terradue 2.0

www.terradue.com

On the Development of Geospatial Ontologies

Jo Walsh

The following essay provides a critical overview of the concepts of "geosemantics", data abstraction and packaging, in the light of recent developments in collaborative information modelling over the Internet. What does it mean to have "semantic interoperability", and how do we get at it?

Short Title	White Paper Template
Prepared by	Jo Walsh
Approved by	Pedro Gonçalves
Reference	T2-Research-07-002-OnOntology
Issue	1
Revision	0
Date of issue	2007-09-07
Status	Final
Document type	White Paper
Distribution	Public
Category	Research
Keywords	metadata, geographic information
Comments	The second document on the Duetopia Series about Geospatial Services and Applications

Table of Contents

Introduction

O Brave New World, full of more and denser data sets, with more and weirder uses for them. We can continue to build world models in isolation, keep data in silos; we can break out and model data collectively over the internet.

The inability to transfer data between many closed, undocumented formats was once a huge blocking issue for the geographic information industry. "Interoperability" of different syntaxes for describing geodata was a Holy Grail. The Geography Markup Language was designed as an attempt to integrate; many different platforms would be able to exchange data cleanly by adopting a common model.

Yet a de facto interoperability was not reached by a mass movement to GML, but brought about through data abstraction libraries, offering a gateway between one format and another. Using a library such as GDAL, GeoTools or FDO, GIS client software can transparently read many different data models and combine them. The resolution of *Syntactic Interoperability* came as small pieces stitched together, not cut from whole cloth.

Having gained one Grail, the geographic information community reaches out for another: that of *Semantic Interoperability*. In GIS there are many reasons to share the semantics of data. A classification system is used to annotate a bitmap with land coverage information; an abstract model is used to identify constraints on a shape representing a road; a polygon model of an object can be chosen for use in a simulator. How do we share models between systems in a way that allows those models can be automatically compared and combined? To align spatial conceptual models across governance and linguistic systems is a great effort. The term "geosemantics" is difficult because semantics are almost everything about GI apart from geometries and the contents of bitmap imagery. The border between 'syntax' and 'semantics' blurs even more when considering non-traditional geospatial data sources.

Thus the concept of "*geosemantics*" and a concern with the "*ontology*" of spatial data models grows in the literature of Spatial Data Infrastructures. "Ontology" itself is a word that causes many to recoil in terror. It is a complex label used for simple and familiar things. Machine-readable descriptions of structured data formats, taxonomies used for classification, the domain model represented by a database schema - all these things are expression of an "ontology". For the rest of this text, "ontology" is to be understood as equivalent to "domain model", and the two terms are used quite interchangeably [domain-model]

The rest of this essay looks at the practises of data abstraction and packaging, in the light of recent developments in collaborative information modelling over the internet, in search of answers to these questions:

- Why should we want "semantic interoperability"?
- How can we get at it?

One reason to fear ontology is that it seems like a totalising effort; building one world model which must fit all. A Standard Upper Ontology constrains all statements to fit within a taxonomical, logical structure. If one uses terms which are not a part of this structure or mappable to it, the statements one makes are invalid, and cannot be understood. Efforts to make smaller domain models and

share them between applications become tarred by association with this "top-down" approach.

To mandate the use of GML was to attempt to create a "fiat interoperability" between geographic data models, by means analogous to imposing a Standard Upper Ontology - an Abstract Feature Model. If a structure cannot be expressed in GML, then it cannot be transmitted, so the standard must expand to include all common structures. The tasks of specifying, reading and re-using the standard become monumental; the task of adding to it becomes controversial.

Yet to define one unifying model offers this promise: that many different discrepant domain models can be translated into it, providing a kind of common carrier, a transport for semantic interoperability.

No Ontology in a Vacuum

An "ontology" of data is created as a by-product of practise - of existing schemas for data storage, rules for manipulating objects, implied the properties and structure of data. The details of the domain model may, and likely do, shift constantly. In accurate copies of such a model, structural integrity must be maintained, so structure must be transmitted somehow. Once the domain model of a data set is "exported" with it into different form, changes to the original need to be kept in currency.

The manager of any given data system will have lots of small domain models already well established. If the model implies taxonomy, a map of descriptions of classes of things, there may be a way to re-map it onto some well known classification scheme. But the full richness of the original domain model is likely to be highly customised and purpose-specific, a lot of the details of internal data management are caught up with the details of the data model itself.

Is any proposed external domain model going to be a good fit for the tool set in current use? That depends on the appropriateness of the original tool set itself for the purpose at hand, and of the general will to change it. It is not realistic to expect all data sources to provide a common interface, a common domain model. The best we can hope is ad-hoc transformation in the manner of data format abstraction libraries, and the ability to convey and understand a growing number of model-conveying formats - be those ESRI Shapefiles, SQL dumps or CSV data.

Having said this, under the heading of "no ontology in a vacuum" why continue to write theories about it rather than to let the organic development of semantic interoperability take its course? The point is pressed here because there is potential for the development of spatial data search engines - interchanging structured data based on the upfront establishment of a common domain model. [GeoDialogue]

In SDI discourse, a new hope is placed in the "spatial semantic web" and the practise of "geosemantics" [Lieberman]. In the abstract it is easy to work on and release formal domain models for others to re-use, either directly or for translation between local models - ontology in a vacuum. Such work, if it is to be of real re-use value, needs to be accompanied by instances of real-world data, and implemented in software used to process and store data.

For a spatial data broker, or a search and distribution service provider, it is of benefit to have a sensible means of sharing domain models with users and contributors. Domain models may change over time, so *versioned* domain models

should help to resolve conflicts which will arise when updates to the same original data set are collected from different sources.

"Conflict resolution" may appear to some a distant problem, yet it will be critical to any effort at "geosyndication". It is a direct consequence of efforts to establish a lot of small purpose specific domain models, and mappings, in order to share and package data in a more re-usable way.

Before going too far with geontological speculations and the view from the space station, let us zoom in to what is on the ground. The semantics of geographic information are expressed in standard geospatial metadata profiles. The GI community works with fiat standards for metadata, containing a common core which most people implement. Agreement on what that core is, is informal, and the prospect of one unifying syndication format for metadata in the shape of ISO19115, causes alarm amongst those with existing technological commitments, working with a lot of legacy data in well known but not necessarily well structured forms. Without a means of linking between indexes of and references to data, it becomes hard to form the bigger picture of what is available globally.

For those working closely with geospatial metadata standards, larger concerns loom behind rough consensus. SDI efforts cross linguistic and governance boundaries, each containing domain specific taxonomies. "Internationalisation" concerns add a new dimension of complexity to the problem. It is important to remember that common classes of, as well as language labels for, spatial things are cultural artefacts. Furthermore, the scope of the things labelled, the meaning of labels and the mappings between them, all change over time. Even if a direct 1-1 translation of a linguistic label is possible, the category that the label represents differs between cultures. Cadastral data in particular, is the direct expression of legal and governance systems that vary immensely across physical space. How can it be possible to model and understand the whole of this complexity?

In any given application, is it necessary or even desirable to understand the whole? One needs to be able to understand the parts that fit inside one's spatial extent, temporal span and conceptual set of interests. One needs to understand where the parts join to the other parts, how they are arranged relative to the general shape of the whole. Producers and users of data have limited concerns, where limit is firmly a good thing.

However the work of a data broker or aggregator - the provider of a spatial search facility - has to be able to provide an intelligible interface to any set of aspects of the whole. Legal frameworks for sharing spatial data (such as Europe's INSPIRE) mandate or assume the existence of central "geoportals" and "one-stop shops" for heterogeneous spatial data sources. Thus "ontologies" are Big in Metadata, as they seemingly offer a key to building better discovery services.

Small models, loosely joined

Domain experts will have a great deal of data modelling expertise on different subjects, but may insist on a level of detail which 90% of use cases do not require. Geoscientists concerned with a wide range of data styles, may go too far, encouraged by software tools, in the attempt to fit them all into a framework.

Focusing on getting small things right – the development of core consensus models in different data domains - helps to work out a "middle way" between the deep understanding of experts and the re-use potential for non-experts. If

experts don't have a broad constituency in mind, their great knowledge will be lost as the availability of intentionally simplified and generic web-based tools allows specialist constituencies to collaborate on providing "drive by data" (Examples of "non-traditional" communities with large holdings of geographic data and tools to ground-truth it, are marine sports and mountain sports communities).

These communities come up with highly specialised models which are unlikely to be reusable in other applications, yet their observations and conclusions may provide new kinds of insight to expert analysts, given the possibility of sharing data semantics consistently. It will help non-experts a lot, if suggested domain models from which they can work, are kept as small and simple as possible.

Meanwhile there is an interest in the development of "consensus" or "emergent" ontologies; not formalised by modelling tools, but growing from open "tagging" keyword or key-value classification of observed things. The Commons of Geographic Data group are optimistic that it will be possible (and desirable) to infer and build structured taxonomies of spatial things through observation of large enough keyword clusters [CGD Research Challenges]. The OpenStreetmap open vector annotation project offers a large body of user-classified spatial things as a collection of "tags". OSM is massively contributory, and a social process for maintaining one central user-maintained global ontology has arisen on its wiki, with a basic voting system on its mailing list.

Problems arise when things are described and that description is re-used in software, but no overt attempt to "ontologise" or transmit a domain model is made. It is hard to maintain ongoing integrity of data over time and across space, as the meanings of labels change. A very open system for data annotation is vulnerable to casual mis-spelling or to wilful abuse of the namespace. Another potential drawback is lack of reusability. It is difficult in software to translate data into and out of different applications, unless there is a clear guideline as to the data's internal structure, its domain model.

In their study on Geographic Categories, Smith and Mark state that "Ontological engineering has been, in fact, an exercise not in ontology at all, but in model-theoretic semantics" [Geo-Cat]. They note that what is carried out in the name of ontology in computer science, is more often the modelling of the innards and behaviours of computer systems, database models, limited and controlled environments with "clean" semantics. Clean model semantics affect what we implement in systems, how we measure things in the world and come to perceive them – but they may not help us address the social and environmental problems that we collect and share spatial data in order to help solve.

A first order logic approach such as the one taken by the W3C's Web Ontology Language (OWL) can demand excessive rigour in an environment where knowing and preserving vagueness can be useful. A "tagging" approach still demands excess description, the use of semantic energy, without offering the benefits of a transmissible domain model. Either way, any classification and processing of raw research data has to be considered as a cost-benefit to the data provider; there is a minimal level at which providing descriptions of data for others' reuse will be really useful.

The next essay considers further a "domain model aware" data distribution and search service; assisting with collection, then with query, then with subsequent correlation and co-citation of data sources. For data sources that are frequently updated, in whole or part, one should be able as a client to register interest in

only receiving updates to parts of a model. Parts then must arrive in appropriate packages for client software, with compatible interfaces, just in order to be viewed, let alone re-used. Such packages then need to be accompanied by credentials as to data integrity and judged accuracy. If a domain model is new or changed, a user needs some way to gauge its "respectability" and potential for wider re-use.

A domain model shared early on in a transaction allows client and service to establish a mutual policy regarding what elements of corresponding data sets may be transferred, connected with other datasets, and redistributed in whole or part.

It seems almost too much like "common sense" to state that knowledge of a data domain corresponds to one's capacity to usefully query the data. One needs to understand the behaviours of objects in a domain model, the implications of their properties and relations to one another.

One can take the approach of overt statement, using languages specially designed to carry domain models around – XML Schema, Web Ontology Language. One can look at the more open, keyword oriented approach to querying geographic information exemplified by the regular expression support in the Open Geospatial Consortium's Web Feature Service - Simple, or Geodata Query Service as it is becoming known [WFS-Simple]. This standard interface to any given domain model can return a list of queryable properties of the data it contains. A more flexible approach regarding standards helps to get partially-spatial data "out there", knowing or hoping that conversion and extension interfaces will be written in the future, as more people find the data and wish to re-use it.

There are many conceptual domains which are close to GIS and could be more connected - CAD, architectural structural design, earth sciences. Software applications in all these fields have their own clearly defined data domain models. What tactics can be recommended or explored to make it easier to share such models, query and reuse them, without having to establish a lot of complex taxonomical structure, logical rules and meta-model mapping?

One aspires to jump through fewer hoops in working with different-yet-similar datasets, either as a publisher, broker or end-user. A functioning domain model, implemented in at least two software systems, is the best expression of practise for others to design parallel applications - no ontology in a vacuum.

In the development of geospatial ontologies, our tools should encourage re-use, yet not enforce it. Existing software systems for partially-spatial data must be accommodated, yet a common core that allows for interoperability with more generic GIS tools needs to be maintained. Support for multiple "levels of detail" is needed to maintain the balance between expert and non-expert use.

How can we build systems that will allow "semantic interoperability" of data sets to arise, without imposing too much cognitive or economic load on data providers? The following tactics are likely to help:

- Develop means - technical and social - to semi-formalise collaboratively developed, open bodies of data description.
- Provide a way to "seed" or "shard" specialised subsets of data from a core, well-understood initial model, bringing together networks of concern from a wide field of potential contributors.

- Create ongoing agreements about shared domain models that are "local" to a community, spatially, linguistically, or conceptually.
- Build data search services that start with an exchange of data, not simply with an exchange of description of data.

But why are we looking for data? *Why do we want semantic interoperability?* The goal is to provide an end-user, an analyst or machine agent, with a package of relevant data that can be reused in an application. If it is not possible to look at the data, only the domain model, then it is not really possible to evaluate it for appropriacy of use - a matter which becomes more critical, not less, if use of the data requires an up-front commitment of financial resources.

Open Questions

Open questions remain, which will influence future efforts to build a a "spatial semantic web". Many of these questions are byproducts of the practise of "open annotation"; collaborative modelling and verification of spatial data, or research data in general.

Data Versioning

How is versioning managed if the process of contributing to a domain model is "open", even among a constrained community? Social arrangements should be supported by tools and not replaced by them. The process of release management in open source software is interesting to watch - software process management tools are accompanied by a high degree of crosstalk and backtalk between developers and maintainers of associated packages. This process, while it can be painful to go through, is desirable in more ways than the production of a stable new release accessible to a wide community. Shared goals and dependencies are restated and refined, developments of features that may threaten to become forks are merged back into a shared core. The management of formal "releases" of data packages across a collaboration network can adapt a lot from this process.

Data Brokering

Is the maintainer of a data repository or processing service also in a good position to provide a brokering service for domain models? GEONGRID, the collaboration and processing network for geoscience data in the US, is beginning an ontology collection aspect to its programme. This essay envisions large collections of small and fine-grained domain models, potentially containing much overlap or conflict. Will a need to collect many models make the job of a data broker, at least in the medium term, exponentially more complex?

Increasing client expressiveness

The interface between a data storage and index service (or a registry/repository) and a client that needs to query it, is at present semantically limited. In sending a small package of sample data to a query service, a client is announcing what it understands, what it will accept, what core data types, geometries and topologies, language encodings and labels are already in its purview. This can provide a search service with a good basis on which to filter and promote initial

suggestions, essentially to walk the user through "more like this" and "less like this".

Packages containing data and domain models can be generated specifically for the client as the result of an interaction. This technique can be particularly useful in an "open" kind of discovery where the search for new and relevant information is speculative, rather than when a user is seeking to rediscover a source which is already well known. If "Interpreting the data semantics is the reverse engineering of the spatial data creation process", what knowledge about processing history can the creator of a data set share with others who want to reuse it? [GeoDialogue]

Vocabulary does not become vocabulary until it is shared; nor does ontology become ontology until it is reused in more than two places. Sometimes we have to talk more in order to be able to communicate less.

References

Writing and Research

- [Geo-Cat] http://www.ncgia.buffalo.edu/ontology/ "Geographic Categories, an Ontological Investigation", B. Smith and M. Mark, December 2000
- [GeoDialogue] http://www.geovista.psu.edu/publications/2006/2006-geoinformatica-cai.pdf "Contextualization of Geospatial Database Semantics for Human-GIS Interaction", G. Cai et al, 2006
- [GD-Vague] http://www.geovista.psu.edu/publications/2003/Cai_COSIT_03.pdf "Communicating Vague Spatial Concepts in Human-GIS Interactions: A Collaborative Dialogue Approach", G. Cai et al, 2003
- [UCGIS] http://www.spatial.maine.edu/~max/UCGIS-Ontologies.pdf "UCGIS Emerging Research Theme: Ontological Foundations for Geographic Information Science", D. Mark et al, 2001
- [Domain-Model] http://en.wikipedia.org/wiki/Domain_model references Fowler's Patters of Enterprise Object Oriented Architecture on the "domain model" concept: http://www.martinfowler.com/eaaCatalog/domainModel.html 'a conceptual model of a system which describes the various entities involved in that system and their relationships'
- [SUO] http://suo.ieee.org/ Standard Upper Ontology Working Group, http://www.cyc.com/cycdoc/vocab/vocab-toc.html OpenCYC Selected Vocabulary and Upper Ontology
- [Sem-Interop] http://www.semantic-conference.com/Presentations_PDF/Lieberman-Joshua.pdf "Geospatial Semantic Web: an Interoperability Experiment" J.Lieberman, et al
- [Sem-Evol] http://www.w3.org/2005/04/FSWS/Submissions/48/GSWS_Position_Paper.html "Semantic Evolution of Geospatial Web Services", J. Lieberman, et al

- [OS] http://www.ordnancesurvey.co.uk/oswebsite/partnerships/research/research/semantic.html Ordnance Survey GeoSemantics research group

Tools and Specifications

- [SUO] http://suo.ieee.org/ Standard Upper Ontology Working Group, http://www.cyc.com/cycdoc/vocab/vocab-toc.html OpenCYC Selected Vocabulary and Upper Ontology
- [GDAL] http://gdal.org/ Geographic Data Abstraction Library
- [GeoTools] http://www.geotools.org/ Java Library for Geographic Data Abstraction
- [OWL] http://www.w3.org/2002/07/owl Web Ontology Language
- [SUO] http://suo.ieee.org/ Standard Upper Ontology Working Group
- [CYC] http://www.cyc.com/cycdoc/vocab/vocab-toc.html OpenCYC Selected Vocabulary and Upper Ontology
- [WFS-Simple] http://www.ogcnetwork.net/wfssimple OGC's WFS Simple "Mass Market" data query specification effort