Exploring The Linked Open Data Cloud to Retrieve Geographic Information – Some Challenges

Rolf Grütter
Swiss Federal Research Institute WSL
Zürcherstrasse 111
8903 Birmensdorf, Switzerland
Phone: +41 44 739 25 09
{rolf.gruetter}@wsl.ch

ABSTRACT

Think of a user looking for a new home around her place of work. In order to inform herself about possible places of residence, she enters a noun, a local expression and a toponym into the form of a search service, for instance `<parish near Dietlikon>`. As a result of the search she expects instances of the class represented by the noun, for instance, a list of URIs referring to the websites of parishes that satisfy the query.

In order to behave in accordance with the intended meaning of the query, the search service requires structures and processes (i.e. algorithms) that assists it in analyzing the input semantically and in retrieving the instances of the (complex) concept resulting from the analysis. Leaving the disambiguation of the noun "parish" out of consideration, semantic analysis consists of checking whether a model meets some condition imposed on nearby places, and it comprises the disambiguation of the toponym "Dietlikon".

Nearby places can be modeled in Euclidean space or in a topological space. Models can be quantitative (e.g. metric) or qualitative (e.g. topological). It has been argued that topology is more critical for the semantics of spatial relations than metric [1]. Accordingly, a logical theory based on regions and connections has been developed in the first place [2]. This engendered the establishment of methods for qualitative spatial reasoning (cf. [3], for instance). A challenging question today is whether and how topological models can be aligned with set-theoretic models, such as (populated) ontologies, while reconciling the associated logical frameworks. Another question is whether reasoning can be kept tractable in a practical application.

To reconcile the above mentioned logical theory on regions and connections with description logics, research has been done, particularly on description logics with concrete domains [4]. Results show that, in order to uphold decidability, expressivity of description languages must be constrained. A computationally tractable logic can be constructed by further restricting the way the spatial domain is accessed from within the logic [5]. There are other approaches that address the mentioned challenges. A detailed discussion is, however, outside the scope of this abstract.

Place names may not be unique.[1] Having on hand powerful gazetteers such as GeoNames[2] or SwissNames[3] one might expect that disambiguation of toponyms is straightforward. In a concrete case, however, some problems need to be solved. Particularly, if places are represented as polygons it can be tricky to align these with the point data of the gazetteers. Even if something like a "central coordinate" is given, a heuristics needs to decide whether two pairs of coordinates describe the same place or not (unless it is explicitly stated that two places are the same, for instance, by using the `owl:sameAs` property). To give an example, the data for the parish "Dietlikon" in DBpedia.org comprise a pair of coordinates in WGS84, namely N 47° 25' 0'' / E 8° 37' 0''. Compared with this, the coordinates N 47° 25' 26''/ E 8° 36' 55'' are provided in GeoNames for the third-order administrative division "Dietlikon". – Do the two pairs of coordinates describe the same place?

A first attempt to retrieve the instances of the concept resulting from query analysis is to feed the names of those parishes that have been evaluated as near Dietlikon into a search engine, as is done in [6], for instance. Proceeding like this deals with the semantic analysis of the local expression "near" and possibly addresses the disambiguation of the toponym "Dietlikon". However, it reintroduces and multiplies the ambiguity problem as the search engine has now to deal with a number of possibly ambiguous place names (i.e. eight in the example). Accordingly, although the authors observed a significant increase in recall of 170 web searches, precision remained below 0.5. Funnily enough, a parish named "Dorf" (in English "Village") caused the engine to return a long list of websites with little or no relation to the query.[4]

A more sophisticated approach to instance retrieval goes beyond query expansion (cf. [7] for an outline). This requires that the search service in question makes use of a spatial index. As in the case of spatial models, there are different kinds of spatial indices, some of which are implemented in today's geographic information

GIRT'12, April 24, 2012, Avignon, Vaucluse, France.
Copyright 2012 ACM 1-58113-000-0/00/0010…$10.00.

[1] With nouns, the problem is even more challenging. Parishes, for instance, are also termed "municipalities" or "communes", such as in DBpedia.org and on the website of the Federal Administration (http://www.bfs.admin.ch). Disambiguation of nouns requires a thesaurus or an ontology.

[2] http://www.geonames.org

[3] http://www.swisstopo.admin.ch/internet/swisstopo/de/home/products/landscape/toponymy.html

[4] Relevance ranking can alleviate the problem: Precision of the n-best results was much better than that of the entire result set [6].

systems (GIS). In the example the service requires a geometric index, listing the URIs of parishes together with a spatial reference. It also requires a topological index stating, for instance, that the parish of Kloten borders the parish of Dietlikon.

Given the growing number of Linked Open Data (LOD) on the web it is tempting to think of the LOD cloud as an index space which can be explored by a search service in order to retrieve the indexed resources. In as much as they are pointing to the resource they describe (notably by a dereferenceable URI), LOD are indeed index items. DBpedia.org, for instance, can be thought of as a webized index of the resources described in Wikipedia.org.[5] DBpedia.org is more than an index, however. Strictly speaking, most LOD items are describing the "things" (cit.) in the data set and not the resources on the web. DBpedia.org – and other data sets in the LOD cloud – are highly self-referred [8].

LOD can be obtained as downloads or queried from SPARQL endpoints. In order to evaluate the example query, a search service will ask the user to mark the geolocation of Dietlikon on a map and retrieve the corresponding LOD based on the name and the coordinates provided. Alternatively, it will use the ambiguous toponym and return a list of LOD items to choose from.[6] The search service will then explore the topological relations between LOD items in order to establish a model of parishes near Dietlikon (note that in [6] nearby parishes are not just neighboring parishes). Finally, the resulting URIs will be dereferenced in order to retrieve the searched resources from the web.

In addition to the already mentioned, this scenario requires that LOD items of a given set, for instance DBpedia.org, be linked to data items of other sets, for instance GeoNames, and, of course, to the resources they describe. This sounds like a commonplace remark. It's not, as the statistics in [8] reveal.

Required are also standardized vocabularies which can be used, for instance, to query LOD from SPARQL endpoints. Such a vocabulary can be referred to by prefixing terms with a variable which is initialized in a namespace declaration, such as `xmlns:ogc="http://www.opengis.net/rdf#"`, where `ogc` is the prefix. In the course of specifying GeoSPARQL, the Open Geospatial Consortium (OGC) is defining a vocabulary to represent features, geometries and their relationships [9].

Other issues of open data, such as provenance and quality also apply to LOD. However, these are not specific to LOD and have to be addressed when dealing with information on the web anyway.

Retrieving geographic information from the web, thus, requires a joint effort of GI scientists, knowledge engineers, experts in natural language processing, user interface designers, database specialists, and standards junkies. Bringing together these people for a common endeavor might be a major challenge.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *query formulation, retrieval models, search process.*

General Terms

Algorithms, Performance, Design, Standardization, Languages, Theory.

Keywords

Geographic Information Retrieval, Linked Open Data, Local Expression.

ACKNOWLEDGMENTS

Research on geographic information retrieval is funded by the Swiss Federal Office for the Environment (FOEN).

REFERENCES

[1] Shariff, A. R., Egenhofer, M. and Mark, D. 1998. Natural-Language Spatial Relations Between Linear and Areal Objects: The Topology and Metric of English-Language Terms. *International Journal of Geographical Information Science* 12, 3, 215–246.

[2] Randell, D. A., Cui, Z. and Cohn, A. G. 1992. A spatial logic based on regions and connection. In *Proceedings of the 3rd International Conference on Knowledge Representation and Reasoning*. Morgan Kaufmann, San Mateo, 165–176.

[3] Renz, J., Ed. 2002. *Qualitative Spatial Reasoning with Topological Information.* Springer, Berlin Heidelberg. LNCS 2293.

[4] Lutz, C., and Miličić, M. 2007. A Tableau Algorithm for Description Logics with Concrete Domains and General TBoxes. *Journal of Automated Reasoning* 38, 1–3, 227–259.

[5] Özçep, Ö. L., and Möller, R. 2012. Computationally Feasible Query Answering over Spatio-thematic Ontologies. In *Proceedings of The Fourth International Conference on Advanced Geographic Information Systems, Applications, and Services*. GEOProcessing 2012.

[6] Grütter, R., Helming, I., Speich, S., and Bernstein, A. 2011. Rewriting Queries for Web Searches That Use Local Expressions. In *RuleML 2011 – Europe*, N. Bassiliades et al. Eds. Springer, Berlin Heidelberg. LNCS 6826, 345–359.

[7] Jones, C.B., Purves, R., Ruas, A., Sanderson, M., Sester, M., van Kreveld, M., and Weibel, R. 2002. Spatial Information Retrieval and Geographical Ontologies: An Overview of the SPIRIT Project. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 387–388.

[8] Bizer, C., Jentzsch, A., and Cyganiak, R. 2011. *State of the LOD Cloud.* Freie Universitaet Berlin. URI= http://www4.wiwiss.fu-berlin.de/lodcloud/state/.

[9] Perry, M. and Herring, J. 2011, January 28. *GeoSPARQL – A geographic query language for RDF data.* A proposal for an OGC Draft Candidate Standard. Version 1.0.4. Reference number OGC 09-157r4. The Open Geospatial Consortium. URI= http://www.w3.org/2011/02/GeoSPARQL.pdf.

[5] In the given context "webized" means that the index items themselves are identified by HTTP (i.e. dereferenceable) URIs.

[6] This is similar to faceted browsing (cf. http://dbpedia.org/fct/).