Multiscale Exploration of Spatial Statistical Datasets: A Linked Data Mashup Approach

Ba-Lam Do, Tuan-Dat Trinh, Peter Wetz, Elmar Kiesling,
Amin Anjomshoaa, and A Min Tjoa

Vienna University of Technology, Vienna, Austria
{ba.do,tuan.trinh,peter.wetz,elmar.kiesling,
amin.anjomshoaa,a.tjoa}@tuwien.ac.at

Abstract. Many national and international organizations today leverage semantic web technologies to make statistical datasets available as Linked Open Data (LOD). A key advantages of this approach is that the data not only becomes publicly available, but also machine-readable and hence suitable for automated discovery and exploration. Whereas this has great potential to support interesting use cases, it remains difficult for end users today to utilize and combine these statistical Linked Data. Three challenges are: (i) directing users to relevant data sources based on a specified location; (ii) facilitating data integration despite a lack of outgoing links between datasets; and (iii) offering flexible means to integrate and aggregate data from various sources. As time and location are highly relevant dimensions in most statistical data, we address the identified challenges by first constructing geographical metadata for statistical sources. Following a mashup approach, we introduce mechanisms to recommend interesting datasets to end users and automatically enable data integration, visualization, and comparisons based on user-defined criteria.

1 Introduction

In recent years, statistical data communities have increasingly adopted semantic web technologies in order to formalize and link their published datasets. In particular, many organizations make use of the W3C RDF Data Cube vocabulary [2] to publish their data as LOD. The data hence becomes accessible in a standardized format. However, means for end users that allow them to utilize and combine statistical LOD datasets on the web are lacking, which may be attributed to the following reasons:

(i) There are a large and growing number of statistical sources covering a wide range of domains. Those are valuable assets, but without an appropriate catalogue, users are unable to discover relevant statistical information.

(ii) Each statistical LOD dataset typically uses a distinct terminology and does not consistently include outgoing links to other sources. Consolidating equivalent entities stored in different sources therefore poses a significant chal-

lenge. For example, Ireland's 2011 census data[1] includes thousands of geographical entities (county, province, town, city, etc.), none of which are linked to any external resources. Other LOD sources refer to Irish cities with different URIs. Collecting all statistical information on a city available as LOD is therefore difficult.

(iii) Current statistical data visualization applications typically focus on displaying a given dataset of a particular granularity (e.g. country level, city level, etc.). They do not, however, provide mechanisms for locating relevant datasets suitable for meaningful comparisons.

This paper addresses these issues, focusing on the geographical dimension which is highly relevant in most statistical datasets. In particular, we use geographical points as input for the data source discovery process. Based on a given location, we can find appropriate datasets that provide statistical information. For instance, a user may locate his/her home on a map and trigger a retrieval process that yields datasets that contain data about the specified location. Using the supplied location, we determine the geographical entities that the location is situated in at different levels and allow users to compare data on the respective areas to appropriate similar areas, e.g., compare the population of their country to that of others, or compare income levels in their district to that in other districts in the same city.

To deal with the first issue of discovering relevant statistical datasets, we analyze available SPARQL endpoints to construct geographical metadata, which connects each dataset to the related locations. We use Google Geocoding API[2] to collect information of a geographical area based on their name.

The second issue, i.e., lacking support for data integration due to missing or inconsistent links, is addressed by using our constructed data catalogue. This catalogue uses a consistent unique identifier for the same location, which can refer to different identifiers in disparate datasets. This approach is possible even in cases where the original location entities do not reference common external entities.

To allow users to automatically integrate data from different statistical sources and overcome the third issue, we make use of the Linked Widget platform [7] – a flexible, interactive mashup platform. This platform implements semantic web principles and allows users to model widgets and data flows. It also provides mechanisms for semantic widget discovery, terminal matching, and automatic widget composition. In this paper, we describe the design of a collection of widgets targeted towards three use cases dealing with statistical data.

The remainder of this paper is organized as follows. In Section 2, we discuss the development of a geographical metadata catalogue. Section 3 illustrates our mashup approach with a collection of widgets and three practical use cases. Section 4 provides pointers to related work and we conclude in Section 5 with an outlook on future research.

[1] `http://data.cso.ie/datasets/index.html`
[2] `https://developers.google.com/maps/documentation/geocoding/`

2 Geographical Metadata for Statistical Data

Countries are typically subdivided into hierarchical administrative units, e.g., *regions*, *provinces*, *cities*, *districts*, *communes*. Current statistical LOD sources usually provide information describing specific characteristics of these administrative units in chronological order. Time and location are therefore typically the most important dimensions for this kind of statistical data.

We therefore develop a catalogue for spatial statistical data. If users provide an address, e.g., *Donaufelder Strasse 54, Austria*, or simply define a point on a map, we detect corresponding administrative areas and use those to select relevant datasets. We then provide suggestions to compare these areas with other areas at the same administrative level. For example, users can compare the population of their *district* to that of other *districts* in their *city*, or contrast the income of their *city* to other *cities* in the *country*; on a higher level, they can compare the happiness index of their *country* to that of other *countries* in the world.

We use SPARQL endpoints as input of the catalogue creation process and perform four steps: (i) identify all datasets of the source; (ii) identify all dimensions and measures for each dataset; (iii) from the set of detected dimensions, determine the location dimension, (iv) for each text value of the dimension location, identify its administrative area and create a unique resource in the catalogue. We follow the results returned by the Google Geocoding API to deal with different geographical hierarchies of countries, ensuring consistent entity classification.

The metadata structure of a dataset is illustrated in Listing 1. A definitive mapping between the administrative area and the resource URI in the metadata catalogue must exist so that same locations from different sources have the same unique URI in the metadata catalogue. This facilitates integration and aggregation. We define the mapping rule and provide a *URI mapping service* to convert an administrative area to the corresponding URI in the catalogue.

```
PREFIX qb: <http://purl.org/linked-data/cube#>
PREFIX lc: <http://location-based-catalogue.ifs.tuwien.ac.at/>

[]   a   lc:SPARQLEndpoint;
     lc:hasDataSet [ lc:describes   [ a     lc:AdministrativeArea ];
                     qb:dimension   [ a     qb:DimensionProperty  ];
                     qb:measure     [ a     qb:MeasureProperty    ];
                     qb:attribute   [ a     qb:AttributeProperty  ]; ]
```

Listing 1: A SPARQL query for terminal matching

3 Widget and Mashup approach

3.1 Linked Widget Platform

To enable users to flexibly select and combine statistical datasets and synthesize desired information, we follow a visual programming paradigm implemented in

the Linked Widgets platform – a widget-based mashup platform. Its key elements are Linked Widgets, an extension of standard web widgets backed by a semantic model that follows Linked Data principles. This model describes data input/output and metadata (such as data provenance and licensing terms) that is useful for widget search and auto composition features.

There are three types of widgets, i.e., data, process, and visualization widgets. The platform provides a graphical interface for creating a data flow and composing various applications by connecting widgets in different ways. Other stakeholders can develop widgets independently and contribute widgets to the platform to extend its functionality. Whereas existing similar platforms are oriented more towards low-level data processing with generic widgets, the platform is more problem-oriented and supports modeling on a higher, semantic level.

3.2 Statistical Widget Collection

In the platform, widgets are grouped into widget collections. Each collection addresses a different problem domain. We developed a collection[3] for statistical data exploration based on spatial context. Each widget can have multiple inputs, but only a single output. The output of a widget can serve as input of another one, if their semantic models can be matched, i.e., they have a certain level of overlap. Our exemplary statistical widget collection consists of the following four widgets:

Spatial Entity Recognizer (data widget): This widget accepts an address text or a user-defined location point as an input and calls the Google Geocoding API service to obtain the corresponding spatial entities at different levels, i.e., country level, administrative area level one to five, and sublocality level. It then uses the *URI mapping service* to return the URIs of these areas from the catalogue. Note that these URIs may not be in the catalogue, because the catalogue does not create instances of all locations in the world; it only contains instances of locations that exist in the analyzed statistical datasets.

Spatial Data Locator (process widget): This widget returns a list of datasets related to the input spatial entity. It contains options to filter the output datasets based on different domains, e.g., census, transportation, or income.

Spatial Comparator (process widget): From the input of two administrative areas, this widget returns a list of datasets that contain information on the two areas that can be compared.

Google Chart (visualization widget): This widget presents the input of a single dataset or a list of datasets in appropriate charts.

All datasets returned by the two process widgets follow the W3C RDF Data Cube vocabulary and represent the data in JSON-LD format.

[3] `http://linkedwidgets.org/MashupPlatform.html?widgetCollectionId=SpatialStatisticalCollection`

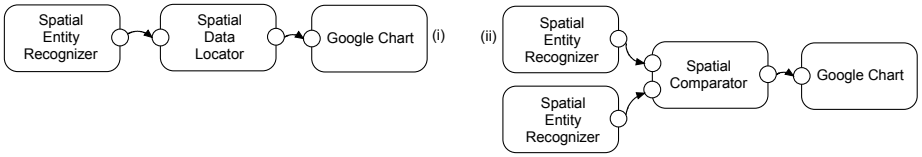

Fig. 1: Sample mashup use cases.

3.3 Available Mashups

Sample mashups created from the four widgets are shown in Fig. 1: (i) discovery of statistical information on an area (at different administrative levels) based on a user's address;[4] (ii) users select different criteria to compare a pair of areas, which have the same administrative level and the same parent area.[5]

4 Related Work

In the research field of Geographic Information Systems (GIS) the problem of spatial dataset discovery and integration has also been addressed. Hariharan et al. [9] create summaries of GIS datasets which are then used for source discovery. They adapt histogram-based techniques that take textual and spatial information into account. Jones et al. [10] present an architecture of a spatially-aware search engine to discover and access geo-datasets on the web. Their so called SPIRIT Spatial Search Engine uses a geo-ontology and spatial indexing to deal with query disambiguation, relevance ranking, and metadata extraction. Li et al. [11] enable intelligent data discovery of geo-referenced data based on the combination of Latent Semantic Analysis and Two-Tier Ranking. This technique allows them to build a semantic search engine called Semantic Indexing and Ranking (SIR), which outperforms existing keyword-matching approaches. However, these approaches do not rely on Linked Data principles.

In the Linked Data domain, with the growing number of statistical SPARQL endpoints available, several useful applications have emerged from the semantic web community. The Linked Data Query Wizard is a web-based tool for displaying, accessing, filtering, exploring, and navigating Linked Data [3]. It aims at providing a tabular interface to cope with users lack of knowledge about Linked Data and its graph structure. CubeViz is a facetted browser for statistical data utilizing the RDF Data Cube vocabulary [6]. Based on a dataset CubeViz generates a facetted browser that can be used to filter observations which in turn can be analyzed, explored, and visualized. Stats.270a.info provides an interface to compare statistical data retrieved from different sources on the web [1]. The application makes use of the RDF Data Cube vocabulary to discover statistical data in a uniform way. Furthermore, federated SPARQL queries are employed to gather data from disparate sources.

[4] `http://linkedwidgets.org?id=MashupSpatialDataLocator`
[5] `http://linkedwidgets.org?id=MashupSpatialDataComparator`

Zapilko et al. [8] discuss issues and challenges for using Linked Open Data for analyzing and enriching statistical data. They reveal problems regarding data integration, link generation, aggregation level of the data, and complexities within the data structures. Kämpgen [5] proposes to use expressive semantic web ontologies to build a conceptual model of statistical Linked Data from disparate sources. However, he also describes challenges that need to be addressed as for instance discovering datasets, or integrating data of different granularity. Kalampokis et al. [4] argue that the availability of previously closed governmental statistical datasets now enables to gain unexpected and unexplored insights into different domains. They support their claim by describing a linked open government data analytics vision along with its technical requirements. Finally they present a use case dealing with UK general election data.

Similar to the related work, we try to provide an easy to use interface for lay users to discover and explore statistical datasets. However, in contrast to current implementations, we want to provide the user with greater control to influence the process of dataset discovery, on the one hand, and data exploration, on the other hand. This is achieved by employing an approach based on widgets. Users can apply widgets to control the data processing flow from the beginning, i.e., dataset discovery, to the end, i.e., data visualization, allowing for the creation of many different scenarios. Furthermore, through creating a metadata catalogue, we offer a flexible means to compare different datasets, which are related via administrative levels, or statistical criteria they offer.

5 Conclusion and Future Work

This paper addresses two crucial issues in statistical Linked Data: (i) Where and how can users discover statistical data? (ii) How can data reconciliation, ontology matching and instance matching be performed with statistical data?

We address these issues by building a data catalogue, describing identified datasets, and finally fostering data discovery by adopting a mashup-based approach. Our approach is built around geographical context information that users provide by specifying a geographical point or an address. In future work, we will focus on the temporal dimension of the data. For instance, users provide a certain timespan they are interested in and our system suggests datasets, which support this timespan. Furthermore, we will support users in discovering datasets based on keyword search.

The current mashups work well with datasets that have a fairly limited number of dimensions. For larger datasets, it is difficult to compare different areas in a single chart. We can standardize locations through available geocoding APIs, but to standardize dimensions or measures that do not have the same meaning is still an open challenge. For instance, the year dimension has different URIs in different statistical data sources. At present, our implementation is applicable for datasets of the same statistical LOD source only and we need to improve this in future.

We currently develop a *Dataset Recommender* widget which accepts a single spatial entity as input and finds all spatial entities that share the same parent. Then, based on the criteria chosen by the user, it returns a list of datasets that contain statistical information that relate to the areas. When connecting its input with the output of *Spatial Entity Recognizer* and its output with the input of *Google Chart*, we have another mashup enabling users to compare their area to appropriate, automatically identified, other areas.

Additionally, we plan to automatically create outer links to DBpedia for the location entities of the metadata catalogue. This would allow other developers to more easily integrate statistical data with the Web of Data.

Finally, it will be necessary to evaluate precision and recall of the location detection process during the metadata construction process in a dedicated experiment.

References

1. Capadisli, S., Auer, S. Riedl, R.: Linked Statistical Data Analysis, ISWC SemStats (2013), http://csarven.ca/linked-statistical-data-analysis
2. Cyganiak, R. et al.: The RDF Data Cube Vocabulary. W3C Recommendation (2014)
3. Hoefler, P. et al.: Linked Data Query Wizard: A Novel Interface for Accessing SPARQL Endpoints. In Proceedings of the Workshop on Linked Data on the Web, CEUR-WS (2014)
4. Kalampokis, E. et al.: Linked Open Government Data Analytics. In: Wimmer, M. et al. (eds.) Electronic Government, pp. 9–110. Springer Berlin Heidelberg (2013)
5. Kämpgen, B.: DC Proposal: Online Analytical Processing of Statistical Linked Data. In: Aroyo, L. et al. (eds.) The Semantic Web – ISWC 2011, pp. 301–308. Springer Berlin Heidelberg (2011)
6. Mader, C. et al.: Facilitating the Exploration and Visualization of Linked Data. In: Auer, S. et al. (eds.) Linked Open Data – Creating Knowledge Out of Interlinked Data, pp. 90–107. Springer Berlin Heidelberg (2014)
7. Trinh, T.-D. et al.: Open Linked Widgets Mashup Platform. Proceedings of the AI Mashup Challenge 2014 ESWC Satellite Event, CEUR-WS (2014)
8. Zapilko, B. et al.: Enriching and Analysing Statistics with Linked Open Data. In NTTS-Conference on New Techniques and Technologies for Statistics, Brüssel. (2011)
9. Hariharan, R. et al.: Discovering GIS Sources on the Web using Summaries. In Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries, pp. 94–103. ACM (2008)
10. Jones, C. B. et al.: The SPIRIT Spatial Search Engine: Architecture, Ontologies and Spatial Indexing. In Geographic Information Science, pp. 125–139, Springer Berlin Heidelberg (2004)
11. Li, W. et al.: Towards Geospatial Semantic Search: Exploiting Latent Semantic Relations in Geospatial Data. In International Journal of Digital Earth, 7(1), pp. 17–37, Taylor & Francis (2014)