Defining and Executing Assessment Tests on Linked Data for Statistical Analysis

Benjamin Zapilko and Brigitte Mathiak

GESIS – Leibniz Institute for the Social Sciences, Knowledge Technologies for the Social Sciences, Bonn, Germany
{benjamin.zapilko, brigitte.mathiak}@gesis.org

Abstract. Currently there is a strong trend for governmental agencies to publish statistical data as linked data (e.g. Eurostat, data.gov.uk). Unfortunately, these published datasets are still very diverse in their structure, making the analysis very complicated and technical. In this paper, we analyze datasets according to defined assessment tests and exploit both domain knowledge and the inherent semantic annotations. Therefore, we scan existing datasets for known patterns that signify e.g., typical numerical data blocks or potential temporal or geographical dimensions. Thus, linked data is made evaluable for a possible usage in standard statistical analysis tools. This allows researchers to use statistical data from diverse linked data sources for analysis with only a minimum of technical expertise used for integration of the data.

Keywords: Knowledge Discovery, Linked Data, Statistical Analysis, Data Assessment, Data Integration

1 Introduction

With the increase of available linked data in the web, the need for a meaningful usage grows that goes beyond the visualization of such data. Especially numerical and statistical data can be used for scientific analysis, e.g. by social scientists [1]. Due to the complex structure of most statistics, the effort to compare and integrate data from different data sources is huge. Linked data can provide more meta-information about the data itself than a traditional relational database, because of connections between distributed data via their URIs and aggregated information about a single element at its own URI.

The tools currently in use are mainly validation services for a general validation of linked data concerning data modeling and logical aspects. When linked data is used for scientific analysis further assessment tests have to be done. Of special interest in this regard is the comparability between heterogeneous datasets and the identification of common characteristics such as time and geographical region of the study. Also the identification of provenance and other circumstance of the study is relevant, such as base population, observation intervals and the nature of the sample used. All this information is relevant for scientists to make an educated decision about which data to use and how.

In this paper we present the definition of assessment tests in order to support researchers during their decision process on how relevant and useful a specific linked data resource might be for a scientific statistical analysis. We have implemented these checks in a web-based prototype application which is capable of extracting information from linked data resources that is relevant to support judgment of usage, like detecting observation values, different dimension, etc. Furthermore, multiple datasets can be analyzed together in order to detect possible similarities or conflicts between them. We have evaluated our implementation with existing linked datasets from Eurostat, ISTAT or data.gov.uk. The results provide not only information on potential usage of the data, but also on differences and difficulties in data modeling aspects.

The rest of the paper is structured as follows. In section 2 we present related work regarding tools and guidelines on publishing valid linked data as well as existing approaches on knowledge extraction and discovery with the focus on the use of data for statistical analysis. Section 3 describes the definition of assessments checks. In section 4 we present the implementation of these checks in a prototype. Section 5 discusses the results of the implementation including the evaluation with existing linked data statistics. In section 6 we conclude and present future work.

2 Related Work

A lot of activities can be identified regarding the validation and meaningful publication of linked data in the web. Beside textual guidelines [2,3] on how modeling and publishing linked data conceptionally and technically, there are also validation tools for RDF or OWL modeling [4,5] available as well as tools like Vapour [6], which has been especially developed for validating linked data according to the linked data principles [7]. An overview on additional tools and validators as well as support in fixing semantic web data can be found at the Pedantic Web Group [8]. There are several vocabularies specifically designed for modeling statistical data as linked data like SCOVO [9] or the RDF Data Cube vocabulary [10], which focuses on multidimensional data.

Recently, there have been approaches on extracting linked data for statistical analysis. The LiDDM system [11] allows the execution of statistical analysis on prior extracted and combined linked data. Our approach differs that we do not provide a tool for assisting manual integration by the user. We propose and implement general assessment tests to secure a possible usage of linked data for scientific statistical analysis based on an automatically integration. A different approach on using linked data for statistical analysis is followed by the SPARQL client [12] for the open source statistics package R Project[1]. This plugin provides the execution of SPARQL [13] queries in a statistical tool and the direct usage of the retrieved results.

[1] The R Project for Statistical Computing: http://www.r-project.org/

3 Linked Data Assessment Tests for Statistical Analysis

Before defining assessment tests general requirements regarding necessary data features have to be examined.

3.1 Basic Data Requirements

In order to use linked data for statistical analysis, the datasets have to fulfill some basic requirements. Obviously, such datasets have to contain observation values and at least one dimension (e.g. time) to which the values correspond to. Furthermore, the dataset should contain information about an indicator or a variable to describe what the data is about. These requirements are necessary, otherwise no meaningful analysis can be made. For analyzing multiple datasets, both need to have at least one matchable dimension, i.e. a dimension on which values of both datasets are comparable, e.g. time or geographical areas. In general, this does not imply that there have to be comparable observation values in this particular dimension. That is due to different purposes and methodologies for analyzing data.

The described requirements for our purpose are kept at a minimum, which is justified in the current state of quality and extent of statistical linked data. This is mostly a result of the extent of the openly published data source, which underlies the RDF representation of statistical data. Only few datasets hold a very detailed description about the data itself, e.g. its attributes, measures and dimensions, or information about acquisition and provenance. Especially the latter information and details about variance and bias in the data are highly relevant for judging the statistical quality and possible usage of the data. In this paper we do not address such issues, but will focus on the pragmatically relevant data items, i.e. observation values and dimensions.

3.2 Defining Assessment Tests

The following assessment tests are based on the data requirements described above. They focus on extracting information from dedicated linked data sources and on detecting matching possibilities between multiple datasets. This builds the basis for statistical analysis on integrated linked data. They can be divided into three stages: (A) identification of data items, (B) analysis of data characteristics and (C) data matching. Each of the stages is subdivided into smaller packages, which deal with specific aspects regarding the extraction of information from linked data and the analysis of using multiple datasets together. Table 1 provides an overview on the defined assessment tests.

Table 1. Overview on defined assessment tests.

Stage	Package	Description
A		Identification of Data Items
	A1	Observation Values
	A2	Indicators / Variables
	A3	Dimensions
B		Analysis of Data Characteristics
	B1	Observation Values
	B2	Dimensions
C		Data Matching
	C1	Detection of Similarities
	C2	Detection of Conflicts

A. Identification of Data Items. The first stage of the checks identifies necessary items in the dataset, which are required in order to perform statistical calculations on the data. Package A1 covers the identification of observation values, where the data is searched for included numbers and digits, which might be suitable as observation values. A2 identifies one or more indicator or variable labels which might fit to the detected values. Dimensions and their labels are extracted from the data in A3. Most statistical data should contain at least one temporal or geographical dimension, even if it is the same for all observation values, e.g. all data collected for a precise year or a precise country. Furthermore additional dimensions like populations, units, etc. are identified in A3. The results of stage A are one-dimensional datasets, each one for observation values as well as for temporal and geographical dimensions.

B. Analysis of Data Characteristics. If the checks in stage A are successful, more detailed examination of the detected values and information can be done. Package B1 analyses the characteristics of observation values, if they are correct and suitable. This is quite tricky as e.g. the number 2005 can be both a year number, but also the number of cities in a country. Currently, we have only shallow sanity checks, but plan to expand on this in the future. The dimensions are analyzed in B2 in detail. Date ranges, e.g. time intervals, time patterns (annually observations, monthly, quarterly, etc) are checked as well as geographical areas.

C. Data Matching. While the first two stages are performed on single datasets, stage C detects similarities and conflicts between at least two datasets. Package C1 is detecting similarities. Dimensions and their values are compared in order to detect matching points i.e. time points, geographical areas, etc. Such similarities have to be identified in at least one dimension in order to match them for a combined analysis. C2 is built on C1 and detects conflicts. Conflicts can arise through differing time ranges or intervals (observation frequencies, e.g. annually or monthly) or different geographical areas or levels of NUTS[2], which cannot be compared with each other. The Nomenclature of Territorial Units for Statistics (NUTS) 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

[2] Nomenclature of Territorial Units for Statistics (NUTS): http://epp.eurostat.ec.europa.eu/portal/page/portal/nuts_nomenclature/introduction

level 3 the smallest subdivision, the districts. As already mentioned, if there are no similarities between two datasets, it does not mean that there is no scientific analysis possible.

We are still working on differentiating between incomparable datasets such as data with only a time dimension vs. data with only a geographical dimension and theoretically comparable, but incompatible datasets, such as annually vs. monthly observation frequency. In that case the datasets could be made comparable, but it would require additional input from the user. The system could make suggestions such as leaving out certain data points or using averages.

4 Implementation

The defined assessment tests have been implemented in JAVA and are accessible in a first experimental web-based prototype[3]. The data is retrieved by an internal SPARQL query service[4], which loads data from the web that is addressed by FROM/FROM NAMED clauses in the query. For the implemented checks all triples from a desired source are queried.

Example SPARQL query.

```
PREFIX sdmx-measure: http://purl.org/linked-
data/sdmx/2009/measure#
PREFIX dcterms: http://purl.org/dc/terms/
PREFIX eus:
http://ontologycentral.com/2009/01/eurostat/ns#
PREFIX rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#
PREFIX qb: http://purl.org/linked-data/cube#
PREFIX rdfs: http://www.w3.org/2000/01/rdf-schema#
SELECT *
FROM http://estatwrap.ontologycentral.com/data/tps00001
FROM http://estatwrap.ontologycentral.com/data/teicp000
WHERE {
?s ?p ?o .
}
```

Additional datasets are retrieved by adding FROM statements to the query. The retrieved data is written into one table, where each column depicts a type from the json result. The rows of the table are then filled with the corresponding values. All of the defined tests are performed on this resulted table. In a first step, information is extracted from the table as much as possible. In a second step the extracted information is analyzed and compared according to the corresponding test.

[3] http://lod.gesis.org/gesis-lod-pilot/stat/structure.jsp
[4] http://qcrumb.com

5 Results and Discussion

The implementation has been tested with several linked data sources, which are all statistical data, but modeled and structured in different ways. Therefore the received results have been very different. In detail, diverse datasets from Eurostat[5], ISTAT[6] and data.gov.uk[7] as well as data from the 2000 US Census[8] has been included into the following evaluation.

5.1 Results of the Assessment Tests

Diverse datasets of the above mentioned data sources have been tested according to the defined assessment tests. The first results of the prototype mirror especially challenges in detecting and identifying characteristics about observation values and dimension. Table 2 presents the results of the stages A and B.

Table 2. Results of the Stages A and B.

Test Package	Eurostat	ISTAT	data.gov.uk	2000 US Census
A1	✔	✔	✔	✔
A2	✔	✔	✔	✔
A3	✔	✔	✘	✔
B1	✔	✔	✘	✔
B2	✔	✘	✘	✘

The table depicts that in case of data from Eurostat, all assessment tests could be performed successfully. Data from ISTAT could not pass package B2. While the temporal dimension was detected correctly (A3), the exact value for the date could not be identified, because it was stated in an URI. This is a general issue of the prototype and will be further discussed in section 5.2. The same issue has an impact on the results for the datasets of data.gov.uk. According to the very complex modeling of the diverse statistics and due to naming conventions of the used URIs, even dimensions could not always be detected as such. The results of data from the 2000 US Census prove that multiple indicators in one single dataset (in this case population and households) can be detected and allocated to their corresponding observation values.

The results of stage C, where two datasets have been analyzed together, in order to detect similarities and conflicts between them, could not be analyzed thoroughly, as they depend on the results of the first stages. Again, Eurostat received the best results. Between two datasets of Eurostat or ISTAT, both similarities and conflicts could be detected in the temporal dimension (e.g. differing time points or intervals) and in the geographical dimension (e.g. both datasets do not comprise the same geographical

[5] http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home/ via http://estatwrap.ontologycentral.com/
[6] http://www.istat.it/ available as linked data via http://www.linkedopendata.it/
[7] http://data.gov.uk/
[8] http://www.rdfabout.com/demo/census/

areas). In mixed tests with Eurostat and other data sources the conflicts were always detected correctly. But there is no counter example as the precise time point had not been detected in the other datasets. This is also the reason for any other conflict detection yet. Conflict detection regarding the geographical dimension suffers not only from the identification of the precise string. In fact, codes (e.g. ISO codes) can be detected from the URIs, but unfortunately different codelists are used by different data providers depending on the geographical scope of their data. In those cases either conflicts or nothing regarding the geographical dimension are detected. These observations and the results regarding a combination of datasets from data.gov.uk with others confirm that stages A and B have to be passed at a minimum in order to receive valuable results. The detailed results between the different datasets are presented in table 3.

Table 3. Results of the Stage C (1= no detection, 2= similarities detected, 3= conflicts detected, 4= similarities and conflicts detected[9]).

Test Package	Eurostat	ISTAT	data.gov.uk	US Census 2000
Eurostat	4	4	3	3
ISTAT	4	4	1	3
data.gov.uk	3	1	1	1
US Census 2000	3	3	1	4

5.2 General Observations and Discussion

During the implementation phase and according to the observed results, some general statements about the assessment of linked data for statistical analysis can be made.

The complexity and extent of modeling data is often very different. Some providers deliver additional information about units, populations, provenance etc, but this is not always the case. In most cases, this is not a problem of the RDF representation of the statistical data, it is often due to the original published data format, which often does not include such information directly.

All examined datasets are modeled according to the linked data principles. Therefore a lot of additional information about dimensions, etc. is encoded in URIs. Currently, the implementation does not query URIs in a dataset in order to retrieve more information. This hinders the full identification of data characteristics as intended in stage B and thus complicates the identification of similarities and conflicts in stage C.

Example: The date is stated as http://data.linkedopendata.it/istat/resource/code-time-2007 at ISTAT. This URI is detected as part of a temporal dimension, but the precise value of the date “2007” is not detected. Querying the URI would deliver the

[9] The detection of similarities and conflicts (4) means that either one of both could be detected in at least one dimension of the participating datasets. In the result table it is not differentiated, if there has been a detection of a conflict and a similarity in one dimension at the same time (e.g. the same annually frequency of observations in two datasets, but different date ranges).

precise string. The use of linked data principles depicts a step beyond traditional relational databases because of the possibility to get further relevant information about a precise element, which is not included in the original data.

Important for the detection of values and dimensions is the naming of the property and class types in a dataset. The more standardized vocabularies (e.g. Data Cube vocabulary, SCOVO, Dublin Core [14]) are used or the naming conventions of the URIs are generic and machine-interpretable, the easier is an automatic detection. A promising approach, especially for finding similarities in package C1, might be the use of link discovery tools (e.g. Silk [15], SERIMI [16]). Such tools might detect linkages between dimensions or precise values of them.

The results in 5.1 have unveiled the challenge that there a sometimes more than one dates in one single observation. For example, data about schools from data.gov.uk includes diverse dates like the opening date or the date of the last welfare visit among others. This complicates the automatically detection of temporal dimensions, because there might be not only one correct solution, because research interests are diverse.

In order to guess possible factors for making datasets comparable, it is necessary that the information on dimensions is very detailed, e.g. the existence of hierarchical structures in a dimension. For example, the structure of NUTS levels may be useful in order to aggregate data between different levels. This can be a solution, if one dataset is available on NUTS level 2 and the other one on NUTS level 1. From a scientific point of view this might be a loss of data quality, but it may support the researcher by getting an initial insight on the data.

6 Conclusion and Future Work

In this paper, we presented a definition of necessary assessment tests for using and integrating statistical linked data for scientific analysis. These tests have been implemented as a prototype and evaluated with a variety of statistical linked data sets. While the results are preliminary and can be further refined, the general approach seems to be a viable way to assist non-expert researchers in combining various data sources with only a small investment in domain specific knowledge.

For the future, we strive to exploit the typical characteristics of linked data more thoroughly. With the URIs additional knowledge can be obtained through the internet. We hope this will provide even more insight into the structure and properties of the data.

References

1. Gregory, A., Vardigan, M.: The Web of Linked Data. Realizing the Potential for the Social Sciences (2010), http://odaf.org/papers/201010_Gregory_Arofan_186.pdf
2. Tom Heath and Christian Bizer (2011) Linked Data: Evolving the Web into a Global Data Space (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136. Morgan & Claypool.

3. Leigh Dodds, Ian Davis: Linked Data Patterns. A pattern catalogue for modelling, publishing, and consuming Linked Data. http://patterns.dataincubator.org/book/
4. W3C RDF Validation Service. http://www.w3.org/RDF/Validator/
5. University of Manchester. OWL Validator. http://owl.cs.manchester.ac.uk/validator/
6. Vapour, a Linked Data validator http://vapour.sourceforge.net/
7. Berners-Lee, T. (2006): Design Issues: Linked Data. http://www.w3.org/DesignIssues/LinkedData.html
8. Pedantic Web Group http://pedantic-web.org/
9. Hausenblas, M., Halb, W., Raimond, Y., Feigenbaum, L., Ayers, D.: SCOVO: Using Statistics on the Web of Data. In: Proceedings of the 6th European Semantic Web Conference: Research and Applications (Heraklion, Crete, Greece) pp. 708--722 (2009)
10. Cyganiak, R., Reynolds, D., Tennison, J.: The RDF Data Cube vocabulary (2011), http://publishing-statistical-data.googlecode.com/svn/trunk/specs/src/main/html/cube.html
11. Venkata Narasimha Pavan Kappara, Ryutaro Ichise, O. P. Vyas: LiDDM: A Data Mining System for Linked Data. In Proceedings of the LDOW2011 (2011)
12. SPARQL client for R: http://cran.r-project.org/web/packages/SPARQL/
13. Prud'hommeaux, E., Seaborne, A.: SPARQL Query Language for RDF. W3C Recommendation (January 2008), http://www.w3.org/TR/rdf-sparql-query/
14. DCMI Metadata Terms, http://dublincore.org/documents/2010/10/11/dcmi-terms/
15. Robert Isele, Anja Jentzsch, Christian Bizer: Silk Server - Adding missing Links while consuming Linked Data. 1st International Workshop on Consuming Linked Data (COLD 2010), Shanghai, November 2010.
16. SERIMI: RDF Interlinking. https://github.com/samuraraujo/SERIMI-RDF-Interlinking