LinkedGeoData – Collaboratively Created Geo-Information for the Semantic Web

Sören Auer and Jens Lehmann

Universität Leipzig, Institute of Computer Science,
Johannisgasse 26, 04103 Leipzig, Germany
`{lastname}@informatik.uni-leipzig.de`
`http://aksw.org`

Abstract. In order to employ the Web as a medium for data and information integration, comprehensive datasets and vocabularies are required as they enable the disambiguation and alignment of other data and information. Many real-life information integration and aggregation tasks are impossible without comprehensive background knowledge related to spatial features of the ways, structures and landscapes surrounding us. OpenStreetMap is a collaboratively created database with a very large and active community providing a rich source of geo information. In this article we present the LinkedGeoData project, in which we transformed OpenStreetMap to RDF and interlinked it with other knowledge bases. We present LGD browser, which provides an interface for humans and allows to browse and edit this information efficiently.

1 Introduction

It is meanwhile widely acknowledged that the Data Web will be an intermediate step on the way to the Semantic Web. The Data Web paradigm combines lightweight knowledge representation techniques (such as RDF, RDF-Schema and simple ontologies) with traditional Web technologies (such as HTTP and REST) for publishing and interlinking data and information.

In order to employ the Web as a medium for data and information integration, comprehensive datasets and vocabularies are required as they enable the disambiguation and alignment of other data and information. With DBpedia [1], a large reference dataset providing encyclopedic knowledge about a multitude of different domains is already available. A number of other datasets tackling domains such as entertainment, bio-medicine or bibliographic data are available in the emerging linked Data Web[1].

Many real-life information integration and aggregation tasks are, however, impossible without comprehensive background knowledge related to spatial features of the ways, structures and landscapes surrounding us. Such tasks include, for example, to depict locally the offerings of the bakery shop next door, to map

---

[1] See, for example, the listing at: `http://esw.w3.org/topic/TaskForces/CommunityProjects/LinkingOpenData/DataSets`

distributed branches of a company or to integrate information about historical sights along a bicycle track.

With the OpenStreetMap (OSM)[2] project, a rich source of spatial data is freely available. It is currently used primarily for rendering various map visualizations, but has the potential to evolve into a crystallization point for spatial Web data integration. In this paper we contribute to the generation of an additional spatial dimension for the Data Web by elaborating on:

- how the OpenStreetMap data can be transformed and represented adhering to the RDF data model,
- how this data can be interlinked with other spatial data sets,
- how it can be made accessible for machines according to the linked data paradigm and for humans by means of a faceted geo-data browser.

The resulting RDF data comprises approximately 2 billion triples. In order to achieve satisfactory querying performance, we have developed a number of optimizations. These include a one-dimensional geo-spatial indexing as well as summary tables for property and property value counts. As a result, querying and analyzing LinkedGeoData is possible in real-time; thus enabling completely new spatial Data Web applications.

The paper is structured as follows: after introducing the OpenStreetMap project in Section 2, we describe how we published this information as RDF in Section 4. We present a mapping to DBpedia in Section 5. In Section 6 we showcase a faceted geo-data browser and editor and conclude in Section 7 with an outlook to future work.

The article is a short version of the original LinkedGeoData article [3]. However, here we focus on the added value of LinkedGeoData for the Semantic Web community. This includes a description of how other application can interact with the data and the presentation of the LGD browser. Many technical details on the transformation process are omitted.

2 The OpenStreetMap Project

OpenStreetMap is a collaborative project to create a free editable map of the world. The maps are created by using data from portable GPS devices, aerial photography and other free sources. Registered users can upload GPS track logs and edit the vector data by using a number of editing tools developed by the OSM community. Both rendered images and the vector dataset are available for downloading under a Creative Commons Attribution-ShareAlike 2.0 license. OpenStreetMap was inspired by the Wiki idea - the map display features a prominent 'Edit' tab and a full revision history is maintained.

Until now the OpenStreetMap project has succeeded in collecting a vast amount of geographical data (cf. Figure 1), which in many regions already surpasses by far the quality of commercial geo-data providers[3]. In other regions,

---

[2] http://openstreetmap.org
[3] Data about the Leipzig Zoo, for example, includes the location and size of different animals' vivariums.

where currently only few volunteers contribute, data is still sparse. The project, however, enjoys a significant growth in both active contributors and daily contributed data so that uncharted territory vanishes gradually. For some regions the project also integrates publicly available data (as with the TIGER data in the U.S.) or data donated by cooperations (as in The Netherlands).

Category	Overall Amount	Daily Additions (avg.)	Monthly Growth in the last year
Users	127,543	200	11%
Uploaded GPS points	915,392,139	1,600,000	10%
Nodes	374,507,436	400,000	5%
Ways	29,533,841	30,000	7%
Relations	136,245	300	6%

Table 1. OSM statistics as of June 2009.

The OSM data is represented by adhering to a relatively simple data model. It comprises three basic types - *nodes*, *ways* and *relations* - each of which are uniquely identified by a numeric id. Nodes basically represent points on earth and have longitude and latitude values. Ways are ordered sequences of nodes. Relations are, finally, groupings of multiple nodes and/or ways. Each individual element can have a number of arbitrary key-value pairs (tags in the OSM terminology). Ways with identical start and end nodes are called closed and are used to represent buildings or land use areas, for example.

3 The LinkedGeoData Ontology

A part of the LGD ontology[4] is derived from the relational representation of OpenStreetMap. It includes (or reuses) classes like `geo-wgs84:SpatialThing` with subclasses `node`, `way`, `relation` and properties such as `geo-wgs84:lat`, `geo-wgs84:lon`, `locatedNear`, `rdfs:label`. A major source of structure, however, are the OSM tags, i.e. attribute-value annotations to nodes, ways and relations.

There are no restrictions whatsoever regarding the use of attributes and attribute values to annotate elements in OSM. Users can create arbitrary attributes and attribute values. This proceeding is deliberate in order to allow new uses as well as to accommodate unforeseen ones. There is, however, a procedure in place to recommend and standardize properties and property values for common uses. This procedure involves a discussion on the OSM mailinglist and after acceptance by the community the documentation of the attribute on the OSM wiki[5].

When we examined the commonly used attributes we noticed that they fall into three categories:

---

[4] The LGD ontology is available at: `http://linkedgeodata.org/vocabulary`

[5] `http://wiki.openstreetmap.org/wiki/Map_Features`

- *classification attributes*, which induce some kind of a class membership for the element they are applied to. Example include: `highway` with values `motorway`, `secondary`, `path` etc. or `barrier` with values `hedge`, `fence`, `wall` etc.
- *description attributes*, which describe the element by attaching to it a value from a predefined set of allowed values. Examples include: `lit` (indicating street lightning) with values `yes/no` or `internet_access` with values `wired`, `wlan`, `terminal` etc.
- *data attributes*, which annotate the element with a free text or data values. Examples include: `opening_hours` or `maxwidth` (indicating the maximal allowed width for vehicles on a certain road).

We employ this distinction to obtain an extensive class hierarchy as well as a large number of object and datatype properties. The class hierarchy is derived from OSM classification attributes. All classification attributes are interpreted as classes and their values are represented as subclasses. Thus `secondary`, `motorway` and `path`, for example, become subclasses of the class `highway`. OSM elements tagged with classification attributes are represented in RDF as instances of the respective attribute value. In some cases the value of classification attributes is just `yes` - indicating that an OSM element is of a certain type, but no sub-type is known. In this case we, assign the element to be an instance of the class derived from the classification attribute. Consequently, a way tagged with `highway=yes` would become an instance of the class `highway`. Description attributes are converted into object properties, the respective values into resources. Data attributes are represented as datatype properties and their values are represented as RDF literals.

The resulting ontology contains roughly 500 classes, 50 object properties and ca. 15,000 datatype properties. Only half of the datatype properties, however, are used more than once and only 15% are used more than 10 times. We aim at making this information timely available to the OSM community so that the coherence and integration of OSM information can be increased.

4   Publishing LinkedGeoData

For publishing the derived geo data, we use Triplify [2]. Triplify is a simplistic but effective approach to publish Linked Data from relational databases. Triplify is based on mapping HTTP-URI requests onto relational database queries. Triplify transforms the resulting relations into RDF statements and publishes the data on the Web in various RDF serializations, in particular as Linked Data. However, in order to retrieve information, the point or way identifiers have to be known, which is usually not the case. A natural entry point for retrieving geo data, however, is the neighborhood around a particular point, possibly filtered by points holding certain attributes or being of a certain type. To support this usage scenario, we have developed a spatial Linked Data extension, which allows to retrieve geo data of a particular circular region. The structure of the URIs used looks as follows:

Longitude Latitude Radius   Property

http://LinkedGeoData.org/near/48.213,16.359/1000/amenity=pub

The linked geo data extension is implemented in Triplify by using a configuration with regular expression URL patterns which extract the geo coordinates, radius and optionally a property with associated value and inject this information into an SQL query for retrieving corresponding points of interest. This allows for efficient access of the data using several SQL optimisations and geographic indices.

The Triplify configuration can be also used to create a complete RDF/N3 export of the LinkedGeoData database. The dump amounts to 16.3 GB file size and 122M RDF triples. The different REST services provided by LinkedGeoData project by means of Triplify are summarized in Table 3. Some performance results for retrieving points-of-interest in different areas on a standard computer are summarized in Table 2.

Location	Radius	Property	Results	Time
Leipzig	1km	-	291	0.05s
Leipzig	5km	amenity=pub	41	0.54s
London	1km	-	259	0.28s
London	5km	amenity=pub	495	0.74s
Amsterdam	1km	-	1811	0.31s
Amsterdam	5km	amenity=pub	64	1.25s

Table 2. Performance results for retrieving points-of-interest in different areas.

5   DBpedia Mapping

Interlinking a knowledge base with other data sources is one of the four key principles for publishing Linked Data according to Tim Berners-Lee[6]. Within the Linking Open Data effort, dozens of data sets have already been connected to each other via `owl:sameAs` links. A central interlinking hub is DBpedia, i.e. if we are able to build links to DBpedia, then we are also connected to data sources such as Geonames, the World Factbook, UMBEL, EuroStat, and YAGO. For this reason, our initial effort consists of matching DBpedia resources with LinkedGeoData. In future work, we may extend this further.

For creating the mapping, we proceeded in two steps: The first one involves a schema mapping between OSM and LGD using the DL-Learner [4] tool. In the second step, we derived a matching heuristic based on three criteria: type information (e.g. checking whether two objects both correspond to cities), spatial distance, and name similarity. We used several optimisations, which allowed us to compute all mappings between DBpedia and LGD in less than two days on

---

[6] `http://www.w3.org/DesignIssues/LinkedData.html`

Description	URL
Points of interest in a **circular area**	`lgd:near/%lat%,%lon%/%radius%`
Example: Points of interest in a 1000m radius around the center of Dresden	`lgd:near/51.033333,13.733333/1000`
Points of interest in a **circular area having a certain property**	`lgd:near/%lat%,%lon%/%radius%/%category%`
Example: Amenities in a 1000m radius around the center of Dresden	`lgd:near/51.033333,13.733333/1000/amenity`
Points of interest in a **circular area having a certain property value**	`lgd:near/%lat%,%lon%/%radius%/%property%=%value%`
Example: Pubs in a 1000m radius around the center of Dresden	`lgd:near/51.033333,13.733333/1000/amenity=pub`
A **particular point** of interest (identified by its OSM id)	`lgd:node/%OSMid%`
Example: The Cafe B'liebig in Dresden	`lgd:node/264695865`
A **particular way** (identified by its OSM id)	`lgd:way/%OSMid%`
Example: Alte Mensa at TU Dresden	`lgd:way/27743320`

Table 3. LinkedGeoData services provided using Triplify.

an average computer. The heuristic is pessimistic, i.e. we rather omit a potential mapping instead of providing false mappings. We evaluated precision and recall of the mapping on a hand labelled fraction of the data. We refer to [3] for details.

Table 4 provides an overview of the quantity of obtained mappings for different types. Despite our strict matching heuristic, the matches cover 53.8% of all DBpedia entities of the given types and can therefore be considered a valuable addition to the Linking Open Data effort. Most of the 53.010 matches found are cities, since they are common in Wikipedia and well tagged in OSM. Many DBpedia entities, which cannot be matched, do either not exist in LGD, are not classified, or misclassified in DBpedia (e.g. the German city Aachen is typed as `dbpedia-owl:Country` since it used to be a country in the Middle Ages).

6   Faceted LinkedGeoData Browser and Editor

In order to showcase the benefits of revealing the structured information in OSM, we developed a facet-based browser and editor for the linked geo data (cf.

Type	#Matches	Rate	Type	#Matches	Rate
city	45729	70.9%	country	160	20.1%
railway station	929	24.8%	island	313	29.8%
university	210	13.3%	mountain	1475	24.5%
school	1483	38.4%	river	677	32.0%
airport	649	8.4%	lighthouse	25	4.3%
lake	1014	22.1%	stadium	346	17.0%

Table 4. Matching Results: The second column is the total number of matches found for this type. The third column is the percentage of DBpedia entities of this type, which now have links to LGD.

Figure 1)[7]. It allows to browse the world by using a slippy map. Once a region is selected, the browser analyzes the descriptions of nodes and ways in that region and generates facets for filtering. Once a facet or a specific facet value has been selected, matching elements are displayed as markers on the map and in a list. If the selected region is changed, these are updated accordingly.
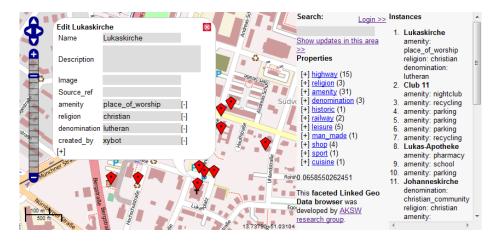

Fig. 1. Faceted Linked Geo Data Browser and Editor.

If a user logs into the application by using her OSM credentials, the displayed elements can directly be edited in the map view. For this, the browser generates a dynamic form based on existing properties. The form also allows to add arbitrary additional properties. In order to encourage reuse of both properties and property values, the editor performs a type-ahead search for existing properties and property values and ranks them according to the usage frequency. When changes are made, these are stored locally and propagated to the main OSM database by using the OSM API.

Performing the facet analysis naively, i.e. counting properties and property values for a certain region based on longitude and latitude, is extremely slow.

---

[7] Available online at: `http://linkedgeodata.org/browser`

This is due to the fact that the database can only use either the longitude or the latitude index. Combining both - longitude and latitude - in one index is also impossible, since, given a certain latitude region, only elements in a relatively small longitude region are sought for.

A possible solution for this indexing problem is to combine longitude and latitude into one binary value, which can be efficiently indexed. The challenge is to find a compound of longitude and latitude, which preserves closeness. This is possible by segmenting the world into a raster of, for example, 2^{32} tiles, whose x/y coordinates can be interleaved into a 32-bit binary value[8]. The resulting tiles are squares with an edge length of about 600m, which is sufficient for most use cases[9]

The 32-bit tile address for a given longitude and latitude value can be efficiently computed by the DBMS using the following formula:

```
(CONV(BIN(FLOOR(0.5 + 65535*(longitude+180)/360)), 4, 10)<<1)
    | CONV(BIN(FLOOR(0.5 + 65535*(latitude+90)/180)), 4, 10)
```

In this formula "`<<1`" symbolizes a bit-shift by one digit to the left, "`|`" is the bitwise "OR" and `CONV` converts the first argument from number base given as second argument to the number base given as third argument.

After elements are associated with the tiles they are located on and after tiles are indexed by the DBMS, elements located on a certain tile can be fairly efficiently retrieved. If the user browses to a certain area, the application has to determine all the tiles encircled by that area. Since co-located tiles are assigned to adjacent tile numbers, a certain area usually consists of a small number of tile ranges, which can be efficiently processed by the DBMS.

Even these indexing optimizations were not yet sufficient to obtain acceptable response times for the faceted browser. In order to further increase the querying performance, we precomputed the counts for all properties on all tiles, as well as the counts of all property values for a set of predefined properties of which we know that they have only a limited number of values. We did that not only for the highest zoom level, but for each zoom level which users are able to select. The lower the zoom level, the more the number of tiles reduces and the faster corresponding property and property value count aggregates can be computed.

7   Conclusions and Future Work

The transformation and publication of the OpenStreetMap data according to the Linked Data principles adds a new dimension to the Data Web: spatial data can be retrieved and interlinked on an unprecedented level of granularity. This enhancement enables a variety of new Linked Data applications such as geo-data syndication or semantic-spatial searches. The dynamic of the OpenStreetMap

---

[8] This is also discussed on `http://wiki.openstreetmap.org/wiki/Quadtile`

[9] In fact, the precision can be increased arbitrarily by using simply a larger number of tiles.

project will ensure a steady growth of the dataset. Furthermore, we established mappings with DBpedia as the central interlinking hub in the Web of Data. We also presented an efficient browser and editor for semantically enriched geo-data. For a discussion of related work in this area, we refer to the original LGD article [3].

In the future, we aim to extend LinkedGeoData with further geographic knowledge bases, e.g. Geonames[10]. One of the benefits will be that the tagging structures in OpenStreetMap will be complemented by the hierarchical structural features in Geonames. We may also integrate the efficient matching methods we have used in ontology matching tools like SILK.

In general, we plan to build a community around LinkedGeoData and encourage people to use the data provided by OpenStreetMap in novel ways through our interfaces, SPARQL endpoint, and Linked Data.

References

1. Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, and Zachary Ives. Dbpedia: A nucleus for a web of open data. In *Proc. of the 6th International Semantic Web Conference*, pages 11–15. Springer, 2007.
2. Sören Auer, Sebastian Dietzold, Jens Lehmann, Sebastian Hellmann, and David Aumueller. Triplify - lightweight linked data publication from relational databases. In *Proceedings of the 17th International Conference on World Wide Web, WWW 2009, Madrid, Spain, April 20-24, 2009*, pages 621–630, 2009.
3. Sren Auer, Jens Lehmann, and Sebastian Hellmann. LinkedGeoData - adding a spatial dimension to the web of data. In *Proc. of 7th International Semantic Web Conference (ISWC)*, 2009.
4. Jens Lehmann. DL-Learner: Learning concepts in description logics. *Journal of Machine Learning Research*, 2009. To appear.

---

[10] `http://geonames.org`