Connecting LinkedGeoData and Geonames in the Spatial Semantic Web

S. Hahmann, D. Burghardt

Dresden University of Technology, Institute for Cartography, Helmholtzstraße 10, 01069 Dresden, Germany
Email: Stefan.Hahmann@tu-dresden.de; Dirk.Burghardt@tu-dresden.de

1. Introduction

Recent developments in cartographic applications employ methods of the Web 2.0 and user generated geoinformation. This leads to a considerable amount of up to date but heterogeneous data. Development of methods for interoperability on a semantic level is required to use these data sources. As a growing amount of information is moved from classical databases to the web, there is an ongoing paradigm shift from the web of documents to the web of data / *Semantic Web*. Consequently this opens new perspectives for cartographic data retrieval. Within the Semantic Web the LinkedGeoData project as an RDF implementation of the OpenStreetMap data set has the capability to serve as a central interlinking hub for geodata.

This paper starts with a short introduction to the Semantic Web followed by a comparison of the current definition of Multiple Representation Databases (MRDB) and Linked Data of the Semantic Web. Ongoing work on connecting LinkedGeoData and Geonames will be described. The presented matching method uses type information, spatial distance and name similarity. Matched features allow an integrated access to both data sets and the validation of the data sets against each other.

2. The Semantic Web

In Berners-Lee et al. (2001), Tim Berners-Lee, one of the inventors of the Internet and today a director of the W3C, introduced the idea of the "Semantic Web". As discussed in Lassila and Swick (1999), the *Semantic Web* aims to make the *World Wide Web* that was initially made for human consumption intelligible not only to humans but also to machines. Though the World Wide Web is "machine-readable" it is not "machine-understandable".

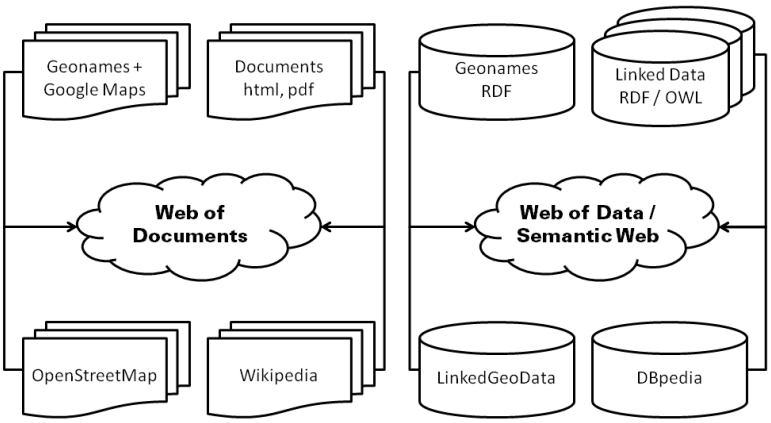

Figure 1. Relationship of documents to the *Web of Documents* compared to the relationship of relational databases and the *Web of Data / Semantic Web*.

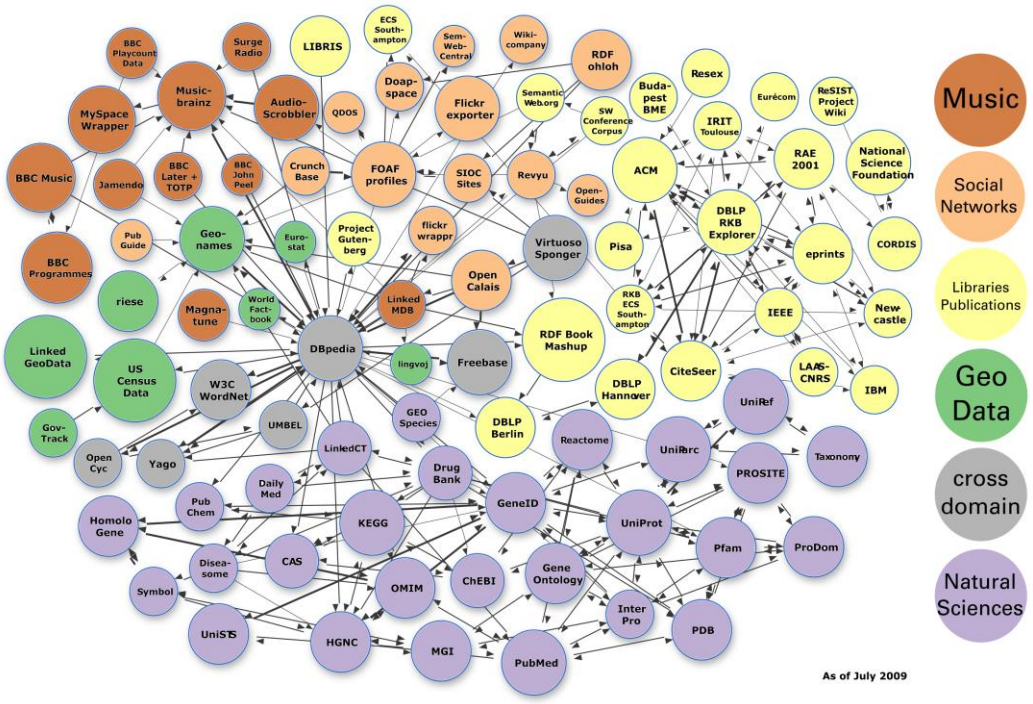

Figure 2. The Linked Open Data cloud of data sets visualizing contents of the Semantic Web. Modified after Cyganiak and Jentzsch (2010).

The semantic description of data can be accomplished by using vocabularies, as presented in the RDF Schema recommendation by the W3C in Brickley and Guha (2004). An extension of the RDF Schema recommendation is the Web Ontology Language (OWL) as presented in van Harmelen (2008). In the decentralized environment of the Semantic Web, there is no single world view and hence no single ontology, which could be used by the Semantic Web. Thus, as Kuhn (2003) states, research on Semantic reference systems is important. Their objective is to generate formalized representations of the meaning of geographic features to achieve interoperability between different domains which use shared vocabularies.

Berners-Lee (1998) compares the data model of the Semantic Web to the data model of relational databases (RDB) and in fact as shown in Figure 1 the relationship of the Semantic Web to databases parallels the relationship of the World Wide Web to documents.

For the Semantic Web, the Resource Description Framework (RDF) is the core technology. As Lassila and Swick (1999) show, RDF enables data providers to publish data and specify the semantics of their data in an interoperable way on the World Wide Web. Once RDF data is published on the web and linked to other data sources, this data is called linked data. The Linked Open Data project, illustrated in Figure 2, aims to list at least all openly accessible linked data. Notably a significant part of the cloud is geo related data. The dbpedia data set, presented by Auer et al. (2007), is a central point within the Linked Open Data cloud.

The OpenStreetMap (OSM) project massively impacts the world of geographic data, because it is the first global, comprehensive and accessible source of geoinformation, which can be used free of charge and free of license restrictions. However, also Geonames, Flickr or even Wikipedia can now be used as geographic information sources. All these projects collect, as Goodchild (2007) terms it, "volunteered geographic information (VGI)".

Linking this information at web scale and making it meaningful for computers to allow automatic processing and reasoning has a big potential to generate new knowledge from interconnected information sources. Berners-Lee (2006) argues that "it is the unexpected re-use of information which is the value added by the web". The fact of accidentally finding information, which is important for some purpose, while looking for something entirely unrelated is also known as serendipity. To make serendipity effects in the geographic domain possible both spatial and semantic linking of information is necessary.

3. Comparison of MRDB and Linked Data

For cartography and geoinformation systems the Semantic Web can be an alternative for retrieval and query of (geo)information. Therefore we will compare the concept of Linked Data to "classical" Multiple Representation Databases (MRDB). We use the definition of MRDB given by Sarjakoski (2007). As Table 1 shows, there are basic similarities between both concepts, especially if the definition of MRDB is extended to not only imply a database but also the web as an underlying structure.

Table 1. Similarities and Differences between MRDB and Linked Data.

	MRDB	Linked Data
Similarities	• a *(database | web)* structure in which several representations of the same geographic entity or phenomenon such as a building or a lake are stored as different objects in a *(database | web)* environment and linked (Sarjakoski 2007) • consist of various representations […], providing a set of different views of the same object (Sarjakoski 2007) • geometry-driven feature matching	
Differences	• focus on different geometric and semantic abstraction levels • Level of Detail strongly considered • persistence and consistency can be supervised by the producer • corresponding objects at different levels are explicitly linked (Sarjakoski 2007) • focus on geometry, attributes and class hierarchies • schema matching • corporate data • authority driven	• focus on different representations of the same entity: different type and content of information • Level of Detail sparsely considered • persistence and consistency cannot be guaranteed by web links • marginal vertical structure of geographic data • focus on formal ontological descriptions • ontology matching between different OWL ontologies • web / distributed data • community-driven

These two concepts differ in their focus on Levels of Detail and in semantic as well as geometric abstraction. As Linked Data contains data that is distributed over the web, it is a more community-driven approach than corporate MRDBs, which are mostly maintained by an authority. This results in another important difference: persistence

and consistency of Semantic Web resources cannot be guaranteed, whereas the producer of an MRDB can supervise his product.

Matching techniques as used for the assignment of homologous features in an MRDB can also be applied for interconnecting Linked Data. Schema matching with community data will be more complex compared to authority data as contributors of community data tend to interpret existing rules less strictly than employees of an authority. Hence community data in general is less consistent than authority data.

4. Matching of LinkedGeoData and Geonames

LinkedGeoData[1] is the implementation of the OSM data for the Semantic Web as presented by Auer et al. (2009). Geonames[2] is a community-driven database which contains place names and points of interest. In this paper we are going to examine a method for linking at least parts of both data sets. The actual benefit of this linking can be that the multilingual place gazetter of the Geonames project will enrich the mostly monolingual tagged points of interests within the OSM data set. A second advantage will be the possibility of data validation through the use of two independent data sets.

As our current work concentrates on residential areas, we are furthermore interested in the explicitly tagged structure of the administrative hierarchy, which is contained in the Geonames data set. Implicitly this information is also included in the OSM data set because of the given geographic extent of the features. Furthermore, using both data sets could lay the ground for more efficient querying.

Table 2. Residential areas in the Geonames data set, 19.04.2010.
Feature codes with no instances in the data set are omitted.

Feature code[3]	Feature description	Number of features
P.PPL	populated place, a city, town, village, or other agglomeration of buildings where people live and work	73990
P.PPLA	seat of a first-order administrative division	15
P.PPLA2	seat of a second-order administrative division	1
P.PPLC	capital of a political entity	1
P.PPLL	populated locality, an area similar to a locality but only with a small group of buildings	2236
P.PPLQ	abandoned populated place	8
P.PPLR	religious populated place, a place whose population is engaged in religious occupations	1
P.PPLS	populated places, cities, towns, villages	1
P.PPLW	destroyed populated place, a village, town or city destroyed by a natural disaster, or by war	1
P.PPLX	section of populated place	2297
Sum		78551

Table 2 and Table 3 show the number of features for each type of residential area and the related sums for both data sets in our test region Germany. The OpenStreetMap features were derived by importing the planet file into a PostGIS spatial database.

[1] http://linkedgeodata.org/
[2] http://www.geonames.org/
[3] http://download.geonames.org/export/dump/featureCodes_en.txt

Throughout the whole Geonames data set, populated places are stored as points, while the OSM project allows these features to be modelled as a point, line or polygon. The total number of residential area features in OSM is by 12.5% less than in the Geonames data set.

Table 3. Residential areas in the OpenStreetMap data set.
Derived using the osm2pgsql tool, 31.03.2010.

Place[4]	point	way (line)	way (polygon)	Sum
City	87	5	10	102
Town	2235	10	52	2297
Village	35743	85	355	36183
Suburb	7694	31	170	7895
Hamlet	21930	122	297	22349
Sum	67689	253	884	68826

Our matching heuristic, which links the data sets LinkedGeoData and Geonames, applies a combination of type information, spatial distance and name similarity.

We use the Levenshtein minimum string distance, which is implemented in the PostgreSQL database[5], as a measurement for the name similarity. The Levenshtein algorithm is described in Levenshtein (1966). Two names are considered similar, if the Levenshtein distance between both names is equal or less than 1. Figure 3 shows an example of possible place names in Geonames that still can be matched to the corresponding OSM feature. Thus the Levenshtein function allows the matching heuristic to be robust against minor spelling mistakes and differences, such as in the case of "Asbach-Süd", "Asbach -Süd", "Asbacc-Süd" and "Asbach Süd".

For the task of the geometric matching, we used the spatial database PostGIS. To take advantage of the built-in spatial indexes we calculated bounding boxes for each point with a size of 0.1 x 0.1 degree. In Germany this translates into search areas around each point with a size of approximately 6 x 10 km.

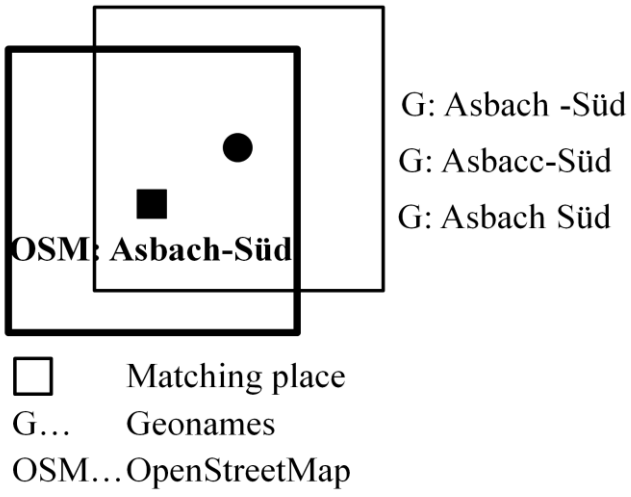

Figure 3. Examples for different names, which are detected as similar by using the Levenshtein distance with a threshold of 1.

[4] http://wiki.openstreetmap.org/wiki/Map_Features#Places
[5] http://www.postgresql.org/docs/8.4/interactive/fuzzystrmatch.html

In our algorithm a place of OSM is matched with a place of Geonames, if the buffering bounding boxes overlap, the Levenshtein distance for both names is equal or less than 1 and if both features are tagged as a residential area feature. Figure 4 shows a configuration where more than one matching Geonames point (G: Mansbach and G: Ransbach) is found for an OSM point by this algorithm. To avoid wrong matches for these cases, the matching method is refined by allowing only exact name matches, if more than one candidate is within the search area of the OSM place.

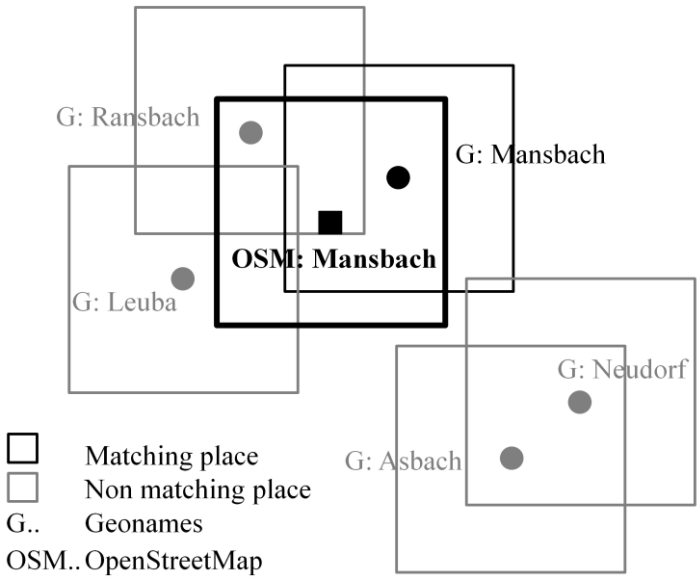

Figure 4. Combination of overlapping buffer bounding boxes and name similarity.

Table 4 and Table 5 show the results of the matching algorithm. It can be seen that the portion of matched features for cities, towns and villages is better than for suburbs and hamlets. The high percentage of matches for cities, towns and villages shows on the one hand that the proposed algorithm is well suited to link place names of both data sets. On the other hand it can be concluded that both data sets already have a good coverage of residential area features in Germany.

Furthermore it can be stated that there is big overlap of both data sets at the current stage. However both data sets contain features that the other data set does not contain. Further investigation is needed to examine the low percentage of matches for the hamlet and suburb features.

Table 4. OSM places and their matches with Geonames features.

Place	Features	Matches with Geonames	Matching percentage (%)
City	102	83	81
Town	2297	1933	84
Village	36183	32098	89
Suburb	7895	5756	73
Hamlet	22349	10816	48
Sum	68826	50686	74

Table 5. Geonames features and their matches with OSM places.

Feature Code	Features	Matches with OSM	Matching percentage (%)
P.PPL	73990	49236	67
P.PPLL	2236	450	20
P.PPLX	2297	1223	53
Sum	78523	50909	65

5. Work in progress

After finishing the matching process, we are going to validate both data sets against each other. Furthermore the results will be published in the Semantic Web using the RDF mapping implementation of the D2RQ tools presented by Bizer and Seaborne (2004). Beside that there is ongoing research work on using the SPARQL query language to query spatial and non-spatial information from Linked Data.

References

Auer S, Bizer C, Lehmann J, Kobilarov G Cyganiak R and Ives Z, 2007, Dbpedia: A nucleus for a web of open data. *Proceedings of ISWC07*, 2007.

Auer S, Lehmann J and Hellmann S, 2009, *LinkedGeoData - Adding a Spatial Dimension to the Web of Data*. Institute of Computer Science, Universität Leipzig.

Berners-Lee T, 1998, *What the Semantic Web can represent.* http://www.w3.org/DesignIssues/RDFnot.html.

Berners-Lee T, Hendler J and Lassila O, 2001, The Semantic Web. *The Scientific American*, 284(5):34-43.

Berners-Lee T, 2006, *Linked Data - Design Issues.* http://www.w3.org/DesignIssues/LinkedData.html.

Bizer C and Seaborne A, 2004, D2RQ – Treating Non-RDF Databases as Virtual RDF Graphs. In: McIlraith SA et al. (eds), *Proceedings of 3rd International Semantic Web Conference (ISWC04)*, Hiroshima, Japan.

Brickley D and Guha RV, 2004, *RDF Vocabulary Description Language 1.0: RDF Schema*, http://www.w3.org/TR/rdf-schema/.

Cyganiak R and Jentzsch A, 2010, *About the Linking Open Data dataset cloud*, http://richard.cyganiak.de/2007/10/lod/.

Goodchild MF, 2007, Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(1):211–221.

Kuhn W, 2003, Semantic Reference Systems. *International Journal of Geographic Information Science*, 17(5):405-409.

Levenshtein VI, 1966, Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics-Doklady*, 10(2):707-710.

Lassila O and Swick RR, 1999, *Resource Description Framework (RDF) Model and Syntax Specification*, http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/.

Sarjakoski LT, 2007, Conceptual Models of Generalisation and Multiple Representation. In: International Cartographic Association (ed), *Generalisation of geographic information: cartographic modelling and applications*. 11–36.

van Harmelen F, 2008, Semantic Web Technologies as the Foundation for the Information Infrastructure. In: P. van Oosterom and S. Zlatanova (eds), *Creating Spatial Information Infrastructures. Towards the Spatial Semantic Web*, Taylor & Franics Group, Boca Raton, USA, 37–52.