Publishing, Harmonizing and Consuming Census Data: the CEDAR Project

Albert Meroño-Peñuela, Christophe Guéret, Ashkan Ashkpour, Andrea Scharnhorst
Data Archiving and Networked Services (DANS)
Den Haag, The Netherlands
albert.merono@dans.knaw.nl

Abstract. This paper discusses the use of Linked Open Data to increase complex tabular datasets quality, machine-processability, and ease of format transformation. We illustrate this application with the historical Dutch censuses: census data is open, but notoriously difficult to compare, aggregate and query in a uniform fashion. We describe an approach to achieve these goals, emphasizing open problems and trade-offs.

1. Introduction

Census data plays an invaluable role in the historical study of society. In the Netherlands, the Dutch historical censuses (1795-1971) are among the most frequently consulted sources of statistics of the Central Bureau of Statistics[1]. During the period they cover, it is the only regular statistical population study performed by the Dutch government, and the only historical data on population characteristics that is not strongly distorted.

The dataset comprises 507 Excel workbooks with 2,288 tables, being a digital representation of a partial subset of the original census books. These books have been translated to several digital formats in subsequent stages: first as scanned images, then as PDF documents, and finally as Excel tabular files. The dataset is archived at DANS' archiving system EASY[2], and publicly available at www.volkstellingen.nl as open data.

Nevertheless, the usability of the dataset by citizens and researchers is severely compromised. The tabular files are not well self-described, and users can spend hours of manual seeking before finding the data they are interested in. Moreover, the richness of variables in the dataset (e.g. age, gender, religions, occupations, locations, housing types...) together with drifting variables as a consequence of time, make automated longitudinal queries impracticable in tabular format. Without these cross queries and a systematic access to fine-grained information, stories from data stay away from citizens and researchers.

In this position paper we analyze our proposal for solving this problem in the framework of the CEDAR project. We expose the issues we have encountered and the implementations we have developed with respect to format transformations, quality checking and combination of this

[1] http://www.cbs.nl
[2] http://easy.dans.knaw.nl

dataset with others, as well as describing open problems still pendant to solve.

2. Census Linked Open Data

The CEDAR project aims at exposing census open data on the Web, making it more accessible, linkable and queryable. In CEDAR we apply a specific web-based data-model (exploiting the Resource Description Framework (RDF) technology) to make census data inter-linkable with other hubs of historical socio-economic and demographic data and beyond. The project will result in generic methods and tools to weave historical and socio-economic datasets into an interlinked semantic data-web.

2.1. Complex Tabular Data to RDF

The census tables are exceptionally messy in their layout and contents. Structural heterogeneity is present in row and column headers that do not follow any recognizable pattern between different tables. Sometimes contents of row and column headers are defined hierarchically, spanning through several columns or rows. Other issues, like unstructured subtotals in the data cells, notes written outside of the table borders, and unstructured provenance in annotations make it even harder to identify a common structure. Additionally, semantic heterogeneity makes the meaning and value of variables incompatible in cross-year comparisons. Traditionally, social historians have solved these problems with *harmonization*, a complex process of data restructuring and alignment performed, until now, by hand.

To gain access to a more fine-grained atoms of these tabular data, we converted the dataset to RDF with our own tool, TabLinker[3]. TabLinker produces an RDF named graph per census table according with the RDF Data Cube Vocabulary[4]. Other csv2rdf scripts were discarded due to the high complexity of the census tables. TabLinker requires prior table formatting with styles that define specific cell bounding boxes.

2.2. Images, PDF, Tables, RDF Data Cube: Formats on-demand

One of the CEDAR premises is to provide data to stakeholders in their favourite formats. Despite our take for RDF in solving harmonization in an automated manner, data should be delivered as requested on-demand by users, including tabular formats, structured formats like XML or JSON, RDF or relational databases (according with some relational schema). Checking whether we can round-trip among these formats is one of our subjects of study.

Additionally, we are aware that some tasks final users, especially researchers, may want to do require not only accessing the data atoms, but also referencing the books, images, PDFs and Excel tables these data come from. In other words: extensive provenance about the data source and the transformations these data may have suffered.

2.3. Quality Assessment: Census Annotations

[3] https://github.com/Data2Semantics/TabLinker

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

TabLinker also produces independent RDF named graphs for annotations attached to data cells. These annotations are represented in the Open Annotation Data Model[5], and they contain useful corrections, interpretations and issue descriptions by census experts. These annotations provide a useful mechanism to receive feedback on the dataset, enabling a correction workflow that improves data quality without harming the original values. These annotations, though, need to be harvested at query time to provide appropriate responses, and how their contents need to be compared, selected or transformed for that purpose is no trivial task.

We also developed a script that checks Benford's Law for the census dataset[6]. This is a very well known rule that censuses have historically met, and the script proves that this is also the case for the Dutch historical censuses. We provide these results as a measure of the quality of the data, and we claim that proof of data quality has to be always provided with open data. Because of that, we generate dataset metadata with additional tools[7], and we are working on how to integrate data quality measures in dataset self-descriptions.

2.4. Longitudinal Queries

Representing tabular data into RDF does not solve the problem of cross querying *per se*. Solving longitudinal queries requires combining the census data with other datasets. For instance, answering a cross query about occupational titles (e.g. *evolution on the number of shoemakers in Amsterdam from 1795 to 1971*) needs a mapping between all cells of all tables that describe properties of shoemakers with the class *Shoemaker* of the Historical International Standard Classification of Occupations[8] (HISCO), that represents all individuals with that historical occupation. Then, that class can be queried explicitly to retrieve shoemakers, no matter the way they are spelled in a specific table.

In CEDAR we need to combine a variety of different data sources to solve harmonized queries. If these data sources come in a non-RDF format (and they usually do), we rely on our conversion scripts (like TabLinker) to express them in standard vocabularies that make them suitable for linking. If no such standard vocabularies exist, we arise the need for creating and standardizing them.

3. Architecture and Interface

We have designed a layered architecture to achieve CEDAR goals. We represent all data coming from a variety of formats in RDF space in a three-tier consisting of *raw data, annotations* and *harmonization*. Any produced graph belongs to one and only one of these. Once queries are solved in RDF space, results are returned in on-demand formats as previously described.

However, the interface that wraps this architecture represents a challenge. On the one hand,

---

[5] http://www.openannotation.org/spec/core/
[6] https://github.com/CEDAR-project/MP2Demo
[7] https://github.com/CEDAR-project/TabExtractor
[8] http://historyofwork.iisg.nl/

citizens want stories from data: they want nice visualizations explaining socio-historical reality, and they need an interface which allows them to build these visualizations themselves, making use of all metadata available. On the other hand, researchers do not want such a guided assistance: they prefer diving into the data and defining their own workflows, in order to discover new stories. Finding the right balance between these two approaches is a trade-off in CEDAR we study carefully.

4. Conclusion

This position paper has presented the CEDAR project and has motivated a framework for effective machine-processing of disparate historical open data. We have discussed a generic architecture that deals with complex table conversion, data quality, provenance, and longitudinal queries. We have discussed an in-out data transformation approach, transforming tabular data into standard RDF data models and providing results back in the format preferred by the user. We have shown a practical application of data combination to solve cross queries in RDF space. Finally, we have exposed the challenge of finding the right balance between different user requirements appropriately, when designing an interface making out stories from open data in RDF-based backends.