Performing Statistical Methods on Linked Data

Benjamin Zapilko
GESIS – Leibniz Institute for the Social Sciences, Germany
benjamin.zapilko@gesis.org

Brigitte Mathiak
GESIS – Leibniz Institute for the Social Sciences, Germany
brigitte.mathiak@gesis.org

Abstract

In recent years, many government agencies have published statistical information as Linked Open Data (e.g. Eurostat, data.gov.uk). Yet, while there are a number of visualization tools, researchers use data for scientific statistical analysis to answer their research questions. Currently, they have to download the statistical data in a table-based format, in order to use their statistics software, unfortunately losing all the benefits Linked Data provides to them like interlinking with other data sets. In this paper, we present an approach specifically designed to help researchers to perform statistical analysis on Linked Data. By combining distributed sources with SPARQL, we are able to apply simple statistical calculations, such as linear regression and present the results to the user. Results of testing these calculations with heterogeneous data sources expose a wide range of typical issues on data integration which have to be aware of when working with heterogeneous statistical data.

Keywords: Linked Data; statistical methods; SPARQL; data integration

1. Introduction

Working with statistics is a bothersome task. First, the statistics have to be gathered from a variety of sources like governmental agencies, archives, libraries and researchers. Second, the data has to be made readable, filtered and cleaned. If using more than one data set, they have to be made comparable. Finally, diverse statistical functions have to be defined and can be executed according to desired analysis methods (Schnell, et al. 2005; King, et al. 1994; Kohler, et al. 2008).

Linked Open Data promises to improve that workflow by making data available at a click, directly usable and already linked in a meaningful way. Yet, while there are many Linked Open Data sets on the Web, it is still an open problem on how to actually use the data, as they do not yet come with an integrated statistics program. The technical problems involved with this task are many-fold (Zapilko, et al. 2011). The Linked Data sources tend to be distributed and have to be fetched in order to be processed. The data sets are typically heterogeneous on many levels. Examples range from different units, e.g. total number of unemployed people vs. percentage of unemployment, different formats, e.g. "11-4-2011" vs. "2011", different weights, sample points, populations, etc. Even if the data sets typically contain the necessary metadata needed for integration, it is often not machine-readable and standardized.

Our approach for this problem is two-fold. First, we implement the statistic functions to be as flexible as possible about the differences in the data, e.g. ignoring differences in format and minimizing the needed data to what is mathematically needed. Second, we use the semantic query language SPARQL to handle distributed data sources and to specify the additional information needed. Also the computational complexity has to be kept in check, as we strive to achieve reasonable response times for the calculations.

The paper is structured as follows: in section 2 we present related work both on Linked Data and on other approaches to consume statistical data on the web. We introduce our method for performing statistical calculations on Linked Data in section 3 and present two example

calculations which have already been implemented. Section 4 covers the results of our implementation as well as first findings, which come along when calculating with Linked Data. We conclude in section 5 and present an outlook on possible future work.

2. Related Work

The main intention behind the idea of Linked Open Data (Bizer, et al. 2009) is a method to expose, share and connect freely available data on the web using Semantic Web standards. The publication of data as Linked Open Data from a technical perspective (Bizer, et al. 2007) is based on common standards and techniques, which have been developed for years and are established worldwide as fundamental formats and interfaces for publishing data on the web, e.g. URIs, HTTP and RDF. With the standardization of SPARQL[1] in 2008, a common technology for querying RDF data has been established.

The paradigm of Linked Open Data was well received in the Semantic Web community and has encouraged organizations to publish data. Work on standardization and discussions about the openness of data generate new impulses. As the term of data is of a very wide range the idea of Linked Open Data addresses many diverse recipients. In recent years a lot statistics and other numerical data has been published as Linked Data by e.g. government agencies, statistical offices or research organizations. There are also activities in the publishing and linking of library data (Coyle, 2010; Dunsire, et al. 2010; Vatant, 2010), like the efforts from the Library Linked Data Incubator Group[2] of the W3C. This paper focuses on numerical Linked Data like statistics and survey data.

Extending SPARQL for new domains is not a new endeavor. An overview of proposed and implemented extensions can be found at the corresponding page in the W3C-Wiki[3]. The most relevant extensions include functions on database management (like inserting or updating RDF data), extended functions (e.g. a free-text search) and data calculations. In context of statistical methods the data calculations provide a good basis. Different frameworks and tools which are using SPARQL have already implemented aggregate functions like MAX, MIN, AVG or SUM. Some of these extensions found recognition and are planned to be included in the next revision of the language, SPARQL 1.1[4]. More complex statistical methods are still missing in the current plans. The SPARQL client[5] for the R Project[6] marks an alternative approach for performing statistical analysis on Linked Data. It provides the possibility to load a SPARQL result into the open source statistic tool R. This client marks the connection between Linked Data and existing and established statistic tools.

Data providers for statistical or survey data are very keen on offering the possibility to browse, analyze and download their data. Examples are ZACAT[7] by GESIS[8] – Leibniz Institute for the Social Sciences or SOEPinfo[9] by the Research Data Centre of the SOEP[10] (Socio-Economic Panel Study). Both portals offer a wide range of tools for processing, analyzing and visualizing data as well as providing different export formats. But both are restricted to the data holdings of their particular organizations. There are no connection points to other external data sources.

[1] http://www.w3.org/TR/rdf-sparql-query/
[2] http://www.w3.org/2005/Incubator/lld/
[3] SPARQL Extensions, http://esw.w3.org/SPARQL/Extensions
[4] SPARQL 1.1, http://www.w3.org/TR/sparql11-query/
[5] http://cran.r-project.org/web/packages/SPARQL/
[6] The R Project for Statistical Computing, http://www.r-project.org/
[7] ZACAT – GESIS Online Study Catalogue, http://zacat.gesis.org/
[8] http://www.gesis.org/
[9] SOEPinfo, http://panel.gsoep.de/soepinfo2009/
[10] SOEP – German Socio-Economic Panel Study, http://www.diw.de/soep

A web-based application without data boundaries is GraphPad QuickCalcs[11], a collection of free online calculators for different analysis purposes. It enables statistical calculations based on data, resp. numbers entered by the user manually. However, calculations are only possible on single numbers and not on complete data, which separates the data from its context and meaning. A combination of different data sources seems to be possible, but a lot of manually work is left to the user.

3. Method

The proposed method in this paper provides the possibility to perform statistical methods on numerical Linked Data on the web. Data is retrieved via SPARQL queries from different data sources. The results of the queries are stored in arrays and statistical calculations are performed on the data. Along with the query mechanism two typical statistical calculations have been implemented: (i) variance and (ii) linear regression. The data sources in the examples are chosen according to realistic research questions raised in the Social Sciences, as was confirmed by domain experts. The following implementation can be accessed online[12], where SPARQL queries of both calculations can easily be edited and executed with other available data sets.

3.1. General Approach

The general method is implemented using the Jena Framework[13] and consists of three steps: (i) performing combined SPARQL queries, (ii) storing SPARQL results in arrays and (iii) performing statistical calculations on values from the arrays.

Step 1: SPARQL as an established query language for RDF is used to retrieve data from Linked Data sources. According to this method and because of the automatism of the calculations in step 3 it is important that the query is formulated in a way that only numerical values are received, which are obviously needed for being used in a calculation. A standard query on distributed data uses the `UNION` operator to merge the results.

Step 2: The retrieved SPARQL query results are stored in arrays alongside with their label. This is used to create more meaningful output.

Step 3: The statistical calculations are performed on the results. For now, these calculations have been implemented in JAVA as a separate step. The alternative would have been to expand the SPARQL syntax with statistical operations and methods. However, the approach is computationally too expensive, as we will discuss in section 4.1.

3.2. Data Sources

The test calculations involve two data sources: (i) official European statistics from Eurostat[14], the statistical office of the European Union, whose data holdings have been published as Linked Data[15] and (ii) an excerpt of cumulated survey data from the German General Social Survey ALLBUS[16], which collects up-to-date data on attitudes, behavior, and social structure in Germany and is archived at GESIS – Leibniz Institute for the Social Sciences. Due to data privacy restrictions a special processed version of a subset of ALLBUS is used, which includes only a small excerpt of the variables surveyed in the original study. Also, it is aggregated from the individual level of single participants in order to be comparable to statistics. The version used for this paper has been created for technical feasibility experiments only, but is available for computational research intentions on inquiry.

[11] GraphPad QuickCalcs, http://www.graphpad.com/quickcalcs/index.cfm
[12] http://lod.gesis.org/gesis-lod-pilot
[13] http://openjena.org/
[14] http://eurostat.ec.europa.eu
[15] http://estatwrap.ontologycentral.com/
[16] http://www.gesis.org/en/allbus/

Both data sets have been exposed as Linked Data using the RDF Data Cube vocabulary[17], a vocabulary for modeling statistical data along with its metadata as Linked Data. It is an RDF representation of SDMX[18], an established exchange format for statistical data. Statistical data is described and organized in the Data Cube vocabulary as follows:

> "A statistical data set comprises a collection of observations made at some points across some logical space. The collection can be characterized by a set of dimensions that define what the observation applies to (e.g. time, area, population) along with metadata describing what has been measured (e.g. economic activity), how it was measured and how the observations are expressed (e.g. units, multipliers, status)." (The RDF Data Cube vocabulary)

One observation of the mentioned ALLBUS data set looks like follows, when exposed as RDF using the Data Cube vocabulary:

```
<qb:Observation>
  <qb:dataset rdf:resource="./ZA4570v590.rdf#ds"/>
  <dcterms:date>2004</dcterms:date>
  <gesis:geo rdf:resource="./geo.rdf#00"/>
  <gesis:variable rdf:resource="./variable.rdf#v590_2"/>
  <sdmx-measure:obsValue>204</sdmx-measure:obsValue>
</qb:Observation>
```

The excerpt above describes an observation done in "2004" in a geographical area with the regional code "00", which can be resolved as "Germany". The observation belongs to "ZA4570v590", which is resolvable as the variable "590" with the label "Afraid of Unemployment, Employees" from the ALLBUS data set "ZA4570". Due to the fact that survey data often organizes possible answers to a surveyed question in a scale, the observation value refers to the scale value "v590_2", which is resolved as "Yes, afraid of becoming unemployed". The example above depicts that 204 participants of the survey have been afraid of becoming unemployed in the observation.

3.3. Implementing the calculations of variance and linear regression

In the following section two example calculations are presented. Both calculations reveal typical problems and challenges when working with data, which are addressed in section 4 in more detail.

Variance: The variance of a set of numbers determines how far they are spread out from each other and how far they lie apart from their mean. Given the variance of an observed set of values the standard deviation can easily be computed, which describes the variability and diversity of a set of numbers. This is often used to measure the confidence in statistical conclusions. Examining the variance and the standard deviation is one of the most basic analyses in statistics. The following example calculates both values for statistics of unemployment from Eurostat with the SPARQL query[19] below:

```
SELECT ?time ?value ?cat
FROM <http://estatwrap.ontologycentral.com/data/lfst_r_lfu3pers>
FROM <http://estatwrap.ontologycentral.com/dic/geo>
WHERE {
  ?s qb:dataset <http://estatwrap.ontologycentral.com/id/lfst_r_lfu3pers#ds> .
```

[17] http://publishing-statistical-data.googlecode.com/svn/trunk/specs/src/main/html/cube.html
[18] Statistical Data and Metadata Exchange (SDMX), http://www.sdmx.org/
[19] Because of their length, the SPARQL queries in this paper are cut off their PREFIX statements. The full queries can be seen and executed at http://lod.gesis.org/gesis-lod-pilot/stat/variance.jsp and http://lod.gesis.org/gesis-lod-pilot/stat/regression.jsp

```
  <http://estatwrap.ontologycentral.com/id/lfst_r_lfu3pers#ds> rdfs:label ?cat
.
  ?s dcterms:date ?time .
  ?s eus:geo <http://estatwrap.ontologycentral.com/dic/geo#DE> .
  ?s sdmx-measure:obsValue ?value .
}
ORDER BY ?time
```

This query retrieves the total values for "Unemployment by sex and age, at NUTS levels 1, 2 and 3 (1000)" from the corresponding URI for the complete available range of time and for the geographical area of "Germany". The queried data set contains the values for sexes and different groups of ages for different so-called NUTS levels. The Nomenclature of Territorial Units for Statistics (NUTS)[20] denotes a common standard for referencing regional areas in the member states of the EU, where the three levels stand for different levels of subdivisions of the countries. For Germany, for example, NUTS level 1 marks the federal states, level 2 government regions and level 3 the smallest subdivision, the districts. The attribute "(1000)" in the indicator label denotes that the numbers are indicated in thousands. The calculation of variance and standard deviation delivers the result depicted in figure 1.

```
Result of the variance calculation of 'Unemployment by sex and age, at NUTS
levels 1, 2 and 3 (1000)' including the following values:
[2002, 153.2]
[2003, 170.0]
[2004, 207.3]
[2005, 311.4]
[2006, 279.7]
[2007, 251.3]
[2008, 226.7]
[2009, 218.7]


Variance: 2455.3735937499996
Standard Deviation: 49.55172644570519
```

FIG. 1. Result of the calculation of variance and standard deviation.

The calculations are executed on the retrieved values from the queried data set. They range from the unemployment in 2002, which has been 153.2 in thousands to the unemployment in 2009, which has been 218.7 in thousands. For researchers, the standard deviation of observed data is used to detect statistical significances. This way, random errors and variations in measurements can be distinguished from casual variation.

Linear Regression: Linear regression is one of the many types of regression analysis, which models the relationship between a scalar variable y (the so-called "dependent" variable) and one or more differing and assumed independent variables X_i. A correlation between X_i and y is supposed. This kind of regression analysis is often performed when examining predictions or forecasts based on an observed data set of y and X. Given multiple X values the strength or grade of relation between a single variable X_j and y can be detected. The example implementation of linear regression covers the research question, whether there is an impact of the unemployment in Germany on the fear of losing the job. The first indicator "Unemployment by sex and age, at NUTS levels 1, 2 and 3 (1000)" serves as the independent variable X in the example and is retrieved from Eurostat. The latter variable "Afraid of Unemployment, Employees" is taken from the ALLBUS and marks the dependent variable y, which is supposed to be influenced by X. The corresponding SPARQL query looks as follows:

[20] http://epp.eurostat.ec.europa.eu/portal/page/portal/nuts_nomenclature/introduction

```
SELECT ?time ?value ?cat
FROM <http://lod.gesis.org/lodpilot/ALLBUS/ZA4570v590.rdf>
FROM <http://lod.gesis.org/lodpilot/ALLBUS/variable.rdf>
FROM <http://estatwrap.ontologycentral.com/data/lfst_r_lfu3pers>
FROM <http://estatwrap.ontologycentral.com/dic/geo>
WHERE { {
  ?s qb:dataset <http://estatwrap.ontologycentral.com/id/lfst_r_lfu3pers#ds> .
  <http://estatwrap.ontologycentral.com/id/lfst_r_lfu3pers#ds> rdfs:label ?cat
.
  ?s dcterms:date ?time .
  ?s eus:geo <http://estatwrap.ontologycentral.com/dic/geo#DE> .
  ?s sdmx-measure:obsValue ?value .
  }
UNION {
  ?s qb:dataset <http://lod.gesis.org/lodpilot/ALLBUS/ZA4570v590.rdf#ds> .
  ?s qb:dataset ?cat .
  ?s dcterms:date ?time .
  ?s gesis:variable <http://lod.gesis.org/lodpilot/ALLBUS/variable.rdf#v590_2>
.
  ?s sdmx-measure:obsValue ?value .
  }
  FILTER (?time = "2004" || ?time = "2006" || ?time = "2008")
}
ORDER BY ?time
```

The query above retrieves again the total numbers for the "Unemployment by sex and age, at NUTS levels 1, 2 and 3 (1000)" for the geographical area "Germany" and for the three observation points in "2004", "2006" and "2008". It also queries the variable "Afraid of Unemployment, Employees", for the same observation points and spatial coverage. Due to different possible scale values for this variable, i.e. "No", "Yes, afraid of becoming unemployed" or "Yes, afraid of having to change my job", which is reasoned in the survey design and the used scale for possible answers, only the values of "Yes, afraid of becoming unemployed" are chosen for a combined calculation with the unemployment statistics. This secures a more meaningful analysis with the given data. Since precise observation pairs are required for linear regression, they are stated by a `FILTER` operator in the query. In the example "2004", "2006" and "2008" are chosen, because there is only overlap for these three observations in both queried data sets.

The calculation of linear regression delivers the following output (see figure 2) including the labels of the queried data sets and the values of each observation regarding the chosen points of time, e.g. the unemployment in 2004, which has been 207.3 in thousands and the number of participants of the survey, who answered that they would have been afraid of becoming unemployed in 2004. That has been 204 participants. After that, correlation coefficient, linear coefficient of determination and the regression line are presented as results.

```
Result of the linear regression between 'Unemployment by sex and age, at NUTS
levels 1, 2 and 3 (1000)' and 'FURCHT: STELLUNGSVERLUST, ARBEITNEHMER' for the
following observation points:
[2004, 207.3, 204]
[2006, 279.7, 255]
[2008, 226.7, 160]


Correlation Coefficient r: 0.7365172451879961
Linear Coefficient of Determination r2: 0.5424576524593148
Regression Line: y= -15.946285326424459 * 0.9343405576282379x
```

FIG. 2. Result of the calculation of linear regression.

The results above indicate whether there is a correlation between the values of unemployment and those of the fear of losing the job, which could be interpreted by domain experts that there are relationships between e.g. an increasing unemployment and an increasing afraid of job loss. Values for the correlation coefficient and the linear coefficient are typically in the range between -1 and 1, where 0 means that there is no correlation. Although the three observation points in the example are too few for a scientifically qualified statement, the result suggest that there might be a correlation between fear of unemployment and actual unemployment, as one would expect. For a scientifically qualified statement of such a correlation more observation points are necessary.

4. Results and Findings

In the following section we present observed results during the implementation of our method and discuss problems which occur when calculating with numerical Linked Data. Some of the covered issues are typical for processing Linked Data (e.g. different modeling of data) others derive from the domain of data analysis and statistics (e.g. different samples and units).

4.1. Technical Issues

Due to performance reasons, the SPARQL queries in the presented method are sent to one central SPARQL endpoint (see figure 3 for an overview of the implemented architecture). The retrieved data is then used for calculations. The runtime of the variance calculation takes 6.1 s from sending the SPARQL query to the delivery of the result. The linear regression takes 8.2 s in the same period. Of course, these results are not comparable to queries sent to external SPARQL endpoints in the web. In that case, the runtimes for network traffic would have to be added. While this seems like a rather long time to wait, the measured runtimes are much better than performing the implemented statistical calculations non-automatically. Especially the prior conversion of data (e.g. from csv, pdf or html) to formats suitable for statistic tools takes much more time when done manually in comparison to process Linked Data directly.

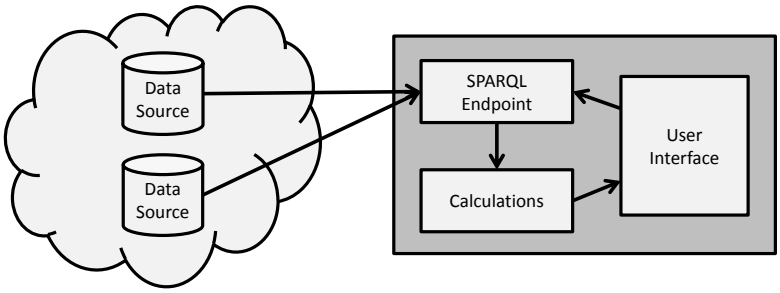

FIG. 3. Overview of the implemented architecture.

Both of the example calculations have been implemented in JAVA. The syntax of SPARQL has not been expanded with mathematical operations, because the resulted queries would be highly complex and expensive. Both factors lower the performance. Another reason for separating SPARQL query and statistical calculations from each other is the possible use of statistic tools like the R Project for Statistical Computing. A lot of statistical methods are highly complex and it is highly beneficial to reuse already existing packages from lively communities. A plugin for the R Project, which can execute SPARQL queries and process their results in the statistic tool, has already been published.

A very important aspect is the structure of the retrieved SPARQL result set and therefore the formulation of the SPARQL query. To leave the implemented calculations as independent and flexible as possible, the queries and results have to be in a very specific form, in which they can be processed further. For the proposed method it is important that there are only numerical values

and their corresponding label in the result set. For additional metadata information from the desired data sets an additional query should be executed, which results are not used for any calculation. In order to formulate precise SPARQL queries not only the structure of the data, but also precise dimensions (e.g. time) and indicators or variables have to be known or queried beforehand. Furthermore, time ranges, time intervals, geographical regions and areas, etc. have to be known for their use in later statistical analysis. This information is not added automatically.

4.2. Data Integration

The modeling of Linked Data remains to be one of the most challenging issues. Although there are massive efforts in developing vocabularies for describing e.g. statistical data (SCOVO (Hausenblas, et al. 2009), RDF Data Cube vocabulary) or survey data (Bosch, et al. 2011), data is modeled differently, due to the nature of the data sources and their structure and domain. Of course, SPARQL queries can be adjusted according to the desired data set, but this complicates all automatically-motivated approaches in processing the data in further implementations. For the implemented linear regression observation pairs of at least two different data sets have to be identified and extracted. The dimension of an observation is most commonly temporal. But the problem thereby is that in some data sets time can be denoted e.g. as "2011" or as "2011-04-27", which can also refer to different time intervals like annual or quarterly observations. Time intervals of observations should be included into the Linked Data representation. The implemented statistical calculation has to be flexible enough to interpret these values correctly. The problem increases when the SPARQL result delivers values as be seen in figure 4, where for some observation points no precise pairs are identifiable.

```
[time, FURCHT: STELLUNGSVERLUST, ARBEITNEHMER, Unemployment by sex and age, at
NUTS levels 1, 2 and 3 (1000)]
[1980, 42, null]
[1991, 344, null]
[1992, 213, null]
[1994, 209, null]
[1996, 260, null]
[1998, 254, null]
[2000, 201, null]
[2002, null, 153.2]
[2003, null, 170.0]
[2004, 204, 207.3]
[2005, null, 311.4]
[2006, 255, 279.7]
[2007, null, 251.3]
[2008, 160, 226.7]
[2009, null, 218.7]
```

FIG. 4. SPARQL result with overlapping time values.

The figure above depicts a SPARQL result, which omits values for some points of time for one of the retrieved indicators. Obvious observation pairs can be detected automatically, but it is not detectable, which ones from the other values could form interesting or relevant observation pairs. Because research interests can differ massively, no defaults can be made. For some scenarios it is common practice to form observation pairs from two adjacent years, like in electoral research when e.g. impact on election votes is analyzed. But, in other scenarios this can decrease the quality of the results. The user has to choose the observation pairs to be examined based on individual preferences. An UI between the second (data storing) and the third (calculating) step of our approach could mitigate the problem and is already planned.

Another problem which has been observed is the availability of data in different units. Calculating linear regression with values for unemployment in absolute numbers causes not only different mathematical results as if values in percentages were used. Obviously, values in percentages hold a different significance due to their projection onto the used sample. Therefore a

scientifically more qualified interpretation can be made. Figure 5 depicts the results of linear regression with the use of "Unemployment rates by sex and age, at NUTS levels 1, 2 and 3 (%)", where percentages are used for the calculation instead of absolute numbers (see section 3.3). In comparison to the results presented in figure 2, this time a higher correlation between both indicators is assumed.

```
Result of the linear regression between 'Unemployment rates by sex and age, at
NUTS levels 1, 2 and 3 (%)' and 'FURCHT: STELLUNGSVERLUST, ARBEITNEHMER' for
the following observation points:
[2004, 10.2, 204]
[2006, 12.9, 255]
[2008, 9.9, 160]


Correlation Coefficient r: 0.9248430777008266 Linear Coefficient of
Determination r2: 0.8553347183711373 Regression Line: y= -86.39560439560402 *
26.611721611721578x
```

FIG. 5. Result of the calculation of linear regression with percentages.

In order to detect units automatically, such information has to be included into the corresponding metadata, because it is important in order to interpret values correctly. The RDF Data Cube vocabulary already provides properties to describe units. The same problem can be observed, when trying to compare data based on different samples or populations (e.g. age groups, regional areas, etc.), which is not unusual in research. Special weightings have to be defined to make data with different samples comparable. Information about necessary weightings is typically available in the documentation for the data sets, but this information and the weightings themselves have to be exposed as Linked Data as well.

5. Future Work

In this paper we presented a method for performing statistical calculations directly on Linked Data. The major obstacle is still the lack of standardization in the data itself. This would much improve, when more data providers would decide to use standardized vocabularies and ontologies such as the RDF Data Cube vocabulary for statistical data. Also these vocabularies should put more emphasize on issues such as comparability of data, especially when different vocabularies are used. This is not addressed as a design goal for the ontology. Still, our method works reasonably well, as long as there is someone with knowledge of SPAQRL to support domain experts formulating their query.

To lower the usability threshold, it is planned to work on a user-friendly interface that allows the user to click together their query, including choice of data set, statistical method and a visualization of the results. From the technical point of view this future work includes the extension of SPARQL to a "Statistical SPARQL", which is capable of extended SPARQL queries that include additional parameters like `STAT` for interpreting the desired statistical method. A possible query could look like this:

```
SELECT ?time ?value
WHERE {
  ?s qb:dataset <http://estatwrap.ontologycentral.com/id/tps00001#ds> .
  ?s dcterms:date ?time .
  ?s eus:geo ?g .
  ?g rdfs:label "Germany (including  former GDR from 1991)"@en .
  ?s sdmx-measure:obsValue ?value .
  FILTER (?time > "2000" && ?time < "2011")
  STAT variance
}
```

Following the arguments given in section 4.1 it is not intended to expand the basic syntax of SPARQL beyond a simple parameter, e.g. via query rewriting (Correndo, et al. 2010). Instead, a connection of the SPAQRL result to open source statistic packages like the R Project seems to be a reasonable direction. Regarding the observed issues on data integration, the further development of "Statistical SPARQL" will form a set of assessments on the possible usage of data sets for statistical analysis. Such assumptions can derive from executed SPARQL queries on the data.

References

Bizer, Chris, Richard Cyganiak, Tom Heath. (2007). How to publish Linked Data on the Web. Retrieved April 29, 2011 from http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/.

Bizer, Chris, Tom Heath, Tim Berners-Lee. (2009). Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems (IJSWIS), Vol. 5(3), pp. 1-22.

Bosch, Thomas, Andias Wira-Alam, Brigitte Mathiak. (2011). Designing an Ontology for the Data Documentation Initiative. In: 8th Extended Semantic Web Conference 2011.

Correndo, Gianluca, Manuel Salvadores, Ian Millard, Hugh Glaser, Nigel Shadbolt. (2010). SPARQL query rewriting for implementing data integration over linked data. Proceedings of the 2010 EDBT/ICDT Workshops (EDBT '10).

Coyle, Karen. (2010). Understanding the Semantic Web: Bibliographic data and metadata. Library Technology Reports, 46(1). Chicago: American Library Association.

Dunsire, Gordon, Mirna Willer. (2010): Initiatives to make standard library metadata models and structures available to the Semantic Web. Proceedings of the IFLA World Library and Information Congress (IFLA'10), Gothenborg.

Hausenblas, Michael, Wolfgang Halb, Yves Raimond, Lee Feigenbaum, Lee, Danny Ayers. (2009). SCOVO: Using Statistics on the Web of Data. Proceedings of the 6th European Semantic Web Conference: Research and Applications. Heraklion, Crete, Greece.

King, Gary, Robert O. Keohane, Sidney Verba. (1994). Designing Social Inquiry: Scientific Inference in Qualitative Research. Princeton University Press. 1994.

Kohler, Ulrich, Frauke Kreuter. (2008). Datenanalyse mit STATA. Oldenbourg. 2008.

Schnell, Rainer, Paul B. Hill, Elke Esser. (2005). Methoden der empirischen Sozialforschung. Oldenbourg. 2005.

Vatant, Bernard. (2010). Porting library vocabularies to the Semantic Web, and back: A win-win round trip, Proceedings of the IFLA World Library and Information Congress (IFLA'10), Gothenborg, August 2010.

Zapilko, Benjamin, Andreas Harth, Brigitte Mathiak. (2011): Enriching and Analysing Statistics with Linked Open Data. Proceedings of the NTTS Conference 2011 (New Techniques and Technologies for Statistics).