Semantics, ontologies and eScience for the Geosciences

Femke Reitsma[1], John Laxton[2], Stuart Ballard[3], Werner Kuhn†, Alia Abdelmoty‡
[1]femke.reitsma@ed.ac.uk: School of Geosciences, Edinburgh University
[2] jll@bgs.ac.uk: British Geological Survey
[3]stuart@niees.ac.uk: The National Institute for Environmental eScience
†kuhn@uni-muenster.de: Institute for Geoinformatics, University of Münster
‡a.i.abdelmoty@cs.cf.ac.uk: Cardiff School of Computer Science, Cardiff University

Key Words: geosemantic; grid; eScience; ontologies; GeoSciML; geoscience; geoinformatics

1. Introduction

Semantics, ontologies and eScience have increasing currency in our present climate of data deluge, information overload, and widely distributed knowledge sources. Geo-anything also seems to be the flavour of the month following technological advances such as the proliferation of mashups, the rise of geography in the public consciousness, and real world problems such as the recognition of the need for global science to deal with environmental problems of global consequence. Bringing these fields together, the European Geoinformatics workshop was run to share the latest research in representing and using semantics and ontologies for eScience in the Geosciences, supporting the crosspollination of ideas from at least two different communities and encouraging future collaborative research.

What follows are some significant current topics, challenges and research directions regarding enhanced geospatial computing involving semantic and grid technologies, which arose from a workshop held at the e-Science Institute (eSI) in Edinburgh from the 7th -9th of March called the European Geoinformatics workshop[1]. The workshop brought together geoinformatics researchers, geoscience data providers, semantic web researchers and software developers, and covered topics of geoscience ontologies and language, shared geoscience data models and developments in the semantic web and GRID technologies relevant to geoscience. It was run in conjunction with an eSI theme called "Spatial Semantics for Automating Geographic Information Processes"[2], which involves research, a series of events and collaboration through a visitor programme.

2. Challenges

[1] For all referenced presentations and material from the workshop, please see http://www.nesc.ac.uk/action/esi/contribution.cfm?Title=712 and the theme events at http://wiki.esi.ac.uk/Spatial_Semantics_for_Automating_Geographic_Information_Processes

[2] http://www.nesc.ac.uk/esi/themes/theme_04/index.htm

There are a number of challenges that are critical to the development and use of eScience applications for the Geosciences that utilise the semantics of information. These challenges include the need to identify appropriate standards and languages for formalising semantics, the limitations of ontologies for representing the meaning, Grid challenges and top down versus bottom up development of geosemantic standards and applications.

2.1 Semantic agreement

One of the main drivers of research in this area is the increased expectation of users of geoscience data to be able to access the data they require digitally and in a standard form from different suppliers, including legacy data. This requires not only the adoption of geoscience data exchange standards but also semantic interoperability, reasoning across heterogeneous definitions of geoscience concepts in order to achieve semantic integration across disciplines.

Geospatial data and geoscientific research standards are needed to support geosemantic interoperability. There are levels of interoperability from a basic interchange between software applications, to the exchange of data and concepts, to a high level of interoperability that involves automated discovery and use of geospatial resources. Geosemantic standards need to be able to handle things such as dynamic geometric representations, discrete and continuous data, spatio-temporal things and processes. Ideas for how this might be accomplished may involve semantic languages such as OWL and extensions such as SWRL and SWSL, building on the work done on existing relevant ontologies such as the GEON project[3].

In Europe the INSPIRE initiative is acting as a driver for standardisation of data structures for data exchange and interoperability, such as the use of ISO/OGC standards to enable the interoperability of both marine and earthquake data. Similar standards-based interoperability initiatives are taking place in the USA, with the coordination of geoinformatics efforts between USGS and state geological surveys, and in Australia with the AuScope program aiming to integrate a wide range of earth science information. Globally, the GeoSciML initiative is developing an internationally agreed geoscience data exchange format, underpinned by an agreed conceptual data model. While this conceptual data model involves developing agreed geoscience concept definitions, the semantics of these definitions are not yet being formalised, as is the case with the majority of other efforts. However the objective is that the work will lead to the development of formal ontologies, which both handle the relationships between concepts and also the relationships between the 'same' concepts defined by different agencies in slightly differing ways.

The development of GeoSciML is leading to the establishment of an internationally agreed structure for the exchange of geoscience data. However in order to achieve useful data exchange, the concept definitions (ontology) that underpin the data also need to be harmonised. A GeoSciML Concept Definitions working group has been set up for this purpose, their first step being the issue of a widespread call for existing concept definitions, which will be collated. Subsequent work will look at the relationship between concept definitions, including partially overlapping definitions.

[3] http://www.geongrid.org/

The aim is to enable mapping from concept definitions in existing data sets to the commonly agreed definitions, which will be made available in different languages.

2.2 The limitations of ontologies and reasoning

Geoscience ontologies as a basis for 'smart' data discovery and use are seen as a solution to semantic interoperability. The GEON project, for example, has developed methods for ontology-based discovery, integration and analysis of geoscience data. A further example is NASA's move from an instrument based to a measurement based data structure, with a view to integrating data from different disciplines to assess, for example, the climatic impact of volcanic eruptions. But ontologies have limitations. Ontologies are more concerned with 'what' than 'why' or 'how', limiting the questions they can address. Furthermore, ontologies are typically static. Ontologies could develop through usage and context with evolving concepts.

Within qualitative spatial reasoning research much work has been done on qualitative spatial representation, but relatively little on world views and the incorporation of context into reasoning. While the end user mixes quantitative and qualitative reasoning within a certain context, the challenge arises about whether we can understand and represent context abstractly in a useable way. A shift in focus is needed that moves from research on pure qualitative reasoning to that within a disciplinary or application context, where application areas such as way finding research are taking the lead. We also need appropriate languages to be able to represent vagueness related to qualitative reasoning, as well as use cases and benchmarks against which to test these, such as spatial decision support applications used in emergency planning.

2.3 Grid technologies for geoscience

Currently a range of grid technologies are being used to overcome issues related to working collaboratively across multiple institutes to fulfil either data sharing or computational requirements. Grid technologies hide the computational complexity from scientists, allowing them to focus on new and exciting research. For example, Grid technologies could support the combination of simulation codes and data sets from different institutes, and the rapid modelling of a variety of situations to produce predictions. Depending on the user's requirements, they can carry on as normal with either command line interfaces and/or portals.

The vision and challenge of grid technologies is to seamlessly harness the power of geographically distributed and controlled heterogeneous computational resources and data, without having any control over those resources and without dedicated computer officer support. Realising this vision would redefine 'cutting edge scientific research'.

2.4 Small "g" versus big "g" geosemantics

Within the Semantic Web community, an early interest in geospatial semantics resulted in a basic RDF vocabulary that provides a namespace for representing latitude (<geo:lat>), longitude (<geo:long>) and altitutde (<geo:alt>) as properties of points (<geo:Point>) using WGS84 as a reference datum[4]. This minimalist ontology provides very little in the way of geospatial semantics and yet has been widely used in applications such as Flickr, giving us “small s” semantics through tagging and the application of microformats for “small g” geosemantics that largely involves geotagging things. The online gazetteer geonames[5] provides another example of expressing the semantics of geospatial information. Within geonames, relationships expressed in OWL defining spatial containment (ChildFeature, ParentFeature), and neighbourhood relations (NeighbouringFeatures, NearbyFeatures), are generated on demand from an underlying database that does the relevant spatial reasoning.

The developments around what is called the "Web 2.0", i.e. the transition of users of the web (and semantic web) into information providers, deeply affects standardization in the geospatial area. Traditional geospatial standards have been designed with a view of geospatial information coming "down" to users from agencies (such as national mapping agencies) or large enterprises (such as navigation data providers). Suddenly, geospatial information pops up in wikis and mash-ups all over the internet, often provided by users rather than authorities. The challenges posed by this "bottom-up" kind of information include: (1) trading off of simplicity and security; (2) designing different kinds of metadata; (3) coping with legal implications and justification processes; (4) transforming the role of agencies as information providers; (5) developing measures of quality based on trust. Addressing these challenges will require coping with an explosion of small standards, making information sources more transparent, tracking identities of providers, lowering entry barriers, adapting trust and reputation models from other domains, and providing simple but useful geospatial processing services. A first step in supporting the transition will be to come up with different types of applications, organized along reliability, trustworthiness, and other axes.

In contrast, “big G” Geosemantics involves the representation of spatial and spatio-temporal theories, relationships, mediations and transformations in order to enhance interoperability, as have partly been developed and used in research on geographic information retrieval. But such semantics have yet to be expressed and used in an appropriate language or formalised in a standard, such as OGC or W3C standards. There exist many standards for geospatial interoperability, which standardise the vocabulary or syntax but not the meaning of concepts and terms. In order to develop a Geospatial Semantic Web we need standards for ontologies, reasoning, discovery and services, to push us up the hierarchy of interoperability from systems and schema-level interoperation to semantic interoperation.

Some of the key challenges in representing and utilising spatial and spatio-temporal semantics are found in the languages we have to express those semantics, particularly in providing calculi which allow a machine to represent and reason qualitatively with spatial entities of higher dimension, without resorting to traditional quantitative techniques. We do not yet have appropriate calculi that encompass the expression of spatio-temporal primitives and their spatio-temporal properties and relationships. Nor are current languages for expressing non-spatial concepts always appropriate (such as

[4] http://www.w3.org/2003/01/geo/
[5] http://www.geonames.org/

OWL), as has been evident in research at the Ordnance Survey for developing a new language, Rabbit, a controlled natural language (English) for ontology authoring.

Despite the fundamental limits of languages to express and reason with spatial and spatio-temporal objects and their relations, work with “small g” geosemantics are forging ahead utilising collective knowledge to develop emergent knowledge systems from structured data. Such collective knowledge systems are typically a mix of structured, machine-readable data and unstructured data derived from human input. A future vision is that geospatial data can provide a backbone for cobbling together different types of data and services. Geosemantics would then enable the meaningful cobbling of data and services, bringing spatial data to a social web context. But there is some skepticism about whether collectively developed ontologies will be useful in a scientific context and a recognition that we need to capture things in an ontology that our current languages cannot express, such as process behavior, spatial/temporal characteristics and data and process relationships. Such future developments in ontologies would better support large scale scientific computing on the Grid, which is another crucial component of a future vision of eScience for geosemantics.

3. Brainstorming a Research Agenda

A central theme that was highlighted at the workshop was the need for semantically rich registries that support discovery and use of georesources such as models, data and knowledge. New ways of representing these georesources are needed that handle different notions of time and space as well as processes. And research needs to be undertaken to clearly define the semantics we need for geoscience research, beyond the limitations of current ontology languages, and to consider appropriate languages for geoscientific reasoning.

As another key item for a research agenda, it is critical that we find a way to bridging the gap between community developed knowledge systems for creating georesources, such as those formed through geotagging and microformats, and standards driven approaches, such as OGC/ISO supported methods.

For these georesources, digital rights management or georights are critical for supporting their appropriate use and for providing trust in their validity and usefulness. Notions of trust are necessary for utilising georesources based on provenance and lineage information in order to automate scientific processes. From the perspective of commercial and government sectors, there is also interest in how we distribute and fund the use of georesources. Clearly we need new business models to cope with the changes in resource creation and use.

And finally, in supporting the development of research in this field, use cases and data sets are needed. More specifically, geoscience uses cases that clearly present the challenges of expressing and utilising semantics, ontologies and eScience for the GeoSciences are needed to help communicate the problem to computer scientists and other researchers developing languages and tools for representing and using the semantics of information.

Acknowledgements: Thanks to the e-Science Institute for their support of the workshop and theme.