Health Informatics Journal

Vol 14(1): 27–45 [1460-4582(200701)14:1; 27–45; DOI: 10.1177/1460458207086333] www.sagepublications.com

A method to map heterogeneity between near but non-equivalent semantic attributes in multiple health data registries

Nadine Schuurman and Agnieszka Leszczynski

Health registries from multiple jurisdictions often include terms that are assumed to be semantically equivalent (e.g. fetal death and stillbirth). Closer examination reveals that such attributes have near – but non-equivalent – semantics. Thus their degree of semantic heterogeneity is an important indicator of uncertainty associated with data integration between registries. We build an OWL-encoded ontology which formalizes the relationships between similar perinatal concepts found in different databases. We also introduce the concept of *ontology-based metadata* **as a means of contextualizing such terms and linking context to the attribute data. This extended metadata are exported as XML from the health registries, and it – along with the OWL ontology – is interfaced via a webz-based GUI accessible to health researchers. The GUI mapping serves as the basis for making** *ad hoc* **comparison and integration decisions. Uncertainty is addressed by precisely mapping semantic heterogeneity between fields.**

Keywords

ontology, OWL, semantic data integration, semantic heterogeneity, Semantic Web

Introduction

Health policy decisions are based increasingly on assessment of spatial and statistical analyses within and between database registries [1–3]. This is part of a movement toward 'evidence-based medicine' which relies on data to inform resource allocation [4]. Comparison and integration of registry data are based on assumptions of semantic equivalence. For example, when perinatal (pregnancy, maternal and infant outcome) statistics are

assembled from multiple jurisdictions, semantic attributes such as stillbirth or pregnancy-induced hypertension often are only superficially similar and are related in terms of hierarchies and partial membership. Closer examination frequently reveals a lack of consistency in either use or interpretation.

Semantic integration (e.g. linkages and comparisons) remains a thorny problem with both spatial and non-spatial data. Since Bishr [5, 6] identified semantics as the most problematic of interoperability problems, much progress has been made. There is an emphasis, however, on feature level integration [7–10]. Moreover many strategies build upon object-oriented databases [8] whilst many businesses and government agencies continue to use simple relational database formats [11]. In this article, we examine this set of problems from the perspective of attribute data – the non-spatial information tied to geographic objects or locations – using examples from population health data registries.

Ontology-based or extended metadata are a means of providing context for variable terms that are otherwise considered transparent but are often interpreted dissimilarly [12]. We use an informatics interpretation of ontology which refers to the total universe of discourse associated with a given attribute (database field). In other words, an ontology encompasses the range of meaning offered by an encoded field, and contextualizes its use. Ontology-based metadata consist of eight additional fields that can be flexibly implemented at the attribute level on a needs basis. Ontology-based metadata can be encoded directly at the database level for selected attributes in additional tables linked to those attributes via relational keys. We propose using XML (eXtensible Markup Language) as a standardized syntax for tagging and exporting these extended metadata fields and illustrating linkage problems between horizontally organized jurisdictions (e.g. provinces) or between vertically organized multiple registries in the same jurisdiction.

We also construct an OWL (Web Ontology Language) ontology mapping of the relationships connecting semantic perinatal health terms (e.g. fetal death related concepts) between provincial-level jurisdictions – in this case, British Columbia (BC) and Nova Scotia (NS). This ontology makes the relationships between concepts explicit, and restricts which concepts can be compared or integrated. For example, pregnancy-induced hypertension is not related to stillbirth, and therefore the two should not be merged under any circumstances. We illustrate the potential for a JAVA-based GUI that creates a link to the OWL-generated ontologies via extended metadata exported in XML. The GUI will provide population health researchers with two critical pieces of information: (1) how two concepts in two separate registries relate – conveyed using OWL markup of property relationships; and (2) detailed descriptions about the semantic terms in the form of the XML-encoded extended metadata.

The use of web-based markup languages differentiates our semi-automated approach to data integration. We argue for the use of web-based encoding formats because they are emerging standards and hence take advantage of Semantic Web developments that will facilitate the realization of the 'Semantic Geospatial Web' (Egenhofer, 2002). They also provide a mechanism for realizing interoperability in distributed, web-based GIS environments linking organizations, spatial databases, metadata, applications and services [13].

Our approach is distinguished from previous attempts that employ markup languages for integration, many of which conform to the automation paradigm (identified below) and are oriented towards seamless spatial object exchange rather than semantic interoperability [8–10].

Context for mapping semantic heterogeneity using web-based tools

Background

The objective of this research is to map differences in semantic meaning when similar – and closely related but not identical – terms are used in different databases by incorporating multiple conceptual frameworks and methodologies. We begin with a review of several relevant literatures that bear on this work including precedents in semantic similarity metrics, traditional approaches to integration in medical informatics and bioinformatics, the need for extended metadata for non-spatial attributes, and Semantic Web technologies for encoding semantic differences.

The push for automated integration

Research at the intersection of biology and medicine, particularly genetics, and computing science has been pivotal in terms of both pioneering and operationalizing computing solutions for semantic data integration. Recent literature in this domain provides a comprehensive overview of traditional approaches – and their associated architectures – for interoperating medical/health databases. The simplest of these is the peer-to-peer [14] or point-to-point [15, 16] model, which consists of direct communication between all participating systems. This approach requires multiple interfaces (one for each connection), and moreover immediate knowledge of every other system's data, structure, and semantics [14, 15]. Rule-based links are based on formal conditions for semantic association, and are often established on the basis of database keys [16]. Data warehouses store all data, or their schemata abstracted versions, in a centralized repository [16–18]. In this model semantics are theoretically handled by schemata, which perform conversions on the basis of semantic (non)equivalence according to predefined semantic mapping rules between participating databases [16–18]. Broker architectures, premised on a middleware component that intercepts both requests for information and their retrieval, consist of a central mediator that handles data conversions between systems [15]. Brokers can be very simple, handling data types alone, or more complex as a component of federated data-bases [17]. The federated database model is one devoid of a central data repository; however, all data must adhere/conform to a common data model, again mandating that semantic translation is performed by automated schemata mapping [18]. Indeed integration in this last scenario is premised on a common data model of the source databases at some level of the integration architecture [16].

Semantic integration has similarly figured at the forefront of cutting-edge research in GIScience for the better part of a decade [8, 19, 20]. Early techniques for automating the integration of semantics are similar to those identified in the health informatics literature above, involving federated data sharing environments [19, 21, 22], schematic resolution of semantics [23], rules for class membership [24] and approaches based on semantic priming.

As in health informatics, these preliminary GIScience attempts have been superseded by efforts to leverage the reasoning capabilities of more sophisticated artificial intelligence formalisms – ontologies – for the encoding of semantic context and relationships at the

machine level [7, 20, 25–34]. Parallel efforts have focused on methods of incorporating multiple ontologies and representing them within a single system [7, 27, 35–39].

Previous (non-ontological) solutions are severely limited in several respects. Arzt [14, 17] asserts that in health informatics, all the identified models and architectures hinge on adherence to standardized data formats, messaging syntaxes (namely HL7), communication protocols, and most importantly, vocabularies/nomenclatures such as SNOMED. Indeed these non-ontological solutions do not really 'handle' semantics at all, but rather facilitate interoperability at the systems and syntactic levels [14, 17]. Gardner moreover emphasizes that while common data models provide a standardized way of moving between data and processes, 'Their success at this depends largely on the semantic richness and granularity of the model that they employ' [16, p. 1003].

Formal ontologies are a mechanism for capturing the granularity of the semantic relationships between concepts in a domain. Relationships between concepts are implicitly represented in the hierarchical structure of computation ontologies. In attempts to integrate data from multiple datasets, the role of semantics thus becomes one of determining degrees of equivalence between concepts: for example, are concepts truly equivalent (describe the same object), or is there a nested relationship between them [16]? Formal statements of relationships between concepts allow a more comprehensive, and 'fully descriptive', representation of knowledge [16, p. 1004]. Ontologies endow encoded content with a semantic structure which makes the gleaning of *context* – what data actually mean – tractable [40].

Realizing automation of ontological semantic integration, however, necessitates the adoption of full object-oriented systems, which in turn requires the re-engineering of existing databases, many of which are relational. There is evidence that few organizations are willing to commit resources toward a wholesale replacement of relational databases (RDBMS) with object-oriented (OO) systems [11]. Operationalizing these solutions furthermore rests on the assumption that semantics can be normalized [11]. This finding is paralleled in Blake and Bult's [40] recent work in bioinformatics involving the integration of the Gene Ontology (GO) with the smaller-scale Mouse Genome Informatics Database (MGI). The GO project defines the semantics of, and relationships between, identified genes. These are subsequently used to annotate – mark up – other data sources using these gene semantics; in this case, the genetic content stored in the MGI database. Semantic normalization is the crux of this integration exercise: the GO provides a 'semantic consistency to functional annotations for mouse genes' [40, p. 315]. In other words, the GO functions as an ontology, but more importantly in this scenario as a common framework for semantics, allowing for integration on the basis of common semantic annotation. Only databases that conform to normalization, however, are candidates for integration.

Ahlqvist describes an alternative to ontologies for semantic integration based on the use of rough fuzzy (RF) sets as a formalism for the representation of *uncertain conceptual spaces* [41, 42]. Two metrics become available for comparing the semantics of concepts: *overlap*, or the proportion of shared features/properties between concepts; and *distance*, which is formally the distance between two fuzzy membership functions, based on the conceptual model of psychological distance between features [41–43]. Ahlqvist's approach is unique in several respects in that it allows users to define concepts – including spatial phenomena – as continuous membership functions rather than discrete entities. This differs from the automated solutions described above, which necessitate an identification of finite objects in space [42]. However these approaches are also based on the assumption that

semantics can be fixed [11, 44]. It is also computationally intensive, furthermore requiring users to interface directly with the output matrices, which are generally unfamiliar to most practitioners and researchers. Making sense of these metrics is a prerequisite for their use in the interoperability equation.

Lord et al. [45] formally establish correlation between protein sequence similarity across databases and the semantic similarity of their annotations via statistical measures of semantic similarity between every possible pair of protein sequences in the GO on the basis of their annotations, which constitute three independent subgraphs of the ontology: molecular function, biological process, and cellular component. These annotations – which occur in the form of either semi-structured or free-text descriptions – comprise a standardized vocabulary across biological databases that support query across multiple resources mapped to the GO. Centred on the awareness that similar sequences will have similar annotations, semantic similarity is premised on the notion of 'information content', which pertains to the consideration that terms used less frequently – i.e. more specialized terms or children further down the hierarchical structure of the ontology – are 'more informative' [45, p. 603]. Semantic similarity is then measured using the semantic context of the metaclasses or parents shared by any two terms. Correlations were calculated over each aspect independently, and not over the entire GO, as each 'aspect' constituted its own data structure.

Lord et al.'s [45] statistical quantification of semantic with structural similarity represents an ontology mapping of sorts in that it establishes a correlation between proteins indexed in various remote data sources mapped to the global GO. Their approach to associating data resources is an instance of local–global ontology alignment as described by Choi et al. [46]. In their review of ontology mapping practices and the tools available for (semi)automating the process, the authors identify ontology mapping as the practice of associating entities in multiple ontologies on the basis of the semantic relations between concepts, which are 'semantically related at a conceptual level' [46, p. 35]. Mapping may be local–global as in the case of Lord et al. [45] where the GO itself is the global ontology to which all other remote, local ontologies are mapped; or local, wherein semantic links are established directly between source and target ontologies such that source concepts become members of the target ontology, an instance of ontology *merging*, which Choi et al. define as 'the process of generating a single, coherent ontology from two or more different ontologies related to the same subject [such that the] merged single coherent ontology includes information from all source ontologies but is more or less unchanged' [46, p. 35]. In contrast, ontology *integration* involves producing a singular ontology from multiple ontologies in disparate domains, whereas *alignment* consists of establishing relational links between ontologies and data sources that remain separate.

These statistical correlations of Lord et al.'s [45] approach constitute a semantic basis for querying proteins across bioinformatics databases via the GO, which comprises a standardized vocabulary whose semantics are fixed. Even where semantics are fixed, however, only a semi-automated solution is available at best with existing technology [46]. The development of requisite processing technologies, intelligent agents, etc. for automation is lagging behind tools for ontology building. Ahlqvist [43] moreover argues that the user is increasingly important in the context of interoperating ontologies. Expert knowledge must not only supplement ontology definitions, but must also be used to guide mapping between ontologies. Indeed many ontology-building platforms which facilitate ontological integration, such as Protégé, require user input in the ontology merging process. Ahlqvist [42] attributes this to unavoidable semantic uncertainty: vagueness of methodologies used to

define concepts; inherent vagueness between classes; and measurement granularity – for example, when a concept is a member of 2+ classes, there is confusion as to which class's measurement technique should be used in assigning spectral range (values and intervals).

Such semantic uncertainty is moreover a function of the often contentious process of standardization which precedes the ontological construction process [47]. Indeed ontology development is a process of collective negotiation; even where it occurs in a singular domain, the formalization of domain knowledge requires consensus on not only concept definitions and nomenclature standards, but also how the products of standardization are effectively interpreted on the ground. This is particularly pronounced in perinatal health, where events of pregnancy are often imbued with social connotations concerning women's reproductive roles. For example, 20 weeks gestation – recognized as the mid-pregnancy mark – provides definitional separation between spontaneous abortion (miscarriage occurring < 20 weeks) and stillbirth (fetal demise at e" 20 weeks). However, in British Columbia, because of the stigma associated with stillbirth, doctors routinely record incidences of fetal demise in the initial period beyond the 20 week gestation demarcation point as a miscarriage rather than as stillbirth. Conversely, the province of Alberta imposed stringent recording standards such that it statistically appeared to have experienced a significant increase in the rate of stillbirth, one moreover disproportionately higher than that reported by any other provincial perinatal body. In yet other Canadian jurisdictions all incidences of fetal demise are considered to be cases of 'fetal death', with no differentiation between therapeutic abortion, miscarriage, and stillbirth in recording.[1] This example attests to the need for social acquiescence even where knowledge representation takes place in a singular domain. Consensus is a prerequisite for the 'specification and transformation of domain knowledge into discourse' amenable to formalization as ontology [47]. Inherently a process of discretization, knowledge representation is thus a constraining process that in many contexts runs counter to the concept of knowledge as fluid, dynamic, and constantly renegotiated; indeed, fluidity is not an option.

The alternative solution introduced in this article parallels Ahlqvist's eschewal of fully automated integration solutions, but uses a formal ontology layer interfaced by a user-friendly GUI. The user need not navigate encoded concept maps – OWL ontologies – which may appear complicated to the untrained user. Furthermore it recognizes that classification systems are taxonomic and models hierarchy explicitly. These hierarchical relationships are preserved in relational database structures but conceptual spaces do not account for hierarchical relationships; overlap is only an indirect indicator of how concepts relate taxonomically, with *is-a* relationships inferred [41, 42]. Formal ontologies, however, *are* explicitly hierarchical – this is what becomes formalized. Thus a method that explicitly accommodates hierarchy is useful for many concepts, especially those encoded in extant relational database models.

In addition, most public registries are unable to commit the resources required to implement the statistical semantic similarity solution; they need something that both is compatible and interfaces with existing relational data models. We argue for mapping semantic heterogeneity using web technologies in the interest of implementation pragmatics. In order to map near but non-equivalent semantic terms for the purpose of integration or comparison, a mechanism must exist to translate the respective semantic terms from each database. Our pragmatic approach couples extended attribute metadata with Semantic Web technologies. We draw examples from population health databases,

where concepts have different definitions but are used for inter-jurisdictional and national-level aggregate comparisons.

The need for contextual information on non-spatial attributes

Extended metadata provide a framework for including context-based metadata for non-spatial attributes as a way of dimensionalizing attributes so that current and future users can assess the suitability of data for interoperability or comparative purposes. Such *ontology-based metadata* also provide historical context for archiving data. Moreover the methodology does not require reformatting of existing relational databases or metadata formats. It simply builds on current metadata formats by extending the fields to include information about methodological issues related to data collection, procedures used for data cleaning, especially those highlighting the derivation of any fields that resulted from data transformations or were otherwise derived, and issues related to limitations on the integration of data across computing platforms.

At present metadata – if included at all – are collected using wizards contained within existing software programs. The limitation of these metadata is that they focus on geometric properties of data such as latitude and longitude and positional accuracy of spatial data. They ignore metadata for any attribute that is not geometric. In this article, we introduce a mechanism for capturing ontological context using eight fields that can be linked to existing variable definitions using Semantic Web technologies.

Looking to the Semantic Web

The Semantic Web is a sprawling initiative for re-engineering the World Wide Web to facilitate the (semi)automated definition, linkage, and processing of web resource content. Resources include web pages, documents, data repositories etc. Ontologies are a critical component of the Semantic Web, functioning as standardized terminologies for communication between agents [48]. Because the Semantic Web effort addresses structuring content and developing technologies for processing it in intelligent ways, emerging Semantic Web standards and technologies can be leveraged to operationalize interoperability for geographic data. Using Semantic Web tools to encode semantics is a way to leverage the extensive, ongoing body of research conducted in the artificial intelligence domain.

We employ OWL (Web Ontology Language) as a means of formalizing relationships between concepts. OWL is a markup language with a formal, logical semantics – in this case, the ontology language Description Logics (DL). DL is more expressive than primitive first-order logics (FOL), allowing the definition of new concepts composed from existing concepts via necessary and sufficient conditions, including restrictions on properties (relationships between concepts) [48, 49]. OWL – whose semantics are 'defined via … translation' to DL [48, p. 13] – allows the formalization of domain knowledge as ontologies in terms of hierarchical relationships between concepts, explicitly supporting the encoding of hyponymic and merenomic relationships in a web environment. OWL furthermore exploits RDF (Resource Description Framework) as serialization syntax [48], and is well-formed XML.

While Agarwal [7] characterizes OWL as inferior to other logical semantics, specifically DAML + OIL, OWL has the advantage that it has been endorsed by the World Wide Web

Consortium (W3C) as a standard for the Semantic Web. Furthermore, it is supported by ontology building platforms such as Protégé (Stanford Medical Informatics) [50].

Methodology

Ontology-based metadata

We introduce 'ontology-based metadata' as a means of storing extended metadata to accompany standard variable definitions. We have developed eight fields – plus anecdotes – to add to existing frameworks that will enable ontological context to travel with the data [12]. These fields, identified in Table 1, are intended for the documentation of non-spatial attributes. The rationale behind their addition is that they provide data users with the pertinent information necessary for evaluation of data appropriateness that is lacking from conventional geographic metadata.

The information to populate extended metadata is gleaned through a technique called *database ethnographies* [12]. This involves conducting in-depth interviews with data producers to elicit details about the logics and methods behind data collection. It is based

Table 1 Ontology-based metadata fields: eight fields and anecdotes

Field	*Description*
Sampling methodologies	Indication of how data were collected. Was it a sample or a complete survey? What was the sampling grid size?
Definition of variable terms	Data definitions, naming conventions etc. used to identify and describe entities and attributes
Measurement specification	Measurement systems; instrumentation; thresholds as well as range (e.g. clarification of the maximum and minimum)
Classification system	Documentation of classification scheme used and taxonomic details. This is the collection of variable terms
Data model	Specification of proprietary DBMS, data structure, data model, and data formats. Also includes data trajectory and data model history – for example, has the data model changed? Were the data migrated from a legacy system?
Intended use	For what purpose were the data originally collected? For and by which domain? What were the logics behind data collection? The collection rationale?
Policy constraints	Legal and other constraints or influences on data collection, classification, and use
Linkages	What other databases, registries, etc. does the dataset link to? How does the entity/attribute relate both hierarchically and semantically to similar entities/attributes in other databases?
Anecdotes	Additional comments pertinent to understanding how to use the database – for example, is the entity or attribute subject to statistical anomalies? This field should not be confused with abstract, an often existing metadata field

on the premise that data collection practices and an understanding of standardized variable terms are unique to organizations, even in the presence of standardized variable terms.

Extended metadata are attached only to fields identified – via a flagging system in the database code or as separate files – as requiring additional, extended metadata. Furthermore only pertinent ontology-based metadata fields need be filled out for flagged attributes. An example of populated ontology-based metadata is provided in Table 2.

The development of ontology-based metadata is profoundly different from the current trend to incorporate ontological context at the model level in GIScience [20, 35, 37, 51]. It is pragmatic, however, in that it presents a vehicle for incorporating use context with data in a manner that is accessible; it requires little re-engineering; and it is intuitively understood by GIS users.

These eight fields are the basis for extracting the necessary information to export individual fields as XML to be used as a boundary object for the purpose of semantic comparison. Database managers have the option of either documenting these extended fields in a separate file in the tradition of conventional metadata, or at the level of the database (Figure 1). The second alternative allows extended metadata to be stored directly with the data themselves in separate tables linked to flagged attributes via relational keys. The advantage is that the descriptive information travels with the data. Both these approaches conform to existing data models: either a separate file sits on top of the database, and does not interrupt it; or the database flagging system integrates seamlessly with relational data structures. In both instances, the extended metadata are encoded as XML: either the metadata are exported as such from the database, or the metadata file is converted/stored in XML format (most spatial metadata editors have an XML option).

Encoding semantic context

We construct OWL ontologies as formal representations of semantic concepts in perinatal databases using Protégé (Stanford Medical Informatics) [50]. Each database is ontologically modelled. Semantic comparison is determined via merging ontologies on the basis of a 1:1 mapping between local ontologies. While ontology development is centralized in our approach, it nevertheless affords the high degree of interoperability 'as mediation between distributed data in [local ontology] environments' [46, p. 35].

We do not, however, operationalize any of the tools identified by Choi et al. [46] for semi-automating the mapping process. Instead mapping is guided by the user – akin to a supervised classification of remotely sensed imagery – and is executed directly in Protégé. This is particularly salient in the context of local ontology mapping, which is identifiably more complicated than that between an integrated global ontology and multiple local ontologies because, in the latter, mapping rules are more easily defined as all mapping is unidirectional to an intermediate context of standardized definitions contained in the global ontology [46]. Moreover, local–local mapping is much more scaleable and thereby more easily streamlined with the web; indeed Choi et al. [46] identify the primary application of local ontology mapping to be the Semantic Web.

Because a global ontology represents a semantic 'least common denominator', however, the immediate limitation is a loss of semantic granularity. Thus in the process specified herein, the user both authors mapping operations and selects which of the proposed operations, suggested by the system on the basis of auto-detected class similarity, to execute. Merging allows identical concepts, such as 'baby' and 'infant' in respective databases, to

Table 2 Ontology-based (extended) metadata for STILLBIRTH and FETAL DEATH fields in perinatal databases for two provincial jurisdictions in Canada: British Columbia (BC) and Nova Scotia (NS). 'Sampling methodologies' and 'measurement specification' were not included as extended metadata fields because they were not applicable in this case (i.e. there was no pertinent information for these fields)

	Stillbirth BC (BC Perinatal Database)	*Fetal death NS (NS Perinatal Database)*
Definition of variable terms	The complete expulsion or extraction from its mother after at least 20 weeks or weighing at least 500 grams, of a product of conception in which, after expulsion or extraction, there are no signs of life (breathing, beating of heart etc.)	Fetal death before birth
Classification system	This definition conforms to ICD-10-CA (Canadian modification)	A 'hybrid' classification including a system (ATLEE database) developed in-house by Nova Scotia Reproductive Care, and ICD-10-CA (Canadian modification) definitions
Data model	Flat file	RDBMS
Intended use	Reproductive care	Reproductive care
Policy constraints	n/a	Termination of pregnancy – therapeutic abortion – affects the reporting of fetal death because technically, under the present definition, any fetal death > 500 g or > 20 weeks is classified as a fetal death
Linkages	n/a	Includes STILLBIRTH (BC) IFF ≥ 500 g or ≥ 20 weeks
Anecdotes	'Stillbirth' is associated with social stigma, resulting in physicians reporting stillbirths as miscarriages (spontaneous abortion) around the 20/21 week or 500 g range. The reporting of this statistic is hence affected by a statistical anomaly	The physician's determination may supersede the data definition: for example, 'recorded as weighing ≥ 500 g OR when documented as a fetal death by the physician'. Hence there is considerable room for misreporting fetal death versus abortion, defined as a pregnancy resulting in a fetus weighing > 500 g or > 20 weeks

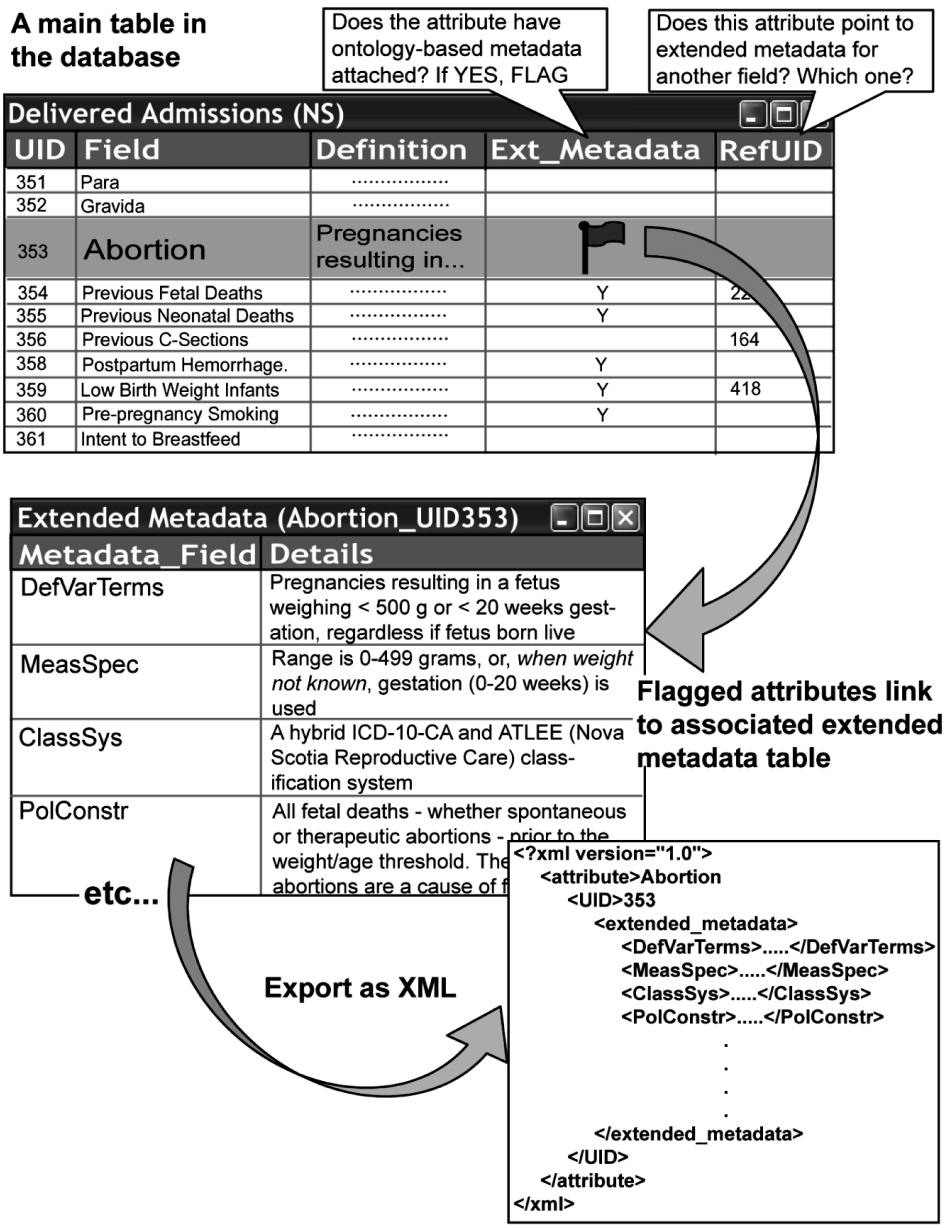

Figure 1 Storing extended metadata directly in the database, and exporting as XML

be *merged*, producing a composite class which inherits the subclasses from both original (input) ontologies.

The end product is a new ontology composed of the merged classes – in other words, a formal mapping of semantic equivalences. Because many of the subclasses inherited by the merged classes are similar but not equivalent, they themselves are not merged.

However, they can be related in the context of the merged ontology via properties identifying hierarchical relationships (*is-a, has-a, kind-of, part-of*) and restrictions on those properties (logical quantifiers and cardinality restrictions) which specify necessary and sufficient conditions for class/concept membership. For example, stillbirth can be identified as a *kind-of* fetal death.

The rationale for using OWL for ontology construction lies in its ability to explicitly encode *relationships* between concepts via properties and property restrictions. OWL – as a DL-based language – is ideally suited for such mapping as it builds upon formalization of concept trees. Indeed named classes alone do not convey meaning. A mechanism to make semantic context explicit is required and this is achieved through the encoding of relationships between classes. OWL makes these representations computationally formal. Property restrictions are furthermore an explicit indication of how the semantics of two concepts relate hierarchically. OWL can provide explicit mapping where, for example, two concepts are equivalent *or* where one concept is subsumed by another *or* where one concept is a partial member of another class.

We map each ontology to every other ontology, 1:1. Because there are a finite (and small) number of perinatal databases in the country, this is manageable as it remains tractable. It is an alternative approach to Ahlqvist's [43] that is most suitable when designing for interoperability between a small number of datasets, and indeed most health linkages. It also involves minimal infrastructure and expertise, and is therefore pragmatic from an implementation perspective.

Because we map semantics explicitly as opposed to using a proxy context or global ontology, both of which establish equivalency between similar yet heterogeneous terms and an intermediate definition which constitutes a lowest-common-denominator type of annotation, we produce a more precise encoding of semantic heterogeneity. Ontological mapping is hence a means of formally encoding uncertainty because it captures concept relationships even where the concepts themselves are vaguely defined. It also serves to mitigate the problems associated with indeterminate conceptual boundaries by recognizing that even though entities are conceptually vague they are related to each other in specific ways.

Bringing together ontology-based metadata and formal ontology

We construct a JAVA API to interface both the ontology-based metadata and the formal ontologies. We chose JAVA because it interfaces well with OWL, and is web-based. This GUI-based application consists of a series of screens that take the user through the semantic comparison process (Figure 2). First, the user selects which two perinatal registries to com-pare from parallel dropdown menus and this indicates which merged ontology will be called up. Next the user selects which two concepts to compare. Property restrictions encoded in the ontology serve to limit which concepts can be compared: only those concepts semantically associated via property restrictions, which indicate that they are indeed related, may be selected. This is critical because it limits the possible comparisons to be made between individual concepts to those which are semantically legitimate. The OWL language contains tags to explicitly code property restrictions as relationships, and hence this is extracted directly from the markup (the ontology). The subsequent

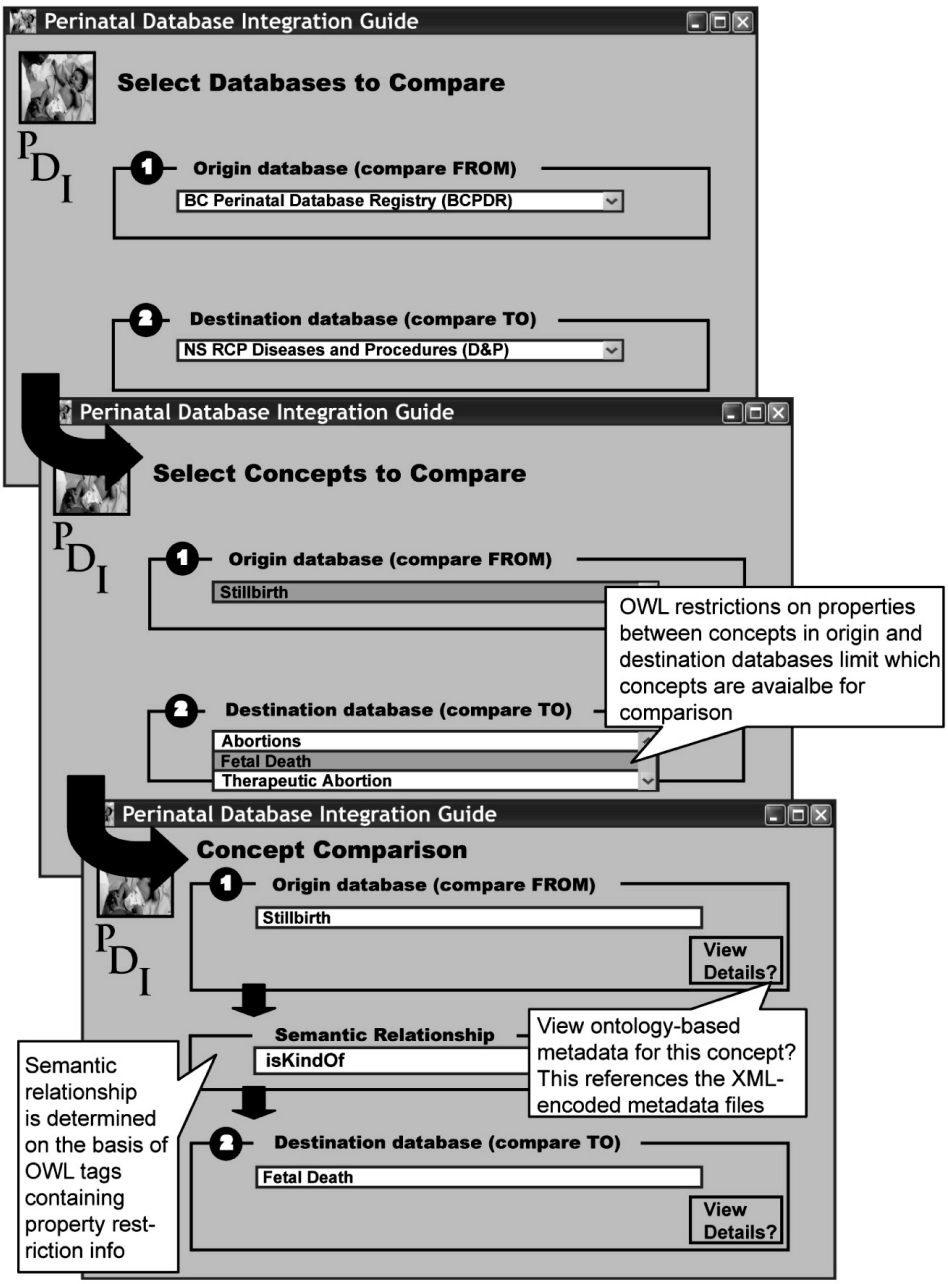

Figure 2 Java-based GUI for comparing perinatal concepts across jurisdictions

screen presents to the user the nature of the semantic relationship between the selected concepts – *stillbirth is-kind-of fetal death*. Using quantifier and has-value restrictions allows us to recursively capture the nature of the relationship between these two closely related but non-equivalent concepts. Imposing the existential (∃) quantifier on the *includes* property allows the explicit statement, 'At least one value of *includes* for *fetal death* must be of type *stillbirth*'; in other words, it is a specification that fetal death includes the concept 'stillbirth'. The inverse of this relationship is likewise formalized via the use of the has-value restriction on the *is-kind-of* property. This renders the expression, 'The *is-kind-of* property for *stillbirth* must have the value *fetal death*'; in natural language, 'stillbirth is a kind of fetal death'. The need to recursively define relationships may seem redundant, but it clarifies semantics such that there is no uncertainty as to whether or not, for example, *all* instances of stillbirth are kinds of *some* fetal death (the intended meaning), versus *only* instances of stillbirth are kinds of *all* fetal death (a formal confusion). This screen also links to the XML ontology-based metadata files providing detailed information – context – for each concept.

Nothing is actually physically 'integrated' through this web-based GUI application. Rather, it functions to guide comparison and integration decisions by providing users with the detailed contextual information needed to make legitimate determinations of semantic similarity on the basis of formal concept mappings. It is also user-friendly, releasing users from dealing with complications of line-based code which may be unfamiliar and intimidating as well as difficult to navigate.

Implementation

We illustrate the implementation of our coupled methodology with an example from population health. We compare fetal death related concepts in the Nova Scotia (NS) perinatal database to those in the British Columbia (BC) registry. Both are provincial jurisdictions and are therefore horizontally equivalent. We began by collecting ontology-based metadata through extensive interview sessions with data stewards, managers, and clinical practitioners and analysts affiliated with both reproductive care programs (BC and NS). These extended metadata were XML encoded.

In terms of formalization, we first constructed two ontologies, each reflecting the structure and organization of concepts in each respective relational database using Protégé [50]. We then performed mappings between – or merged – the two ontologies using the Protégé [50] plug-in (Figure 3). The result was a new ontology containing the merged classes and all the original superclasses. Subsequently we related concepts inherited from the two main merged superclasses – mother and baby (infant) – in the resulting ontology via the use of property restrictions. Protégé is a frame-based GUI representation of OWL/DL, which allows users to interact with concepts directly in a visual environment while simultaneously generating the corresponding OWL code. Figure 4 shows both the visual representation of concepts related via property restrictions, and the associated OWL code. We then implemented our interface on top as a means of serving the information to users. This mapping represents a selective example using a small number of concepts but effectively illustrates the potential for mapping semantic difference of near-identical but non-equivalent concepts and their hierarchical relationship using OWL.

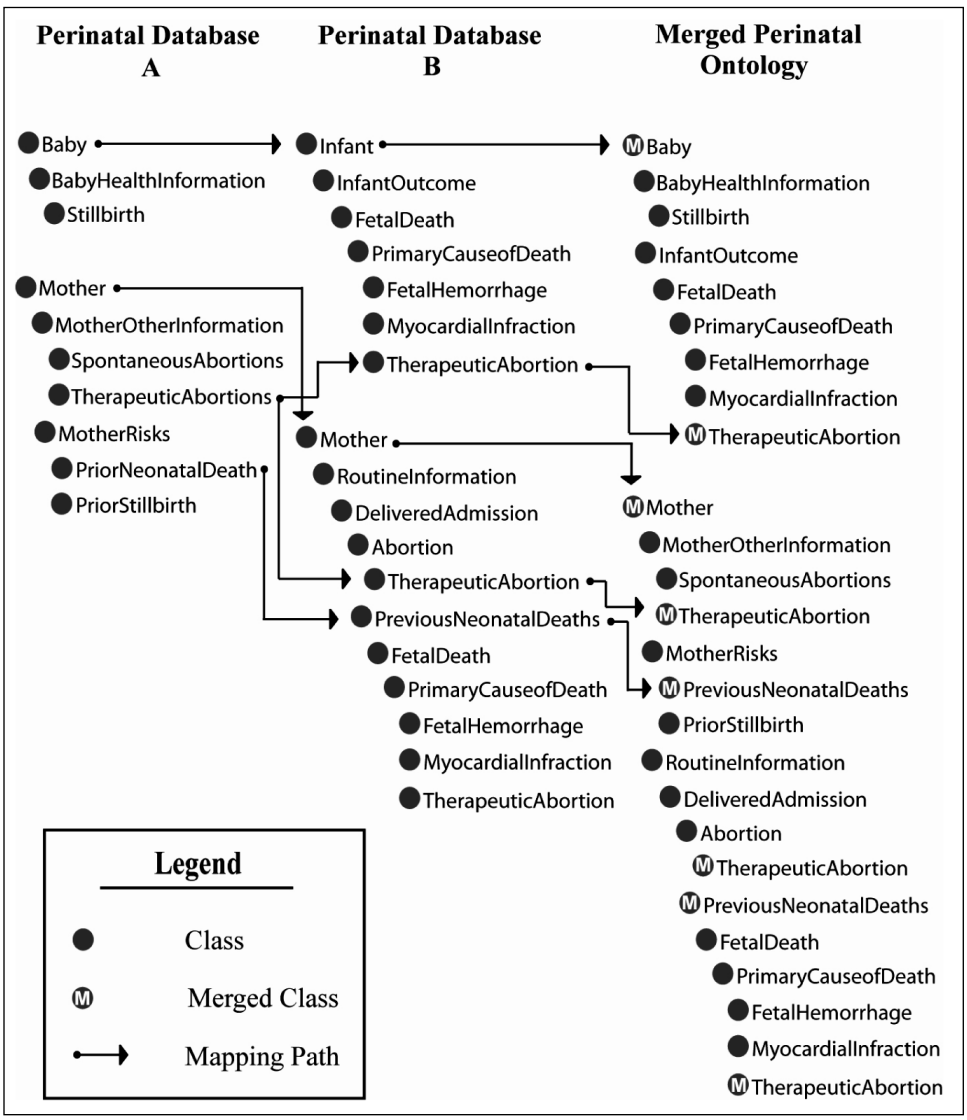

Figure 3 Mapping fetal-death concepts between the BC and NS perinatal database registries; merged concepts are marked in the output ontology

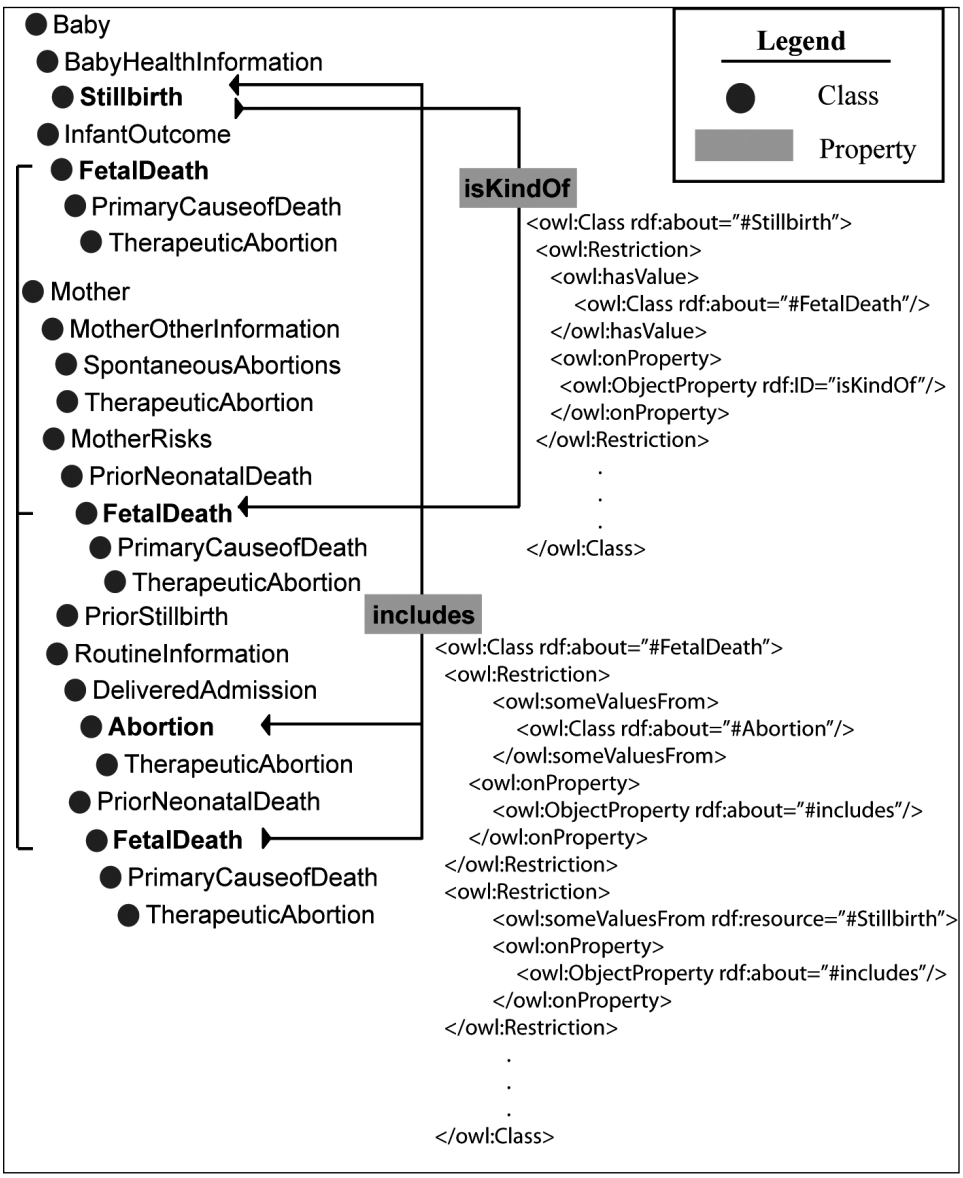

Figure 4 Relating fetal death and abortion from the Nova Scotia database to stillbirth found in the BC registry. This is accomplished via properties, and restrictions for class membership imposed on those properties. Note the OWL formal representation of the graphical relationships

Conclusion

There is increasing demand for integration of similar but non-identical non-spatial attributes between jurisdictions (e.g. provinces) and between multiple related databases within jurisdictions (e.g. linkages between perinatal and diabetes registries) in population health.

As we have argued, semi-automated integration of related semantic fields from multiple databases can be facilitated by mapping their relationships in a formal manner. In this article, we reviewed other methods of creating the necessary equivalences with focus on automated solutions and their limitations. Early integration approaches can be characterized as distinctly non-ontological; they were premised on simple architectures and protocols, such as peer-to-peer data sharing, where semantic integration is enabled in the form of rule-based links, schematic resolution of semantics, mediation, broker architectures, and semantic priming. While the majority of current efforts at realizing interoperability involve operationalizing formal ontologies, other methods pursue statistical metrics of concept and semantic similarity.

Whilst some of this work offers conceptually superior solutions, we find that for linking a relatively small number of databases, a Semantic Web solution using OWL is more attractive from a pragmatic implementation perspective. We introduced the concept of ontology-based metadata as a mechanism to 'hold' extended context for non-spatial attributes. These metadata can be encoded or exported in XML and subsequently incorporated into OWL. We used OWL as a recognized Semantic Web technology to map semantic heterogeneity between concepts as well as create a merged class concept. This technology can be implemented in relational databases using a user-friendly GUI and stands to facilitate understanding of semantic difference among data users and analysts.

Notes

1 Information gleaned through extensive interviews with the data stewards at the British Columbia Reproductive Care Program (BCRCP) and the Nova Scotia Reproductive Care Program (NS RCP).

References

1 Barnard D K, Hu W. The Populath Health Approach: health GIS as a bridge from theory to practice. *International Journal of Health Geographics* 2005; **4**; 9 pp.

2 Ricketts T C. Geographic information systems and public health. *Annual Review of Public Health* 2003; **24** (1); 1–6.

3 Solbring H R. Metadata and the reintegration of clinical knowledge: ISO 11179. *MD Computing* 2000; **17**; 25–8.

4 Moynihan R, Cassels A. *Selling Sickness: How the World's Biggest Pharmaceutical Companies Are Turning Us All into Patients*. Vancouver: Greystone, 2005.

5 Bishr Y. Semantic aspects of interoperable GIS. PhD thesis. Enschede, The Netherlands: International Institute for Aerospace Survey and Earth Sciences (ITC), 1997.

6 Bishr Y. Overcoming the semantic and other barriers to GIS interoperability. *International Journal of Geographical Information Science* 1998; **12** (4); 299–314.

7 Agarwal P. Ontological considerations in GIScience. *International Journal of Geographical Information Science* 2005; **19** (5); 501–36.

8 Peng Z-R. A proposed framework for feature-level geospatial data sharing: a case study for transportation network data. *International Journal of Geographical Information Science* 2005; **19** (4); 459–81.

9 Peng Z-R, Zhang C. The roles of geography markup language (GML), scalable vector graphics (SVG), and Web feature service (WFS) specification in the development of Internet geographic information systems (GIS). *Journal of Geographical Systems* 2004; **6** (2); 95–116.

10 Tsou M-H. An operational metadata framework for searching, indexing, and retrieving distributed geographic information services on the Internet. In Egenhofer M J, Mark D M eds *GIScience 2002*, Boulder, CO, 313–32. New York: Springer, 2002.

11 Schuurman N. Flexible standardization: making interoperability accessible to agencies with limited resources. *Cartography and Geographic Information Science* 2002; **29** (4); 343–53.

12 Schuurman N, Leszczynski A. Ontology-based metadata. *Transactions in GIS* 2006; **10** (5); 709–26.

13 Nogueras-Iso J, Zarazaga-Soria F J, Lacasta J, Bejar R, Muro-Medrano P R. Metadata standard interoperability: application in the geographic information domain. *Computers, Environment and Urban Systems* 2004; **28** (6); 611–34.

14 Arzt N H. The new alphabet soup: models of data integration. Part 1. *Journal of Healthcare Information Management* 2005; **20** (1); 15–18.

15 Wangler B, Ahlfeldt R-M, Perjons E. Process oriented information systems architectures in healthcare. *Health Informatics Journal* 2003; **9** (4); 253–65.

16 Gardner S P. Ontologies and semantic data integration. *Drug Discovery Today* 2005; **10** (14); 1001–7.

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

18 Wang K, Tarczy-Hornoch P, Shaker R, Mork P, Brinkley J. BioMediator data integration: beyond genomics to neuroscience data. *AMIA Annual Symposium 2005* 779–83. AMIA, 2005.

19 Abel D J, Ooi B C, Tan K-L, Tan S H. Towards integrated geographical information processing. *International Journal of Geographical Information Science* 1998; **12** (4); 353–71.

20 Kuhn W. Modelling the semantics of geographic categories through conceptual integration. In Egenhofer M J, Mark D M eds *GIScience 2002* Boulder, CO, 108–18. New York: Springer, 2002.

21 Devogele T, Parent C, Spaccapietra S. On spatial database integration. *International Journal of Geographical Information Science* 1998; **12** (4); 335–52.

22 Sheth A P. Changing focus on interoperability in information systems: from system, syntax, structure to semantics. In Goodchild M F, Egenhofer M J, Fegeas R, Kottman C eds *Interoperating Geographic Information Systems* 5–29. Boston: Kluwer, 1999.

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

24 Stock K, Pullar D. Identifying semantically similar elements in heterogeneous spatial databases using predicate logic expressions. In Vckovski A, Brassel K E, Schek H-J eds *INTEROP'99* 231–52. Zurich: Springer, 1999.

25 Agarwal P. Contested nature of place: knowledge mapping for resolving ontological distinctions between geographical objects. In Egenhofer M J, Freska C, Miller H J eds *GIScience 2004* Adelphi, MD, 1–21. New York: Springer, 2004.

26 Bittner T, Edwards G. Towards an ontology for geomatics. *Geomatica* 2001; **55** (4); 475–90.

27 Duckman M, Worboys M. An algebraic approach to automated geospatial information fusion. *International Journal of Geographical Information Science* 2005; **19** (5); 537–57.

28 Fabrikant S I, Buttenfield B. Formalizing semantic spaces for information access. *Annals of the Association of American Geographers* 2001; **91** (3); 563–80.

29 Frank A U. Tiers of ontology and consistency constraints in geographical information systems. *International Journal of Geographical Information Science* 2001; **15** (7); 667–78.

30 Kuhn W. Geospatial semantics: why, of what, and how? *Journal of Data Semantics III* 2005; LNCS 3534; 1–24.

31 Rodriguez M A, Egenhofer M J. Determining semantic similarity among entity classes from different ontologies. *IEEE Transactions on Knowledge and Data Engineering* 2003; **15** (2); 442–56.

32 Smith B, Mark D M. Geographical categories: an ontological investigation. *International Journal of Geographical Information Science* 2001; **15** (7); 591–612.

33 Stoimenov L, Djordjevic-Kajan S. An architecture for interoperable GIS use in a local community environment. *Computers & Geosciences* 2005; **31** (2); 211–20.

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

35 Brodeur J, Bedard Y, Edwards G, Moulin B. Revisiting the concept of geospatial data interoperability within the scope of human communication processes. *Transactions in GIS* 2003; **7** (2); 243–65.
36 Cruz I F, Sunna W, Chaudhry A. Semi-automatic ontology alignment for geospatial data integration. In Egenhofer M J, Freska C, Miller H J eds *GIScience 2004* Adelphi, MD, 55–61. New York: Springer, 2004.
37 Fonseca F, Davis C, Camara G. Bridging ontologies and conceptual schemas in geographic information integration. *GeoInformatica* 2003; **7** (4); 355–78.
38 Kokla M, Kavouras M. Semantic information in geo-ontologies: extraction, comparison, and reconciliation. *Journal on Data Semantics III* 2005; LNCS 3353; 125–42.
39 Winter S, Nittel S. Formal information modelling for standardisation in the spatial domain. *International Journal of Geographical Information Science* 2003; **17** (8); 721–41.
40 Blake J A, Bult C J. Beyond the data deluge: data integration and bio-ontologies. *Journal of Biomedical Informatics* 2006; **39** (3); 314–20.
41 Ahlqvist O. A parametrized representation of uncertain conceptual spaces. *Transactions in GIS* 2004; **8** (4); 493–514.
42 Ahlqvist O. Using uncertain conceptual spaces to translate between land cover categories. *International Journal of Geographical Information Science* 2005; **19** (7); 831–57.
43 Ahlqvist O. Using semantic similarity metrics to uncover category and land cover change. In Rodriguez M A, Cruz I F, Egenhofer M J, Levashkin S eds *GeoSpatial Semantics (GeoS 2005)* Mexico City, 107–19. Berlin: Springer, 2005.
44 Schuurman N. Why formalization matters: critical GIS and ontology research. *Annals of the Association of American Geographers* 2006; **96** (4); 726–39.
45 Lord P W, Stevens R D, Brass A, Goble C A. Semantic similarity measures as tools for exploring the gene ontology. *Pacific Symposium Biocomputing, 3–7 January 2003, Lihue, Hawaii* 601–12.
46 Choi N, Song I-Y, Hyoil H. A survey on ontology mapping. *SIGMOD Record* 2006; **35** (3); 34–41.
47 Ribes D, Bowker G C. A learning trajectory for ontologies. In *4th Annual Knowledge and Organizations Conference, 2005, Long Beach, CA*.
48 Baader F, Horrocks I, Sattler U. Description logics. In Staab S, Studer R eds *Handbook on Ontologies* 3–28. Berlin: Springer, 2004.
49 Tu S. *Protege OWL Short Course*. Palo Alto, CA: Stanford Medical Informatics, 2005. 14 pp.
50 Informatics S S M. *The Protege Ontology Editor and Knowledge Acquisition System*. 2005.
51 Guarino N, Masola C, Vetere G. Onto-Seek: content-based access to the web. *IEEE Intelligent Systems and their Applications* 1999; **14** (3); 70–80.

Correspondence to: Nadine Schuurman

Nadine Schuurman PhD,
Associate Professor
Department of Geography, Simon Fraser University, RCB 7123, 8888 University Drive, Burnaby BC, Canada, V5A 1S6
Tel: 1 604 291 3320
Fax: 1 604 291 5841
E-mail: nadine@sfu.ca

Agnieszka Leszczynski
Department of Geography, Simon Fraser University, RCB 7123, 8888 University Drive, Burnaby BC, Canada, V5A 1S6