Semantic Integration of Authoritative and Volunteered Geographic Information (VGI) using Ontologies

Jimena Martínez Ramos  
Memorial University of Newfoundland  
NL A1C 5S7  
St. John's, Canada  
jmr125@mun.ca

Arnaud Vandecasteele  
Memorial University of Newfoundland  
NL A1C 5S7  
St. John's, Canada  
a.vandecasteele@mun.ca

Rodolphe Devillers  
Memorial University of Newfoundland  
NL A1C 5S7  
St. John's, Canada  
rdeville@mun.ca

Abstract

Volunteered Geographic Information (VGI) projects such as OpenStreetMap (OSM) have been growing in users and edits in the past few years, resulting in more accurate, complete and up-to-date datasets. As a result, governments are now interested in VGI as a source of information that could help to improve their own datasets. Integrating geographic information requires solving semantic heterogeneity in a way that allows the identification of features in different datasets representing the same real phenomena.

This paper presents a method that uses ontologies for identifying sets of homologous features in VGI and authoritative geographic sources. The method is based on two steps. First, a domain ontology is used as a common knowledge among datasets. Second, direct mapping from the datasets to the domain ontology is done, resulting in new datasets using the same vocabulary.

Keywords: VGI, authoritative geographic information, semantic heterogeneity, geographic information integration, ontology, interoperability.

1 Introduction

The integration of geographic information coming from different sources has been studied over the last two decades in order to effectively reuse and exchange increasing amounts of data and meet the increasing need of geographic information in different applications and fields [1, 2].

The integration of geographic information is also gaining importance with the recent development of Volunteered Geographic Information (VGI), since many National Mapping Agencies (NMAs) are looking at integrating VGI from projects like OpenStreetMap (OSM) to improve the freshness of their datasets, their processes, and possibly reduce their production costs [3].

Integrating two or more geospatial datasets involves making the correspondence between objects of different datasets at the geometric and semantic levels [4]. Semantically, the integration of two datasets is possible when both datasets share a common knowledge (commonness) through which they can communicate [5].

One of the main obstacles to a seamless integration of geographic information is semantic heterogeneity [6, 7], which refers to the different ways datasets conceptualize real phenomena [8]. Because of semantic differences, two feature classes named the same in two datasets may not represent the same phenomenon, while feature classes named differently in two datasets may refer to the same phenomenon.

While semantic heterogeneity makes the integration of geographic datasets challenging, this issue becomes even more difficult in the context of VGI datasets, due to their flexible and dynamic data models [9, 10].

Although the OSM data model has a common set of features and attributes combined in tags and recommendations on how to use them, users from all over the world can freely add new tags to the map [2, 9]. This freedom, which creates richer datasets semantically, is also what makes the integration of OSM datasets more challenging. Semantic heterogeneity may occur inside OSM datasets, since one real phenomenon could be represented with different tags for different users, and two users could use the same tag to represent different phenomena. This means that, while common knowledge between two authoritative datasets is static because of their fixed data models (Figure 1, left), common knowledge of OSM and authoritative datasets is dynamic, since users can add their knowledge to the OSM dataset (Figure 1, right).

Figure 1: Representation of the common vocabulary used by authoritative and VGI (OSM) datasets. Circles represent the knowledge used by each dataset or OSM user and darker grey areas illustrate the common knowledge.
domain ontology, following recent directions from standardization bodies like ISO/TC 211 and Open Geospatial Consortium (OGC). The domain ontology becomes a common knowledge onto which authoritative and VGI datasets can interface. Direct mappings between the datasets and this domain ontology are then performed, without having to build an ontology for each dataset.

Section 2 presents studies that have used ontologies to handle semantic heterogeneities and studies that compared VGI and authoritative datasets. Section 3 describes the proposed method for semantically integrating authoritative and VGI datasets. Conclusions and future works are discussed in Section 4.

2 Literature review

Interoperability can be defined as the ability of two or more systems to exchange and understand information [11]. To be interoperable, systems need to solve heterogeneities between them.

Such heterogeneities have been classified into four levels [6] and, applied to geographic information [5], they are: ‘system’, which refers to different networks where databases are stored; ‘syntactic’, which refers to the data formats (e.g., shapefile) and data types (e.g., raster, vector); ‘structural’, which refers to the ways data are modelled in each database, and ‘semantic heterogeneity’, which refers to differences of meaning between concepts.

Most research and standardization efforts have focused on the first three levels (i.e., system, syntactic and structural), while solving semantic heterogeneity (i.e., the fourth level) remains a challenge to effectively exchange or reuse geographic information and allow interoperability [5, 6, 12, 13].

Ontologies are one of the most interesting approaches to handle semantic heterogeneity [1] because of their capacity to link different conceptualizations of the same reality [8]. Studer [14] describes an ontology as being “a formal, explicit specification of a shared conceptualization”. Ontologies represent a domain of knowledge using five components: classes, relations, functions, axioms and instances.

Guarino [15] identifies four types of ontologies based on their level of generality: ‘top-level’ ontologies, which describe very general concepts independently of a particular domain, ‘domain’ and ‘task’ ontologies, referring respectively to the vocabulary of generic domain or generic tasks or activities, and ‘application’ ontologies, describing concepts depending on a particular domain and task.

Ontology matching is one of the most common approaches used by studies handling semantic heterogeneities between geographic datasets [13, 16]. This approach creates mappings between related concepts of two different ontologies [17] built in advance from each dataset. To match concepts from these two ontologies, different semantic similarity techniques are often used [5, 16], based on measures of differences between concepts or their descriptions.

A general problem with these methods is that the user or provider of the data has to build a new application ontology for each new dataset [13]. In addition, semantic similarity techniques require more research [10] as they are still unable to solve the matching between ontologies, due to different descriptions of concepts in different ontologies [8]. Besides, most studies have focused on the semantic interoperability of authoritative datasets, and no solution is presented for the case of VGI datasets as OSM.

There are different approaches and motivations to the integration of authoritative and VGI datasets. Several studies have compared the two in order to assess the quality of VGI [18, 19]. Other studies have explored the use of OSM datasets to improve NMAs production or updating processes [20, 21], and some studies looked at the integration of both datasets to obtain the best information from each one [9, 22].

In these studies, semantic heterogeneities between VGI datasets and authoritative datasets were usually resolved manually by identifying similar feature classes between datasets for their names or descriptions. A quality assessment of the VGI dataset was then performed, assuming that the semantic matching between classes was correct.

However, this way of solving semantic heterogeneity can lead to two types of problems. First, some of the mistakes detected could result from an incorrect matching of the features. Second, this approach only works for the two datasets used in the study, so the matching cannot be reused and needs to be recreated for each new dataset.

A recent study working with semantic similarity measures to handle semantic heterogeneity between VGI and authoritative geographic information concludes that semantic similarity alone is not sufficient to solve the problem and that more research is needed [2].

3 Method

The proposed method tries to overcome two main problems found in other methods handling semantic heterogeneity in the context of VGI.

The first problem is that techniques for matching classes from two datasets may not always present a satisfactory solution for VGI. As discussed in Section 2, VGI datasets can have different conceptualizations within a single dataset. A more general and upper conceptualization is needed to facilitate the matching of datasets that have dynamic data models like OSM. We propose the use of a domain ontology as a general and superior conceptualization of geographic information.

The second problem is that the approach used by most studies, building an application ontology for each dataset, leads to two difficulties: users need to know how ontologies work, and no reuse of the matching is allowed, requiring a new matching to integrate each pair of datasets.

To solve these difficulties, we propose the use of mappings from datasets to the domain ontology. This way, datasets can be used in their original formats and no knowledge about ontologies is needed from the user or provider of the data. Besides, mappings from one dataset to the domain ontology could be reused. The results of these mappings, as new datasets semantically interoperable, can be used for quality comparison purposes, or queried for more specific results.

Section 3.1 describes how to build the domain ontology, and once it is built, Section 3.2 explains how to match semantically the two datasets throughout direct mappings.
3.1 Using a domain ontology

The proposed approach for building a domain ontology is based on previous studies [17, 23] using a pivot for making geospatial datasets semantically interoperable. Cruz [17] uses a global ontology onto which other local ontologies are linked and Boucher and Zimányi [23] use a pivot ontology for solving syntactic heterogeneity (i.e., differences in formats) between datasets. This ontology will serve as a pivot allowing matching between different datasets.

Although there are many ontologies describing geographic information, as GeoRSS [24], the US National Map topographic ontology [25], the NASA Semantic Web for Earth and Environmental Terminology (SWEET)¹, the DBPedia², the Ordnance Survey topographic ontology [26], an OSM ontology (OSMonto) [27]), there is no standard to describe geographic features.

Because standards provide the fundamental structure for sharing and integrating geographic information [5], the proposed domain ontology should be a standard, allowing interoperability between different types of datasets, and reusing mappings from static data models. The domain ontology should hence contain not only meta-concepts as geographic features or properties, but also the concepts themselves (e.g., river, road). In this sense, the General Feature Model (GFM) [28] is the starting point and the development of the domain ontology should follow its structure.

GFM is based on two main features, GF_FeatureType and GF_PropertyType, both are meta-classes that could represent, respectively, feature classes and their properties.

3.2 Identification of key concepts

There are always different possible alternatives when modelling a domain and the best solution will depend on the application the ontology will have. This domain ontology aims to allow semantic interoperability of a broad range of geographic information sources, from authoritative to VGI sources, with different data models and structures.

In this context, classes or concepts in the ontology must be clearly defined with some level of generality and must be close to objects (geographic phenomena) [7]. Relationships must be defined in the general case and should be close to properties (attributes in the datasets) [29]. This will allow the matching between features and attributes from different datasets to the concepts of the domain ontology.

Following a top-down approach, general concepts have to be defined first, then more specific concepts can be created as sub-classes. Finally, relationships between concepts must be established.

Following this approach, and based on the GFM, two superclasses are proposed (Figure 2):

- **GeographyFeature**, refers to every geographic phenomenon.
- **FeatureProperty**, refers to every property or attribute and their values.

![Figure 2: Structure of the domain ontology: superclasses, classes and sub-classes and their relationships through object properties. Example with concepts and properties related to Transportation.](image)

3.2.1 Classification of concepts and sub-concepts

Concepts are organized hierarchically into the ontology with concepts of a higher level having sub-concepts attached to them. In our example, themes of features (e.g., transportation, energy, hydrology) are represented as classes under the superclass GeographyFeature, and feature types (e.g., road, power line, river) are sub-classes of these classes. Then, more specific classes can be described when required. For instance, Transportation becomes a class of the superclass GeographyFeature, and Road becomes a sub-class of Transportation. Attributes become classes of the superclass FeatureProperty and values of the attributes become sub-classes. In our example, Status becomes a class under FeatureProperty superclass, with Functionally and UnderConstruction being sub-classes (Figure 2).

For the definition of specific sub-concepts under the same theme, as hydrology for instance, existing ontologies should be studied and reused, as it is the case with hydrOntology which intends to serve as a harmonization framework for hydrology [30].

3.2.2 Identification of relations between concepts

Relationships between geographic features and their properties should be defined by using object properties (Figure 2). Object properties relate one or more classes (domain) to other or more classes (range), implying that the domain has the property that states the range. Relationships must be made in the general case and including in the relation as much classes and sub-classes as possible, leading to a more effective and non redundant design.

For instance, the object property hasStatus has Transportation as a domain and Status as a range. It means

---

¹http://sweet.jpl.nasa.gov/ontology
²http://dbpedia.org
that the class *Transportation* has a property (*hasStatus*) with values stated in the class *Status*. Thereby, every sub-class of the class *Transportation* (as *Road* or *Motorway*) will have the same connection to the object property *hasStatus* and to the class *Status*, due to the inheritance principles of ontologies.

Such structure for the ontology allows matching both geographic features (e.g., motorway) and feature properties (e.g., under construction) to concepts in the ontology, avoiding redundancy and allowing to easily understand, maintain and extend the ontology with new concepts.

If there is a need to add a new class (e.g., *NationalRoad*), it can become a new sub-class of *Road* and would inherit from the relationship with the class *Status* through the object property *hasStatus*.

### 3.3 Semantic matching

Based on the domain ontology, it is possible to match common concepts between the datasets and the ontology, a step that only has to be done once for each type of datasets (Figure 3). While the creation of a domain ontology is a significant task that is beyond the scope of this paper, a proof of concept for roads and railroads has been developed and used to test the method.

#### 3.3.1 Mapping from datasets to the domain ontology

The mappings between a dataset and the domain ontology is performed using R2RML [31], a World Wide Web Consortium (W3C) standard that can be used to express customized mappings from relational databases to RDF (Resource Description Framework), a W3C standard for data interchange on the Web [32].

A mapping file, based on this standard, is used to specify the mappings between each feature class or subset of features to the concepts of the domain ontology. Features from datasets (e.g., rows of tables in databases, or features in OSM files) become individuals in the ontology, populating the class identified in the mapping file.

This way of making the mapping allows more flexibility because R2RML language allows querying the dataset, so whole feature classes or subsets of features could be mapped to the domain ontology, making also possible mappings from VGI datasets. Figure 3 shows an example of the method with two datasets (i.e., authoritative dataset and OSM dataset) that have different classification of motorways. In the example, roads named “A-7” and “AP-7” use a same tag in OSM (highway=motorway), while they belong to two different classes in the authoritative dataset (freeway vs. toll-highway). The two datasets use different classifications, words and descriptions but represent the same real phenomena. Using mappings to the domain ontology, each feature can be mapped to the same class in the domain ontology (i.e., *Motorway*) and as a result, both output files will represent the same phenomena under the same conceptualization.

Doing that with two or more datasets leads to the semantic integration of these datasets.

The mapping results in two files populated with the individuals that originally were features in the source datasets. After the semantic matching, individuals in the output files are representing the same real phenomena. Once mappings are defined, the output files are semantically interoperable.

Making the mappings manually is currently a task that involves knowing how R2RML standard works. Since there are some studies that have developed graphic interfaces for ontology matching [17, 33], it is likely that there will be implementations in the future allowing R2RML graphic mappings between datasets and an ontology, in a more user-friendly interface.

#### 3.3.2 Using and reusing the mappings

The files obtained from the previous step can be used in different ways, such as for quality comparison purposes or for querying the results in order to obtain more specific subsets of features.

Mapping files could be extended and reused. In the case of OSM datasets, which have a dynamic data model, one mapping file could be reused and extended with the extra mappings needed. In the case of authoritative datasets, mappings to the domain ontology only need to be done once for each source, because of the more static nature of the data models. Taking two existing mapping files from different sources, the semantic integration could be automatic, because both of them would have been matched to the same domain ontology.
4 Discussion and conclusions

This paper has presented a method for handling semantic heterogeneities when integrating VGI datasets with authoritative geographic information. VGI datasets are characterized, unlike most authoritative datasets, by a flexible and dynamic data model.

The proposed method introduces novelties compared to other techniques handling semantic heterogeneities. First, it proposes using a domain ontology based on standards as the common knowledge allowing matching of different geographic information sources, including VGI. Second, it suggests using R2RML mappings from the datasets to the domain ontology, instead of semantic similarities or ontology matching techniques. With such an approach, users do not have to manage ontologies and mappings from datasets to the domain ontology can be reused.

This paper has presented a proof of concept for the domain ontology but more work will be needed to create a complete domain ontology. Mappings using R2RML have been developed for some features from both official and OSM datasets to the domain ontology, showing that a simple process can be implemented once the domain ontology is built.

Creating the mappings manually between the datasets and the domain ontology is a task that involves knowledge about R2RML standard. Although it is not a difficult standard to understand, further work could explore the use of graphic mapping approaches that could be easier to manipulate by users.

Acknowledgements

Authors want to thank Dr. Jean Brodeur and the three anonymous reviewers who provided valuable comments and ideas for improving the manuscript. We also would want to thank the AGILE Grant Committee, EuroSDR/AGILE Project and the Natural Sciences and Engineering Research Council of Canada (NSERC) for their funding support.

References


