DC Proposal: Online Analytical Processing of Statistical Linked Data

Benedikt Kämpgen

Institute AIFB, Karlsruhe Institute of Technology, 76128 Karlsruhe, Germany
benedikt.kaempgen@kit.edu

Abstract. The amount of Linked Data containing statistics is increasing; and so is the need for concepts of analysing these statistics. Yet, there are challenges, e.g., discovering datasets, integrating data of different granularities, or selecting mathematical functions. To automatically, flexibly, and scalable integrate statistical Linked Data for expressive and reliable analysis, we propose to use expressive Semantic Web ontologies to build and evolve a well-interlinked conceptual model of statistical data for Online Analytical Processing.

1 Introduction

An important part of the Semantic Web comprises statistical Linked Data (SLD). Typically, SLD contain dimensions of metadata some of which hold temporal properties, e.g., for time-series. Also, most SLD contain numerical values that may represent aggregations from raw operational data and that often are further aggregated for analysis. According to the Linked Data principles[1] SLD should use unambiguous URIs for all relevant entities, e.g., datasets; entity URIs should be resolvable using the HTTP protocol to offer useful metadata, e.g., the location where data points of a dataset can be found; metadata should use Semantic Web standards such as RDF and SPARQL to be understandable to machines; and data should be reusing URIs from other datasources, e.g., so that relationships between datasets can be discovered.

Using Linked Data principles for publishing and consuming statistical data for decision support bears advantages such as easier integration and enrichment with other datasources. Efforts such as the Linking Open Data project, data.gov, and data.gov.uk have resulted in the release of useful SLD. Such open data is not restricted to any specific usage, its full values can be unlocked, driving innovation. First projects have demonstrated useful consumption of SLD, e.g. the Open Data Challenge.

Thus, we assume that the amount of Linked Data containing statistical information will be increasing; and so will the need for concepts of consuming SLD. There are still open questions of how to publish SLD. For instance, the W3C Working Group on Government Linked Data[2] is working on best practices and

[1] http://www.w3.org/DesignIssues/LinkedData
[2] http://www.w3.org/2011/gld/charter.html

L. Aroyo et al. (Eds.): ISWC 2011, Part II, LNCS 7032, pp. 301–308, 2011.

standard vocabularies to publish SLD from governmental institutions. Although publication issues will be inherently important for our work, we want to focus on concepts of consumption, namely the analysis of SLD.

Consider the task of comparing metrics that quantify a country's well-being with numbers that describe employees' perceived satisfaction at work. Here, we want to integrate, for instance, the European Commission's publication of the Gross Domestic Product growth of all European countries per year as provided by Eurostat, and a dataset with survey data about employees' fear of unemployment in the last few years, also published as Linked Data.

Research Question. When we try to fulfil similar scenarios of analysing SLD, we encounter specific challenges:

Distributed Datasources. Single information pieces about datasets may be distributed over servers and files and published by different parties and in different formats. Permanent availability and performance is not guaranteed.

Heterogeneous Datasets. Several heterogeneous ontologies for describing SLD are in use. There is no common agreement on how to make important aspects of statistical data self-descriptive, e.g., hierarchies of categorisations and conversion or aggregation functions.

Varying Data Quality. SLD may be incomplete, inaccurate, sparse, imprecise and uncertain; best-effort answers may be required. Generally, data may not be self-descriptive enough to aid machines in interpretation and analysis. Again, there is no common agreement on how to attach sufficient provenance information to data. Aggregating data with varying quality should be transparent for users. Also, data may imply access restrictions that need to be considered.

Scale of Linked Data. Datasets may be large and need to be explored iteratively and interactively; direct querying and analysis of SLD using ad-hoc queries on the SLD will not scale. Data warehouses are needed to (temporarily) store, pre-process, and analyse the data. Also, values may need to be pre-computed for fast look-ups of calculations. However, SLD are dynamic and may be updated or refined continuously; this requires an automatic approach to building and evolving data warehouses.

Based on these challenges, we want to approach the following research question: How to automatically (with few manual effort), flexibly (integrating many heterogeneous datasources), and fast (in comparison to other integration systems) integrate distributed SLD to one conceptual model for expressive (e.g., aggregations by hierarchies, conversions of metrics, complex calculations) and reliable (e.g., transparent best-effort answers) analysis.

Approach. To sufficiently fulfil our research question, we propose to exploit expressive Semantic Web ontologies to automatically map SLD to a well-interlinked conceptual model, so that the data can be analysed using Online Analytical Processing (OLAP).

OLAP is a commonly used decision support analysis method characterized by a multidimensional view of data and interactive exploration of data using simple to understand but data-complex queries, e.g., selection, drill-down/roll-up, and slice/dice[17]. The necessary Multidimensional Model (MDM) describes statistical data by data points, Facts, in a coordination system forming a Hypercube (Cube) of n axes, or Dimensions. Dimension Values can be grouped along Hierarchies of one or more Levels. Dimensions also can be Measures. If subsuming sets of Facts, Measures are aggregated using aggregation functions, e.g. *sum*. Cubes that share Dimensions and Values are put together into Multicubes.

Figure 1 shows an MDM that fulfils our scenario. Here, the average GDP and the cumulated number of answers given in the employment survey are made comparable in a Multicube with shared geographic and temporal Dimensions.

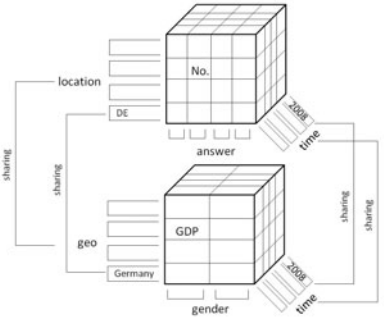

Fig. 1. Example of a Multicube

For OLAP, typically, data is extracted from heterogeneous data sources, transformed into a well-interlinked MDM, and more or less temporarily loaded into a data warehouse. Common problems[14] include: Manual effort needed for developing and maintaining such ETL pipelines; semantic gap between conceptual model and logical implementation; and inflexibility to change. Semantic Web ontologies allow to make data self-descriptive; to represent consensus about the meaning of data; to find implicit knowledge and inconsistencies; and to ease the integration effort. Although usage of Semantic Web ontologies is not explicitly required by the Linked Data principles, we assume that expressive ontological structures will make it possible to overcome the challenges of SLD analysis[10].

2 Related Work

Our research question has been addressed by roughly two kinds of work: approaches to integrate and analyse statistical data from the Web; and approaches to apply Semantic Web concepts to data warehousing and OLAP.

Publishing and consuming statistical data over the Web often is based on XML[12]. There are XML standards to transfer statistical information, e.g.,

XBRL, SDMX, and DDI. However, these approaches have problems with integrating heterogeneous datasources. They lack the concept of semantically describing statistical data. Efforts to apply Semantic Web concepts to such standards are at an early stage (e.g., SDMX[1]). Other related approaches retrieve statistical information from the Web, automatically integrate the data and let the user analyse it: Google Squared, Google Refine, and Needlebase use keyword searches and structured background information to structure data from the Web in tables. They rely more on concepts and techniques from Information Retrieval, Machine Learning, NLP and Pattern Matching, and less on ontologies and Linked Data. Google Public Data Explorer allows expressive analyses. There is work on analysing Linked Data about sensors, however, it does not allow expressive queries[13].

Niinimäki and Niemi[9] describe an ETL approach to first transform data into an ontology for multidimensional models and then serialise the ontology for use with an Online Analytical Processing server for expressive analysis. They put much focus on their specific ontology, which directly models a multidimensional model and which needs to be deployed manually for the statistical data at hand. With SLD, manually mapping the data to a conceptual model is not an option. We intend to use SLD that is sufficiently semantically described to be automatically mapped to a meaningful conceptual model. There is recent work on creating data warehouses using general ontologies[15,7]; however, they do not deal with the problem of integrating datasets described by heterogeneous ontologies. Nebot et al.[6] do so, however, they limit their work to static datasources, which is not realistic with Linked Data. Also, they require the user to manually control the building of a conceptual model; our work focuses on automatically retrieving a valid conceptual model from SLD. Much work regarding consumption of SLD has been done for Semantic Sensor data. There is work on OLAP for Semantic Web ontologies describing sensor data, however, it is not dealing with challenges of Linked Data[16].

3 Research Plan

In this section, we describe in more detail our approach of analysing SLD, and also, how we intend to measure our success. There are several Multidimensional Models (MDM)[11] with different expressivity and focus. So far, no MDM has been adopted as a standard[14]. Similarly, there are several ontologies but no commonly agreed standard to describe SLD[18].

In previous work[2], we have developed a proof-of-concept mapping between a basic MDM and the RDF Data Cube vocabulary, an ontology that is already used by some publishers of SLD. We have implemented this mapping and used the prototype in experiments with real world data for a preliminary evaluation. This approach, however, does not cope with our mentioned challenges of SLD. In the following, we describe our plan to extend our approach towards this. Mostly, this will require to automatically build and evolve more expressive MDMs from SLD.

Distributed Datasources. require us to find, select and retrieve datasets. In our current system, an analysis is started from URIs of datasets to be integrated and analysed. An analysis could also start from a business question, e.g., a multidimensional query. The system then would automatically look for suitable datasets that can answer the query. Datasets and ontologies can be found in repositories and catalogs such as CKAN, or by Semantic Search engines such as Sindice and be automacally matched to users' information needs.

For instance, if we want to compare the GDP with survey results measuring the people's fear of becoming unemployed, datasets containing such metrics could be automatically added to the MDM. Also, if one dataset only contains the relevant measures for one country, additional datasets covering other countries could be recommended.

URIs of datasets are resolved to retrieve information about the datasets. This may provide new URIs, which are resolved, iteratively. At the moment, URIs are not distinguished; if once collected, they are every time used for querying the datasets, resulting in longer query times as actually needed. Also, at the moment, we do not consider data that is available in a form other than plain RDF.

Heterogeneous Datasets. require us to integrate SLD using various ontologies. At the moment, our mapping only supports a certain ontology. Datasets may even be described without any specific ontology for statistical data but still bear interesting statistics. For instance, datasets describing large numbers of people or institutions – such as the Billion Triple Challenge dataset – contain useful statistics, e.g., for each pair of institutions the number of people that know each other. Similarly of interest may be Linked Data and ontologies for geo-spatial, sensor and social-network data. In order to integrate information from different datasources a more complex MDM may be needed. E.g., many MDMs only support many-to-one relationships between Facts and Values of one specific Dimension. In real world scenarios, many-to-many relationships are possible, e.g., a patient having several diagnoses at the same time.

There are many possible heterogeneity issues when integrating SLD, e.g., how to handle time aspects such as the notion of "now". Or how to handle special-purpose values such as "unknown", "explicitly not inserted", and "not applicable". Also, different levels of granularity regarding hierarchies and calculations may need to be aligned for integration:

Hierarchies aid the user to retrieve correct and useful information. At the moment, most Dimensions only have one Hierarchy and Level, represented by *rdfs:label* from the Dimension Value. Only for time dimensions, we consider the natural time hierarchy of year, month and day. More complex Hierarchies may be useful[4]. For instance a Cube of people having their Hierarchy of supervisors as one Dimension. The supervisor Dimension is asymmetric as it may contain varying numbers of Levels, depending on the person. Linked Data provides ontologies to explicitly describe hierarchies, however, current datasets do not make use of this. Hierarchies may be contained implicitly and retrieved and enriched automatically using other sources[5].

Calculations are implicitly contained in Measures queried from an MDM. It is still an open problem how to automatically retrieve useful aggregation functions. At the moment, we use a simple heuristic to determine aggregation functions for a Measure: For each possible aggregation function we create a Measure; e.g., sum, avg, min, max, count, count, and distinct count for numerical values. However, not all aggregation functions make sense, e.g., to use as aggregation function the sum operator for a Measure giving the current stock of a product in a certain period of time. This is known as summarizability problem. Ontological structures can be of use, here[7]. Similarly, there is no common agreement on how to represent and convert between heterogeneous representations of mathematical information[18]. For instance, this would require to state how Measures were created and to uniquely represent Measure attributes such as units. Another open issue is to represent and share complex Measures over heterogeneous datasets, e.g., *precision* and *recall* in one dataset to analyse another.

Varying Data Quality. requires us to integrate incomplete, inaccurate, sparse, imprecise and uncertain information and to give best-effort answers to business questions. For instance, missing values could be filled with most probable values or values from another, less trusted source. Yet, the process of automatically selecting values to be integrated and aggregated should be transparent and comprehensible. Also, data may imply access restrictions that need to be considered. For that, the MDM could be enriched with information describing users, privileges and policies that serve to articulate an access control and audit (ACA) policy[10].

Scale of Linked Data. requires us to incrementally build and update data warehouses. In our current system no versioning of the MDM is done. However, SLD and ontologies may be continuously changing, e.g., new Facts added, Measures corrected, and Dimensions modified. Queries over such changes may be interesting, e.g., whether a Fact Dimension Value has been modified several times (known as "slowly changing dimensions"[11]). User queries may allow to restrict the search space of possibly useful MDMs[8].

For evaluation, we will implement an information system and analyse its capability to automatically, flexibly, and scalable allow expressive and reliable analysis of SLD. See Figure 2 for our planned integration system architecture which we have generalized from an early prototype[2].

The user queries for answers from SLD using an OLAP client (1). To answer this query, an RDF/SPARQL engine finds, selects, and retrieves SLD (2). In an integration engine an MDM is built or updated as a common conceptual model for the retrieved data (3). Our work will be centered around this integration engine. The MDM may be serialized in an OLAP server (4.1). Then, either this OLAP server (4.2) or the MDM directly gives the answer (5).

To overcome the challenges in analysing SLD, we intend to represent the MDM as an expressive Semantic Web ontology and to make use of concepts and techniques from fields such as Ontology Engineering and Matching, as well as Reasoning.

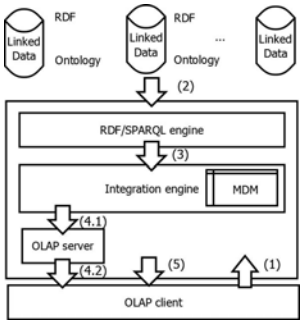

Fig. 2. Integration System Architecture

We do not intend to evaluate our concepts using qualitative usage analyses. Still, we plan to use the system in real-world use cases and compare its suitability with other systems. More concretely, for an analysis task, we plan to compare the amount of manual effort, the amount of data available, the expressivity of possible business questions, the performance of the system, and the quality of given answers. Also, we consider benchmarks such as the Business Intelligence Use Case from the Berlin SPARQL Benchmark and general quality criteria for MDMs [3].

4 Conclusion

We have proposed to exploit expressive Semantic Web ontologies to automatically create a well-interlinked conceptual model from various sources of statistical Linked Data that can be interactively and reliably analysed using Online Analytical Processing (OLAP). We have described challenges that we plan to work on, e.g., discovering datasets, integrating data of different granularities, or selecting mathematical functions. For evaluation, we intend to implement an information system suitable for real-world scenarios.

Acknowledgements. This work was supported by the German Ministry of Education and Research (BMBF) within the SMART project (Ref. 02WM0800) and the European Community's Seventh Framework Programme FP7/2007-2013 (PlanetData, Grant 257641). I thank Andreas Harth, Elena Simperl, and Denny Vrandečić for guidance and insights.

References

1. Cyganiak, R., Field, S., Gregory, A., Halb, W., Tennison, J.: Semantic Statistics: Bringing Together SDMX and SCOVO. In: Proceedings of the WWW 2010 Workshop on Linked Data on the Web, pp. 2–6 (2010)

2. Kämpgen, B., Harth, A.: Transforming Statistical Linked Data for Use in OLAP Systems. In: Proceedings of the 7th International Conference on Semantic Systems. I-SEMANTICS 2011. ACM (2011)
3. Lechtenbörger, J., Vossen, G.: Multidimensional normal forms for data warehouse design. Information Systems Journal 28(5), 415–434 (2003)
4. Malinowski, E., Zimányi, E.: Hierarchies in a multidimensional model: From conceptual modeling to logical representation. Data Knowl. Eng. 59, 348–377 (2006)
5. Mazón, J.N., Trujillo, J., Serrano, M., Piattini, M.: Improving the Development of Data Warehouses by Enriching Dimension Hierarchies with WordNet. In: Collard, M. (ed.) ODBIS 2005/2006. LNCS, vol. 4623, pp. 85–101. Springer, Heidelberg (2007)
6. Nebot, V., Berlanga, R., Pérez, J.M., Aramburu, M.J., Pedersen, T.B.: Multidimensional Integrated Ontologies: A Framework for Designing Semantic Data Warehouses. In: Spaccapietra, S., Zimányi, E., Song, I.-Y. (eds.) Journal on Data Semantics XIII. LNCS, vol. 5530, pp. 1–36. Springer, Heidelberg (2009)
7. Niemi, T., Niinimäki, M.: Ontologies and summarizability in OLAP. In: Proceedings of the 2010 ACM Symposium on Applied Computing SAC 2010, p. 1349 (2010)
8. Niemi, T., Nummenmaa, J., Thanisch, P.: Constructing OLAP cubes based on queries. In: Proceedings of the 4th ACM international workshop on Data warehousing and OLAP. DOLAP 2001. ACM (2001)
9. Niinimäki, M., Niemi, T.: An ETL Process for OLAP Using RDF/OWL Ontologies. In: Spaccapietra, S., Zimányi, E., Song, I.-Y. (eds.) Journal on Data Semantics XIII. LNCS, vol. 5530, pp. 97–119. Springer, Heidelberg (2009)
10. Pardillo, J., Mazón, J.N.: Using Ontologies for the Design of Data Warehouses. Journal of Database Management 3(2), 73–87 (2011)
11. Pedersen, T.B., Jensen, C., Dyreson, C.E.: A foundation for capturing and querying complex multidimensional data. Information Systems Journal 26, 383–423 (2001)
12. Perez, J.M., Berlanga, R., Aramburu, M.J., Pedersen, T.B.: Integrating Data Warehouses with Web Data: A Survey. IEEE Transactions on Knowledge and Data Engineering 20, 940–955 (2008)
13. Phuoc, D.L., Hauswirth, M.: Linked Open Data in Sensor Data Mashups. In: Proceedings of the 2nd International Workshop on Semantic Sensor Networks (SSN 2009) in conjunction with ISWC 2009 (2009)
14. Rizzi, S., Abelló, A., Lechtenbörger, J., Trujillo, J.: Research in data warehouse modeling and design: dead or alive? In: Proceedings of the 9th ACM International Workshop on Data Warehousing and OLAP, pp. 3–10 (2006)
15. Romero, O., Abelló, A.: Automating Multidimensional Design from Ontologies. In: Proceedings of the ACM Tenth International Workshop on Data Warehousing and OLAP DOLAP 2007 (2007)
16. Shah, N., Tsai, C.F., Marinov, M., Cooper, J., Vitliemov, P., Chao, K.M.: Ontological On-line Analytical Processing for Integrating Energy Sensor Data. Iete Technical Review 26, 375 (2009)
17. Vassiliadis, P.: Modeling Multidimensional Databases, Cubes and Cube Operations. In: Proc. of the 10th SSDBM Conference. pp. 53–62 (1998)
18. Vrandečić, D., Lange, C., Hausenblas, M., Bao, J., Ding, L.: Semantics of Governmental Statistics Data. In: Proceedings of the WebSci 2010 (2010)