Social Perspectives on Semantic Interoperability: Constraints on Geographical Knowledge from a Data Perspective

Nadine Schuurman
Department of Geography / Simon Fraser University / Burnaby / BC / Canada

Abstract

Much attention has been paid by government agencies and GIS researchers to standardization of data and interoperability of systems. Many of these efforts, however, have focused narrowly on technical hurdles while ignoring the social and political contexts that influence interoperability decisions. This article illustrates how social factors influence interoperability along three axes: classification, ontologies of data models, and government policy. Extant research approaches to interoperability of GIS are discussed and their strengths and weaknesses assessed. The article begins with definitions of what interoperability is, why it is important to academic users and policy makers, and its influence on geographical knowledge in a digital age. Exploration of social influences, as an alternative analytical approach to interoperability, begins with a discussion of the roles of classification and scale. The dangers of maintaining inflexible ontologies associated with specific data models are illustrated as a technical limitation with profound social implications for the construction of knowledge. Finally, policy at the multiple levels of governance with respect to interoperability is explored as an infrastructural constraint – and a diminishing influence.

Keywords: data standardization, GIS, interoperability, semantic interoperability, spatial databases

Résumé

Les organismes gouvernementaux et les chercheurs dans le domaine des SIG ont porté une grande attention à la normalisation des données et à l’interopérabilité des systèmes. Toutefois, la plupart de ces efforts ont été particulièrement concentrés sur les difficultés techniques sans tenir compte des contextes sociaux et politiques qui influent sur les décisions en matière d’interopérabilité. Dans l’article, on peut voir comment les facteurs sociaux influent sur trois aspects de l’interopérabilité : la classification, les ontologies des modèles de données et les politiques gouvernementales. Les démarches qui existent encore en matière d’interopérabilité des SIG sont passées en revue, avec leurs points forts et leurs points faibles. Au début de l’article, on définit l’interopérabilité, son importance pour les utilisateurs en milieu universitaire et les décideurs, et ses effets sur la connaissance géographique à l’ère numérique. L’exploration des influences sociales, en tant qu’autre approche analytique en matière d’interopérabilité, commence par une discussion sur les rôles de la classification et des échelles. Les dangers associés au maintien d’ontologies inflexibles liées à des modèles de données précis représentent une limite technique qui a des conséquences sociales profondes sur l’approfondissement des connaissances. De plus, l’auteur explore les politiques sur divers plans de gouvernance pour ce qui est de l’interopérabilité en tant que contrainte dans l’infrastructure, et en tant que sphère d’influence qui perd de l’importance.

Mots clés: normalisation des données, SIG, interopérabilité, interopérabilité sémantique, bases de données spatiales

Data exist in a folded space and time, torqued by practices and problematics of local scientific communities that produce them. (Bowker 2000)
Nadine Schuurman

Introduction

GIS interoperability and semantic standardization are the sorts of topics that halt conversations and cause people’s eyes to glaze over. The potential social, political, and economic effects of data integration are not exciting at first glance, yet a closer look reveals the important impact data integration has on how we construct geographical reality in information systems. Every time information about the world is recorded, classified, scaled, or modelled, selections about what is important have been made. The process of selection creates a representational reality, expressed through data models and maps, that influences decisions in domains from health care to subsurface waste management. Policies that constrain or enable a diversity of expression in databases ultimately affect the construction of social and physical knowledge. How data are classified and structured is a reflection of political and social priorities, and these decisions have implications for establishing policy priorities. Social geographers, on the one hand, might agree that knowledge is socially constructed but, with few exceptions, cannot pinpoint these influences in a technical context (Demeritt 1996, 2001; Hacking 1999). By contrast, such influences may be implicitly acknowledged in the technical realm of GIS interoperability research, but explicit reference to them is absent from methodological and technical literature.

There is a gap between scholars who focus on social implications of data structuring (of which interoperability is one form) and those who develop technical implementations for data interoperability. Social scholars of science have done an excellent job of historicizing classification, standardization attempts, and interoperability (Bowker 2000; Hacking 1999; Waterton 2002). They have drawn attention to the vagaries of each of these processes and the extent to which they reflect current cultural context. Terminology, categories, and the bases for classification are socially negotiated (Bryant 2000; Hacking 1999; Mitchell 2001). The certainty that classification, standardization, and interoperability are imperfect methods of structuring information should not paralyse us, nor should it constrain the development of much-needed criteria and methods for merging data from multiple sources. Indeed, contemporary research on interoperability within GIScience seeks to find ways of integrating spatial attributes based on their “true” meaning and context – though specifications do frequently allow provision for extending standards to comply with user needs (OpenGIS Consortium 1999).

This article will tease out intersections between social and technical approaches to interoperability by addressing social or theoretical issues that have clear and discernible implications for semantic interoperability in GIS. Three areas in particular will be discussed: (1) classification and scale; (2) data models; and (3) governmental data and interoperability policies. The article begins, necessarily, with an explanation of interoperability as a research area in GIScience.

What is interoperability, and why does it matter to geographers?

Interoperability is the pursuit of a means of combining data sets based on a common spatial grammar. It is a rubric for negotiation between systems and information at many levels of GIS, including systems, syntax, and semantics. Semantics are shorthand for attributes connected with geographical features or objects, but they also incorporate a worldview, an interpretation of reality, that is variable and difficult to pinpoint even within a single domain. Semantics also have to do the extra work of describing multiple phenomena that share the same classification or category. The term “interoperability” refers to the ability to share data and analysis between multiple users on different platforms. Yaser Bishr (1998) identifies six components of interoperability: network protocols, hardware and OS, spatial data files, database management systems, data models, and application (domain) semantics. Michael Goodchild, Max Egenhofer, and Robin Fugeas (1997) have identified eight pertinent elements. The dimensionality of interoperability, however, is not the chief focus here. Rather, our focus is the belief that conformity among these components will enable seamless sharing of data across jurisdictions and borders. Though this is changing, a sense that the interoperability problem can be “solved” through conscientious programming efforts has prevailed (Cuthbert 1999; Laurini 1998; Sheth 1999; Vckovski 1999b). More recent work emphasizes the dynamism of language and the need for interoperability research to develop methods for self-updating of semantic meaning – through automated solutions (Kuhn 2001, 2003; Mennis 2002).

Interoperability promises greater conceptual as well as software stability for GIS (Bishr 1998; Brodeur and others 2003). Perceived benefits include reduced costs and time to transform and manage data. The “existence of clear principles for the internal logical consistency of GIS databases” is perceived as a benefit of interoperability (Goodchild and others 1997). Geodata standards are touted as an “economic weapon” that will lead to more open global markets (Salgé 1999). In this scenario, standardized inventory systems will lead to frictionless trade between nations. GIS interoperability promotes shared organizational and analytical structures and promises to make application development easier and to increase the lifespan of applications. Like the open architecture plan, introduced by IBM with the first PCs...
in the 1980s, and the current promise of Linux, interoperability harbours the aspiration that common specifications will seed and enhance the market. In principle, interoperability will allow data to be used across platforms without any loss of integrity (Kemp and Vckovski 1998). It also promises to reduce the costs of data acquisition and management. According to proponents, interoperability is a means of encouraging communication and information sharing across communities of discourse (Schuurman 2002).

Interoperability between systems and data is a goal that has been recognized since the 1970s (Arctur and others 1998) but pursued aggressively only since the early 1990s. The 1994 US presidential order that brought the Federal Geographic Data Committee (FGDC) into being was the harbinger of a new era of analysis (Guptill 1994). The subsequent proposal for the Spatial Data Transfer Standard (SDTS) was intended as a means of implementing spatial data sharing. Other countries followed suit. In Canada, for instance, Geoconnections was established in 1999 by the federal government, along with industry partners, to oversee the development of the Canadian Geospatial Data Infrastructure (CGDI). As one of its five policy thrusts the CGDI is to oversee the development of geospatial standards. Its policy goals are closely aligned with those of international spatial data standards, so that Canadians can share information with other nations and Canadian businesses can sell geospatial information products to compete internationally. Like so many other initiatives, the CGDI was born out of the logistic imperative: that information is the key to a future of borderless data (Albrecht 1999; O’Donnell and Penton 1997; Tosta 1997).

Optimism has generally marked early forays into standardization and interoperability for GIS; integration standards are hailed as a means of allowing information products to compete internationally. Like so many scientific and technical problems, the logistics of the enterprise have proved more difficult than was initially imagined (Vckovski 1999a). Early researchers assumed that the chief problems of interoperability would revolve around data structuring techniques and networking protocols. Ironically, it is semantic interoperability that has proven the most difficult (Vckovski 1999a). Semantics are the language in which variables and attributes are expressed in databases, and they are subject to all the vagaries of communication that have plagued linguists and social theorists for decades (Haraway 1994; Kitcher 1998; Kwan 2002; Poster 1996; Sardar 2000). The response among GIS and systems researchers has been to focus on technical aspects of the problem. Interoperability has been treated like a systems engineering problem.

Social Perspectives on Semantic Interoperability: Constraints on Geographical Knowledge from a Data Perspective

Traditional Research Paradigms and the Algorithmic Mantra

The terms “interoperability” and “standardization” have sometimes been used interchangeably, though they should be differentiated. There are multiple visions of interoperability (Bishr 1998; Sheth 1999). There is general agreement, however, that interoperability is a broad rubric and semantic standardization is one component. Semantic standardization has also been recognized as one of the most difficult aspects of interoperability to resolve (Bishr and others 1999; Kottman 1999; Vckovski 1999a). This article focuses primarily on semantic standardization while accepting that different data schematics associated with classification and data models and structures affect semantics (Bishr 1998).

Semantic standardization is, broadly, the problem of calling the same thing different names or slightly different things the same name. Most terms are contextual: “urban” means different things depending on the place and time the term is used. Different disciplines give similar entities different names or give the same names to different entities. The word “range,” for instance, 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 practised; or a verb meaning “to roam.” Even a semantic term as banal as “road” can be understood very differently in different institutional contexts. The Ministry of Forests and the Ministry of Sustainable Resource Management in British Columbia use the same provincial “feature code” dictionary to define what constitutes a “road.” Yet the interpretation of this feature is dramatically different, to the extent that 1:20,000 maps of the same area illustrate non-congruent road networks (Schuurman 2002). Barry Smith and David Mark (1998) pose the quintessential semantic interoperability question: Where does the mouth of the volcano end and the mountain start? In the discipline of geography, ideas such as “cultural space” and “knowledge maps” employ a complex and non-linear transformation to the physical space, and the two are ultimately “unmappable” (C.J. Keylock, personal communication, 2 March 1999). This lack of semantic congruity leads to misunderstandings even between closely related groups of researchers and data users.

Science does attempt to mitigate this problem by insisting on explicit definitions and standardization. The problem with this approach is that nailing down meaning is fraught. Negotiation is involved in the development of categories. The medical category of “low birth weight,” for instance, is differently defined in multiple jurisdictions precisely because it is a contentious category with implications for statistics on the efficacy of national health care systems. The politics of classification, however,
are forgotten once the system is in place (Bowker 2000). The process of formalization used in information technologies attempts to tie users to precise categories. From a user perspective, this can entail determining whether the terms “clay” and “at risk” mean the same thing in a New Brunswick database as they do in a British Columbia (BC) database. The standardization of geospatial data is, in theory, a means of ensuring that databases are versatile and transferable between applications and institutions. It does, however, restrict the meaning of certain terms (e.g., “range,” “road,” or “low birth weight”) in the interests of ensuring a common “language” for geospatial databases and thus expanding the scope of geographic enquiry. Semantic diversity, or the range of discourse that can be represented in a GIS, is linked to the categories that are supported and their complexity. Thus the way that semantic components of databases are structured or ordered has a direct impact on GIS. How a certain characteristic is defined (and interpreted) has real effects on the resulting spatial entities used for queries and display. Users of data tend to interpret semantics based on their own history and context, rather than on the context in which the data were collected.

Scientists search for accuracy in the classification of every phenomenon of study. Qualitative social scientists, likewise, conduct ethnographic and idiographic studies in order to avoid voicing over differences muted by totalizing theories or narratives. Geography has hosted a number of debates about the suitability of nomothetic versus idiographic research (Hess 1997; Walton 1995), and there are clear divisions in the discipline over which approach is more appropriate. Standardization is stifling for some researchers. It is perceived as using a legalistic and mechanistic approach to interoperability and data sharing. Each information community, such as geology or population health, has a vocabulary. Each has rules and classification systems to create categories that organize that vocabulary. Success comes from developing standardization in a manner that is acceptable to members of specific intellectual and user communities.

There has been considerable research in GISScience on ways of achieving semantic interoperability, but most of this research has focused on technical solutions (Bishr 1998; Kottman 1999; Laurini 1998). Several strands of technical interoperability and standardization research in GIS are relevant to the problem of semantic standardization. Each is united by an algorithmic approach to semantic standardization. One such approach focuses on federated data sharing environments as one means of facilitating semantic standardization (Abel and others 1998; Devogele, Parent, and Spaccapietra 1998; Sheth 1999). In one iteration of this model, a mediator shares data among federated databases, supporting scalable information and multiple ontologies (Sheth 1999). Another version of the mediated system is proposed by Frederico Fonseca and others (2002), who envisage a GIS architecture for enabling integration of semantic terms that incorporates ontologies. In this instance, ontology is understood as a component of a database, with different ontologies constituting different classes (Fonseca and others 2002). The advantage of this ontology-based system is that it surpasses the spatial or map metaphor that has dominated other technical approaches to spatial data.

Another tactic finds semantic similarities between data sets. There are several ways to achieve this, but each approach is linked by the fundamental goal of identifying entities with similar descriptions in different databases (Kashyap and Sheth 1996). One strategy is to develop rules that stipulate conditions for inclusion in a particular semantic category (Stock and Pullar 1999). Another linguistic approach uses the principle of “semantic proximity” between spatial objects. A more pragmatic method is being developed by the Open GIS Consortium, an affiliation of universities, industry, and government partners. Open GIS uses existing technologies that can be harnessed to facilitate data sharing using a component architecture approach (Cuthbert 1999; Kottman 1999; Voisard and Schweppe 1998). The advantage of designing interoperability around distributed computing platforms is that components can be designed to work between existing software environments (Hardie 1998; Ungerer and Goodchild 2002). A major advantage of Open GIS is that it has the support of multiple institutions, an important social influence in the implementation of interoperability. To date, however, Open GIS efforts have not focused specifically on semantic interoperability.

Most of these technically oriented solutions to semantic heterogeneity arise at the intersection between computing science and GIS. Certainly this research has contributed to increased awareness of the problems of integrating semantically heterogeneous data, but it is based on the assumption that the problems with semantic heterogeneity arise from technical constraints and remain unrelated to classification, scale, or policy. There is general recognition that communities of discourse exist, but it has been difficult to develop means of representing them while enabling automated data sharing (Bishr and others 1999; Bittner and Edwards 2001; Kuhn 2003; Sheth 1999).

Technical approaches would be considerably strengthened by closer attention to institutional practices that contribute to the development of unique communities of discourse – which are subsequently embedded in data semantics. Until very recently, a disjuncture has existed between semantic standardization theory and the on-the-ground reality of navigating local data collection, reporting, and institutional and governmental practices. Research within the last five years has increasingly recognized unique and separate communities of
GIS across a range of issues including epistemology, in the past decade, human geographers have critiqued requires a hybrid approach that straddles human effects. Speaking to both social and technical issues established in the absence of consensus from many GIS specifications and protocols at the semantic level may be social scientists. As a result, technical standardization remains the domain of technical researchers as opposed to constituency across geography and other research areas; it time, interoperability research has failed to attract a wide and gaining support for interoperability. At the same however, for a host of social issues as well as political associations with digital systems. It cannot account, thereby raising the question of whether GIS are ultimately technical in their social and scientific have never been separate; there has only been a refusal to acknowledge their relationship (Latour 1993). Interoperability challenges the traditional division between social and technical realms, partly because locating social influences requires some understanding of the technical.

**Show Me the Social**

Research to date on interoperability (which includes standardization) has been concerned with the coordination of applications and systems (Bishr and others 1999; Buehler and McKee 2000; Egenhofer 1999b; Gahegan 1999; Kottman 1999; Sheth 1999; Vekovski 1999a, 1999b). This approach is attractive and practical for software engineers, giving the requirements for formalization associated with digital systems. It cannot account, however, for a host of social issues as well as political players involved in the process of creating specifications and gaining support for interoperability. At the same time, interoperability research has failed to attract a wide constituency across geography and other research areas; it remains the domain of technical researchers as opposed to social scientists. As a result, technical standardization specifications and protocols at the semantic level may be established in the absence of consensus from many GIS users who will later be concerned about representational effects. Speaking to both social and technical issues requires a hybrid approach that straddles human geography and GIS.

In the past decade, human geographers have critiqued GIS across a range of issues including epistemology, militarism, and ethics (Pickles 1993, 1997; Sheppard 1993, 1995; Smith 1992). Many of these critics were legitimately concerned with the prospect of a GIS that ignored social dimensions, but they were unable to articulate their apprehensions in a manner that allowed GIS researchers to address the issues. Geographers concerned with social issues in GIS have only recently started to communicate meaningfully with researchers in GIScience (Sheppard 2005). These efforts have coincided with a period of increasing recognition for science and technology studies (STS), which highlight interaction between science and society. STS researchers ask for recognition of the complicated interplay between social and scientific players in fashioning scientific and technical realities. Recognition of how and when GIS problems are social is crucial for developing appropriate technical solutions (Schuurman 1999).

In order to develop technical solutions for interoperability, one must understand the politics that were involved in its negotiation. Francis Harvey (1999a) uses the concept of boundary objects to describe the relationship between politics, technology, and society in fashioning specific iterations of technology. Boundary objects are those that are recognized by more than one group, though they may have different semantic designations. They act as communication emissaries between groups (Harvey 1999b). Such social apparatuses are key elements in illustrating social dimensions of GIS. Recognition that a GIS concept or artefact is created through social or political influences should not detract from its importance. Indeed, it should structure or instruct technical solutions.

The most difficult and broadest of these is the issue of classification, as it relates both to scale and data models. Domain is usually regarded as the chief influence in the development of categories, but there are multiple variables involved, some with particular implications for geospatial data. Related to classification is scale. Indeed, the two are inseparable: as scale shifts, so do classification categories. Though a number of human geographers have begun to examine the effects of scale in analysis (Marston 2000; Sheppard and McMaster 2004; Towers 2000), interoperability researchers have largely failed to include scale as an obstacle to integrating data between domains. Interoperability stabilizes not only classification categories but also data models – or the digital representation of categories. Semantic categories are partly shaped by data models, but more importantly they are the expression of the range of ontologies permitted by specific data models. There are, for instance, specific ontological repercussions associated with object-oriented data models (Burrough and Frank 1996; Couclelis 1996). This is a fair litany of social implications for standardization and interoperability, and this article only begins to identify them.
Semantic Heterogeneity and the Classification of Spatial Data

The most difficult aspect of data integration is not based on network protocols or database schematics (Bishr 1997; Harvey 1999a). Rather, it is the difficulty of creating equivalence between attributes or characteristics of spatial data (Vckovski, Brassel, and Schek 1999). In the Vancouver region, for example, air quality is reported using a numerical index. The index is a composite of types of atmospheric pollution, but it is formulated differently for each station. Figure 1 illustrates the difficulty of comparing index figures for different points in the metropolitan region, as they are not based on the same pollutants. The casual user may not realize that such data are not equivalent, and little guidance is provided. Databases seldom include detailed meaning about either attributes or their intended use. Even when attribute interpretation seems self-evident, semantic interpretation varies widely (Schuurman 2004).

Semantic interpretation often assumes a commensurability between attributes that is not justified. As a result, data are merged based on irreconcilable meanings. Conversely, researchers may exclude data based on an assumption that differently spelled or articulated fields are not equivalent. Figure 2 illustrates a well-log data file. Note the multiple terms for clay. Automated integration systems flounder when faced with such data.

Figure 1. 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 calculation, yet both stations publish figures linked to the same index. Variation in attributes is common for all monitoring stations in the area.

These problems are exacerbated by the curt, often truncated field descriptions typical of institutional databases where meaning is implicit. Cryptic field headers such as “%F” are not uncommon. Disparities in interpretation of semantics lead to error and uncertainty in analysis that is not easily detected by non-GIS experts or even by data analysts. Semantic incommensurability is illustrated by the use of the word “pond” in Newfoundland and Labrador to refer to a large lake. In mainland Canada, “pond” refers to a small, often undrained body of water while “lake” refers to larger bodies of water (Mark 1993). More complex attributes invite even greater misinterpretation. These issues extend to all disciplines. Population health researchers who work across jurisdictions, for example, must deal with the problem of seemingly identical categories being variously interpreted. Categories such as “at risk” and “low birth weight,” for example, are defined differently in each Canadian province, making interprovincial comparison difficult.

There have been a number of entreaties among interoperability researchers to establish a standard language that limits semantic ambiguity (Frank and Kuhn 1999; Vckovski 1999a; Vckovski, Brassel, and Schek 1999). Indeed, the desire for a more tightly defined vocabulary – semantic structures compatible with computer architecture – has been expressed in various corners of GIS (Hornsby and Egenhofer 1998; Kuhn 2001; Ruas 1998a, 1998b; Weibel and Dutton 1998). The reality of
domain influence on semantics, however, has tempered this goal. A tension remains between proponents of strict vocabularies and those who view this ambition as folly. Researchers in favour of building a vocabulary that is consistent with digital representation argue that our semantic structures have been too loose to allow efficient representation or generalization of objects and their complex interrelationships. They contend that in order to translate meaning and methods into digital terms, we need a stricter vocabulary that allows us to impart contextual knowledge. This maxim stresses that it is necessary to move to a semantic system that conforms to machine architecture.

Other researchers have been at pains to build semantic bases for operations and geographic phenomena based on the argument that the traditional representation of digital data as $(x, y)$, plus a semantic code, has limited the amount of contextual knowledge available for making generalization and other decisions about the interrelationships between objects (Hornsby and Egenhofer 1998; Ruas 1998b). Strict semantic systems are fiercely contested because of their premise that language can be absolute.

Dianne Richardson, a senior researcher at the Canada Centre for Remote Sensing (CCRS), bases her opposition to rigid classification on a Canadian experience with semantic standardization:

The conclusion that has been drawn, at least here in Canada, is that it is very difficult to impose some kind of standard for semantic classification. In some form or another, the actual semantic classifications need to be translated and you can do it through data dictionaries or look-up tables for an alternative approach. Now, an example is in my own work where I have come up against that problem particularly in transportation networks, hydrography, land-use coverage. Canada as a whole spent a huge amount of time and money trying to standardize the classifications so that the semantics associated with it were the same across all the provinces. That has just wasted many people's time, a lot of money, and a lot of effort. So rather than investing a lot of time on getting agreement on semantics, Good heavens, it is just such an extraordinary task. Why would people waste their time on it? (D. Richardson, personal interview, 26 October 1998)

Richardson's view is gaining support, especially as evidence accumulates that rigorous semantic structures fail to enhance communication and data sharing (Brodeur and others 2003; Harvey 2003). An increasing number of researchers in GIScience have focused on incorporating...
multiple ontologies with their own semantic usage into GIS (Bittner and Edwards 2001; Brodaric and Hastings 2002; Cruz and others 2002; Fonseca and others 2002; Frank 2001; Kuhn 2002; Raubal 2001; Smith and Mark 1998; Winter 2001). Acknowledgement is emerging among some interoperability researchers that there will never be a standard set of semantics deployed for database entry (Brodaric and Hastings 2002; Forer 1999). Instead, researchers will have to work with a federated database system in which translation between databases from different domains is mediated at a meta-level (Abel and others 1998; Sheth 1999). Virtual or temporary databases will be generated through the aid of an administrator. Though there are high hopes for a semi-automated system (Devogele and others 1998), current practice involves building translation tables in a painstaking way that creates equivalences between every semantic designation in every database. Translators in a federated database system will continue to face the same problem that confronts those who try to create a unified vocabulary. Moreover, automated translators are notoriously cumbersome and narrow in scope.

Classification can be described as “a spatial, temporal or spatio-temporal segmentation of the world... A classification system is a set of boxes, metaphorical or not, into which things can be put in order to then do some kind of work – bureaucratic or knowledge production” (Bowker and Star 1996). Classification systems are, by definition, not complete. They aim to include a “statistically significant” level of occurrences of a given phenomenon. Depending on the resolution of the study, different phenomena will constitute a statistically significant category. Classification remains a domain-specific undertaking, and one fraught with social and political influence. The impetus for important standards is the desire for global analysis of spatial data. Thus, it is important to recognize that decisions about classification are decisions about the extent of study areas. Clearly, the higher the resolution of the study, the greater the number of categories that emerge. Both scientific and social pressures bear on the choice of scale associated with research.

Scale has been a minor concern for researchers examining interoperability, in part because emphasis has focused on more immediate problems (Arctur and others 1998). The problem of semantic heterogeneity is essentially a problem of scale, not just, as is generally assumed, one of domain. The damming of the Yangtze River in China provides a metaphor for the relationship between scale, classification, and standardization. It also clarifies the scale-dependent nature of naming and domain knowledge. The dam will obliterate many small villages, agricultural areas, subsistence farms, and bicycle and footpaths to enable hydroelectric production and limit flooding. There is understandable opposition from those whose livelihoods, homes, and fields will be flooded, just as state bureaucrats and engineers envision a great power source to solve China’s energy problems. Similarly, the question in GIS is how to retain those local knowledges and artefacts while allowing the larger project to go ahead.

The translation between classification and standardization is fundamentally an issue of scientific and governmental communication (Bowker 2000). Not unlike nation-states, scientific and social domains have their own conceptual models. Certain domains may be better able to inscribe their needs on emerging specifications from government and industry. Geology, for instance, uses three-dimensional surfaces. Each community of interest will need to appeal to have its domain models incorporated into the new vocabulary, and three-dimensional data, for instance, will not be everyone’s priority (Jacquez 1999). Geologists will petition to have the standard language include geologic terms and processes. But the ontologies of all communities need to be addressed (Peuquet, Smith, and Brogaard 1998). Clearly this is an example of the merging of the political with the technical. And, as in every political negotiation, there will be winners and losers. Interoperability does not serve everyone equally; it serves some knowledge domains better than others. For instance, models and terms used in natural-gas exploration are standardized by the petroleum industry using a well-established and accepted data model (PPDM); but community mapping groups (re)invent a vocabulary with every project.

The creation of classification systems for data is, in effect, the creation of a standardization infrastructure, a process mediated by political pressures as well as scientific motivations. Little research in geography has been conducted on how raw data get converted into database fields, but politics often determine what will be recorded in the database (and what is excluded). As a result, systems may become data driven rather than process driven. Many women physicians, for example, spend more units of time with each patient than their male counterparts. As a result, they see their patients less frequently but for longer intervals (Bowker and Star 1996). A data-driven classification system cannot account for relationships outside of the classification system; it may ostensibly appear that women are less efficient physicians than men. An example from my own research with urban geological datasets indicates that subsurface groundwater records contain information about lithology but not about the relationships between subsurface layers or the processes that affect (shape) multiple layers. Field notes associated with these processes are frequently written in the record margins, but they do not become part of the official record: that information is lost in the analysis. A process-centred model would be better able to model relationships among data. A key goal for semantic standardization is the retention of information related to processes associated with data. Even if individual domains
are able to record and incorporate information about process, the utility of that information is limited by the choice of data models.

**Ontologies and Data Models: An Argument against Inflexible Semantic Standardization**

Ontology researchers over the past decade have recognized that the data models we are forced to choose between represent very different ways of seeing the world (Campari 1991; Couclelis 1992, 1996; Frank 1996; Kemp and Vckovski 1998; Mark 1993; Peuquet and others 1998; Rodriguez, Arnada, and Navarro 1996; Schatzki 1991; Smith and Mark 1998; Volta and Egenhofer 1993). Object data models epitomize the idea of discrete, separate entities in a frictionless neutral space. In the process of their definition, objects make a strong ontological commitment. Field models, on the other hand, discretize areas of geographical space but make a lesser ontological commitment (Couclelis 1999). Both fields and objects can be represented using either raster or vector, though objects tend to be associated with vector representation and fields with raster. Both field and object models ultimately rely on a Newtonian view of the world coupled with Euclidean geometry; neutral space has traditionally been assumed (Couclelis 1999). This view of space, however, is increasingly being questioned (Massey 1999; Raper 2000). Neither fields nor objects allow the characterization of complex interrelated geographic entities or processes (Couclelis 1996; Goodchild 1992). Each is a simplistic characterization of a complex geographical reality.

Ontologies, ideally, should not be restricted to the digital primitives available in current GIS, but discussions of ontologies have led inexorably back to existing data models. "They substitute," Egenhofer has commented, "for the fundamental elements of geographic thought" – in a database model of the world (Egenhofer 1999a). Discussion about ontologies is taking place precisely because there is increasingly suspicion that the data models that underlie our primitives are not ideal surrogates for spatial objects and their relations. Nevertheless, a recursivity between available data models and the framework for exploring geographic ontologies persists. I have argued elsewhere that the field/object dichotomy in GIS speaks to a tradition of binarism rather than to any inherent duality in the geographical world (Schuurman 1999).

Indeed, fields and objects are self-reinforcing. Current thinking is often caught up in a false binary of either/or. These models are our current alternatives, and, as a result, all conversations end up revolving around the two conceptualizations of space as either a set of objects or an infinite field. This conundrum has led to an extreme view in which there is a duality to geographical space, akin to the light and wave models of light used in physics (Peuquet 1988). It would be more productive to envisage current data models as provisional – in the absence of more complete technologies for representing the complexities of geographical features and relationships. Ontological implications of data models have been widely discussed in various sectors of the GIS community over the past decade, but these conversations were absent until recently from the interoperability literature. An exception is Amit Sheth (1999), who argues that the present (third) generation of interoperability research must support multiple ontologies associated with different domains. More recently, ontologies have been closely linked to standardization. Smith and Mark (2001) illustrate that non-experts conceptualize geospatial phenomena using a simple, fairly consistent set of spatial categories. The implication is that ontological terms such as "object" or "entity" hold common meanings that can be naturalized. This naturalization of perception and classification is problematic, however, because it is based on a clear separation between ontology and epistemology that is difficult to support, given that data models are used in different domains, many with varying epistemologies. One way to acknowledge that geospatial data include various levels of inconsistency, as well as varying reliability in terms of correspondence to the "real world," is to include a meta-layer of information about the observer and observations that documents information about measurement techniques as well as subjective influences on the data (Frank 2001). This tactic has been attempted in the geosciences with the development of the North American Geological Data Model (NADM). NADM provides a means of documenting recognized institutional epistemologies that are invariably manifest in the choice of terms used to describe geographical phenomena (Brodaric and Hastings 2002). NADM is innovative in that it attempts to resolves the traditional problems of separating ontologies and epistemologies. The two are inevitably intertwined, and one of the major contributions of post-structuralism has been to demonstrate this fact (Haraway 2000).

Even if terminology and processes are successfully standardized, the risk remains that standards will be altered after their imposition to suit various environments. Since the advent of the first shared standards between US agencies, for example, intra-agency additions and revisions have continually appeared as institutions attempt to incorporate local context (Guptill 1994). Data are information with contextual frameworks that interoperability standards may fail to accommodate. Institutions are quite skilled in altering standards to suit their data and analysis needs – thus the dictum that "God must have loved standards; she made so many." Jochen Albrecht echoes this irony in his summary article on the multiple standards presently in use or under
Vanishing or burgeoning? Policy, Government, and Interoperability

Depending on who is talking, national mapping organizations (NMOs), such as the United States Geological Survey (USGS) or the Geological Survey of Canada (GSC), are integral to achieving semantic standardization or institutions that are losing their grip on data production as privatization proliferates (Forer 1999). In a trenchant analysis of changes in the role of NMOs, Forer identifies that, from the perspective of the NMOs, their role has never been more critical as they prepare to lead their respective countries into a world of seamless data interoperability. Depending on the country, of course, the strategy is different. Canada, New Zealand, and Australia sing the praises of cost recovery, while the better proven strategy of the United States in providing free data as a means of seeding the market is offered as the democratic alternative (Mooney and Grant 1997; Morrison 1997; O'Donnell and Penton 1997; Robertson and Gartner 1997). Representatives of NMOs believe that interoperability will emerge from cartographic and bureaucratic traditions based on limited ontologies for “agreed upon geographic objects” at well-defined scales (Forer 1999, 3). This view fails to take into account the changes wrought by digital data, which are fluid, unstable, and mobile – indeed, they share many of the qualities of global capital. Forer argues that NMOs will, even in an optimistic scenario, retain only the ability to influence the infrastructure upon which data will accumulate. Data production itself will go the way of globalization, becoming disparate and privatized. In this scenario, unless NMOs rapidly establish the basis for data infrastructures, they will be overtaken by data producers of many stripes who may care little for protocol and even less for the historical role of NMOs in setting the standards for scale, ontologies, and formats of data.
ontological implications for the geosciences as well as for GIS (Schuurman 2002). Coordinating multiple stakeholders is one prong of a broader project to find ways of integrating large, multiple-source data sets. What is noteworthy, from a social and policy perspective, is that it is likely that a national classification system for subsurface lithologies will be developed and implemented outside of a national spatial data infrastructure. The repercussions of an organic classification that results from an anarchic set of interested parties are varied. Lack of a supporting national data infrastructure is one, though this may be offset by the flexibility of possible approaches to classification and standardization. It does point, however, to the importance of social networks in developing science.

What might an alternative look like? Database Ethnographies and Ontology-Based Metadata

I have outlined the myriad challenges and obstacles to semantic interoperability, with an emphasis on the homogenizing effects of standardization, the inability of policy-based national infrastructures to build viable frameworks for interoperability, and the vagaries of classification that lead to skeletal understandings of complex phenomena. A very gloomy portrait. There are alternatives, however. One is being developed by my research group at the Institute for Health Research and Education (IHRE) at Simon Fraser University. We are developing an ontology-based metadata protocol for population health data that builds upon ISO 19115. Rather than include only traditional metadata, such as lineage, scale, projection, boundary coordinates, meaning of attribute codes, and data authorship, we are collecting "ontology-based" metadata that include measurement techniques, rationale for collection, and usage within specific communities of discourse. These data are being collected from data stewards of existing population health databases as well as from data contributors, using a method of database ethnography.

Database ethnographies draw on research from science and technology studies in which scientists and their practices are interrogated in order to reveal the contingent and contextual nature of scientific practices (Latour 1987; Law 1994; Law and Hassard 1999). The initial pilot project is being conducted with data from the Vital Statistics branch of the Ministry of Health for the British Columbia provincial government. Vital Statistics collects mortality and morbidity data for BC based on the International Classification of Diseases (ICD) 10 codes. Our hypothesis, however, is that the ICD codes are used very differently in the field than is understood by the Ministry of Health. The codes are, in effect, boundary objects that have different meanings in specific communities of discourse but act as semantic emissaries between groups. Interviews with coroners and AIDS physicians as well as street nurses are designed to corroborate this hypothesis by illustrating that their use of the ICD classification terms to categorize HIV/AIDS deaths varies depending on their relationship with the patient, the deceased’s family context, and the degree to which the person responsible for assigning the category is comfortable with it. Varied usage of the same ICD categories illustrates that semantics have unique ontological or use contexts that form an important dimension of metadata. Moreover, context-based metadata provide the basis for making defensible decisions about data integration.

Clearly, it is not practical to reverse engineer all data to include more expansive metadata. Recognition of the value of dimensionalized metadata, however, is the basis for including them in future data collection projects. The database ethnographies project recognizes that communities of use (e.g., population health researchers in Canada) must be consulted and engaged with such efforts in order to ensure success— and, indeed, to elicit the necessary, contextual information. Ontology-based metadata are the most potent available venue for contextualizing data and providing a basis for making better decisions about data usage and interoperability. Boundary objects are more clearly understood between communities when they are contextualized.

Future Research: Semantic Heterogeneity in the World of Interoperability

I have illustrated here that there is a range of concerns about semantic standardization: (1) that semantic heterogeneity is as much a function of classification and scale as of domain; (2) that there is a need to retain ontological flexibility associated with semantic standardization; and (3) that the infrastructural role of NMOs is shifting as the privatization of data proliferates, and this will affect the ontologies of geographical objects. Although there have been a number of proposed and preliminary technical solutions to semantic standardization, including the stabilization of categories and federated databases, these (1) do not necessarily include process information and (2) often rely on exact classification, a set of well-defined rule for membership, or a specific data model. This critical approach to existing interoperability efforts should be interpreted not as a voice against algorithmic solutions to interoperability but as a plea for incorporation of social and qualitative input— in the form of metadata. We need to continue to solve technical problems while finding ways to work with complex, local, and variable geographical data. Despite the cumulative caution expressed thus far, interoperability and standardization remain the best hope we have for encouraging analysis and scientific cooperation at a global scale based on a growing pool of geospatial data.
Acknowledgements

I am indebted to Francis Harvey, who offered a number of lucid insights into the complex relationships between semantics, interoperability, and data infrastructures. This research was partially funded by Social Sciences and Humanities Research Council (SSHRC) grant #401-2001-1497 and by a grant from the Canadian Institutes of Health Research (CIHR), #116338. A much earlier draft of this article was presented at a workshop on GIS in a Changing Society coordinated by Duane Marble and held at the Center for Mapping, Ohio State University, in May 2001.

Author Information

Nadine Schuurman, Department of Geography, Simon Fraser University, 8888 University Drive, Burnaby, BC V5A 1S6 Canada. Tel.: (604) 291-3320. E-mail: nadine@sfu.ca.

References


Social Perspectives on Semantic Interoperability: Constraints on Geographical Knowledge from a Data Perspective


Social Perspectives on Semantic Interoperability: Constraints on Geographical Knowledge from a Data Perspective


cartographica

the international journal for geographic information and geovisualization