News Fact-checking: One Practical Application of Linked Statistics

Tatiana Tarasova

SpazioDati S.r.l. HQ Via del Brennero 52, 38122, Trento (TN)
tarasova@spaziodati.eu [*]

Abstract. This paper presents an example of practical application of Linked Statistics to the problem of checking facts in news articles. We consider a use case of verifying a news article that exploits statistical facts from the Italian National Institute of Statistics (ISTAT). To realise the use case, we publish a subset of ISTAT as Linked Data[1], thus, contributing to the presence of statistics on the Web of Data. We discuss how links from the news article to LinkedSTAT can enable creation of automated tools and services for fact-checking. We received a positive evaluation of our demo use case from ISTAT experts. With our work we hope to promote the value that could be obtained by publishing statistics as Linked Data by official statistical agencies and organisations.

Keywords: news fact-checking, Linked Statistics, LinkedSTAT

1 Introduction

Official statistics are highly structured and well described data about our everyday life, such as demographic, economic conditions, health, education, government spending, culture, sport, etc. Interest of the Semantic Web community to such a vital source of information is evidenced by plenty of statistical datasets that have been published as Linked Data by third parties: the World Factbook[2], Eurostat[3], Spanish public datasets [4] and statistics about immigration in Italy [5]. Development of the automated converting tool, the SDMX-to-RDF converter [8], enabled massive publication of statistics as Linked Data, among which are the Organisation for Economic Co-operation and Development[4] and the Food and Agriculture Organization of the United Nations[5].

The next objective for the Semantic Web community is to motivate appropriate agencies and organisations to start publishing their data as Linked Data. For this, benefits of the Linked Statistics have to be clarified to the statistics community. Related works [3,9,10,11] tackle this issue by demonstrating how Linked Data facilitates interlinking and analysis of heterogeneous statistics. [12] discusses how Linked Data can be used to solve a problem of concept drift detection in statistics. Our work is similar to

[*] We would like to thank Monica Scannapieco and Vincenzo Patruno from ISTAT for collaboration and valuable feedback to our work.

[1] http://linkedstat.spaziodati.eu

[2] http://wifo5-03.informatik.uni-mannheim.de/factbook/

[3] http://wifo5-03.informatik.uni-mannheim.de/eurostat/

[4] http://oecd.270a.info/

[5] http://fao.270a.info/

the latter effort in that it showcases benefits of Linked Statistics in application to a real-world problem, news fact-checking in journalism.

We introduce the problem of fact-checking and define a use case in Section 2. In Section 3 we briefly report on publication of the LinkedSTAT collection and discuss how it can be used to implement the demo use case in Section 3.1

2 News Fact-checking

Checking facts in journalism is the process of verifying accuracy of facts in publications, such as newspapers, magazines, and other periodical literature[13]. For example, consider the following excerpt[6] from a news article published by the Italian newspaper "L'Adige"[7]:

...employment rate of the resident population aged between 15 and 64 years was at 65.6 percent in Trentino and 71.5 percent in Alto Adige. ...

The author of the article analyses the problem of unemployment in the Italian region Trentino-Alto Adige, and supports his text with statistical facts taken from the Italian National Institute of Statistics (ISTAT). A fact-checker of this article has to verify these statistical facts by finding them the official website of ISTAT (`http://www.istat.it/`). Similar process must be carried out for all other facts presented in the article.

Fact-checking is a resource and time consuming task. Only big publications can afford employing fact-checkers, others at best provide sources that were used in their articles and put the burden of checking facts on the readers. Made by publishers or readers, manual fact-checking remains a tedious and error-prone process. Automated tolls to facilitate this process may be built on links from *facts* in the news article to the *facts* in the source statistics. But to realise such links, we need to have statistics published in a way that enables identification of and reference to single statistical observations.

3 LinkedSTAT

The LinkedSTAT project publishes a subset of ISTAT data[8] as Linked Data. The project is maintained by SpazioDati (`www.spaziodati.eu`) in collaboration with ISTAT. LinkedSTAT collection is composed of 233 datasets from 19 different statistical themes and is publicly available.

The workflow for publishing LinkedSTAT consists of four steps: data retrieval, data transformation, data storage and data publication[9].

Data was retrieved by querying the ISTAT SDMX Web Service[10] and transformed to RDF using the SDMX-to-RDF converter [8]. The vocabularies and the URI scheme

[6] The quoted text is the author's translation of *"...il tasso di occupazione della popolazione residente di età compresa tra 15 e 64 anni si è assestato al 65,6 per cento in Trentino e al 71,5 per cento in Alto Adige. ..."*

[7] The article is available at `http://altoadige.gelocal.it/bolzano/cronaca/2014/06/18/news/giovani-il-16-non-lavora-e-non-studia-1.9449341?ref=search`

[8] We consider ISTAT data available in the Statistical Data and Metadata Exchange (SDMX) format [7]

[9] All the queries and scripts produced during the LinkedSTAT project are available at `https://www.assembla.com/spaces/linked-istat/`.

[10] a dedicated test Web Service established by ISTAT for LinkedSTAT

used to encode ISTAT in RDF were dictated by the converter[11]. Among the vocabularies are the RDF Data Cube vocabulary[12], the PROV-O Ontology[13], SKOS[14] and SDMX-RDF[15]. Virtuoso OSE v6.01.3[16] was used to store the resulting RDF collection. All the LinkedSTAT URIs are dereferenceable both for machines (RDF) and people (HTML pages). The whole LinkedSTAT collection can be queried via the public SPARQL endpoint `http://linkedstat.spaziodati.eu/sparql`.

3.1 Fact-Checking with LinkedSTAT

In Section 2, we introduced a use case of checking facts in a news article. A fact-checker wants to verify the truth of the following statistical fact *... employment rate of the resident population aged between 15 and 64 years was at 65.6 percent in Trentino and 71.5 percent in Alto Adige. ...*. Below we describe how to link the article to LinkedSTAT.

The underlying statistical fact is described in the following dimensions: *territory*, *statistical indicator* and *age*. In LinkedSTAT these dimensions are encoded through the properties: `linked-istat-property:TIME_PERIOD`, `linked-istat-property:REF_AREA`, `linked-istat-property:AGGR` and `linked-istat-property:AGE` (respectively).

The following SPARQL query can be used to retrieve observations with the values of the dimensions given in the article:

```
SELECT DISTINCT * {
?obs linked-istat-property:REF_AREA
   <http://linkedstat.spaziodati.eu/code/1.1/CL_REFAREA/ITD2> .
?obs linked-istat-property:AGGR
   <http://linkedstat.spaziodati.eu/code/1.0/CL_LAV_INDICATOR/05> .
?obs linked-istat-property:AGE
   <http://linkedstat.spaziodati.eu/code/1.2/CL_ETA1/Y15-64> .}
```

The result of the query contains a list of relevant observations that can be attached to the statistical fact in the article to provide “proof” links to the official source. This is just a simple example of how to define links from the news article to LinkedSTAT manually. It is important to stress, that LinkedSTAT provides machine-readable description of ISTAT data. Such description can be used to create automated services for further refinement of the resulting list of observations. We can retrieve the datasets of the observations by adding the following triple pattern to our query `?obs qb:dataSet ?dataset .` and selecting `DISTINCT ?dataset`. Knowing the dataset, one can retrieve its structure, a set of dimensions, attributes and measures, and build a faceted search on top of them. With the help of such faceted search, a user can quicker select the required values of the dimensions and build a more precise query for a statistical fact to check in LinkedSTAT.

[11] More details can be found at `https://github.com/csarven/linked-sdmx/wiki`

[12] `http://www.w3.org/TR/vocab-data-cube/`

[13] `http://www.w3.org/TR/prov-o/`

[14] `http://www.w3.org/2009/08/skos-reference/skos.html`

[15] `http://publishing-statistical-data.googlecode.com/svn/trunk/specs/src/main/vocab/sdmx.ttl`

[16] `http://virtuoso.openlinksw.com/dataspace/doc/dav/wiki/Main/`

4 Conclusion

We presented an example of using Linked Statistics for fact-checking. We published SMDX ISTAT statistics as Linked Data and made the resulting LinkedSTAT collection publicly available, e.g., via a SPARQL endpoint. We demonstrated how to link a news article to LinkedSTAT and argued that such links can be used to build automated tools and services to facilitate the process of fact-checking.

We leave for future the challenge of finding a right set of dimension/value pairs for a given fact to construct queries for it. A possible solution could be running an entity extraction algorithm on a news article and extracting concepts from it. Such concepts can further be analysed and paired with the existing dimensions based on what values they can take at ISTAT.

Our demonstration use case received a positive evaluation from ISTAT as "an interesting use case that shows the peculiarity of linking Official statistics data". Our collaborators at ISTAT found promising new possibilities provided by LinkedSTAT and thought whether, as a Data Steward, ISTAT should publish as Linked Data other datasets (micro and macro) in order to facilitate data certification.

References

1. SDMX self-learning tutorial: Information Model Tab.8 - Attachment levels, p.13 `https://webgate.ec.europa.eu/fpfis/mwikis/sdmx/index.php/Self_Learning_Tutorial:_Information_Model`
2. Eurostat Self Learning Tutorial: SDMX-ML Messages `https://webgate.ec.europa.eu/fpfis/mwikis/sdmx/index.php/Self_Learning_Tutorial:_SDMX-ML_Messages`
3. Zapilko, B., Harth, A., Mathiak, B.: Enriching and Analysing Statistics with Linked Open Data. Proceedings of the NTTS Conference 2011.
4. De Leon, A., Saquicela, V., M. Vilches-Blazquez, L., Villazon-Terrazas, B., Priyatna, F. and Corcho, O.: Geographical linked data: a spanish use case. In I-SEMANTICS 6th International Conference on Semantic Systems (2010)
5. Webpage of the Linked Open Data Italia associatio. `http://www.linkedopendata.it/`
6. Halb, W., Raimond, Y. and Hausenblas M.: Building linked data for both humans and machines. WWW 2008 Workshop Linked Data on the Web LDOW2008 Beijing China (2007)
7. Sdmx initiative webpage. `http://sdmx.org/?page_id=11`
8. Capadisli, S., Auer, S. Ngonga Ngomo, A.-C., Linked SDMX Data, Semantic Web Journal (2013), `http://csarven.ca/linked-sdmx-data`
9. Zapilko, B., Mathiak, B.: Performing Statistical Methods on Linked Data, Proc. Int'l Conf. on Dublin Core and Metadata Applications (2011), `http://dcevents.dublincore.org/IntConf/dc-2011/paper/download/27/16`
10. Zapilko, B., Mathiak, B.: Defining and Executing Assessment Tests on Linked Data for Statistical Analysis, COLD, ISWC (2011), `http://iswc2011.semanticweb.org/fileadmin/iswc/Papers/Workshops/COLD/cold2011_submission_13.pdf`
11. Capadisli, S., Auer, S., Riedl., R.: Linked Statistical Data Analysis `http://csarven.ca/linked-statistical-data-analysis`
12. Meronno-Pennuela, A., Gueret, C., Hoekstra, R. and Schlobach, S.: Detecting and Reporting Extensional Concept Drift in Statistical Linked Data 1st International Workshop on Semantic Statistics ISWC 2013
13. Smith, S.H.: The Fact Checker's Bible: A Guide to Getting It Right. Knopf Doubleday Publishing Group (2007)