Improving Open Data Usability through Semantics

PhD research proposal

Sebastian Neumaier[*]

Vienna University of Economics and Business, Vienna, Austria
sebastian.neumaier@wu.ac.at

Abstract. With the success of Open Data a huge amount of tabular data become available that could potentially be mapped and linked into the Web of (Linked) Data. The use of semantic web technologies would then allow to explore related content and enhanced search functionalities across data portals. However, existing linkage and labeling approaches mainly rely on mappings of textual information to classes or properties in knowledge bases. In this work we outline methods to recover the semantics of tabular Open Data and to identify related content which allows a mapping and automated integration/categorization of Open Data resources and improves the overall usability and quality of Open Data.

1 Introduction

The Open Data movement has become a driver for publicly available data on the Web. More and more data – from governments, public institutions but also from the private sector – is made available online and is mainly published in so called Open Data portals. However, with the increasing number of resources, there are a number of concerns with regards to the quality of the data sources and the corresponding metadata, which compromise the searchability, discoverability and usability of resources [6, 13, 16].

In [2] Berners-Lee defines the quality of Linked Open Data by a 5-star rating: (1) data is available on the web with an open license, (2) available as machine-readable structured data (e.g., Excel instead of image scan of a table), (3) in a non-proprietary format (e.g., CSV), (4) the use of URIs to denote things, and (5) linked to other data in order to provide context. Yet, most of the data published on Open Data portals cannot be considered as Linked (Open) Data. In fact, the findings in [16] show that current Open Data sources mainly publish 1- to 3-star data (i.e., openly licensed data, available in machine-readable and non-proprietary formats): most of the resources in 82 monitored portals are CSV files (27%, i.e., *3-star* data), 12% are Excel tables (*2-star*), and 10% are PDF documents (*1-star*).[1]

The fact that a considerable large amount of Open Data is available in partially structured and tabular form motivates a further investigation of the potential of Semantic Web technologies to integrate and interlink this data, in order to improve the overall quality and usability, and to bring structure to these resources.

[*] Supervised by Axel Polleres and Jürgen Umbrich. Further thanks to Josiane Xavier Parreira for the recent successful cooperation.

[1] Note, that these numbers are based on the metadata descriptions of the datasets and therefore do not include diverging spellings or missing format descriptions (16%).

2 Problem Statement

The main idea of the here proposed work is the use of existing Semantic Web technologies to (partially) improve existing 3-star Open Data found on Open Data portals to 5-star data (according to [2]). In this thesis we focus primarily on tabular data which is currently the predominant format in Open Data portals.

Achieving this overall objective involves the following sub-problems:

(i) *Recovering the semantics of structured/tabular data sources:*
In contrast to highly-structured formats like RDF, tabular data lack a defined vocabulary, definite schemata, and semantic labels. In particular, Open Data tables are frequently created manually and the inherent structure can be hard to detect and understand (e.g., multiple header rows, numerical content, additional comment lines). We will investigate and propose mapping/labeling techniques tailored to the Open Data domain which allow us to semantically describe these tables.

(ii) *Cross-data portal classification and categorization of data sets:*
At the moment most of the (governmental) Open Data portals define and use their own taxonomies for their data. There is no commonly shared categorization and vocabulary [16].[2] Addressing problem (i) will support the development of a semantic classification schema for Open Data, in order to enable cross-data portal search and integration.

(iii) *Identifying related and relevant content:*
Current Open Data portals do not provide recommendation of related content (which potentially serve as candidates to automatically integrate and link resources). We will review existing relatedness measures for tabular data sources and, if possible, adopt or extend these methods. This will be supported by the outcome of the previously mentioned description and categorization approaches.

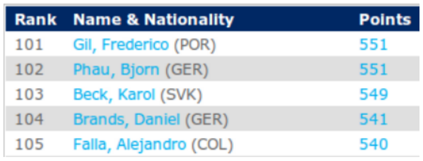
2.1 Limited applicability of existing methods

Existing research in the area of label annotation, exploration of relatedness and linkage of entities is not fully applicable to typical tables found on Open Data portals:

Relatedness of Tables. In [4] Das Sarma et al. consider and define the two most common types of related tables: *Entity Complement* (i.e., different selection over the same set of attributes) and *Schema Complement* (same set of entities for a different and yet semantically related set of attributes).

The authors tackled the problem of finding complementary entities by using named-entity recognition techniques [18, 3]. For instance, given Table 1 (from [4]), the column "`Name & Nationality`" gets the label *Tennis Player* assigned, based on the shared classes of the named-entities in the column. This information is then used to find columns with (different) entities of the same classes.

[2] With DCAT-APP (`https://joinup.ec.europa.eu/asset/dcat_application_profile`) there is an existing vocabulary and application profile for the use of metadata keys in Open Data portals, however, the use of tags and categories remains very heterogeneous.



Rank	Name & Nationality	Points
101	Gil, Frederico (POR)	551
102	Phau, Bjorn (GER)	551
103	Beck, Karol (SVK)	549
104	Brands, Daniel (GER)	541
105	Falla, Alejandro (COL)	540

Table 1: Example from [4].

Wohnungen in Wien, 2011				
Dwellings in Vienna, 2011				
NUTS1	NUTS2	NUTS3	DISTRICT_CODE	WHG_TOTAL
AT1	AT13	AT130	90101	3004
AT1	AT13	AT130	90102	1049
AT1	AT13	AT130	90103	1389
AT1	AT13	...	...	...

Table 2: Table on `data.gv.at`.

Most of the existing work on relatedness of tables use Web/HTML tables as their corpus (e.g., tables found on Wikipedia) [4, 20]. However, typical in Open Data portals (e.g., `data.gov`, `data.uk.gov`, ...) many data sources exist where such textual descriptions (such as column headers or cell labels) are missing or cannot be mapped straightforwardly to known concepts or properties using linguistic approaches, particularly when tables contain many numerical columns for which we cannot establish a semantic mapping in such manner.

Indeed, a major part of the datasets published in Open Data portals comprise tabular data containing many numerical columns with missing or non human-readable headers (organizational identifiers, sensor codes, internal abbreviations for attributes like "population count", or geo-coding systems for areas instead of their names, e.g. for districts, etc.) [7]. We verified this observation by inspecting 1200 tables collected from the European Open Data portal and the Austrian Government Open Data Portal and attempted to map the header values using the BabelNet service (`http://babelnet.org`): on average, half of the columns in CSV files served on these portals contain numerical values, only around 20% of which the header labels could be mapped with the BabelNet services to known terms and concepts.[3]

For instance, Table 2 shows the exemplary content of a CSV file found on `data.gv.at`[4] and clearly highlights the limitations of existing (entity-recognition-based) approaches: the content is mainly numerical, the headers are non-descriptive (e.g., the abbreviation "`WHG_TOTAL`" stands for *total number of dwellings*), there are hardly any named-entities in the document, and there exist (non-standardized) comment lines giving additional information.

Linking of Tables. Connecting CSV data to the Web of Linked Data involves typically two steps, that is, (i) transforming tabular data to RDF and (ii) mapping, i.e. linking the columns (which adhere to different arbitrary schemata) and contents (cell values) of such tabular data sources to existing RDF knowledge bases. While a recent W3C standard [15] provides a straightforward canonical solution for (i), the mapping step (ii) though remains difficult.

Mapping involves linking column headers or cell values to either properties or classes in ontologies or instances in knowledge bases. These techniques work well e.g. for HTML/Web tables which have rich textual descriptions, but again they are not applicable to many Open Data CSVs. The large amount of numerical columns require new techniques in order to semantically label numerical values.

[3] We pre-processed the header by splitting the label on underscores and camel-case. We then consider a header as mapped if we retrieved at least one BabelNet entry.

[4] `http://www.wien.gv.at/statistik/ogd/vie_404.csv`

3 Relevancy

We can identify many areas where Open Data is used and valuable, e.g., by governments to increase transparency and democratic control, or by private companies to encourage innovative use of their data. However, in the current Open Data landscape we observe the risk of isolated "data silos" due to missing data integration and decreasing data quality within the catalogs [16]. Manually scanning these data silos, and the data itself respectively, to locate relevant data sources requires substantial amount of time.

Improving the quality of Open Data resources by recommending related content and providing additional semantical information for tables would allow consumer to find relevant data for their needs and would support an automated integration and linkage to other resources. In fact, this is one of the main objectives of the ADEQUATe project, an Austrian research project in which we are involved.[5]

Further, due to missing semantic information current data portals lack complex search functionalities over datasets (e.g., by types or labels/categories of columns). A richer semantic description and relatedness measure for tabular Open Data would allow a search and recommendation engine for Open Data resources which would also provide search across portals (independently of the underlying portal language). For instance, such an engine would enable geo-location based search functionalities, e.g., by labeling a column as "postal code" and mapping the corresponding values to geo-names.

The importance to (semantically) describe CSV data is also recognized by the W3C in the CSV on the Web Working Group [15]. The objective of this group was to define a metadata standard which allows the automatic generation of RDF out of CSV files.

4 Related Work

There exists an extensive body of research to derive semantic labels for attributes in structured data sources (such as columns in tables) which are used to (i) map the schema of the data source to ontologies or existing semantic models or (ii) categorize the content of a data source (e.g., a table talking about politicians). The majority of these approaches [14, 11, 18, 1, 12, 19, 5] assume well-formed relational tables, rely on textual information, such as table headers and string cell values in the data sources, and apply common, text-based entity linkage techniques for the mapping (see [21] for a good survey). Moreover, typical approaches for semantic labeling such as [18, 1, 19] recover the semantics of Web tables by considering as additional information the, again textual, "surrounding" (section headers, paragraphs) of the table and leverage a database of class labels and relationships automatically extracted from the Web.

In [4] the authors define different types of relatedness of tables on the Web and propose and evaluate their algorithms (cf. section 2). This work is based on previous work on semantic labeling of (textual) columns [18].

5 Research Questions & Hypotheses

The research questions of this PhD proposal directly derive from the problem statement and can be stated as follows:

[5] `http://www.adequate.at`, aims at improving the data quality of Austrian Open Data.

Q1 *How far do we get using existing Semantic Web technologies and what are the limitations and upcoming challenges?*
Existing work on semantic enrichment of tables mainly assume different data sources and domains; therefore the techniques may not be directly applicable.

Q2 *How to assign semantic labels to numerical columns (lacking textual information)?*
A semantic labeling of numerical values based on descriptive features and the distribution of the values is currently not part of existing labeling approaches for tables.

Q3 *How to define and measure relatedness of tabular Open Data in a meaningful way?*
Current Open Data portals lack recommendation systems, which would significantly increase the usability of such portals.

To address these research questions we have to test the following main hypothesis:

H Given a corpus of tabular Open Data resources, the usability can be increased by semantically analyzing Open Data CSVs, assigning semantic labels to CSV columns, and ideally generating 5-star linked data.

This implicitly includes testing the following sub-hypotheses (which relate to research questions Q1, Q2, and Q3):

H1 A report and analysis of current tabular Open Data resources allows us to select (and also filter out) existing mapping/linking methods which can be applied in the later steps of this work.

H2 The labeling of numerical data sources is currently not addressed in the literature. However, based on preliminary results of H1, we know that Open Data tables contain many numerical columns with non human-readable headers. We propose the construction of a background knowledge graph which can be used for labeling columns based on the distribution of their values.

H3 Open Data tables are only partially similar to HTML/Web tables (e.g., found on Wikipedia). By defining a suitable measure for relatedness of Open Data tables we will achieve better results for search and recommendation of related content. Expecting full mappings is unrealistic for many CSVs due to the lack of structure, however, finding partial mappings of columns will already allow a categorization and relatedness measure for resources and therefore enable improved search functionalities.

Consistent with James A. Hendler's hypothesis "*a little semantics goes a long way*",[6] this work intends to *significantly* increase the usability of Open Data by *partially* enriching and relating tabular resources using Semantic Web technologies.

6 Preliminary Results

Initial results include a large-scale quality assessment and monitoring of (meta)data published on Open Data portals [17, 16] and the profiling and analysis of CSVs available on these portals [8]. A more extensive journal version of this quality assessment work is to appear [10]. Regrading the labeling of numerical columns in Open Data tables, we got a research track paper accepted at this year's ISWC [9] (see section 7 for a detailed description). Further, we actively contributed to the W3C's CSV on the Web working group and provided an implementation of the recent W3C standard.[7]

[6] http://www.cs.rpi.edu/~hendler/LittleSemanticsWeb.html

[7] https://github.com/sebneu/csvw-parser

7 Approach

We propose to approach the problem in 4 steps:

(i) *Monitoring and analyzing Open Data:*
In order to identify possible improvement/integration strategies, we periodically measure and assess what actual information is available; for instance, we assess the distribution of column types (e.g., by XSD data types, date formats, tokens) or the readability of the headers.

(ii) *Evaluate the applicability of existing entity linkage techniques:*
To tackle hypothesis **H1** - applicable technologies to label and link tabular Open Data - we review and evaluate existing methods in the literature. Due to the differing characteristics of Web tables and Open Data tables (cf. section 2), we have to identify which approaches are applicable to our corpus of data. For instance, considering Table 2, the column headers and cell values cannot be used for named-entity recognition systems (as used in [4]).

(iii) *Labeling and classification of numerical values:*

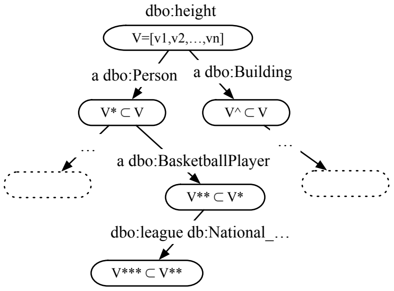


Fig. 1: Hierarchical background knowledge

Open Data tables typically contain a large portion of numerical columns and/or non-textual headers; therefore solutions that solely focus on textual "cues" are only partially applicable for mapping such data sources. As a means of confirming hypothesis **H2**, we develop a method to find and rank candidates of semantic context descriptions for a given bag of numerical values [9]. For instance, given table 2, we want to label column 4 by its corresponding semantic label, i.e., *district code/postal code*, based on similar distributed values found in a background knowledge base. To this end, we apply a hierarchical clustering over information taken from DBpedia to build a background knowledge graph of possible "semantic contexts" for bags of numerical values. For instance, considering Figure 1, we do not only want to label a bag of numerical values as *height*, but instead we want to identify that the values represent the *heights of basketball players who played in the NBA*, or that the values represent the *heights of buildings*. We assign the most likely contexts by performing a k-nearest neighbor search to rank the most likely candidates in our knowledge graph.

(iv) *Open Data tables recommendation and linkage:*
We verify hypothesis **H3** - a relatedness measure for Open Data tables - by incorporating the results of (ii) and (iii) into a recommendation and linkage/integration system which allows an automatic enrichment of resources published on Open Data portals. For instance, considering again table 2, we want to be able to recommend content which describes data for the same regions, based on the NUTS identifiers[8] in the document, but also based on the distribution of the district codes.

[8] http://ec.europa.eu/eurostat/web/nuts/overview

8 Evaluation Plan

Most commonly evaluations for semantic label annotation and linking of tables are based on experiments over a (manually created) gold standard datasets, which is either the result of a crawl of the Web [20], or a domain specific dataset [18], e.g. IMDB or MusicBrainz. Regarding the relatedness of tables, [4] evaluates the experimental results by manual user ratings.

In [9] we evaluate our labeling of numerical values by cross-validating over a sample of DBpedia data generated from the most widely used numeric properties and their associated domain concepts: the evaluation shows that this approach can assign fine-grained semantic labels, when there is enough supporting evidence in the background knowledge graph. In other cases, our approach can nevertheless assign high level contexts to the data, which could potentially be used in combination with other approaches to narrow down the search space of possible labels. Additionally, we tested our approach "in the wild" on tabular data extracted from Open Data portals and reported valuable insights and upcoming challenges which we have to tackle in order to successfully label numerical data from the Open Data domain.

As a test-data corpus for our further evaluations serves the set of tables monitored by our Open Data Portal Watch framework [16], which currently monitors over 200 data portals worldwide and therefore allows large-scale analyses and evaluations.

9 Reflections

As we understand and know the challenges of automated integration and linkage, we do not believe that it is possible to fully map all CSV tables and generate high quality RDF out of them (e.g., prevented by the absence of appropriate ontologies). However, we are convinced that partial mappings already will have an high impact on the usability of Open Data. In fact, current Open Data hold a substantial potential for improvements by semantic web technologies: resources found on Open Data portals are typically created and curated by "non-technicians", e.g., office employees in public administration. Therefore, the principles for producing linked open data sources are possibly ignored (or rather not known) even though the data would allow high quality linked/integrated data. This hypothesis is supported by early results of the ADEQUATe project, in which we are currently involved. In the course of this project we collected feedback by users and providers regarding the current state of Open Data: we identified the potential (but also the demand) for standardization and machine-processable data and a high interest in the outcome of this work.

References

1. Adelfio, M.D., Samet, H.: Schema extraction for tabular data on the web. Proceedings of the VLDB Endowment 6(6), 421–432 (2013)
2. Berners-Lee, T.: Linked data, 2006. http://www.w3.org/DesignIssues/LinkedData.html (2006)
3. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J.: Freebase: A collaboratively created graph database for structuring human knowledge. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. pp. 1247–1250. SIGMOD '08, ACM, New York, NY, USA (2008)

4. Das Sarma, A., Fang, L., Gupta, N., Halevy, A., Lee, H., Wu, F., Xin, R., Yu, C.: Finding related tables. In: Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data. pp. 817–828. ACM (2012)
5. Ermilov, I., Auer, S., Stadler, C.: User-driven semantic mapping of tabular data. In: Proceedings of the 9th International Conference on Semantic Systems. pp. 105–112. I-SEMANTICS '13, ACM, New York, NY, USA (2013)
6. Kucera, J., Chlapek, D., Necaský, M.: Open government data catalogs: Current approaches and quality perspective. In: Technology-Enabled Innovation for Democracy, Government and Governance - Second Joint International Conference on Electronic Government and the Information Systems Perspective, and Electronic Democracy, EGOVIS/EDEM 2013, Prague, Czech Republic. pp. 152–166 (2013)
7. Lopez, V., Kotoulas, S., Sbodio, M.L., Stephenson, M., Gkoulalas-Divanis, A., Aonghusa, P.M.: Queriocity: A linked data platform for urban information management. In: The Semantic Web - ISWC 2012. pp. 148–163 (2012)
8. Mitlöhner, J., Neumaier, S., Umbrich, J., Polleres, A.: Characteristics of Open Data CSV Files. In: 2nd International Conference on Open and Big Data (August 2016), invited paper
9. Neumaier, S., Umbrich, J., Parreira, J.X., Polleres, A.: Multi-level semantic labelling of numerical values. In: The 15th International Semantic Web Conference. Kobe, Japan (October 2016)
10. Neumaier, S., Umbrich, J., Polleres, A.: Automated quality assessment of metadata across open data portals. ACM Journal of Data and Information Quality (JDIQ) (2016)
11. Ramnandan, S.K., Mittal, A., Knoblock, C.A., Szekely, P.A.: Assigning semantic labels to data sources. In: ESWC 2015. pp. 403–417 (2015)
12. Rastan, R.: Towards generic framework for tabular data extraction and management in documents. In: Proceedings of the Sixth Workshop on Ph.D. Students in Information and Knowledge Management. pp. 3–10. PIKM '13, ACM, New York, NY, USA (2013)
13. Reiche, K.J., Höfig, E., Schieferdecker, I.: Assessment and Visualization of Metadata Quality for Open Government Data. In: Proceedings of the International Conference for E-Democracy and Open Government, CeDEM14, 2014, Krems, Austria, May 21-23, 2014 (2014)
14. Taheriyan, M., Knoblock, C.A., Szekely, P., Ambite, J.L.: A Scalable Approach to Learn Semantic Models of Structured Sources. In: Proceedings of the 8th IEEE International Conference on Semantic Computing (ICSC 2014) (2014)
15. Tandy, J., Herman, I., Kellogg, G.: Generating RDF from Tabular Data on the Web (Dec 2015), `https://www.w3.org/TR/csv2rdf/`, W3C Recommendation
16. Umbrich, J., Neumaier, S., Polleres, A.: Quality assessment & evolution of open data portals. In: The International Conference on Open and Big Data. pp. 404–411. IEEE, Rome, Italy (August 2015)
17. Umbrich, J., Neumaier, S., Polleres, A.: Towards assessing the quality evolution of open data portals. In: Proceedings of ODQ2015: Open Data Quality: from Theory to Practice Workshop, Munich, Germany (2015)
18. Venetis, P., Halevy, A.Y., Madhavan, J., Pasca, M., Shen, W., Wu, F., Miao, G., Wu, C.: Recovering semantics of tables on the web. PVLDB 4(9), 528–538 (2011)
19. Wang, J., Wang, H., Wang, Z., Zhu, K.Q.: Understanding tables on the web. In: Conceptual Modeling - 31st International Conference ER 2012, Florence, Italy, October 15-18, 2012. Proceedings. pp. 141–155 (2012)
20. Zhang, Z.: Start small, build complete: Effective and efficient semantic table interpretation using tableminer. Semantic Web Journal (2014)
21. Zhang, Z.: Towards efficient and effective semantic table interpretation. In: ISWC 2014, Lecture Notes in Computer Science, vol. 8796, pp. 487–502. Springer International Publishing (2014)