Exploring Linked Statistical Data Using Linked Widgets

A Min Tjoa, Ba-Lam Do, Amin Anjomshoaa, Tuan-Dat Trinh, Peter Wetz, Elmar Kiesling
Institute of Software Technology and Interactive Systems
Vienna University of Technology
Favoritenstrasse 9-11, Vienna, Austria
{amin, lam, anjomshoaa, dat, peter.wetz, e.kiesling}@ifs.tuwien.ac.at

ABSTRACT
The Open Data movement has gained momentum among governments, in the business world, and in the public sector in recent years. This movement has resulted in a growing number of open and accessible datasets that have established a solid basis for enhanced service offerings and improved experiences for citizens and businesses. Statistical data, which embodies a big portion of Open Data, comprises a wide range of domains including finance, demographics, transportation, employment, etc. Statistical data plays an important role in public policy formation and as a facilitator for informed decision-making in the private sector. Linked Statistical Data is an evolving concept that combines the richness of Linked Data (a set of best practices for publishing and connecting structured data on the Web) with the descriptiveness of statistical data to integrate data from multiple sources and put it in a semantic context. In this short paper, Linked Statistical Data limitations and challenges are explored before introducing Linked Widgets as an innovative approach.

1. INTRODUCTION
Today, in line with the increasing adoption of Open Data policies, the amount of data published on the web by governments such as the U.S., the U.K., and the Austrian government, as well as international organizations such as the World Bank, is growing rapidly [1]. Statistical data in particular has attracted significant interest. It comprises a wide range of domains including finance, demographics, transportation, and employment and plays an increasingly important role in public policy formation and as a facilitator for informed decision-making in the private sector. Therefore, efficient data exploration and data integration approaches are necessary to allow knowledge workers to make use of the fast growing amounts of statistical data available. Data-centric solutions for the integration of statistical datasets may support end users in a number of ways:

i. They can complement incomplete data and allow end users to obtain a more comprehensive view. Using World Bank data as a source, we can obtain economic indicator - *GDP per capita* of a country in the period from 1890 to 2013. This data source, however, does not include forecasts for the following years, which can, however, be found in other data sources such as the International Monetary Fund (IMF). Furthermore, the relationship between this indicator and other indicators such as *Inflation* from World Bank or *Corruption perceptions index* from Transparency International need to be considered. Therefore, an appropriate and semantically meaningful combination of related datasets from a single (e.g., World Bank) or multiple sources (e.g., World Bank, IMF, Transparency International) facilitates end users to explore data in a more detailed manner, which a single data source cannot provide.

ii. It allows users to make comparisons between related datasets and provide policy-makers and other users sound foundations for decision-making. For example, Vienna Open Data provides various public transport datasets such as annual number of passengers and offered seats for bus, tram, and metro vehicles. By comparing these datasets, policy-makers, for instance, can identify appropriate investment for each type of vehicle in upcoming years.

In order to efficiently explore and integrate data, statistical data should be published as Linked Data, which allows to consume and manipulate without proprietary tools. A Linked Data approach allows users to combine data from different sources in order to gain new insights and to obtain higher data quality, completeness, and level of detail. A large number of organizations and governments such as the European Commission, the United Kingdom Department for Communities and Local Government, and the Scottish Government have already published their data sources as Linked Data. This proliferation of available statistical data has created enormous potential for interesting applications, but it has so far resulted only in limited adoption by end users, including developers and knowledge workers [2]. These users need appropriate tools to analyze, combine, remix, visualize and make sense of the data.

2. CHALLENGES
This section discusses challenges that arise in the exploration and integration of statistical datasets.

2.1 Acquisition and consolidation of data
As a first step toward effective exploration of statistical data, we create a common statistical knowledge base that collects data from a variety of statistical sources and represents them in a consistent format and terminology. At present, the majority of data is published in raw formats such as HTML, PDF, Excel, or CSV. This may be attributed to the simple structure and universal applicability of these data formats for various kinds of data. However, these formats generally do not convey semantics, and thus prevent automatic linking to other datasets. Therefore,

SoICT '14, December 04 - 05 2014, Hanoi, Viet Nam

http://dx.doi.org/10.1145/2676585.2676589

several organizations and governments such as the European Commission or the Scottish Government have made use of semantic technologies to annotate and publish their statistical data. Even though their data are accessible through SPARQL queries, they cannot be integrated directly because of inconsistent terminology [3].

In order to create a high-quality statistical knowledge base (i) datasets in raw formats must be semantically modelled, and (ii) datasets that are already published in RDF have to be adapted to match other datasets.

2.1.1 Modelling statistical datasets in RDF format

Currently, the RDF Data Cube vocabulary [4] is widely adopted by many organizations and governments for the modeling of statistical datasets in raw formats. This vocabulary is a W3C standard for the publication of multi-dimensional data on the web. It provides the terminology for describing the structure and components of a dataset, e.g., dimensions, or measures. In order to describe the meaning of individual components and values, this vocabulary is often used in combination with domain ontologies.

Because the RDF Data Cube vocabulary is already an established standard, the focus should now be on building appropriate domain ontologies (e.g., for finance, demographics, and transportation). Then, given a statistical dataset as the input, automatic identification of an appropriate domain ontology for this dataset could greatly improve automatic mechanisms for integration. Furthermore, we need rules to match the labels used in the dataset to the terminology of selected ontology. For example, we consolidate different labels such as "männlich", "m", and "1" with sdmx-code:sex-M.

2.1.2 Adapting RDF statistical datasets

In order to consolidate published RDF datasets of different organizations to adhere to the modelled datasets from the previous step, we can either (i) use domain ontologies to generate new RDF datasets which are corresponding to existing RDF datasets, or (ii) build an adapter for each existing RDF dataset.

Although the former provides a convenient way to manage different datasets, it may increase the data volume by orders of magnitude. Furthermore, we may not have the most current data if they are updated. Therefore, the second approach, which makes use of adapters to interconnect existing RDF datasets as well as modelled datasets, seems to be more appropriate. Each adapter has two tasks: (i) convert queries using the uniform terminology to appropriate queries for the dataset, and (ii) convert results to a new representation, which uses the consolidated terminology.

2.2 Exploration and Integration of Statistical data

Upon completion of the data acquisition and consolidation steps, we obtain high-quality data that should be easy to consume for end users. In fact, users often do not have deep knowledge of available technologies and standards. They need appropriate tools to search, combine, and visualize data without the need for a technical background. Existing approaches allow users to search data based on keywords [5], combine given datasets [6], and visually display the data [7].

We conduct research on context-based search, query generation, related datasets ranking, and data-driven visualization. The first two research areas, which make use of the actual meaning of keywords and complete specification of users' need respectively, support users in looking for data. The latter research areas recommend users appropriate combinations for each given dataset, and then represent data in a readily interpretable way.

2.2.1 Context-based search

Typically, a keyword-based search query cannot completely describe the need of users [8]. Therefore, to understand users' demand better, we need to make use of the search context, which can be query features, or user's domains of interests and background [9].

The query features [9] support expansion of search keywords. For instance, the keyword "income Vienna" is expanded to "GDP per capita Austria". To this end, it is necessary to identify: (i) meanings of keywords e.g., that the "income" keyword has a relation with the "GDP per capita" measure, or that "Vienna" is the capital of "Austria", and (ii) a list of possible relation expansions. We use a hierarchy of geographic regions (e.g. via the Google Geocoding service) in combination with appropriate domain ontologies to detect measures and automatically extract term relations based on the analysis of datasets in the same domain.

2.2.2 Query generation

At present, users cannot create complex queries such as "*Find expenditure on Education of countries which are among the 10% in terms of GDP per capita*". Keyword-based search or context-based search technologies are based on a compact specification of users' need, hence, they cannot be applied for a complex query which combine multiple datasets of different sources (e.g. Education dataset of UNESCO and GPD per capita dataset of World Bank) or have additional query conditions (e.g. top 10%).

We can represent such a query by either (i) using natural language or (ii) modelling the query based on predefined and available components/concepts. Although the former provides a quick and easy way for users to express their requirements, it is typically not possible to directly translate it into a query. In the latter case, user can use semantics of the data to perform stepwise suggestions of appropriate relationships to users.

2.2.3 Related datasets ranking

Combining multiple related datasets allows users to develop new knowledge and facilitate informed decision-making. Due to the variety of statistical data, the number of datasets relating to a given dataset can be extremely large. It is necessary to propose a method to compute the ranking of each dataset.

In order to evaluate the relevance of datasets, specific features of datasets (e.g., domain, dimensions, meaning of values in each dimension) may provide a ranking for each related dataset. For example by starting from a specific dataset that relates to Vienna, the appropriate dataset ranking could be other cities of Austria, or datasets of capital of other countries.

2.2.4 Data-driven visualization

The variety, complexity, and volume of statistical data pose a substantial challenge in data visualization. Users, therefore, need appropriate charts for each dataset. In addition, it is necessary to find ways to allow users to easily understand a complex dataset as well as to visually display parts of a large volume dataset.

Each visual chart has a distinct role in conveying meaning of data. In order to highlight features of a dataset, we need to represent data via appropriate chart types. For example, to compare the change of data over time, we can use line charts. Datasets having

single or few measures in a few periods are suitable for column or bar charts. To show the relationship between dimensions, we can use scatter charts (for two dimensions), or bubble charts (for three dimensions). Compositional data may be visualized via stacked charts or pie charts.

To reduce the complexity of statistical datasets induced by the large number of dimensions and measures, we need to fix values for some dimensions and allow users to choose measures for visualization. On the one hand, this helps users to focus on a limited number of dimensions and measures. This allows them to evaluate the impact of these dimensions on a specific measure. On the other hand, users can select parts of a dataset. Therefore, it facilitates the visualization of large volume datasets.

3. LINKED WIDGET APPROACH

This section introduces the Linked Widgets concept as a potential solution to these challenges in the domain of statistical data.

3.1 Concept of Linked Widget

The W3C Widget specification defines a "widget" as *"an interactive single purpose application for displaying and/or updating local data or data on the Web, packaged in a way to allow a single download and installation on a user's machine or mobile device"* [1]. We have extended this specification for Linked Data by including a semantic model [1]. The semantic model describes data input/output and metadata such as provenance and license. We distinguish three types of widgets:

i. *Data widgets* are used as data feeds to other widget types and generate data in a specific format based on a given set of parameters. Examples of this widget type include data summaries or SPARQL widgets.

ii. *Process widgets* receive a dataset as input and generate the output based on a customized process. Formatters, filters, and simple merge widgets are examples of process widgets.

iii. *Presentation widgets* generate visual output based on a given set of data at runtime. Examples include diagrams, tables, and information visualization.

3.2 Advantages of Linked Widgets approach

From a collection of widgets, end users can create mashups, i.e., "user-driven micro-integration of web accessible data" [10]. The Linked Widgets approach provides three salient advantages: openness, connectedness, and reusability.

i. The Linked Widgets benefits from crowd-sourcing of widgets. Widget developers can contribute widgets about their statistical resources of interest to the platform. This enables the end users to explore these resources via employing the appropriate widgets. As a result, we will have an open platform for communication between developers who provide the widgets and end users such as knowledge workers who benefit from various widgets to accomplish tasks such as data analysis and data visualization.

ii. Each widget is equipped with a semantic model that describes widget's input and output graph as well as a comprehensive set of metadata such as provenance information and required data resources. Therefore, each widget can identify which other widgets can provide the required input or consume the resulted output data. This feature supports end users in connecting semantically compatible widgets in order to create data integration solutions.

iii. End users can reuse and connect available widgets to collect, integrate, and combine data from different sources in a dynamic and creative manner. In order to promote the sharing and reuse of Linked Widgets, their models are published on the web following Linked Data principles. This makes them first-class citizens of the Linked Data Cloud and allows users to easily find and reuse them.

[1] http://www.w3.org/TR/widgets-reqs/

4. CONCLUSIONS

This paper highlights some existing limitations in the context of the processing and integration of statistical datasets and introduces an approach to tackle them in order to facilitate data exploration and data integration. Furthermore, based on the idea of Linked Data, which fosters connecting and reusing of data, we propose a novel approach to provide end users with efficient mechanisms to analyze, combine, remix, visualize, and make sense of statistical data available.

5. REFERENCES

[1] Tuan-Dat Trinh, Ba-Lam Do, Peter Wetz, Amin Anjomshoaa, A Min Tjoa. 2013. *Linked widgets an approach to exploit open government data*. In International Conference on Information Integration and Web-based Application & Services (IIWAS)

[2] Ba-Lam Do, Tuan-Dat Trinh, Peter Wetz, Amin Anjomshoaa, Elmar Kiesling, A Min Tjoa. 2014. *Widget-based Exploration of Linked Statistical Data Spaces*. In 3rd International Conference on Data Management Technologies and Applications (DATA).

[3] Christian Bizer, Tom Heath, Tim Berners-Lee. 2009. *Linked data - the story so far*. Int. Journal on Semantic Web and Information Systems, 5 (3).

[4] Richard Cyganiak, Dave Reynolds. 2011. *The RDF Data Cube Vocabulary*. URL http://www.w3.org/TR/vocab-data-cube/

[5] Patrick Hoefler, Michael Granitzer, Eduardo Veas, Christin Seifert. 2014. *Linked Data Query Wizard: A Novel Interface for Accessing SPARQL Endpoints*. In Linked Data on the Web Workshop (LDOW2014)

[6] Sarven Capadisli, Sören Auer, Reinhard Riedl. 2013. *Linked Statistical Data Analysis.* In International Workshop on Semantic Satistics.

[7] Percy E. Rivera Salas, Fernando Maia Da Mota, Michael Martin, Sören Auer, Karin Breitman, Marco Antonio Casanova. 2012. *Publishing Statistical Data on the Web*. In International Semantic Web Conference (ISWC).

[8] W. Bruce Croft, Xing Wei. 2005. *Context-based topic models for query modification*. CIIR Technical Report, University of Massachusetts.

[9] Jing Bai, Jian-Yun Nie, Guihong Cao. 2007. *Using query contexts in information retrieval*. In International ACM SIGIR conference on Research and development in information retrieval (SIGIR)

[10] JackBe Corporation. 2008. *A Business Guide to Enterprise Mashups*.