to provide appropriate information about the data such that users can make informed decisions about which data is most suited to their analysis?
- How to enable users to understand the limitations of the results of any analysis using that data?
- How do we link uncertainty and data quality assessments to do this?

In the next section the outputs relating to metadata and data quality assessments from the REVIGIS project and of a workshop on metadata for spatial data at the National Institute for Environmental eScience are reported. These outputs identify the need for tools to populate and to mine metadata data for relative measures of data quality.

5. RECOMMENDATIONS

As a preamble to this section on recommendations for activating metadata it is useful to summarise the arguments that have already been articulated. Discord between datasets is endemic, even for repeat surveys, as a result of changes in scientific understanding, technological advances and shifts in political objectives. This means that it is difficult to talk about one dataset being 'better' than another in any absolute sense. Rather measures of data quality are relative to the question in hand (it may be more appropriate to investigate a particular hypothesis using *Dataset_A* rather than *Dataset_B*) as are any associated uncertainties. Currently metadata specifications are concerned with the production of static descriptions pertaining to some quality of the data (*NB* this is **not** data quality). This does not help users assess whether that data is suitable for their problem. There are a number of ways that metadata could be more relevant to data users. These have been identified in the course of the REVIGIS work and during the

Activating Metadata workshop held at NIEeS in the summer of 2005. First, by expanding what we consider to be metadata. Second by developing tools that will activate the as yet unused metadata slots / elements already included in metadata specifications. Third by developing novel approaches to metadata mining and analysis. All of these approaches have been implemented by the authors (as well as other workers) and have been shown to be effective.

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

Documentation such as interim reports and records from steering group meetings contain information describing the process of negotiation amongst individuals and institutions involved in data specification. Data commissioning includes a legitimising activity that involves the major data users such as environment agencies and NGOs to ensure that the product specification fulfils their requirements. 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. Martin (2000) shows how actors may be grouped into categories relating to their influence Input Resources, Accountable Actors and Recipients. The links between these groups of actors can be described in terms of the influence they exert on the process of product specification. Influences include Control, Skills/Abilities, Money and Information – these are all factors that relatively easy to identify. 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

For users wishing to identify the uncertainties associated with using a particular dataset for a specific problem it would be useful to be directed to academic papers that describe the dataset. These papers could either be in the form of a critique of the data or describe their application to a specific problem. Academic papers would provide an independent opinion of the quality and fitness for use. For example, as well as the papers published by the authors describing the uncertainties associated with using different UK land cover maps for an (application) change mapping, Robinson et al. (2005) provide a critique of the data quality.

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 theose of another. This allows measures of (external) data inconsistency to be generated which can be used as weights for applications. 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). In these cases the expert's opinion of the nature and relationships between the classes improves our own understanding of those classes and of the data.

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. Slots (elements) in metadata taxonomies exist to hold such information and it may be that different user communities require different slots. There may be legal implications for such information (especially negative comments) that have to be overcome and there is a question of whether the holder of such metadata needs to be thought of being independent by all – an independent metadata brokerage. 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 (described in Section 3) include elements for free text slots – "Descriptions" in the Dublin Core and "Generic" and "Extra" in the NERC DataGrid specifications. These are to be populated with either producer or user community perspectives. Currently these are not extensively used. Historically, information describing the full complexity of a dataset was published in a survey memoir. The map or dataset was seen as a window onto the wealth of information contained in the report. In respect to the dataset such information is metadata and is supported by further levels of metadata in survey manuals and disciplinary textbooks. 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. Descriptions do not have to use a single structured format or rely on specific agreed nomenclatures. It may be that text mining tools are needed to populate such metadata elements.

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 (2005, 2006, submitted) and Comber (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. Jeansoulin and Wilson (2002) describe the problem of matching user needs to data specifications as relating the "problem ontology" (constraints, definitions) to the "product ontology" (data specifications, data quality). 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.

6. Discussion

The relevance of the proposals suggested for expanding and activating metadata is provided by three contexts: the increased importance of spatial data and geographical information, the increased numbers of data users and the increased distance between user and the data producers.

Importance of spatial data and geographical information

A recent NIEeS workshop on metadata concluded that the case for current metadata standards (FGDC, ISO, OGC, etc) is not yet proven in relation to geographical information. This is in part because of the institutionally orientated and socially constructed nature of much geographical information (Harvey and Chrisman, 1998; Comber et al. 2003; Comber et al, 2005a), but also because of changing context for the use of geographical information: there are many more applications and analyses using incorporating geographical information, there are many more users and there are many more sources spatial data

available to the user (Comber et al. 2005b). The number of users is predicted to increase further. A 2004 Department of Labor 2004 report stated that a "worldwide market for geospatial technologies, estimated at $5 billion in 2001, is expected to have annual revenues of $30 billion by 2005". A recent article in Nature identified GIS technologies as "one of the three most important emerging and evolving fields along with biotechnology and nanotechnology" (Gewin, 2004). Given that many different representations of the same real world feature are possible and equally valid it is important that users understand the data they integrate into their analyses. However, very often the real world objects in geographical information and spatial data are poorly defined. For instance vegetation mapping suffers because it is not always possible to separate attribute (thematic) accuracy from spatial accuracy. Uncertainty arises both from the separation of 2 vegetation types as well locating the boundary between them.

Spatial data use and the user

There are a number of initiatives that promote increased access to and usage of spatial data by new communities of users originating from different areas policy and manifesting themselves at different levels in the political hierarchy. These include the Strategic Environmental Assessment directive (2001/42/EC), the Environmental Impact Assessment directive (85/337/EEC) and the Aarhus Convention on Access to Information. These initiatives promote deeper involvement of the public and NGOs in environmental decision making processes. There is an increasing awareness amongst publicly funded environmental and conservation agencies of the need to be able to provide information about the environment. These are being augmented by a number of recent policy initiatives that promote notions of "environmental justice" by facilitating public and stakeholder access to environmental data. In the EU initiatives such as INSPIRE (Infrastructure for Spatial Information in Europe) and the UK Research Councils through e-science and Grid initiatives are encouraging seamless access to data over the Internet. Comber et al. (2005a) noted a number of aspects relating to the expansion of the spatial data user base and the demise of the traditional monograph:

- Users will not understand the precise *meaning* of the data;
- Users will accept that the land information presented is appropriate for their analysis;
- Users will treat (derived) information as fact.

This reflects the shift in the "balance of power" between the graphical (map) and textual (monograph). Crucially the user may be led to wrongly treat the map as if it were data (a measurement) and not as information (an interpretation).

Greater distance between the user and the data producer

Producers and consumers of data have become far removed from each other as a result of spatial data infrastructure initiatives like the INSPIRE. These are designed to increase the re-use of all types of data and information. However one of the consequences is to decrease the interaction between producer and user and to increase the distance between them. Essentially through increasing user access to spatial data these initiatives eliminate the need for dialogue with the data producer. 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, but on the other hand they know less about the meaning behind that data. Compliance with modern data quality and metadata standards could, in theory, help users become acquainted with the meaning in the data, but in practice they offer the user little information about the meaning behind the categories (classes) and especially the relationship between categories (Comber *et al.* 2006). Any potential user is left struggling to understand a classes mean means in a particular geographical context. For these reasons Schuurman & Leszczynski (2006) and Comber *et al.* (2005b) have

argued that metadata ought to include more than documentation of the technical aspects of data production.

7. CONCLUDING COMMENTS

This paper has sought to reposition metadata away from data producers and towards the users of the data. Specifically, it has been argued that metadata needs to provide information about the *usefulness* of the data rather than its usability. However as any measure or qualification of usefulness (a form of data quality assessment) can only be relative to the intended application what are needed are new forms of metadata and tools them.

The case for current metadata paradigms as specified in standards is not yet proven for spatial data and geographical information. To the authors' knowledge, there are no examples where standard metadata elements have been successfully applied to allow users to assess the usefulness of someone else's data for their problem, to quantify and to manage the associated uncertainties. DeBruin and Hunter (2003) assess the value of different decisions relating to agricultural payment to farmers from from the time stamp of remote sensed data. DeBruin et al. (2001) describe an application that assesses the value of two DEMs with respect to the extent they influence an "expected value of control" relating to an error process for determining the volume of sand required to build a new port area. In neither of these examples is the quality of the information expressed in terms of any standard metadata concepts.

The current metadata for spatial data situation is analogous to buy a car just on the basis of that it has a certificate of legality often relating to safety (this is the MoT in the UK, TÜV in Germany). Standards have been created (each car must safety certificate) and producers ensure that their cars conform. Mediators (the car dealer) want to show the car to be of merchantable quality and use the safety certificate to do so. The user (buyer) is reassured that the car is safe and conforms to the law. However, under this scenario the buyer is not considering the appropriateness of the car for their circumstances. In the real world car buyers use some form of independent assessment (a Which report, motor magazines, experiences of friends and neighbours, etc) to identify *usefulness* of different vehicles relative to their needs (e.g. numbers of children, grandmothers and pets). At the moment there are not tools or independent assessments available to users. The proposals in this paper for expanding and activating metadata seek to address this.

ACKNOWLEDGMENTS

Ideas in this paper were first presented in Vienna at Spatial Data Handling 2006 and originate from work within the REVIGIS project funded by the European Commission, Project Number IST-1999-14189, from discussions with Bérengère Vasseur and during a workshop sponsored by National Institute for Environmental eScience in 2005.

BIBLIOGRAPHY

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., 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., Wadsworth, R. and Fisher, P, (2006). Reasoning methods for handling uncertain information in land cover mapping. In *Fundamentals of Spatial Data Quality*, edited by R.Devillers and R.Jeansoulin, ISTE, London. pp123-139

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.J., Fisher, P.F., Wadsworth, R.A., (2005a). What is land cover? *Environment and Planning B: Planning and Design*, 32:199-209.

Comber, A.J., Fisher, P.F., Wadsworth, R.A., (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

Comber, A.J., Fisher, P.F., Wadsworth, R.A., (2005c). A comparison of statistical and expert approaches to data integration. *Journal of Environmental Management*, 77: 47-55.

Comber. A.J., (submitted). The identification of data primitives to separate the concepts and semantics of land cover and land use: the example of 'forest'. Paper submitted to *International Journal of Land Use Science*

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

DeBruin, S, Hunter, GJ., (2003). Making the trade-off between decision quality and information cost. *Photogrammetric Engineering and Remote Sensing* 69 (1): 91-98

DeBruin, S., A.Bregt, and M. van de Ven. 2001. Assessing fitness for use: the expected value of spatial data sets. *International Journal of Geographical Information Science* vol 15 (5), pp. 457-471.

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.

Fisher, P.F. 1998. Improved Modeling of Elevation Error with Geostatistics. *GeoInformatica* vol 2 (3), pp 215-233

Fisher, P.F. 2003. Multimedia Reporting of the Results of Natural Resource Surveys, *Transactions in GIS*, 7 309-324.

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

Gewin, V., (2004). Mapping opportunities. *Nature* 427, 376 - 377 (2004)

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

Guptill, S.C., Morrison, J.L., 1995. *Elements of Spatial Data Quality*, Oxford: Pergamon Press, p.202.

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

Hunter, G.J., 2001. Spatial Data Quality Revisited, Proceedings of GeoInfo 2001, 04–05 October, Rio de Janeiro, Brazil, pp. 1–7.

Jeansoulin, R., and Wilson, N. (2002). Quality of Geographic Information: Ontological approach and Artificial Intelligence Tools in the REV!GIS project, 8th EC-GI&GIS Workshop, Dublin.

Martin, E., 2000. Actor–networks and implementation: examples from conservation GIS in Ecuador. *International Journal of Geographical Information Science* 14 (8), 715–737.

Robinson, P., Fisher, P., and Smith, G. (2005). Evaluating object-based data quality attributes in the Land Cover Map 2000 of the United Kingdom. *Photogrammetric Engineering & Remote Sensing*, 71(3): 269-276

Salgé, F. 1995. Semantic Accuracy. In *Elements of Spatial Data Quality* edited by S.C.Guptill and J.L.Morrison. (Elsevier, Oxford), pp. 139-151

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, vol.44(7), 1740-1746.

Spatial data Transfer Standard (SDTS) (1998). New York: American National Standards Institute.

Veregin, H.,(1998) Data Quality Measurement and Assessment, *NCGIA Core Curriculum in GIScience*, http://www.ncgia.ucsb.edu/giscc/units/u100/u100.html, posted March 23, 1998.

Visser U, Stuckenschmidt H, Schuster G, Vogele T., (2002). Ontologies for geographic information processing. Computers & Geosciences, 28 (1): 103-117

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*.

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

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