Enriching and Analysing Statistics with Linked Open Data

Benjamin Zapilko[1], Andreas Harth[2], Brigitte Mathiak[1]
[1]GESIS – Leibniz Institute for the Social Sciences, Bonn, Germany, e-mail: benjamin.zapilko@gesis.org, brigitte.mathiak@gesis.org
[2]Institute AIFB, Karlsruhe Institute of Technology, Karlsruhe, Germany, e-mail: harth@kit.edu

Abstract

Scientists often need to analyse heterogeneous and distributed datasets for secondary purposes, for example to verify prior assumptions or to detect correlations between datasets. In most cases the time-consuming effort for converting the desired data into formats for statistical tools is not justified by the information need itself. With the emerging paradigm of Linked Open Data, existing information infrastructures for processing research data can be opened up for the use of external data published at sources on the web. Thus, scientists can conduct statistical analyses on a broader range of available data sets but can also save time-consuming data conversion because of the use of widely established standards and interfaces. This paper discusses issues and challenges when using Linked Open Data for analysing and enriching statistics illustrated by a real life use case scenario.

Keywords: Data Integration, Linked Open Data, Statistics

1. Introduction

A tenet of research in the social sciences is the study of social phenomena via analysing quantitative evidence. Scientists typically need to perform major and complex analyses on statistical data, and with the ever-increasing amount of available digital data, new approaches for data analysis become possible. As part of their main tasks, researchers often require tedious secondary examinations on heterogeneous and distributed datasets, for example to quickly verify prior or referenced assumptions or to detect correlations between two or more datasets (Schnell, et al. 2005; King, et al. 1994; Kohler, et al. 2008). A lot of tools already exist which support researchers in processing and analysing their data, for example SPSS[1] or STATA[2]. Sizeable amounts of data used by scientists are attainable through the web, however, the data is published in a large variety of data formats. To process and analyse data, scientists often need to convert the raw data to particular data formats and integrate data from multiple sources. In general, data

[1] SPSS Statistics, IBM, http://www.spss.com/de/
[2] STATA Data Analysis and Statistical Software, http://www.stata.com

conversion and integration is not a technical barrier, but the effort spent for conversion is a nuisance, especially for necessary but tedious routine tasks or in cases where the expected research gain is minor.

As much research data is being attainable through the web, advanced web technologies may facilitate the access to and integration of such data. In 2006 Tim Berners-Lee articulated four principles for publishing data on the web as so-called Linked Data. Linked Data provides a technical basis for exposing, sharing and linking data on the web, based on the established web architecture comprising standardised formats and interfaces. Spurred by the Linking Open Data project[3] – which aims at making Linked Data available under liberal copyright licenses – a large amount of datasets of scientific interest have been published, e.g. life science data, governmental data and official statistical data covering a broad range of domains and countries. The popularity of the Linked Open Data idea raises expectations that more data providers will publish their data in standardised ways. For researchers, Linked Open Data holds the possibility to gain access to a lot of data which has not been available or easily accessible in the past. To conduct scientific analyses, Linked Data still has to be converted to the formats of the tools scientists are using. However, with data accessible via a single standardised interface, time-consuming transformation processes could be automated and analysis tasks could be performed more easily and thus at lower cost.

The information infrastructure for working with research data are often too strictly adjusted to data sources traditionally used or to specific domains or purposes. Considering the Linked Open Data movement, a method for performing mostly secondarily research tasks is needed, to allow for statistical queries and calculations on standardised datasets. If these standards would be of a broader range than those of for example SPSS or STATA for statistical data, additional datasets which lie outside the researchers' facilities could be used automatically for combined analyses. This would also expand the usage and easy-to-perform analyses of published data to other interested user groups, for instance to journalists.

In this paper, we propose a method which provides the capability of conducting such statistical analyses on Linked Data resources to support common tasks that researchers encounter when working with heterogeneous and distributed datasets. We give a brief overview on Linked Open Data in Section 2 and present a real-world use case from the domain of the social sciences in Section 3 where heterogeneous datasets are linked for further combined analyses. The technical implementation of the use case is presented in Section 4 and will cover four issues: (i) the publication of data as Linked Data, (ii) the integration and processing of multiple datasets, (iii) the processing of statistical methods and calculations on such datasets and (iv) the combined visualisation of multiple data sets in a line graph. In Section 5 we present related work in the field of statistical online analyses of data. We conclude in Section 6 with a discussion on observed problems occurred during the use case development and its implementation and give an outlook on future work.

[3] http://linkeddata.org/

2. Linked Open Data

The main intention behind the recent 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. Linked Data is based on the unique identification of each thing, such as metadata elements or certain entities. The representation of information about these things can be combined with connections to other relevant or associated things on the web. Tim Berners-Lee outlined four principles for Linked Data (Berners-Lee 2006):

1. Use URIs[4] to identify things.
2. Use HTTP[5] URIs so that these things can be referred to and looked up ("dereferenced") by people and user agents.
3. Provide useful information about the thing when its URI is dereferenced, using standard formats such as RDF[6].
4. Include links to other, related URIs in the exposed data to improve discovery of other related information on the web.

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. RDF is a graph-structured data format. With the standardisation of SPARQL[7] in 2008, a key technology for querying RDF data has been established. The later described use case consists partly of a processed version of the German General Social Survey ALLBUS[8] for North Rhine Westphalia, a RDF representation of this dataset looks as follows:

<http://lod.gesis.org/lod.pilot/ALLBUS/ZA4570agg.rdf>
rdfs:label "ZA4570 Cumulated ALLBUS / GGSS 1980-2008";
gesis:geo <http://lod.gesis.org/vocab#D-NRW>;
rdfs:seeAlso <http://www.gesis.org/dienstleistungen/daten/umfragedaten/allbus/>.

<http://lod.gesis.org/vocab#D-NRW>
owl:sameAs <http://dbpedia.org/resource/North_Rhine-Westphalia>.

RDF is structured like Subject-Predicate-Object sentences. The above example reads: The entry *D-NRW* in the *gesis* vocabulary is the *same as* the entry *North_Rhine-Westphalia* in DBpedia[9].
The paradigm of Linked Open Data was well received in the Semantic Web community and has encouraged more organisations to publish data. Research organisations, archives, universities and media corporations publish and link their data and participate in the

[4] Uniform Resource Identifier
[5] Hypertext Transfer Protocol
[6] Resource Description Framework, http://www.w3.org/RDF/
[7] SPARQL Protocol and RDF Query Language (SPARQL), http://www.w3.org/TR/rdf-sparql-protocol/
[8] http://www.gesis.org/en/services/data/survey-data/allbus/
[9] http://dbpedia.org/

movement. Work on standardisation and discussions about the openness of data generates new impulses, so even governments all over the world are beginning to deal with the publication of official government data in the web. This strengthens the transparency of governmental agencies as well as the collaboration with and the participation of citizens in diverse aspects.
The linkage of distributed and apparently unrelated data sources holds potentials for both data providers and users. Data providers are in the position to enrich their own data with external sources from the web which prove to be relevant for their users or at least to be interesting additional information complementary to their data. Developers can in turn use the published and linked data for inclusion in own tools and applications which can in turn be made available to users. Therefore the real advantages of Linked Open Data for end users, e.g. scientists, will be evident when there are enough tools and applications which use linked data sources in a practical and useful way.

3. Use Case

In this section, we present a use case scenario for enriching and analysing statistics based on Linked Data which covers typical research tasks of social scientists: the detection and analysis of correlations between heterogeneous data sources and further calculations on these data sources. Our use case revolves around the following data sources:

- The German General Social Survey ALLBUS, which collects up-to-date data on attitudes, behaviour, and social structure in Germany and is archived at GESIS – Leibniz Institute for the Social Sciences[10]. Due to data privacy restrictions we use a special processed version of a subset of ALLBUS. The used version in this paper has been explicitly created for technical feasibility experiments only.
- Election statistics from the German federal state North Rhine Westphalia provided by IT.NRW[11], the statistical office and IT service provider of the federal state.
- Official European statistics from Eurostat[12], the statistical office of the European Union, whose data holdings are already available as Linked Data[13].

Our research interests in the use case include the detection and analysis of possible correlations between these datasets, e.g. by examining dependencies between the subjective attitude towards the personal economic situation, official election results and unemployment rates. Due to the data sets of this use case, where most of the raw data determine frequencies of observation points according to different variables and dimension, performing regression analyses is a relevant method for examining correlations between a dependent variable and one or more independent variables. The possibility of performing the therefore needed statistical calculations on Linked Open Data resources provides the openness to include other freely available data sources from

[10] GESIS – Leibniz Institute for the Social Sciences, http://www.gesis.org/
[11] http://www.it.nrw.de/
[12] http://eurostat.ec.europa.eu
[13] http://estatwrap.ontologycentral.com/

the web. The process of data integration which determines the basis for such calculations can easier be implemented when built on standardised formats and interfaces, such as Linked Open Data has.
The process of analysing heterogeneous datasets and performing statistical calculations onto these data can be realised easily if all of the participating datasets are available in the same data formats and are on the same level of aggregation. If this is not the case, there are additional barriers to be aware of when linking e.g. individual data (survey data) with aggregated data (statistics). The source which is available on an individual level has to be aggregated to the aggregation level of the other data source.
Two further aspects have to be considered for such a multi-level analysis. First the raw data of the particular observed variables has to be weighted to be compliant to the statistics. The data can then be aggregated and at last, transformations have to be done on the results according to the research context they are used for. What sounds feasible from a technical point of view is not trivial from a researcher's perspective. Weightings and potential transformations vary according to the intended research query, which means that there exists a wide range of possible calculations and a certain degree of openness in the way of using survey data. Another issue which is often raised by scientists is that laymen can easily be led to false conclusions when not exactly knowing how to interpret the data and the weighted and transformed results. The problems between different aggregation levels determine a key challenge in integrating heterogeneous datasets for a combined analysis.
To avoid such difficulties, our use case is restricted to data sources which are available on the same level of aggregation, as the mentioned official statistics and the social survey data from ALLBUS, which has been pre-processed and aggregated by researchers prior to the technical implementation. On this basis we will address the following challenges concerning data integration:

- Consistent data modelling by the use of standardized formats and vocabularies defined for the representation of statistical data.
- Mapping of dimensions and measures to allow the processing of combined queries and calculations on the data as well as for combined visualisations.

The following section describes the technical implementation of the use case, where all three data sources are published and integrated as Linked Data and are included in a combined visualisation with the option of performing further statistical calculations on the data layer.

4. Technical Implementation

The technical implementation of our use case covers four aspects: (i) the publication of data as Linked Data, (ii) the integration and processing of multiple datasets, (iii) the processing of statistical methods and calculations on such datasets and (iv) the combined visualisation of multiple datasets in a line graph.

4.1 Publication of data as Linked Data

The first step of the technical implementation is the exposition of the data as Linked Data. The datasets are exposed as RDF and accessible via HTTP, which allows for querying and integrating the data sources with technologies commonly used in the area of Linked Data. Considering the representation of data as RDF it is important to be aware of re-using already existing and widely-used identifiers from ontologies, vocabularies and instances, because re-use preserves the compatibility to and enables integration with other published data sets.
Currently, there are a number of competing formats available and in development, which focus on the semantic modelling of statistical data, e.g. SCOVO (Hausenblas, et al. 2009), SDMX[14] as RDF (Cyganiak, et al. 2010) or the RDF Data Cube Vocabulary[15] which could be beneficially integrated and re-used. Especially the latter one allows the modelling of more complex (e.g. multi-level) data structures and sets in RDF and complex possibilities in linking to other data sources among the others. On the instance level, DBpedia is a common reference point for linking and thus for enriching data with additional metadata information, which can support data integration and processing. The most important issue when modelling data as RDF is the preservation of the meaning of the data. Metadata has to be transformed to RDF without losing any relevant information Otherwise the underlying raw data cannot be interpreted correctly which leads to research outcomes of questionable validity.
The data sets of our use case consist of a high complexity according to the number of observed dimensions and measures. Especially the ALLBUS data set holds hundreds of variables which determine questions which have been asked during surveys. Although belonging to one observation (one polled person) we decided to represent each variable which has been asked (e.g. the estimation of the personal economic situation) as a separate observation. This is justified by the used Data Cube Vocabulary which advises to split up multiple measures of one observation to separate observations according to better querying possibilities. One observation to a precise variable (the voting results for a German political party) from the IT.NRW data looks as follows:

qb:Observation
qb:dataset <./data?code=14111#ds>;
dcterms:date "2009-09-27";
geo <geo.rdf#051>;
partei <./parteien.rdf#CDU>;
sdmx-measure:obsValue "845318".

One of the most relevant steps when exposing data as Linked Open Data is to establish links between different data sets. According to our use case, these linkages can be of a different kind. They can be built on a more abstract level e.g. between metadata (the observed "universe" in dataset A is the same element as "country" in dataset B or "variable" in dataset A is the same element as "characteristic" in dataset B), but also on a more concrete level like "Germany" in dataset A is equal to "Federal Republic of

[14] Statistical Data and Metadata Exchange (SDMX), http://www.sdmx.org/
[15] http://publishing-statistical-data.googlecode.com/svn/trunk/specs/src/main/html/cube.html

Germany" in dataset B. These linkages between the datasets establish the basis for performing the statistical methods and calculations described in our use case.

4.2 Integration and processing of multiple datasets

Although publication and access to Linked Data resources is standardised and common vocabularies and formats are used, challenges remain when querying combined datasets. Combined SPARQL queries are very individual according to the attributes, dimensions and measures of the integrated data sets. For an integrated view on heterogeneous data, for each dimension, a valid link between the data sets has to be identified. Furthermore, for the processing of the data, the different measures in which the data has been surveyed, e.g. values in percentages versus absolute numbers or different temporal intervals, are important. To solve this problem, links to external sources have to support internal calculations on the raw data, e.g. the calculation from absolute numbers of election votes in particular cities to the turnout of voters. This is not only relevant for further calculations, but also for combined visualisations, for example the combined representation of different data sets in a line graph where the axes have to be aligned on each other.

4.3 Statistical methods on Linked Open Data

The calculations for weighting and transforming the data as well as the statistical methods applied afterwards (e.g., regression analysis) is performed on the level of the integrated data corpus. Since the use case consists of precise tasks and we do not claim to replace statistics tools, our prototype provides the possibility of performing basic secondary analyses based on a small set of implemented functions. For implementing more complex statistical calculations existing sources from the R Project for Statistical Computing[16] can be re-used. An alternative is the extension of the SPARQL query language by query rewriting in order to transform SPARQL queries to particular statistical functions. To allow for more comprehensive analyses, the system provides export capabilities to standard tools such as SPSS and STATA. Our scenario focuses on the processing of data of the same hierarchical (aggregation) levels. Please note that the data integration layer is virtual, e.g., the integration layer provides access to data which remains at their original source.
In a first version the technical implementation the possibility to perform simple regression analyses onto the data is provided. There will be an additional small set of functions implemented at a later stage. When performing those calculations it is important to be aware of different data measures as noted in section 4.2.

[16] http://www.r-project.org/

4.4 Visualising combined query results

Results of the combined queries can currently be visualised in a line chart, where each observation of a variable is presented. Figure 1 shows a visualisation of a combined query on data from IT.NRW and ALLBUS. The election results for the biggest German parties are depicted together with survey results on the subjective attitude towards the personal economic situation.

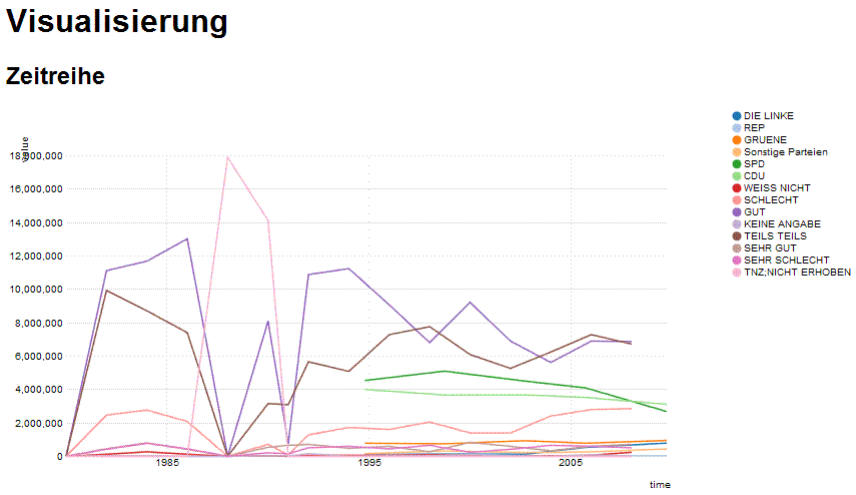

Figure 1: Visualisation of a combined query on different data sets.

In a later version the above described implementations will be included in a web-based prototype, where users can access Linked Data resources for visualising and analysing them for research purposes. Users will be able to choose the data sources to work with and which dimensions they are interested in. Additionally, users will also have the opportunity to weight and transform raw data according to individual criteria based on the research intention of the use case.

5. Related Work

Statistical analyses on Linked Data resources mark a research field which gains in importance recently, especially because more data providers and government agencies have been starting on publishing their data or plan activities in this direction. Since SPARQL marks the key query language for Linked Open Data resources, most activities centre on the syntax of the language which is working directly on the RDF data structures. The lively community of developers working with SPARQL detected a lot of areas for extending the language when processing real-life data. An overview of proposed and implemented extensions can be found at the corresponding page in the W3C-Wiki[17]. The most striking extensions include functions on database management and data calculations. Especially in context of statistical methods the latter ones are relevant. Different frameworks and tools which are using SPARQL have already implemented

[17] SPARQL Extensions, http://esw.w3.org/SPARQL/Extensions

such aggregate functions for selecting and returning compliant functions of multiple result values, e.g. functions like MAX, MIN, AVG or SUM are already possible. A lot of these extensions found recognition among others to be included in the new revision of the language, SPARQL 1.1[18]. But more complex statistical methods are still missing in the current plans.
Currently there exist online portals which provide the possibility to browse, analyse and download social science data like ZACAT[19] by GESIS or SOEPinfo[20] by the Research Data Centre of the SOEP[21] (Socio-Economic Panel Study). Both portals offer a wide range of tools for working with the data, analysing and visualizing it as well as providing different export formats. But both are restricted to the data holdings of their particular organisations. There seems to be no connection point to other external data sources except the user exports the data and loads it in an extra application for further processing and calculations.
A web-based application which is more open to the data which should be analysed is GraphPad QuickCalcs[22], a collection of free online calculators for a broad range of different purposes. It is offered by GraphPad Software[23] and provides the performing of statistical calculations based on data, resp. numbers entered by the user himself. However, calculations are only possible on single numbers and not on datasets as a whole, so that this tool 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.

6. Discussion and conclusion

The technical implementation of our use case revealed problems on data integration which have not been addressed extensively yet in the Linked Data community, although they depict typical challenges in the area of data processing. Establishing links between different and quite heterogeneous data sources has proved not to be trivial due to the highly complex structure of statistical data and survey data. Multiple dimensions and measures even within one observation have to be recognized. The representation of data in RDF and the resulted mappings between the data sets remark two of the keys to perform combined queries on different data sets which is the foundation for the processing of statistical calculations on these data sets. Another issue lies in the aggregation level of data. Individual data cannot apparently be combined with aggregated data. Thus, the desired data sets have to pre-processed, before data integration is possible.
The conversion of data and the use of Linked Data for statistical analyses could strengthen the overall participation in the movement of Linked Data, because standardisation and especially conversion work would not remain on few people and achieved results could be re-used in a broader, more dynamic way.

[18] SPARQL 1.1, http://www.w3.org/TR/sparql11-query/
[19] ZACAT – GESIS Online Study Catalogue, http://zacat.gesis.org/
[20] SOEPinfo, http://panel.gsoep.de/soepinfo2009/
[21] SOEP – German Socio-Economic Panel Study, http://www.diw.de/soep
[22] GraphPad QuickCalcs, http://www.graphpad.com/quickcalcs/index.cfm
[23] GraphPad Software, http://www.graphpad.com/

An important addition to our project will be the implementation of a graphical user interface and a visualization layer. The described implementations in Section 4 will be included in a web-based application. Since the links between the data sets can be quite complex, it is very important to provide the end user with an easy-to-use interface that masks the complexity. The same is true for the visualization. The web-based visualization tool Vizgr[24] offers interconnected diagrams and is already integrated with Linked Data. It could therefore be used as a standard means for representing the data, without having to abandon the already established links between it. An alternative for visualisations can be the Google Visualisation API[25], which offers a wide range of different graphical representations and easy-to-use interfaces.

In summary, we argue that Linked Data is a suitable means for facilitating scientists' access to data – statistical data in particular – and provide evidence supporting that claim via a prototypical implementation solving commonly required tasks in data analysis in the social sciences.

References

Berners-Lee, T. (2006). Linked Data - Design Issues. Retrieved 2010-10-25 from *http://www.w3.org/DesignIssues/LinkedData.html*

Bizer, C., Cyganiak, R., and Heath, T. (2007). How to publish Linked Data on the Web. Retrieved 2010-10-25 from *http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/*

Bizer, C., Heath, T., and Berners-Lee, T. (2009). Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems (IJSWIS), Vol. 5(3), Pages 1-22. DOI: 10.4018/jswis.2009081901

Cyganiak, R., Dollin, C., and Reynolds, D. (2010). Expressing Statistical Data in RDF with SDMX-RDF. Retrieved 2010-10-25 from *http://publishing-statistical-data.googlecode.com/svn/trunk/specs/src/main/html/index.html*

Harth, A., Hose, K., Karnstedt, M., Polleres, A., Sattler, K., and Umbrich, J. (2010). Data Summaries for On-Demand Queries over Linked Data. Proceedings of the 19th World Wide Web Conference (WWW2010). Pages 411-420.

Hausenblas, M., Halb, W., Raimond, Y., Feigenbaum, L., and Ayers, D. (2009). SCOVO: Using Statistics on the Web of Data. Proceedings of the 6th European Semantic Web Conference on the Semantic Web: Research and Applications (Heraklion, Crete, Greece, 2009). Pages 708-722.

King, G., Keohane, R., and Verba, S. (1994). *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton University Press. 1994.

Kohler, U., and Kreuter, F. (2008). *Datenanalyse mit STATA*. Oldenbourg. 2008.

Schnell, R., Hill, P., and Esser, E. (2005). *Methoden der empirischen Sozialforschung*. Oldenbourg. 2005.

[24] www.vizgr.com

[25] http://code.google.com/intl/de-DE/apis/charttools/index.html