Linked Statistical Data Analysis

Sarven Capadisli[1], Sören Auer[2], Reinhard Riedl[3]

[1]Universität Leipzig, Institut für Informatik, AKSW, Leipzig, Germany, [2]University of Bonn and Fraunhofer IAIS, Bonn, Germany, [3]Bern University of Applied Sciences, Bern, Switzerland
[1]info@csarven.ca, [2]auer@cs.uni-bonn.de, [3]reinhard.riedl@bfh.ch

Document ID: http://csarven.ca/linked-statistical-data-analysis

Abstract. Linked Data principles are increasingly employed to publish high-fidelity, heterogeneous statistical datasets in a distributed way. Currently, there exists no simple way for researchers, journalists and interested people to compare statistical data retrieved from different data stores on the Web. Given that the RDF Data Cube vocabulary is used to describe statistical data, its use makes it possible to discover and identify statistical data artifacts in a uniform way. In this article, the design and implementation of an application and service is presented, which utilizes federated SPARQL queries to gather statistical data from distributed data stores. The R language for statistical computing is employed to perform statistical analyses and visualizations. The Shiny application and server bridges the front-end Web user interface with R on the server-side in order to compare statistical macrodata, and stores analyses results in RDF for future research. As a result, distributed linked statistical data can be more easily explored and analysed.

Keywords: Linked Data, SDMX, Statistics, Statistical database, Data integration, Regression analysis, User interface

1 Introduction

Statistical data artifacts and the analyses conducted on the data are fundamental to testing scientific theories about our societies. In order for the society to tract and learn from its own vast knowledge about events and things, it needs to be able to gather statistical information from heterogeneous and distributed sources. This is to uncover insights, make predictions, or build smarter systems that the society needs to progress. This brings us to the core of our research challenge; how do we reliably acquire statistical data in a uniform way and conduct well-formed analyses that is accessible to different types of data consumers and users?

This article presents an approach towards this challenge with its contributions using statistical linked dataspaces. The work herein offers a Web based user-interface for researchers, journalists, or interested people to compare statistical data from different sources against each other without having any knowledge of the technology underneath or the expertise to develop themselves. The service, which we built, proceeds with running decentralized (federated) structured queries to retrieve data from various endpoints, runs an analysis on the data, and provides the analysis back to the user. For future research, analysis is stored so that it can be searched for and reused.

2 Background

As pointed out in Statistical Linked Dataspaces [1], what linked statistics provide, and in fact enable, are queries across datasets: Given that the dimension concepts are interlinked, one can learn from a certain observation's dimension value, and enable the automation of cross-dataset queries.

The RDF Data Cube vocabulary [2] is used to describe multi-dimensional statistical data, and SDMX-RDF as one of the statistical information models. It makes it possible to represent significant amounts of heterogeneous statistical data as Linked Data where they can be discovered and identified in a uniform way. The statistical artifacts that are produced, and which use this vocabulary, are invaluable for statisticians, researchers, and developers.

Linked SDMX Data [3] provided templates and tooling to transform SDMX-ML data from statistical agencies to RDF/XML, resulting in linked statistical datasets at 270a.info [4] using the RDF Data Cube vocabulary. In addition to semantically uplifting the original data, information pertaining provenance was kept track using the PROV Ontology [5] at transformation time, while incorporating retrieval time provenance data.

3 Analysis platform for Linked Statistical Data

The analysis platform is focused on two goals: 1) a Web user interface for researchers to compare macrodata observations and to view plots and analysis results, 2) caching and storage of that analysis for future research and reuse. Here, we describe the platform at stats.270a.info [20].

3.1 User interface

A web application was created to provide users with a simple interface to conduct regression analysis and display of scatter plot(s). The interface presents three drop-down selection areas for the user: an independent variable, a dependent variable, and a time series. Both, the independent and dependent variables are composed of a list of datasets with observations, and time series are composed of reference periods of those observations. Upon selecting and

submitting datasets to compare, the interface then presents a scatter plot with the best line of best fit from a list of linear models that is tested. The points in the scatter plot represent locations, in this case, countries, which happen to have a measure value for both variables, as well as the reference period that was selected by the user. Below the scatter-plot, a table of analysis results is presented. Figure [1] is a screenshot of the user interface.

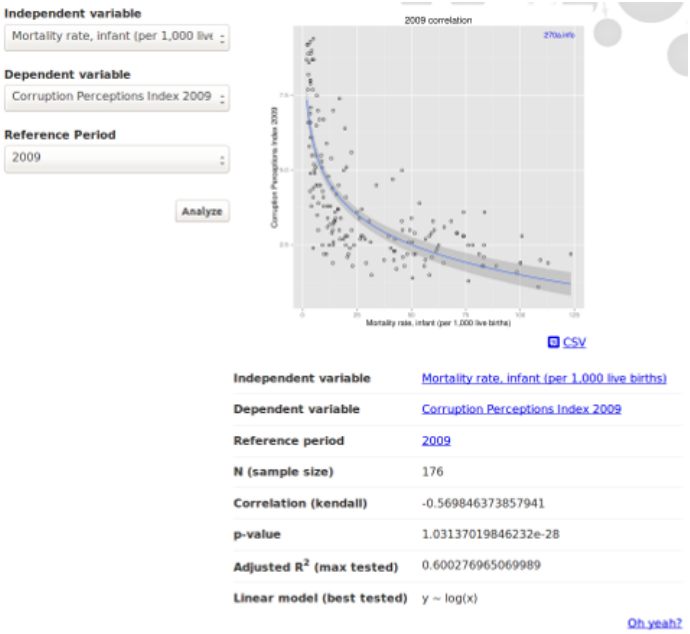

Figure 1: stats.270a.info analysis user interface

The datasets are compiled by gathering `qb:DataSet`s (an RDF Data Cube class for datasets) from each statistical dataspace at 270a.info. Similarly, the reference periods are derived from calendar intervals e.g., `YYYY`, `YYYY-MM-DD`, `YYYY-QQ`.

3.2 Data Requirements

Our expectation from the data is that it is modeled using the RDF Data Cube vocabulary and is well-formed. Specifically, it needs to pass some of the integrity constraints as outlined by the vocabulary specification. For our application, some of the essential checks are that: a unique data structure definition (DSD) is used for a dataset, DSD includes measure (value of each observation), Concept dimensions have code lists, Codes from code list.

To compare variables from two datasets, there needs to be an agreement on the concepts that are being matched for in respective observations. Here, the primary concern is about reference areas (locations), and making sure that the

comparison made for the observations from $dataset_x$ (independent variable) and $dataset_y$ (dependent variable) are using concepts that are interlinked (using the property `skos:exactMatch`). Practically, a concept e.g. Switzerland, from at least one of the dataset's code lists should have an arc to the other. It ensures that there is a reliable degree of confidence that the particular concept is interchangeable. Hence, the measure corresponding to the phenomenon being observed, is about the same location in both datasets. Figure [2] shows available outbound interlinks for the datasets at *.270a.info/.

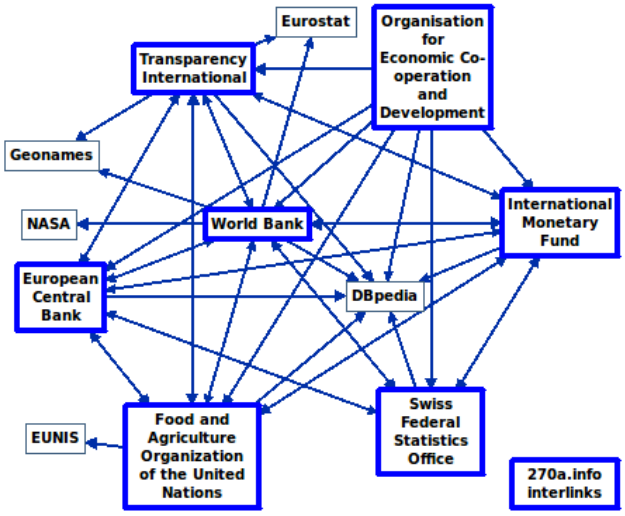

Figure 2: Outbound interlinks for 270a.info datasets

One additional requirement from the datasets is that the RDF Data Cube component properties (e.g., dimensions, measures) either use `sdmx-dimension:refArea`, `sdmx-dimension:refPeriod`, `sdmx-measure:obsValue` directly or are `rdfs:subPropertyOf`s. Given decentralized mappings of the statistical datasets (published as SDMX-ML), their commonality is expected to be the use, or a reference to SDMX-RDF properties in order to achieve generalized federated queries without having complete knowledge of the structures of the datasets, but rather only the essential bits.

In order to proceed with the analysis, we use the selections made by the user: $dataset_x$ and $dataset_y$, reference period, and then gather all observations with corresponding reference areas, and measures (observation values). Only the observations whose values for the reference areas with interlinked concepts area are retained in the final result.

3.3 Application

Shiny [22], an R package, along with Shiny server [23] is used to build an interactive web application. A Shiny application was built to essentially allow an interaction between the front-end Web application and R. User inputs are set to trigger an event which is sent to the Shiny server and handled by the application written in R.

The application assembles a SPARQL query using the input values and then sends them to stats.270a.info/sparql endpoint which dispatches federated queries to two SPARQL endpoints where the datasets are located. The SPARQL query request is handled by the SPARQL client for R. The query results are retrieved and given to R for statistical data analysis. R generates a scatter plot containing the (in)dependent variables, where each point in the chart is a reference area (e.g., country) for that particular reference period selection. Regression analysis is done where correlation, p-value, and the line of best fit is determined after testing several linear models, and shown in the user interface.

3.4 Federated Queries

The challenge in federated queries was compromising between what should be processed remotely and sent over the wire versus handling some of that workload by the parent endpoint. Since one of the requirements was to ensure that the concepts are interlinked at either one of the endpoints, each endpoint had to include each observation's reference area as well as its interlinked concept. The result from both endpoints was first joined and then filtered in order to avoid false negatives. That is, either concept_x has a `skos:exactMatch` relationship to concept_y, or vice versa, or concept_x and concept_y are the same. One quick and simple way to minimize the number of results was to filter out exact matches at each endpoint which did not contain the other dataset's domain name. Hence, minimizing the number of *join* operations which had to be handled by the parent endpoint.

The anatomy of the query is shown in detail at http://csarven.ca/linked-statistical-data-analysis. Essentially, the SPARQL Endpoint URI and the dataset URI are the only requirements.

3.5 Analysis caching and storing

In order to optimize application reactivity for all users, previously user selected options for analysis are cached in the Shiny server session.

In addition to a cache that is closest to the user, results from the federated queries as well as the R analysis, which was previously conducted, is stored back into the RDF store with a SPARQL Update. This serves multiple purposes. In the event that the Shiny server is restarted and the cache is no longer available, previously calculated results in the store can be reused, which is still more cost efficient than making new federated queries.

Another reason for storing the results back in the RDF store is to offer them over the stats.270a.info SPARQL endpoint for additional discovery and reuse of

analysis for researchers. Interesting use cases from this approach emerge immediately. For instance, a researcher or journalist can investigate analysis that meets their criteria.

3.6 URI patterns

The design pattern for the analysis URIs which refer to the data and the analysis is aimed to keep the length as minimal as possible, while leaving a trace to encourage self exploration and reuse. As URIs for both independent and dependent variable are based on datasets, and the reference period is codified, their prefixed names are used instead in the analysis URI to keep them short and friendly: `http://stats.270a.info/analysis/{prefix}:{dataset}/{prefix}:{dataset}/{prefix}:{refPeriod}`

3.7 Vocabularies

Besides the common vocabularies: RDF, RDFS, XSD, OWL, the RDF Data Cube vocabulary is used to describe multi-dimensional statistical data, and SDMX-RDF for the statistical information model. PROV-O is used for provenance coverage. A statistical vocabulary[37] is created to describe analyses. It contains classes for analyses, summaries and each data row that is retrieved. Some of the properties include: graph (e.g., scatter plot), independent and dependent variables, reference period, sample size, p-value, correlation value, correlation method that is used.

4 Conclusions

We believe that the presented work here and the prior Linked SDMX Data effort contributed towards strengthening the relationship between Semantic Web / Linked Data and statistical communities.

The availability of the analysis results in a separate storage system makes it possible for interested parties to take advantage of it variety ways. For instance, a JSON serialization of an analysis or the cached scatter plot in SVG format, is ideal for webpage widgets. Analysis can also be dynamically used in articles or wiki pages with all references intact.

The reuse of Linked analyses artifacts as well as the approach to collect data from different sources can help us build smarter systems. It can be employed in fact-checking scenarios as well as uncovering decision-making processes, where knowledge from different sources is put to their potential use when combined.

References

1. Capadisli, S.: Statistical Linked Dataspaces. Master's thesis, National University of Ireland (2012), http://csarven.ca/statistical-linked-dataspaces

2. The RDF Data Cube vocabulary, http://www.w3.org/TR/vocab-data-cube/
3. Capadisli, S., Auer, S. Ngonga Ngomo, A.-C., Linked SDMX Data, Semantic Web Journal (2013), http://csarven.ca/linked-sdmx-data
4. 270a.info, http://270a.info/
5. The PROV Ontology, http://www.w3.org/TR/prov-o/
6. stats.270a.info, http://stats.270a.info/
7. Shiny, http://www.rstudio.com/shiny/
8. Shiny server, https://github.com/rstudio/shiny-server
9. Stats Vocab, http://stats.270a.info/vocab

Appendix

Minimum Requirements

- *The application has to be an end-user application* – The stats analysis system at **http://stats.270a.info/** is intended for researchers, journalists or interested people. They interact with the Web UI to check for correlation between variables across different statistical datasets.
- *Should be under diverse ownership or control* – The original data in SDMX-ML from statistical agencies: WB (World Bank), TI (Transparency International), ECB (European Central Bank), FAO (Food and Agriculture Organization of the United Nations), OECD (Organisation for Economic Co-operation and Development), and IMF (International Monetary Fund), first had to go through *ETL* process and then made available at their respective linked dataspaces at *.270a.info with their own SPARQL endpoints.
- *Should be heterogeneous* – The original data from all of the sources are in three formats: propriety XML, CSV, and SDMX-ML. All of the data and metadata was composed of varying degree of quality and consistency.
- *Should contain substantial quantities of real world data* – All of the dataspaces contain both, data and metadata close to 800 million triples in total; consisting of dataset observations, concept schemes, codelists, interlinks, VoID, LODStats and other metadata. There are 68 million triples of observations across all datasets.
- *Meaning must be represented using Semantic Web technologies* – The statistical data is modeled using the RDF Data Cube vocabulary, which describes multi-dimensional statistical data using the SDMX-RDF as its statistical information model. The regression analysis data is expressed using a custom vocabulary for information like sample size, p-value, correlation value and method that was used, and adjusted R-squared for the independent and dependent variables.
- *Data must be manipulated/processed in interesting ways to derive useful information* – The observation data that was retrieved from two dataspaces was

used to conduct a regression analysis. The results of the analysis are stored in a separate RDF store and can be queried at a SPARQL endpoint. This information is stored for researchers and journalists so that they can discover information that's of interest to them.

- *Semantic information processing has to play a central role in achieving things that alternative technologies cannot do as well, or at all;* – Given that the statistical dimension concepts across linked dataspaces are interlinked, one can learn from a certain observation's dimension value, and enable the automation of cross-dataset queries in a uniform way with RDF and SPARQL.

Additional Desirable Features

- *The application provides an attractive and functional Web interface (for human users)* – Users only need to make three selections from a drop-down form: independent variable, dependent variable, and the reference period. A scatter plot with the regression line is shown along with the analysis data. Provenance information on the analysis is also provided for the users with the "Oh yeah?" link.
- *The application should be scalable* – The application uses a decent size of statistical linked data that is in the LOD cloud. The application can use more data sources provided that the data is modelled using RDF Data Cube vocabulary and passes integrity checks, and is available over a SPARQL endpoint.
- *Rigorous evaluations have taken place that demonstrate the benefits of semantic technologies, or validate the results obtained.* – Interlinking of concept identifiers from different dataspaces on the Web, and the availability of federated queries.
- *Novelty, in applying semantic technology to a domain or task that have not been considered before* – The application conducts federated queries to different statistical linked dataspaces, gathers the necessary information and proceeding with regression analysis. Storing the analysis results for future use makes it possible for interested parties to discover statistically interesting bits in a uniform way using Semantic Web technologies.
- *Functionality is different from or goes beyond pure information retrieval* – New data is always created based on user-selections.
- *The application has clear commercial potential and/or large existing user base* – The application can be used for teaching, support researchers and journalists to discover previously calculated analysis, displaying in external sites, in natural language searches.
- *There is a use of dynamic data, perhaps in combination with static information* – Data which was not previously analyzed is always retrieved in real-time from federated SPARQL endpoints.
- *The results should be as accurate as possible* – Federated query retrieves all required observation data, and the regression analysis is conducted by R.