Geography Compass 2/5 (2008): 1529–1548, 10.1111/j.1749-8198.2008.00150.x

Database Ethnographies Using Social Science Methodologies to Enhance Data Analysis and Interpretation

Nadine Schuurman*
Simon Fraser University

Abstract
Data are the basis for many decisions ranging from assessing credit applications, to determining societal risk of criminals to adjudicating grant applications. Data collection and use constitute social practices, yet once data are placed in tables, their social lineage is forgotten. Database ethnographies are a unique means of using insights from science and technology studies and practices from the social sciences to enhance data analysis. The goal of this methodology is to elicit information from data stewards about the data in multiple-use databases in order to provide an archive that describes the context and meaning of the data at a particular point in time. This article provides a review of a composite literature that contributed to the concept and implementation of database ethnographies. In addition, it illustrates how database ethnographies contribute to more nuanced metadata and act as the basis for informed decision-making involving data from multiple sources.

Introduction: The Minimalism of Data

We have moved into a new era of recording – one in which information and knowledge is parsed into tabular form and stored in lists held by arrays that are, in turn, stored in computer registers. All that can be coded also can be recorded. Databases constitute, in effect, a minimum dataset. They should be conducive to creating and extending knowledge, but, in fact, their minimalist configuration limits the dimensionality of information. There are two limitations to the data packaging offered by database: not everything can be encoded; and coding frequently annihilates nuances of meaning. While printing presses provided a means of 'packing' knowledge onto the page (Bowker 2005, 102), databases parse it even more concisely.

Data collection and use constitute social practices, yet once data are ensconced in tables, their social lineage usually is forgotten. Bowker (2000b) identifies a new model of scientific production in which the database, rather than scientific results *per se*, is the goal (Bowker 2000b). In the past, data were collected for the purposes of a scientific experiment or survey.

Once the results were published, the data were quietly filed in cabinets and seldom retrieved. The Human Genome Initiative is the archetype of this new science in which the database is the goal, but biodiversity and ecological databases are also growing in scope and scale (Bisby 2000; Bowker 2000a,b). As these database archives are built, the issue of semantic heterogeneity emerges. That is, the same term is interpreted differently in different contexts. For example, the European Union has developed a vegetation classification system based on biotypes. Many conservationists note that this classification does not match UK or Irish vegetation types well. Indeed, some Irish vegetation types are unclassifiable using the system (Wateron 2002). Irish conservationists do not share Russian ecologists' strong epistemology of vegetation types. These classification systems reflect not only different frames of reference, but also were developed under different vegetative and climatic regimes. Both reflect unique institutional settings.

Scientists do recognize that data are interpreted differently among scientific communities and that context dictates use (Bowker 2000; Frodeman 2000). Yet, there has been no concerted effort to link context to data. Rather, data sit in isolated boxes in databases with semantic interpretation left largely in the hands of users. Biodiversity research has a recent tradition of drawing on diverse epistemologies associated with wide-ranging scientific understandings of biomes, phyla, floras and faunas (Bowker 2000a,b; Brodaric and Hastings 2002). Geologists have also examined the issue of how they divide the world into parts (Frodeman 2000). Likewise, human geography follows the social sciences in a tradition of questioning perspective and its relationship to objects on the earth's surface (Gregory 1978, 1994, 2000). None of these fields has, however, found a way to express differentiation among multiple epistemologies and ontologies in a way that can be piggybacked with the data.

Databases are a substitute for memory in the sciences and social sciences (Bowker 2005). However, they need to account for what happened before we meet the field or attribute in its box in the data table. Mark Poster (1996) draws attention to digital persona or the skeletal set of data that describes every individual in Equifax or other credit rating databases. To assume that this representation – based primarily on our age, profession, income and credit history – is adequate is absurd. We are slighted reduced by the shallowness of this digital description, yet routinely accept it in the form of data representing social and physical phenomena (Harvey 2005). The creation of a contextual archive for data, a means of dimensionalizing description, is the goal of database ethnographies.

Database Ethnographies

The article proceeds by reviewing relevant literatures that bear on the implementation of database ethnographies and the storage infrastructure necessary to support them.

Boundary objects have been used by science and technology study (STS) researchers to refer to scientific objects and concepts that have different meanings between disciplines, but are used nevertheless as a kind of *lingua franca* that permits communications among researchers working within different paradigms (Chrisman 1999; Fujimura 1992; Harvey and Chrisman 1998; Hess 1997; Star and Griesemer 1989). Boundary objects are, in effect, what geographic information science (GIScience) researchers use when developing methods for *semantic integration*. One of the chief shortcomings of integration research in the information sciences is that it seeks to impose an automated approach onto language – ignoring the fact that language is dynamic and will always elude fixity. Recently, we introduced *ontology-based metadata* (or data about data) as a means of attaching contextual information to attributes in databases (Schuurman and Leszczynski 2006). Ontology-based metadata requires, however, a means of eliciting information about data units such that they are contexutalized. That is the purpose of *database ethnographies*. A review of ethnographic literature in human geography and the social sciences is a guide to the development of this methodology for contextualizing databases.

It is outside the purview of this article to systematically review all literatures that bear on ethnography. The intention is rather to identify the many influences on the concept of database ethnography. This wide-ranging literature review serves as a background to the description of how database ethnographies have been gathered for a project to link two health registry databases. Database ethnographies are a unique means of using insights from social theory and practices from the social sciences to enhance science.

Literatures That Bear on Database Ethnographies

THE PROLIFERATION OF BOUNDARY OBJECTS

A most useful STS contribution has been identification of boundary objects or objects that are understood differently in different communities of discourse. First, introduced by Star and Griesemer (1989), boundary objects are a means of discussing near but non-equivalent scientific objects among different disciplines and sub-disciplines. They constitute an implicit acknowledgement of the heterogeneity of scientific conventions, naming and practices among disciplines (Star and Griesemer 1989). Boundary objects are 'standardized packages' or grey boxes that are adopted by members of different social/scientific domains to temporarily stabilize a definition (Chrisman 1999; Fujimura 1992; Harvey and Chrisman 1998). Boundary objects act as mediators that allow communication between disparate groups. In effect, they are compromises about what something really means that are created in order to work between different use–context situations. They allow for agreements while allowing for continuation of different uses and agenda with respect to scientific objects.

In the case of data, boundary objects are called the same thing (e.g. low birth weight of babies) but mean different things in their respective communities or contexts. Even more familiar concepts have different meanings in different contexts: 'range' can refer to scope; the habitat of animal; a stove top used for cooking; a spread of values; a series of mountains in a line; a place where shooting is practiced; and a verb meaning 'to roam'. Clearly, as linguists have argued, language makes sense in context, and sense making is an iterative process in which the person receiving a message often must make sense of the message that they heard in relation to other contextual variables (Hill 1997).

Still, decisions for merging data are often based on the superficial equivalency of such terms in the absence of just such contextual information. Databases seldom include detailed information about attributes or their intended use. Even when attribute interpretation seems self-evident, semantic interpretation varies widely (Schuurman 2002) and meanings vary.

Data integration and linkage decisions, therefore, while frequently justified on the basis of the potential that merged data sets hold for decision-making, rest too often on a faulty core assumption: that the semantic variability of data does not exist.

SEMANTIC DATA INTEROPERABILITY AND THE VAGARIES OF LANGUAGE

Semantic data interoperability entails drawing on multiple data collected and designed for a specific purpose in different contexts to analyse a particular phenomenon. If data are appropriately comparable, *boundary objects* should serve as emissaries between these different use–context communities. This is, however, seldom the case. Gambling, for instance, is linked to higher rates of suicide. In the Canadian province of Alberta in 2003, there were 140 suicides linked to gambling debts. In Ontario in 2003, there were 15 deaths classified as suicides that were linked to gambling (CBC News, 4 November 2004). Ontario's population is four times that of Alberta, and presumably there are proportionally more gambling venues. Should one conclude that gambling is more profitable in Ontario than Alberta and thus leads to fewer cases of depression linked to debt? A more likely scenario is that the criteria for linking suicides to gambling are more stringent in Ontario than Alberta. A similar problem occurred with birth weights which were historically defined differently in different jurisdictions.

When data are collected and stored in computer datasets, additional semantic complexity may be introduced. Computer systems often require users to enter values into delimited fields (such as the time a patient arrives in an emergency room or the time the patient is seen by a physician). To the extent that computer systems vary, the specifics of the data collected may also vary (e.g. one system may request time of arrival in an emergency room; another may request time seen by a triage nurse). However, both fields may be called arrival time. The semantic content of the database is superficially comparable but a closer look reveals critical differences.

Language is far more intractable than technology, and semantics are the lynchpin of interoperability. There have been a number of efforts in the GIScience research community to address data interoperability (Bisby 2000; Bishr 1997, 1998; Bishr et al. 1999; Brodeur et al. 2003; Cuthbert 1999; Fonesca et al. 2002; Harvey et al. 1999; Kottman 1999; Kuhn 2002; Laurini 1998; Sheth 1999; Vckovski 1999). Approaches to semantic data interoperability have, however, primarily stressed automated solutions (Schuurman 2002). These have focused on federated data sharing environments, semantic similarities between data sets and ways of using existing software platforms to accommodate data sharing through a component architecture approach as outlined by Kottman (1999) or Schuurman (2002).

A problem with technical approaches to interoperability is that they treat language as a form of mathematics and convert semantic relationships to graphs and networks in which proximity on the graph structure indicates closer meaning (Brodeur et al. 2003; Fabrikant 2000; Peuquet 1983). Using metric symbology to describe relationships between semantics does not address the larger problem of understanding how language is used differently in different situations and contexts. Using geometrics to describe relationships between semantics does not address the larger problem of stabilizing language for the purposes of mapping the graph structure. Mapping language using vertices (nodes) and arcs linking the nodes (more distance = greater distance in meaning), assume that the semantics are static and that their meaning is interpreted similarly in different contexts. Moreover, it implies that language is detached from communication. The limitations of these assumptions are seen in the above examples, where even between provincial jurisdictions, and there is a difference in meaning of the term *gambling-related suicide*.

THE ROLE OF TACIT KNOWLEDGE IN SCIENTIFIC COMMUNITIES

Much scientific knowledge is not explicit but tacit, and confined to the group or laboratory doing the research and its networks. Collins (2001) describes tacit knowledge as that which can be learnt only through participation or direct observation because a vocabulary for formalizing such knowledge has not (as of yet) been developed. He identifies a range of tacit knowledge including: (i) unrevealed (e.g. kept secret on purpose); (ii) lack of communication (e.g. two researchers are working on different parts of the same problem in the same group, and one has not told the other important information); (iii) hard to convey knowledge (e.g. information that cannot be conveyed through text but rather by feeling or demonstration); (iv) unrecognized knowledge (e.g. ways of doing things which contain implicit actions that the researcher does not think to make explicit); and (v) inarticulate knowledge (e.g. abilities or skills that can only be transmitted through watching someone work) (Collins 2001). Collins developed this analysis of tacit knowledge by studying several groups of scientists who were working on the Q *value of Sapphire*. The longer a bell or material resonates, the higher its Q value. Vladimir Braginsky at Moscow State University measured quality factors (Qs) of sapphires, and was able to achieve a very high number of 4×10^8. No other research group was able to replicate this work until a Glasgow-based research group visited Moscow for a week and learned to measure similar Q values. Collins employed the concepts of tacit knowledge to explain the implicit transfer of information that led to the replication of the results.

His analysis of the knowledge transfer between the two laboratories is one illustration of the way that social contact between researchers spreads tacit knowledge through a scientific community. The goal of database ethnographies is to gather information from data stewards about tacit knowledge, so that it can be linked explicitly rather than implicitly with databases.

ETHNOGRAPHY

Social scientists and STS scholars have long recognized that the nuances of meaning associated with practices are best understood using a heuristic methodology (Britten et al. 2002; Habermas 1970; Herbert 2000; Hyndman 2000). In 2000, Steve Herbert argued that ethnography is an undervalued technique in geography. His analysis of ethnography stressed that there is a distinct and recognizable connection between the goals/agenda of a group or institution and the 'geographic world they construct' (Herbert 2000, 551). For example, ethnography can help to explain male-on-male violence in public situations as a consequence of an insult in a public situation eliciting rebuke and the situation escalating. It is hard to understand what triggers soccer or hockey riots using quantitative data. Ethnography can also reveal the ways in which systems of meaning are generated and reinforced by interactions in social and institutional settings. This is particularly relevant for database ethnographies as data collection and interpretation occur within specific use contexts that vary considerably (Forsythe 2001).

The goal of ethnography 'is improvement of theory' (Herbert 2000, 560), so it needs to generate conceptual tools that are useful to others. Herbert (2000) address the three chief criticisms of ethnography, namely: (i) ethnography relies on interpretation and it is not sufficiently scientific; (ii) an intense focus on a single or limited number of situations make it difficult to generalize; and (iii) representation is clouded by the failure of researchers to problematize their methods. Herbert points out, however, that all science relies on interpretation, often of empirical evidence – a position corroborated by scientists themselves (Pickering 1995). Likewise, data are not naïve, and are themselves discerned and interpreted through social processes (Bowker and Star 2000). The concern regarding the appropriateness of generalization can likewise be countered by noting that studies of single institutions have provided insights into their working. For example, Latour (1987) provides documentation of the processes, social and scientific, that guide scientific research. Extrapolating from these singular representations may have certain limitations, but this approach has made proven contributions to furthering our understanding of social phenomena (Demeritt 2001; Haraway 2000; Harley 1992).

In geography, Jennifer Hyndman (2000) has used ethnography to *study up* or analyse and theorize institutions, organizations and bodies that govern human relations rather than to study the governed themselves. This ethnographic approach differs from most traditional anthropology in that the effect of institutions on envisioning populations is taken into account (Hyndman 2000). This work adapted the work of earlier geographers and anthropologists (Abu-Lughod 1991; Pred and Watts 1992) as a way of repositioning the academic gaze in order to focus on institutional culture rather than those affected. It acknowledges that there is a large gap between the world 'out there' and the world that is recorded in academic contexts.

Eric Laurier has worked on ethnomethodology or meta-methodology which examines order and how it is produced in different contexts (Laurier 2003). Founded in 1960s by American sociologist Harold Garfinkel, ethnomethodology is the sociological perspective that analyses the methods people use in order to understand or make sense of their social world (Garfinkel 1967). While traditional sociological paradigms share the assumption that there exists an objective social reality and that it is essentially orderly, ethnomethodology purports that social order is illusive, it only exists as a mental construct and that social life, though often appearing orderly, may potentially be chaotic (Poore 2000; Smith 2000; Wikipedia 2006). Using the 'documentary method', people extract social facts from any given situation that appears to be representative of a pattern, and then utilize this pattern to interpret the meaning of these facts.

Ethnomethodological studies have focused on the construction, organization and temporal order of social activities and interactions and, therefore, provided a greater understanding of the methods used to accomplish work. Ethnomethodological analysis of work and work settings has often been used as a framework for the design of computer-supported collaborative work systems, and throughout the 1990s these ideas were increasingly applied to human-computer interaction study as well (Button and Dourish 1996). Although initially utilized in design critique, this approach has been progressively applied to design practice (Dourish and Button 1998). It is now acknowledged that real-world failures of complex technology are often not attributable to technological shortcomings but to the failure to recognize the methods in which work and communication are actually organized. The fusion of ethnomethodology and computing sciences has been coined, 'technomethodology' by Button and Dourish (1996) to emphasize that, although drawing from both fields, this is a novel approach and a dramatic departure from both founding disciplines.

Ethnomethodology clarifies the gap between theory and practice by asking seemingly trivial questions that may inadvertently topple presumed internal logic. It attempts to avoid judgments of true or false or moral stances in favour of observation (Laurier 2003). For instance, Hester Parr uses ethnomethodology to interrogate how context shapes representation of mental health (Parr 1998). She argues that conceptualization regimes on the part of the researchers (e.g. psychoanalytical frameworks) influence diagnosis. As a result, Parr (1998) advocates avoidance of binary differentiations such as 'sane' or 'not sane'. If ethnomethodology is committed to showing how the internal logics of, in this instance, moral judgment and diagnosis of mental health are established, then database ethnography is committed to describing how particular data establish their status as 'truth' – or at least as the basis for policy-relevant analysis.

For data drawn from different contexts to have value and applicability, they must be considered to be true – that is, their validity derives from their acceptances as 'facts'. Actor-network theory (ANT) is a well-established means of understanding how certain practices and scientific 'facts' are established (Barnes 1998; Bowker and Star 2001; Hess 1997; Kitcher 1998; Latour 1999; Law 1999; Law and Hassard 1999; Law and Mol 2001). It uses the metaphor of a network to point to the social and physical interconnectedness between social relations, scientific and technological research and the non-living actants that are involved. It assumes that all facts and paradigms emerge from a network of relations that can be mapped. ANT is radical inasmuch as it assumes that actor networks – and their scientific products – look very different in different contexts. Nothing is predetermined. Facts are localized, and *immutable mobiles* hold specific configuration of facts together and enable them to qualify as science (Law and Mol 2001).

Some recent critiques of ANT emphasize that the theory presupposes an ontological vacuum (Law 1999). ANT is about relationships rather than absolute entities. ANT also introduces uncertainty into descriptions of entity relations by enabling the consideration of performativity. In the context of performing, entities might behave and present differently in different performances (Harvey 2005; Law 1999). Another problem with ANT is that the usage of 'network' changed with the introduction of the World Wide Web (Latour 1999). It used to imply transformations and now indicates direct communication and fixed links. In summary, ANT is one of many 'anti-essentialist' movements that did an end-run around social constructivism (Latour 1999).

It is important to understand that facts or truths are not static. As understandings of disease change through time, for example, data take on different meanings. This transformation has been difficult to accommodate as classification systems have calcified with the advent of computing and rigid data entry fields. Where previously there was more room for anecdotal information, now we have binary systems. Susan Leigh Star (1997) used a pragmatic approach to analyse how social networks influence medical classification systems. She argues that medical classification does not tell a story; rather, it records a binary: disease or not disease. By ignoring nuance, standard medical classifications fail to reveal the temporal complexity of disease. Cases of polio, for example, that were considered 'cured' often take a toll on the spinal cells of patients 50 years later (Star 1997). This begs the question whether they were cured or the disease was mitigated and lay latent. This level of consideration applies equally to any manner of data and indicates the extent to which the extraction of data from its context of origin can undermine the very analysis you seek to make.

This is exacerbated by institutional forgetting which results in a morphing of use for common terms (Bowker 2005). The problem with ethnography is that you must deal with time slices despite the fact that institutional memory is accreted over long periods. For example, when an elderly physician in rural Nova Scotia leaves the hospital, he may take with him the last of a particular set of antiquated coding practices. It is interesting to note that coding clerks frequently compensate for the perceived ineptitude of doctors' practices, but this is invisible work unless teased out by ethnography.

A further example of the dangers of ignoring context-specific meaning and nuance may be found in Thompson and Gifford's (2000) ethnographic exploration of sugar's use and effect on an Aboriginal group in Australia. Sugar in this context acts as a boundary object. Sugar is viewed as both comforting and linked not only to food served at family gatherings but also to the imbalance of life associated with being disenfranchized from the 'land'. Multiple meanings of the word *sugar* make the research more complex. Complex and conflicting meanings were drawn out through narrative interviews conducted in Aboriginal communities. The research used a cultural ecological perspective in which 'causal factors that impact on health and disease are embedded within the social and cultural processes and the wider context of social life' (Thompson and Gifford 2000). The authors stressed need for more gestalt view of risk for diabetes among Aboriginals in Australia – one that incorporates their cultural view of the 'sugar sickness'. The sugar that is linked to diabetes is also associated with Whites and their upset of traditional lifestyle (Thompson and Gifford 2000).

Meaning, as Barwise and Perry (1999) argue, is embedded in situations. Francis Harvey draws on work by Hubert Dreyfus to argue that humans do not think or make associations in constrained predictable ways (Harvey 2003). Although such associations might be apparent to the inculcated, they would be foreign to those outside that particular culture. For Aboriginal Australians, for instance, sugar has a different set of meaning and associations than for Europeans. *A priori* knowledge and unique associations combine to change the meaning of even very narrowly defined terms. Harvey (2003) explains that 'situatedness' (p. 536) allows us to construct and associate technologies (or data) with particular uses and problem areas. But when those data, for example, are removed from their context, their meaning is frequently shifted – although the new meaning is not marked. The implication for database ethnography is that we make a set of rich associations with particular terms or practices that are ensconced in institutional cultures of data collection and use – and if those associations are not carried over, we are reducing the data's true comprehensive value to our analysis and subsequent decision-making.

If the goal of database ethnographies is to gather crucial contextual information from data stewards about tacit knowledge, then ontology-based metadata is a vehicle for their transmission (Schuurman and Leszczynski 2006). Ontology-based metadata is a means of making explicit that which is unexpressed at the level of the database. It accounts for the truism that the vast majority of data users remain uncritical of data and unaware of their more nuanced narratives.

ONTOLOGIES IN AID OF METADATA

Database ethnographies must be captured in table form and categorized like the data they report on – if they are to be used to extend metadata. There are a number of documented problems with data collected from other sources (e.g. Census data) that are, for the most part, ignored in contemporary science and social science. Indeed, the mere location of data in tables seems to confer authority on it. Some of the problems associated with data are outlined by Bowker (2000) and include:

1. Sampling methodology is often outdated with respect to time and location: for example, street networks files in the Canada Census are often outdated after only a 5-year interval (Schuurman et al. 2006).
2. Institutional changes in terminology and/or use are seldom documented.
3. Measurement context is unclear (e.g. are measuring scales standardized?).
4. Measuring devices have changed (e.g. move to global positioning system lat/long locations from more vague address data).
5. Implicit data in one context become explicit in another. For example, two women living together in 1940 were not likely construed as a couple, but in 2006 they are explicitly recorded as a couple.
6. Data coding rationales and guides to abbreviation are frequently lost, making data indecipherable.

Metadata is a means of storing data about interpretation and implementation of database fields. It is an extant mechanism for conveying ontological information about semantic data. Metadata has the advantage of being institutionally and structurally ensconced in geographic information system (GIS) and other information science contexts. At present, however, they lack fields to express information beyond the technical and geometric domain. Non-locational attributes are often simply ignored. A typical metadata wizard from a GIS program is illustrated in Figure 1.

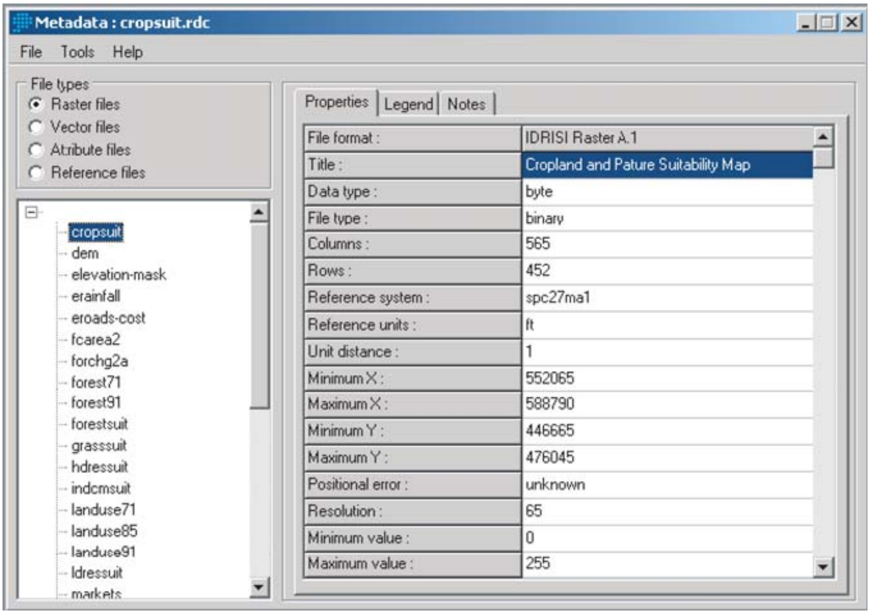

Fig. 1. A metadata screen shot from a popular GIS program IDRISI. Note that the emphasis is on locational data rather than contextual information. 'Crop suitability' metadata provides no information on criteria, rationale for collection or even measurement techniques. Rather, the user is expected to accept such data as authoritative.

The ontology-based metadata system draws on the concept of *ontology*. Ontology in a social science context refers to the essence of being of an object or phenomenon (Gregory 1994). In a computing context, ontology refers to a fixed universe of discourse (Gruber 1995; Smith and Mark 2001). In the latter, an ontology might be the complete set of all objects that can be defined in a GIS – and the possible set of relations between them. Or an ontology could simply be thought of as a classification system such as a map legend (Brodaric and Gahegan 2001). Using ontology in information science is a means of recognizing different contexts and meanings between potentially equivalent semantics. Indeed, ontology research is a well-researched area of GIScience at present (Agarwal 2005; Brodaric and Gahegan 2001; Cruz et al. 2002; Duckham et al. 2001; Egenhofer 2002; Fonseca et al. 2002; Kuhn 2002, 2003; Rodriguez and Egenhofer 2003; Smith and Mark 2001; Winter 2001). Recognition of the importance of ontologies has become the basis for many efforts towards data integration.

The efforts of GIScience to improve data integration have been dominated by efforts to leverage the reasoning capabilities of artificial intelligence formalisms – ontologies – for the encoding of semantic context and relationships at the machine level (Agarwal 2005; Bittner and Edwards 2001; Brodeur et al. 2003; Cruz et al. 2004; Duckham and Worboys 2005;

Fabrikant and Buttenfield 2001; Frank 2001; Fonseca et al. 2003; Kokla and Kavouras 2005; Kuhn 2002, 2003; Rodriguez and Egenhofer 2003; Sheth 1999; Smith and Mark 2001; Stoimenov and Djordjevic-Kajan 2005; Wilson 2004; Winter and Nittel 2003). Different approaches include peer-to-peer models which consist of direct communication between all participating systems (Arzt 2005; Gardner 2005; Wangler et al. 2003), involving federated data sharing environments (Kuhn 2002, Sheth 1999), schematic resolution of semantics (Kashyap and Sheth 1996), and rules for class membership based on similar semantics (Stock and Pullar 1999). Each of these advances in both GIScience and information sciences has focused primarily on technical solutions to the problem of integrating heterogeneous semantic terms.

Ontology research remains a very computationally intense means of linking data (Agarwal 2005). While it does recognize fundamental differences among semantic terms and classification systems – even when superficially equivalent – it requires massive commitment to a particular set of technologies (e.g. object-oriented databases), which few institutions are willing to adopt unless there is a decided market thrust in that direction (Schuurman 2002). It also assumes that ontologies (or context) can be computationally organized and linked in the interest of data integration. This common assumption in GIScience ascribes rather more confidence to technical solutions than has been warranted to date. Technical solutions do not include a means of extracting contextual information about the objects in order to contextualize the data for current and future users. That is the role of database ethnography.

Where Are Database Ethnographies Needed? Three Examples?

The world of data is teeming with examples of data that are assumed to be equivalent but are, in fact, quite different. Semantics heterogeneity is everywhere when you start looking. In Metro Vancouver, for example, air quality is reported from several geographically distributed measuring stations using a numerical index. The index is a composite of several types of atmospheric pollution. What is hidden from the index values is that each air quality monitoring station develops its index using different pollutants (Schuurman 2005). Figure 2 illustrates the different composition of the air quality indices from two disparate points in the metropolitan region. Few observers would expect that these data are not comparable – especially as the indices are reported using the same metrics. Databases seldom include detailed meaning about attributes nor their intended use. Even when attribute interpretation – as in the case of this air quality index – seems self-evident, semantic interpretation varies widely (Schuurman 2005). This is an excellent example of an instance in which database ethnography – in this case, a detailed examination of the attribute composition of the index and data elements would prove salutary.

Air Quality Monitoring Station

Monitoring Station Identity

Number: T001
Location: Downtown Vancouver, Robson and Hornby St.
Latitude: N 49°16'58"
Longitude: W 123°07'14"
Elevation: 56m
Date: March 20 2003

Air Quality Index

Time	AQI	Air Quality
00:00	6	Good
06:00	12	Good
12:00	15	Good
18:00	11	Good

Air Quality Sub-Index

Time	O_2	CO	NO_2	SO_2
00:00	5	0	6	0
06:00	12	0	2	0
12:00	15	0	5	0
18:00	11	0	5	0

Air Quality Monitoring Station

Monitoring Station Identity

Number: T020
Location: Pitt Meadows, 188477 Dewdny Trunk Rd.
Latitude: N 49°14'43"
Longitude: W 122°42'33"
Elevation: 20m
Date: March 20 2003

Air Quality Index

Time	AQI	Air Quality
00:00	17	Good
06:00	19	Good
12:00	17	Good
18:00	18	Good

Air Quality Sub-Index

Time	O_2	CO	NO_2	SO_2	PM10
00:00	17	1	1	0	8
06:00	19	0	0	0	6
12:00	17	0	0	0	6
18:00	18	1	1	0	6

Fig. 2. Air quality indices for two monitoring stations in the Greater Vancouver area. Note that the Pitt Meadows station includes a fifth attribute (PM10) in its calculations, yet both stations publish figures linked to the same numeric index. Variation in attributes is common for all monitoring stations in the area. Note: this illustration was published previously in Cartographica (Schuurman 2005).

Another example of the non-transparency of semantically similar data emerges from forestry (Schuurman 2004). Wildlife biologists and foresters have notoriously different epistemological lenses. While the two groups both have extensive education and knowledge, their agenda are different. Wildlife biologists classify forest data with an emphasis on vegetation available for animals (i.e. open-range area or low-density forest for grazing). Foresters, on the other hand, are chiefly interested in commercially viable tree species. Classification systems developed by the two groups for the same forest tract might contain similar features (open-range land and stands of old-growth forest), yet the interpretation of the similar features could be quite different (Schuurman 2004). While wildlife biologist measure attributes such as crown closure of trees, foresters are more likely to document individual species and their height and diameter. In this instance, the biologists and foresters have different epistemologies that result in divergent ontologies of the forest. Database ethnography is a useful tool in this instance to contextualize forest attributes and to determine if they are comparable.

A third example illustrates the importance of ontological context for ground water data. As water resources become more scarcer, there is increasingly pressure to map the sub-surface with emphasis on lithological structures such as aquifers and aquitards (layers that do not transport water). The location and extent of aquifers and aquitards is important, because they are used to determine where it is agriculture will thrive, where it is safe to bury waste or store it in landfills, as well as many other environmental decisions (Schuurman 2005).

Private water drillers provide most groundwater data in Canada and the USA. Most drillers, however, receive little training in the identification of geological materials. As a result, they might recognize rocks and other materials – but call them something different than would a geologist – or even another water driller. As a result, many names may be recorded for a given geological material (Schuurman 2005). The example of British Columbia where over 120 terms for lithological materials are found in the well logs is indicative of the extent of data heterogeneity. In such cases, semantic standardization is necessary just to be able to use data. This process is aided by interviewing different drillers and geoscientists in order to determine how variations in naming conventions were developed. These database ethnographies are the basis for semantic integration of widely divergent terms.

Database ethnographies offer a means of data collection for extended or ontology-based metadata. These, in turn, can be used to articulate semantic differences between near but non-equivalent semantic terms, such as *air quality*, *crown closure* or *aquifer*. For purposes of data comparison and integration, such semantic differentiation is invaluable, as it provides a mechanism for determining levels of equivalence. Moreover, it demonstrates that the current standard practice of relying on skeletal data frequently provides insufficient grounds on which to make complex decisions.

Conclusion

Data are the basis for information generation and knowledge creation. They carry, however, an authority that frequently exceeds their quality. Many informatics-supported decisions are based on data of questionable quality (Balka 2004; Longley et al. 2001). A more insidious problem is the use of data that are simply not suitable for the analysis at hand (Schuurman 2005). Relevance is one of the key issues that emerge as more and more data are secondary, that is, collected for one purpose and used for another. Data are after all representations of the world, not the world itself; they carry with them political assumptions and agenda that are invisible to users (Bowker and Star 2000; Woods 1992). And as data are moved between contexts for different analysis purposes, they lose implicit knowledge about their meaning. To date, there is no systematic means of imbuing data with context in order to determine relevance for the task at hand. Indeed, data are commonly considered appropriate by virtue of their existence (Schuurman 2004).

The system of eliciting information about data described in this article was developed based on a unique mix of literatures. The resultant database ethnographies do not prevent semantic instability. Indeed, they recognize it as an ongoing problem. Phenomena that are in early stages of study – for example, adverse events related to medical technology or changing water tables – present unique semantic challenges. The meaning of such terms has often not attained semantic stability within a single context, much less across disciplines, although there may be a strong desire to measure such phenomena. More often, semantic instability is indicative of philosophical or epistemological differences in how a phenomenon is viewed. For example, different provinces use different methods to describe aquifers and groundwater levels, as well as a host of other attributes. The contribution of database ethnographies is to provide a background to the institutional and cultural use of key terms at a particular point in time. This background – encoded as ontology-based metadata – is a means of enabling administrators, future data users, policymakers and researchers to discern semantic proximity and/or distance between terms that are mean different things in different contexts.

Database ethnographies build on STS work by enlisting ethnographic techniques not to *study* existing technologies, but to participate in their construction. They are a means of collecting and using qualitative data – from data stewards – to enrich extant data by providing ontological context for semantic terms used to describe attributes. This use–context information becomes part of an extended metadata format that is compatible with ISO 19115 standards for communicating metadata for the purposes of data interoperability. The research seeks to balance qualitative enhancement of data and attendant attention to idiographic details with the insistence on minimalism in informatics.

Database ethnographies are ultimately a means of unsettling the aura of authority that accompanies data. They flesh out the skeletal remains of events that rest in databases in order to provide them with a history and context. The story of the data is, of course, not an authoritative narrative that can be fixed. It is not the 'right story' because everything that is known about an event such as 'flash flood' or 'elevated sea levels' cannot be forced into one story. These ethnographies, encoded as metadata, serve to remind us that semantics are not uniform across time or contexts. They, nonetheless, provide simple maps for enabling defensible comparisons across jurisdictions. They are a contextualization at one point in time that make data more useful in the present and may allow data to remain useful in the future. These ethnographies and their accompanying metadata do not simplify database technologies. They are, rather, messy complications that remind us that human beings invent data based on local stories.

Acknowledgements

I would like to thank the Canadian Institutes of Health Research for generously funding this research through grant no. 116338. I would also like to thank Ellen Randall who kindly proofed the final version of this paper.

Short Biography

Nadine Schuurman is an Associate Professor in the Department of Geography at the Simon Fraser University and holds a PhD from University of British Columbia (2000). Her research is at the intersection of human geography and information science. Her primary expertise is in geographic information science and health informatics. She has also conducted research on critical GIScience, hidden homelessness, data integration and standardization, ontologies and metadata.

Note

* Correspondence address: Nadine Schuurman, Department of Geography, Simon Fraser University, RCB 7123, Burnaby, BC, Canada V5A 1S6. E-mail: nadine@sfu.ca.

References

Abu-Lughod, L. (1991). Writing against culture. In: Fox, R. G. (ed.) *Recapturing anthropology: working in the present*. Santa Fe, NM: School for American Research, pp. 137–162.

Agarwal, P. (2005). Ontological considerations in GIScience. *International Journal of Geographical Information Science* 19 (5), pp. 501–536.

Arzt, N. H. (2005). The new alphabet soup: models of data integration, Part 2. *Journal of Healthcare Information Management* 20, pp. 9–11.

Balka, E. (2004). *Fragmented views from the field: reflections on field observations of computerized health sector work*. Proceedings, Work Practice Research and Information Technology Development in Hospital and Community Health Care Settings. Workshop held in conjunction with Computer Supported Cooperative Work Conference, Chicago, IL, November 7, 2004, pp. 7–12.

Barnes, T. J. (1998). A history of regression: actors, networks, machines and numbers. *Environment and Planning A* 30 (2), pp. 203–224.

Barwise, J., and John, P. (1999). *Situations and attitudes*. Stanford: CSLI Publications.

Bisby, F. A. (2000). The quiet revolution: biodiversity informatics and the internet. *Science* 289, pp. 2309–2312.

Bishr, Y. A. (1997). *Semantic aspects of interoperable GIS*. Ph.D., Spatial Information, International Institute for Aerospace Survey and Earth Sciences (ITC), Enschede, The Netherlands.

——. (1998). Overcoming the semantic and other barriers to GIS interoperability. *International Journal of Geographical Information Science* 12 (4), pp. 299–314.

Bishr, Y. A., et al. (1999). Probing the concept of information communities – a first step toward semantic interoperability. In: Goodchild, M. F., et al. (eds) *Interoperating geographic information systems*. Boston, MA: Kluwer Academic Press, pp. 39–54.

Bittner, T., and Edwards, G. (2001). Towards an ontology for geomatics. *Geomatica* 55(4), pp. 475–490.

Bowker, G. C. (2000a). Mapping biodiversity. *International Journal of Geographical Information Science* 14 (8), pp. 739–754.

——. (2000b). Biodiversity data diversity. *Social Studies of Science* 30 (5), pp. 643–683.

——. (2005). *Memory practices in the sciences*. Cambridge, MA: MIT Press.

Bowker, G. C., and Star, S. L. (2000). *Sorting things out: classification and its consequences*. Cambridge, MA: MIT Press.

——. (2001). *How thing (actor-net)work: classification, magic and the ubiquity of standards, 1996*. [online]. Retrieved on 25 January 2001 from http://www.cse.ucsd.edu/users/goguen/courses/268D/actnet.html

Britten, N., et al. (2002). Using meta ethnography to synthesize qualitative research: a worked example. *Journal of Health Services Research & Policy* 7 (4), pp. 209–215.

Brodaric, B., and Gahegan, M. (2001). *Learning geoscience categories in situ: implications for geographic knowledge representation*. Paper read at Ninth ACM International Symposium on Advances in Geographic Information Systems, Atlanta, GA, November 9–10, 2001.

Brodaric, B., and Hastings, J. (2002). *An object model for geologic map information*. Paper read at Joint International Symposium on Geospatial Theory, Processing and Applications, Ottawa, Canada, July 9–12, 2002.

Brodeur, J., et al. (2003). Revisiting the concept of geospatial data interoperability within the scope of the human communication processes. *Transactions in GIS* 7 (2), pp. 243–265.

Button, G., and Dourish, P. (1996). *Technomethodology: paradoxes and possibilities*. Paper read at CHI Conference on Human Factors and Computing, Vancouver, Canada.

Chrisman, N. (1999). *Trading zones of boundary objects: understanding incomplete translations of technical expertise*. Paper read at Intersections: Society, Technology, and Geographic Thought, an invited workshop on GIS, Rough River Resort, Kentucky, April 1999.

Collins, H. (2001). Tacit knowledge, trust, and the Q of sapphire. *Social Studies of Science* 31 (1), pp. 71–85.

Cruz, I. F., et al. (2002). Handling semantic heterogeneities using declarative agreements. Paper read at ACM International Symposium on Advances in GIS, Mclean, VA, November 8–9, 2002.

Cruz, I. F., Sunna, W., and Chaudhry A. (2004). Semi-automatic ontology alignment for geospatial data integration. In: Egenhofer, M. J., Freska, C. and Miller, H. J. (eds) *GIScience 2004*. Adelphi, MD: Springer, pp. 55–61.

Cuthbert, A. (1999). OpenGIS: tales from a small market town. In: Vckovski, A., Brassel, K. E. and Schek, H. J. (eds) *Interoperating geographic information systems*. Zurich, Switzerland: Spinger, pp. 17–28.

Demeritt, D. (2001). The construction of global warming and the politics of science. *Annals of the Association of American Geographers* 91 (2), pp. 307–337.

Dourish, P., and Button, G. (1998). On 'technomethodology': foundational relationships between ethnomethodology and system design. *Human-Computer Interaction* 13, pp. 395–432.

Duckham, M., et al. (2001). A formal approach to imperfection in geographic information. *Computers, Environment and Urban Systems* 25, pp. 89–103.

Duckman M., and Worboys, M. (2005). An algebraic approach to automated geospatial information fusion. *International Journal of Geographical Information Science* 19 (5), pp. 537–557.

Egenhofer, M. J. (2002). Toward the semantic geospatial web. In: Egenhofer, M. J. (ed.) *GIS 2002*. McLean, VA: ACM, pp. 1–14.

Fabrikant, S. I. (2000). Spatialized browsing in large data archives. *Transactions in GIS* 4 (1), pp. 65–78.

Fabrikant, S. I., and Buttenfield, B. (2001). Formalizing semantic spaces for information access. *Annals of the Association of American Geographers* 91(3), pp. 563–580.

Fonseca, F., Davis, C., and Camara, G. (2003). Bridging ontologies and conceptual schemas in geographic information integration. *GeoInformatica* 7 (4), pp. 355–378.

Fonesca, F., et al. (2002). Using ontologies for integrated information systems. *Transactions in GIS* 6 (3), pp. 231–257.

Forsythe, D. E. (2001). *Studying those who study us: an anthropologist in the world of artifical intelligence*. Stanford, CA: Stanford University Press.

Frank, A. U. (2001). Tiers of ontology and consistency constraints in geographical information systems. *International Journal of Geographical Information Science* 15 (7), pp. 667–678.

Frodeman, R. (2000). A sense of the whole: toward an understanding of acid mine drainage in the west. In: Frodeman, R. (ed.) *Earth matters: the earth sciences, philosophy, and the claims of community*. Upper Saddle River, NJ: Prentice Hall, pp. 119–140.

Fujimura, J. (1992). Crafting science: standardized packages, boundary objects, and 'translation'. In: Pickering, A. (ed.) *Science as practice and culture*. Chicago, IL: University of Chicago Press, pp. 168–211.

Gardner, S. P. (2005). Ontologies and semantic data integration. *Drug Discovery Today* 10 (14), pp. 1001–1007.

Garfinkel, H. (1967). *Studies in ethnomethodology*. Englewood Cliff, NJ: Prentice-Hall.

Gregory, D. (1978). *Ideology, science and human geography*. New York: St. Martin's Press.

——. (1994). Ontology. In: Johnston, R. J., Gregory, D. and Smith, D. M. (eds) *The dictionary of human geography*. Oxford, UK: Blackwell, p. 429.

——. (2000). Ontology. In: Johnston, R. J., et al. (eds) *The dictionary of human geography*. Oxford, UK: Blackwell, pp. 561–564.

Gruber, T. (1995). Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies* 43, pp. 907–928.

Habermas, J. (1970). *Toward a rational society: student protest, science and politics*. Boston, MA: Beacon Press.

Haraway, D. J. (2000). *How like a leaf: an interview with Thyrza Nichols Goodeve/Donna J. Haraway*. New York: Routledge.

Harley, B. (1992). Deconstructing the map. In: Barnes, T. J. and Duncan, J. S. (eds) *Writing Worlds: discourse, text & metaphor in the representation of landscape*. London: Routledge, pp. 231–247.

Harvey, F. (2003). Knowledge and geography's technology – politics, ontologies, representations in the changing ways we know. In: Anderson, K., et al. (eds) *Handbook of cultural geography*. London: Sage, pp. 532–543.

——. (2005). The linguistic trading zones of semantic interoperability. In: Fisher, P. and Unwin, D. J. (eds) *Re-presenting GIS*. London: John Wiley & Sons Inc., pp. 43–55.

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

Harvey, F., et al. (1999). Semantic interoperability: a central issue for sharing geographic information. *Annals of Regional Science* 33, pp. 213–232.

Herbert, S. (2000). For ethnography. *Progress in Human Geography* 24 (4), pp. 550–568.

Hess, D. (1997). *Science studies: an advanced introduction*. New York: New York University Press.

Hill, P. (1997). *Lacan for beginners*. New York: Writers and Readers Publishing Inc.

Hyndman, J. (2000). *Managing displacment: refugees and the politics of humanitarianism*. Minneapolis, MN: University of Minnesota Press.

Kashyap, V., and Sheth, A. P. (1996). Semantic and schematic similarities between database objects: a context-based approach. *VLDB Journal* 5(4), pp. 276–304.

Kitcher, P. (1998). A plea for science studies. In: Koertge, N. (eds) *A house built on sand: exposing postmodernist myths about science*. New York: Oxford University Press, pp. 32–56.
Kokla, M., and Kavouras, M. (2005). Semantic information in geo-ontologies: extraction, comparison, and reconciliation. *Journal on Data Semantics* 3534, pp. 125–142.
Kottman, C. (1999). The open GIS consortium and progress toward interoperability in GIS. In: Goodchild, M., et al. (eds) *Interoperating geographic information systems*. Boston, MA: Kluwer Academic Publishers.
Kuhn, W. (2002). Modeling the semantics of geographic categories through conceptual integration. In: Egenhofer, M. J. and Mark, D. M. (eds) *Geographic Information Science*. Berlin, Germany: Springer-Verlag, pp. 108–118.
——. (2003). Semantic reference systems. *International Journal of Geographical Information Science* 17 (5), pp. 405–409.
Latour, B. (1987). *Science in action*. Cambridge, MA: Harvard University Press.
——. (1999). On recalling ANT. In: Law, J. and Hassard, J. (eds) *Actor network theory and after*. Oxford, UK: Blackwell Publishers/The Sociological Review, pp. 15–25.
Laurier, E. (2003). Guest editorial: technology and mobility. *Environment and Planning A* 35, pp. 1521–1527.
Laurini, R. (1998). Spatial multi-database topological continuity and indexing: a step towards seamless GIS data interoperability. *Geographical Information Science* 12 (4), pp. 373–402.
Law, J. (1999). After ANT: complexity, naming and topology. In: Law, J. and Hassard, J. (eds) *Actor network theory and after*. Oxford, UK: Blackwell Publishers/The Sociological Review, pp. 1–14.
Law, J., and Hassard, J. (eds) (1999). *Actor-network theory and after*. London: Blackwell.
Law, J., and Mol, A. (2001). Situating technoscience: an inquiry into spatialities. *Environment and planning D: society and space* 19, pp. 609–621.
Longley, P. A., et al. (eds) (2001). *Geographical information systems and science*. New York: John Wiley & Sons Inc.
Parr, H. (1998). The politics of methodology in 'post-medical geography': mental health research and the interview. *Health & Place* 4 (4), pp. 341–353.
Peuquet, D. (1983). A hybrid structure for the storage and manipulation of very large spatial data sets. *Computer Vision, Graphics, and Image Processing* 24, pp. 14–27.
Pickering, A. (1995). *The mangle of practice: time, agency, & science*. Chicago, IL: University of Chicago Press.
Poore, S. (2000). *Sociology at Hewett. Ethnomethodology – an introduction*. [online]. Retrieved April 2007 from http://www.hewett.norfolk.sch.uk/CURRIC/soc/ethno/intro.htm
Poster, M. (1996). Databases as discourse, or electronic interpellations. In: Lyon, D. and Zureik, E. (eds) *Computers, surveillance and privacy*. Minneapolis, MN: University of Minnesota Press.
Pred, A., and Watts, M. J. (1992). *Reworking modernity: capitalisms and symbolic discontent*. New Brunswick, NJ: Rutgers University Press.
Rodriguez, M. A., and Egenhofer, M. J. (2003). Determining semantic similarity among entity classes from different ontologies. *IEEE Transactions on Knowledge and Data Engineering* 15 (2), pp. 442–456.
Schuurman, N. (2002). Flexible standardization: making interoperability accessible to agencies with limited resources. *Cartography and Geographic Information Science* 29 (4), pp. 343–353.
——. (2004). *GIS: a short introduction*. Oxford, UK: Blackwell.
——. (2005). Social perspectives on semantic interoperability: constraints to geographical knowledge from a database perspective. *Cartographica* 40 (4), pp. 47–61.
Schuurman, N., and Leszczynski, A. (2006). Ontology-based metadata. *Tranactions in Geographic Information Science* 10 (5), pp. 709–726.
Schuurman, N., et al. (2006). Spatial/temporal mismatch: a conflation protocol for Canada Census files. *Canadian Geographer* 50 (1), pp. 74–84.
Sheth, A. P. (1999). Changing focus on interoperability in information systems: from system, syntax, structure to semantics. In: Goodchild, M., et al. (eds) *Interoperating geographic information systems*. Boston, MA: Kluwer Academic Publishers, pp. 5–30.
Smith, B., and David Mark, M. (2001). Geographical categories: an ontological investigation. *International Journal of Geographical Information Science* 15 (7), pp. 591–612.

Smith, S. (2000). Ethnomethodology. In: Johnston, R. J., et al. (eds) *The dictionary of human geography*. Oxford, UK: Blackwell, pp. 239–240.

Star, S. L. (1997). Of lungs and lungers: the classified story of tuberculosis. *Mind, Culture, and Activity: an International Journal* 4 (1), pp. 3–23.

Star, S. L., and Griesemer, J. R. (1989). Institutional ecology, 'translations', and boundary objects: amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907–1939. *Social Studies of Science* 19 (3), pp. 387–420.

Stock, K., and Pullar, D. (1999). Identifying semantically similar elements in heterogeneous spatial databases using predicate logic expressions. In: Vckovski, A., Brassel, K. E. and Schek, H.-J. (eds) *INTEROP'99*. Zurich, Switzerland: Springer, pp. 231–252.

Stoimenov, L., and Djordjevic-Kajan, S. (2005). An architecture for interoperable GIS use in a local community environment. *Computers & Geosciences* 31 (2), pp. 211–220.

Thompson, S. J., and Sandra, G. M. (2000). Trying to keep a balance: the meaning of health and diabetes in an urban aboriginal community. *Social Science and Medicine* 51, pp. 1457–1472.

Vckovski, A. (1999). Interoperability and spatial information theory. In: Goodchild, M., et al. (eds) *Interoperating geographic information systems*. Boston, MA: Kluwer Academic Publishers, pp. 31–37.

Wangler, B., Ahlfeldt, R. M., and Perjons, E. (2003). Process oriented information systems architectures in healthcare. *Health Informatics Journal* 9 (4), pp. 253–265.

Wateron, C. (2002). From field to fantasy: classifying nature, constructing Europe. *Social Studies of Science* 32 (2), pp. 177–204.

Wikipedia. (2006). *Ethnomethodology*. [online]. Retrieved April 2006 from http://www.en.wikipedia.org/wiki/Ethnomethodology

Wilson, N. (2004). The beginnings of logical semantics framework for the integration of thematic map data. *International Journal of Geographical Information Science* 18 (4), pp. 389–415.

Winter, S. (2001). Ontology: buzzword or paradigm shift in GI science? *International Journal of Geographical Information Science* 15 (7), pp. 587–590.

Winter, S., and Nittel, S. (2003). Information modelling for standardisation in the spatial domain. *International Journal of Geographical Information Science* 17 (8), pp. 721–741.

Woods, D. (1992). *The power of maps*. New York: Guildford Press.