

# So what's so special about spatial?

Glen Hart, Catherine Dolbear  
Ordnance Survey of Great Britain, Romsey Road, Maybush, Southampton,  
Hampshire, SO16 4GU, UK  
[glen.hart,catherine.dolbear]@ordnancesurvey.co.uk

© Crown Copyright 2006 Reproduced by permission of Ordnance Survey

***Abstract.** Geospatial information can act as a thread that can be used to integrate information from heterogeneous sources. It does so by exploiting common location information components that often exist across different domains. As such it has the potential to be a valuable resource in the implementation of the Semantic Web. This paper examines the challenges of adding a geospatial component to the Web, with particular reference to doing so in a way that also supports the current initiatives to semantically enable the Web. It identifies those that are largely peculiar to geography and those which, whilst issues within geography, are also likely to occur in many other domains.*

This article has been prepared for information purposes only. It is not designed to constitute definitive advice on the topics covered and any reliance placed on the contents of this article is at the sole risk of the reader

## 1 Introduction

There is a long lived though unattributed belief that eighty percent of all information has a geographical component: by this it is meant that a significant proportion of databases contain information either directly or indirectly referenced to physical locations. Such information will include obvious candidates such as digital mapping, environmental information and planning information. It also covers information from other domains such as marketing, insurance and so on. Any information that makes reference to a postal address can be considered geospatial.

In most cases the geospatial component is not dominant; it plays a supporting role rather than being central to the business objective. An insurance company will be interested in the location of insured properties to determine flood and subsidence risk but its fundamental interest will be in the overall risk factors and property value. The geospatial nature of the information helps the general information integration problem, as the geospatial component is often one of the most commonly shared information types between different datasets. So the insurance company is able to calculate flood risk through the intersection of property location, digital map information containing river location and perhaps meteorological information for the area.

If it is accepted that geospatial information can form an important element in information integration, it is quite apparent that it will also serve an important role in the development of a more sophisticated World Wide Web. That location is important is certainly recognised by the search engine giants: Microsoft<sup>®</sup> with its Live Local, Google<sup>™</sup>, Google Maps and Google Earth and Yahoo!<sup>®</sup> Local. These initiatives have all been created in response to user needs for location based information services. None is able to fully exploit the geospatial aspects of the information being searched, because of the Web's weak representation of location. Another weakness is the lack of semantics: not only is the Web unable to fully appreciate the spatial relationships that exists between the cities of Southampton and Portsmouth for example, but it also lacks understanding of what a city is in the first place. Thus there are strong arguments for both semantically and geospatially enabling the Web. And indeed, the role of the Semantic Web as a vehicle for information integration is explicitly recognised by its creator: Tim Berners-Lee (Shadbolt, Hall, Berners-Lee, 2006). This extends beyond just web pages. Commenting on the future of the semantic web, Tim Berners-Lee recently asserted that "we need to look at existing databases and the data in them" (Runciman, 2006). Although research interest in this 'deep web' is increasing, the process of linking an ontology to a legacy relational database raises many semantic issues which have, to date, largely been ignored. Of particular interest to Ordnance Survey, the national mapping agency of Great Britain, is the case of mapping an ontology to a spatial database, and how to combine spatial and description logic queries and modelling paradigms for efficient performance.

This paper examines the challenges of adding a geospatial component to the Semantic Web. It identifies issues which are largely peculiar to geography as well as those which occur not only in geography but also in many other domains.

## **2 The geospatially peculiar**

The geospatial domain shares many issues with other domains when considering how to represent and exploit domain knowledge using semantic web technologies. However certain issues are far more relevant to the geospatial domain than others (although many may be shared with the more general spatial domain). These peculiarities are examined in this section.

### **2.1 Spatial Relationships**

Given that what makes geospatial information "different" is the concept of location, it is very obvious that spatial relationships between objects that have location are very important. Central to all Geographic Information Systems (GIS), and more lately to spatially enabled databases systems, is the support of various spatial operators designed to determine the relationship between geospatial objects. For example, all will support (under various names) algorithms to determine containment, overlap, (spatial) disjointness and so on. Significant work has been done to formalise these relationships, such as the 9 intersection model (Shariff, Egenhofer and

Mark, 1998) and Regional Connection Calculus (RCC8) (Randell, Cui and Cohn, 1992).

Due to their formal expression, these calculi can be expressed as topological relationships using description logics (W3C, 2004) and hence in languages such as the Web Ontology Language, enabling reasoners to perform inference over spatial information based on topology. As an example, RCC8 has been implemented using OWL to support an experimental ontology editor (Smart, Abdelmoty and Jones, 2006). However, geospatial information rarely contains explicit topological information. It is more usual for these systems to determine specific topological relationships through geometrical calculation on information with positional information. For example it is unlikely for a geospatial database to explicitly contain the topological relationship that a specific house is “contained within” a specific garden. Rather the containment operator will be used to test whether the building is indeed contained within the garden. Such approaches are necessary because it would be impractical to explicitly pre-compute all topological relationships between all the objects in a geospatial database. However, this contingency means that existing reasoners are unable to perform topological inference on the majority of geospatial information as they are unable to compute the necessary topology. Furthermore, as has been identified by Lemmens (Lemmens et al) relationships may be indirectly expressed through other means such as postal addresses. Thus reasoning is limited to qualitative reasoning over the ontology rather than quantitative reasoning over instances held in the database. Currently there are investigations into the possibility of separating spatial queries into spatial and aspatial components (Dolbear and Hart, 2006). The spatial component is executed first and the results set is presented to the reasoner which completes the query. Although a pragmatic solution given the current technologies, such solutions can never be complete answers since the reasoner may discover the need to investigate further spatial relationships. To arrive at a complete solution there are two possible paths. Either languages such as OWL will need to be modified to explicitly support spatial relationships and will in turn need to be supported by appropriately enhanced reasoners or the reasoners will need to be modified to enable certain class properties to be mapped to database functions or web services. This will not only enable the mapping of spatial properties but also other properties too, enabling a more general solution to be arrived at.

## **2.2 Vague and Uncertain Location**

The spatial relations described by the 9 intersection model, RCC8 and those implemented within spatial databases and GIS are based around “crisp” geometry. It can be precisely determined that a house is or is not contained within a garden because both have a precise (or crisp) geometry and location. However, this is not true of all geospatial objects. Consider an area such as the Lake District in the United Kingdom; it has no well defined boundary and as a result it is a vague object. It is possible to know with certainty that some things are within the Lake District such as Windermere because as a lake it is one of the things that define the Lake District. Precisely where the Lake District ends is uncertain because no one has ever defined

a boundary. Impreciseness in recording geospatial information can also result in uncertainty. Consider road traffic accident information. Typically this will be recorded against particular stretches of road but will not record precisely where individual accidents took place. Thus, although an accident (a spatial event) took place at a precise position along the road, the recorded information cannot tell us where this was. We can only know that it took place on a certain stretch of road and are uncertain as to exactly where.

Existing geospatial implementations enforce some form of crispness on information that is vague or uncertain. Artificial boundaries may be created for vague objects and road traffic statistics may be treated as if they are point objects. These are no proper solutions; merely work-rounds given the limitations of existing technology. Attention has been directed towards developing better solutions to these problems, for example Super-valuation Semantics (Cohn and Gotts, 1997) have been applied by Bennett to represent the vagueness that is implicit in the notion of a forest (Bennett, 2001). Although boundaries are created, Super-valuation semantics differentiates between areas that are known to be part of a vague object such as a forest, those areas that are definitely not part of the forest and those areas that might be. Furthermore using super-valuation semantics it is possible to delay making a decision as to where these boundaries might be until a user's context is known.

Another approach is Anchor Theory (Galton and Hood, 2005) which allows the query results to transmit the notion of uncertainty, rather than enforce some arbitrary boundary or collapse information to a point. For example, say we have information on a road accident that occurred along a stretch of road that runs across the counties of Devon and Somerset but do not know precisely where it occurred. Anchor theory represents this with an "anchored in" relationship. That is to say the location of accident is "anchored somewhere within" the road stretch. A spatial query asking if the accident occurred in England would produce the answer "Yes" and within Wales "No" since it is known that the road stretch lies within Somerset and Devon both of which are within England. If asked whether the accident occurred in Somerset the answer would be "Maybe" thus preserving the uncertainty.

As with the crisp spatial relationships, relationships to handle vagueness and uncertainty, such as those indicated above, are also required and need to be implemented as a fundamental part of any semantic technology.

### **2.3 Vague Relations**

The discussion so far has concentrated on the nature of geospatial objects in terms of the crispness of their location. However, certain spatial relationships can themselves have components that give them vague properties. Consider the relation "Near". Whether something is near or not is very dependent on the context in which the question is asked and the nature of the objects being compared. The scale of distance and thus what is meant by near will vary enormously: consider the distances that are judged appropriate when comparing the results of a question such as "which pubs are near to my house?" with "which airports are near to London?".

In the first case “near” typically refers to a distance of a mile or two, in the latter, airports within 30, 40 and perhaps 50 miles of London might all be considered near.

At present no solution to this problem exists. Indeed, in an absolute sense there can be no perfectly correct solution. Some very elaborate solutions have been devised (Gahegan, 1995). Such proposals have tended to be overly complex in terms of the number of contextual factors taken into account which render practical solutions impossible to implement as it is never possible to accurately quantify these factors. Dolbear is investigating the development of a much more simple algorithm that will attempt to approximate the results of human reasoning (Dolbear, Hart 2006). The approach is to determine nearness through a combination of object footprint size and frequency of occurrence. The conjecture is that these will serve as measures that can determine the likely range for which “near” will be deemed to operate. Footprint size is assumed to be proportionate to distance and population density inversely so.

A relationship like “next to” may be equally problematic. Although both the 9 intersection model and RCC8 have well defined and predictable “next to” relationships, a “badly behaved brother” also exists. Given a situation where a house is surrounded by a garden which in turn is physically next to a foot path running by the side of a road: if asked “is the house next to the road?” most people would answer “yes”. A GIS would say no, since the house is not *physically* located next to the road. Thus a version of “next to” exists where geospatial objects that are deemed to be insignificant are filtered out.

As with near it is likely that an algorithm can be devised that will approximate the results of human thought. A possible solution is to filter out objects that are typically considered less important or are physically less significant. In effect an importance hierarchy could be constructed that would place house and building higher in the structure than gardens, fences and paths.

### **3 Shared Issues**

Whilst not attempting to be a comprehensive identification of geospatial issues that are (or are likely to be) shared with other domains, this short section attempts to highlight some of the more important issues.

#### **3.1 Conceptual fuzziness**

The world is full of objects that don’t quite fit into neat conceptual boxes and in very many cases, particularly in the geospatial world, these are natural things. Consider rivers and streams. Both are bodies of flowing water and indeed there is little physical distinction between them other than size. The problem is where to draw the line between their descriptions and does it matter if a line is drawn? The problem cannot be easily resolved except by enforcing what could be arbitrary distinctions. Given this problem has existed ever since people began to differentiate between rivers and streams, it has been managed, if not solved, by local definitions

being applied. Thus context of use is an important factor and an issue for the whole semantic web.

### **3.2 Troublesome Homonyms**

The geospatial domain is filled with homonyms. Even within a narrow domain, such as a topographical interpretation of inland hydrology, words such as channel, pool and bank can have multiple meanings. To an extent the problem can be mitigated by the use of multiple namespaces. In our own work we have used the `rdf:label` annotation property to provide a label for each term corresponding to what the domain expert would normally call the concept. Homonyms share identical annotations but these are then mapped to class names that are made unique through the artifice of appending a modifier. So for example, pool can be either something that is a type of pond or an area of still water in a river. In both cases the annotation will be `Pond` but the class name for the latter will be `Pond.inRiver`. Whilst this technique enables the homonyms to be represented, the fundamentals of the issue can only be resolved if the exact context of use is known. Disambiguation can then be performed when the ontology is applied.

### **3.3 Weak concepts**

Weak concepts are concepts that are sometimes used as a means to gather together other concepts, but which are themselves not so much poorly defined as poorly thought through. They are probably best explained through example. A water body is a good example as it is often used within geographical ontologies to group together items such as rivers, lakes and ponds. The problem is that although the sub-classes may be well defined, the water body itself really only exists as a means to loosely group together these concepts. In practical terms within the domain they are typically not used. Furthermore because the concept is itself weak, it often occurs in other domains (or even the same domain) with a slightly different usage or position in the taxonomic hierarchy. So someone else may include seas as water bodies, another person may introduce the notion of other weak concepts such as flowing (rivers, streams, lakes) and non-flowing water bodies (ponds, canals), before grouping them in turn under water body. There is a clear and obvious solution – if they are not adding anything to the ontology, don't include them. Or if they must be added, they should not be included in the structure as a superclass, but through the properties that all qualifying members share. The sub-super-class relationship can then be supported through inference. This in turn leads to both good design and more re-usable ontologies and concept definitions. This example has been included as a demonstration that it is not just technology that needs to develop, but the way we in which use the technology. The desire to build complex hierarchies that contain weak concepts (in cartography the completely useless distinction between natural and man-made is another very popular weak concept) can be quite hard to resist. But resist we must, if we wish others to reuse what we have worked so hard to produce.

There is one good reason where it is acceptable to include weak classes. This is where they are used to enable disjoint classes to be efficiently defined, in turn leading to significant performance increases in the performance of the reasoners used to interpret the ontology. In effect the use of such classes is as an optimisation technique. And so we are currently left with the problem of having to balance reusability against performance. Our own method which expresses the ontology in both a structured English form and in OWL offers some way forward. It offers the potential for the structured English to become the visible representational form developed in a way to promote reused and the OWL form acting as an assembler code compiled and optimised from the high level structured English.

## 4 Conclusions

This paper has identified that the most important issue involved in geospatially enabling the semantic web is the incorporation of spatial relationships into semantic web technologies. Where the geography is crisp, there are well founded and formally defined models that may be used such as RCC8. At least some categories of vagueness and uncertainty can be managed through solutions such as super-valuation semantics and anchor theory. Super-valuation semantics are to a certain extent reliant on context of use; thus they touch on a more general and thorny issue of how to incorporate user context into semantic web technologies. Within the field of ontologies, less work has been conducted on vague relationships. Whilst it is possible to implement elements such as RCC8 using the existing language constructs of OWL, this will only enable qualitative reasoning. All these relationship models will require modifications to reasoners, and potentially to languages such as OWL as well, to enable both qualitative and quantitative spatial reasoning.

The paper has also identified certain issues such as conceptual fuzziness, homonyms and weak concepts, which whilst frequently occurring in the geographical domain are also likely to be common to other domains too.

This article has been prepared for information purposes only. It is not designed to constitute definitive advice on the topics covered and any reliance placed on the contents of this article is at sole risk of the reader.

## References

- Nigel Shadbolt, Tim Berners-Lee, and Wendy Hall, The Semantic Web Revisited, *IEEE Intelligent Systems*, vol. 21, no. 3, 2006, pp. 96-101
- Runciman, B, Interview with Tim Berners-Lee, ITNOW, British Computer Society, March 2006.
- Shariff A, R Egenhofer M J and Mark D M, Natural Language spatial relations between linear and areal objects: The topology and metric of English Language terms. *Int Journal of Geographical Information Science* 1998, 12(3): 215-246

- D.A. Randell, Z. Cui and A. G. Cohn, A spatial logic based on regions and connections. In Proc. 3rd International Conference on Knowledge Representation and Reasoning 1992 pp165-176 San Mateo, , Morgan Kaufman
- W3C, OWL Web Ontology Language Guide, W3C Recommendation 10 February 2004, <http://www.w3.org/TR/owl-guide/>
- Lutz, C and Wolter, F. Modal Logics of Topological Relations. In Proceedings of Advances in Modal Logic 2004 (AiML-2004).
- Smart, P, Abdelmoty and Jones, C. A Visual Editor for Validating Geo-ontologies in OWL, Conference Proceedings GISRUK 2006
- Lemmens, R, Wytzisk A, de By, R, Granell, C, Gould, M, van Oosterom, P, Integrating Semantic and Syntactic Descriptions for Chaining Geographic Services, IEEE Internet Computing Sep-Oct 2006
- Gahegan, M, Proximity Operators for Qualitative Spatial Reasoning, 1995, Spatial Information Theory, International Conference COSIT '95, Springer, Pages 31-44, ISBN 3-540-60392-1
- Dolbear C, Hart G, R2D2: combining spatial and semantic queries into spatial databases 2006, Technical Paper, <http://www.ordnancesurvey.co.uk/research>
- Cohn A G, Gotts N M. The 'egg-yolk' representation of regions with indeterminate boundaries in: Burrough, P & Frank, A M (editors) Proceedings GISDATA Specialist Meeting on Spatial Objects with Undetermined Boundaries, 1996, pp. 171-187 Francis Taylor..
- Bennet, B. Application of Supervaluation Semantics to Vaguely Defined Spatial Concepts, Spatial Information Theory: Foundations of Geographic Information Science; Proceedings of COSIT'01, edited by D.R. Montello, LNCS, 2205 , 2001, pp 108-123, Springer, Morro Bay, .
- Galton, A and Hood, J. Anchoring: A new approach to handling indeterminate location in GIS. In Anthony G. Cohn and David M. Mark (eds.), 2005, Spatial Information Theory: Proceedings of International Conference COSIT 2005, Springer pages 1-13. ISBN 3-540-28964-X