UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Linked Humanities Data: The Next Frontier? A Case-study in Historical Census Data

Meroño-Peñuela, A.; Ashkpour, A.; Rietveld, L.; Hoekstra, R.; Schlobach, S.

Publication date
2012

Document Version
Final published version

Published in
LISC 2012 : Linked Science 2012 - Tackling Big Data

Link to publication

Citation for published version (APA):
Meroño-Peñuela, A., Ashkpour, A., Rietveld, L., Hoekstra, R., & Schlobach, S. (2012). Linked Humanities Data: The Next Frontier? A Case-study in Historical Census Data. In T. Kauppinen, L. C. Pouchard, & C. Keßler (Eds.), *LISC 2012 : Linked Science 2012 - Tackling Big Data: Proceedings of the Second International Workshop on Linked Science 2012 - Tackling Big Data, in conjunction with the International Semantic Web Conference (ISWC2012) : Boston, MA, USA, November 12, 2012* [3] (CEUR Workshop Proceedings; Vol. 951). CEUR-WS. http://ceur-ws.org/Vol-951/paper3.pdf

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:10 Mar 2023

Linked Humanities Data: The Next Frontier?
A Case-study in Historical Census Data

Albert Meroño-Peñuela[1,2], Ashkan Ashkpour[3], Laurens Rietveld[1], Rinke Hoekstra[1], and Stefan Schlobach[1]

[1] Department of Computer Science, VU University Amsterdam, NL
albert.merono@vu.nl
[2] Data Archiving and Networked Services, KNAW, NL
[3] International Institute of Social History, Amsterdam, NL

Abstract. This paper discusses the use of Linked Data to harmonize the Dutch censuses (1795-1971). Due to the long period they cover, census data is notoriously difficult to compare, aggregate and query in a uniform fashion. In social history, harmonization is the (manual) process of restructuring, interpreting and correcting original data sources to make a comparison possible. We describe a harmonization methodology based on standard Linked Data principles, illustrate how the size and complexity of the resulting linked data source poses new challenges for Semantic Web technology, and discuss potential solutions.

Keywords: Census data, Linked Data, Harmonization, Big Data, Data Heterogeneity, Linked Science

1 Introduction

Census data plays an invaluable role in the historical study of society. In the Netherlands, Dutch census data are among the most frequently consulted sources of statistics of the Netherlands Central Bureau of Statistics[4] [1]. The Dutch census data set is unique in that it covers a period of almost two centuries (1795-1971). During that period, it is the only regular statistical population study performed by the Dutch government. For the period before the 20th century, the census is the only historical data on population characteristics that is not strongly distorted [12].

Census data is both large and complex [1,9,8]. Collected over centuries, it is a rich source of scientific insights into social history. Qualitative information about the development of professions, habitation and religion are up for grabs: if only they were accessible in a uniform, consistent and systematic way. Efforts on census analysis are hampered by the heterogeneity of the data. Data is not only available in different formats, but internally it uses different classification systems, metadata schemes, inconsistent (and incorrect) representations, temporal invariance, various variable coding schemes or differences in the semantics

[4] See http://www.cbs.nl/.

of annotations. Comparative studies are unfeasible in this context. To solve this, social historians follow a (manual) process of *harmonization*: the creation of a unified, consistent data series from disparate census samples [2].

Linked Data appears to be a natural fit, and we decided to publish harmonized census data as RDF. RDF offers a uniform, schema-agnostic representation format that allows a transparent mapping across censuses, while maintaining a link to the original source. However, social scientists and historians require that the process by which the original source data is transformed into the harmonized data remains explicit: the original data and intermediate steps must always be traceable from the result. Therefore, we choose to use RDF to represent census data faithfully, without committing to a particular conceptualization and applying Linked Data principles from the very start of the harmonization process. We analyze how, when and where data problems arise; and study how harmonization practices can be generalized and applied in a Linked Data context.

This paper contributes a description of the use of Linked Data to represent and harmonize statistical data in the socio-historical domain, describing a harmonization methodology based on standard Linked Data principles. We identify the gaps in this approach, and illustrate how the size and complexity of the resulting graphs poses new challenges for Semantic Web technology. The paper is structured as follows. Section 2 describes the Dutch census dataset in more detail, introduces its problems and summarizes how social historians solve them. Section 3 introduces the RDF model for linked census data, how we generated it from the original data, and how we evaluated its correctness. In section 4 we describe the problems we ecountered applying standard Linked Data practices to the census due to its size and complexity, illustrate new challenges for Semantic Web technology to tackle these problems, and discuss potential solutions.

2 Census data and harmonization

The Dutch historical census dataset comprises 507 Excel workbooks containing 2,288 tables, being a digital representation of the original census books. These books have been digitized as images and manually translated into Excel workbooks. A version of the dataset is available at `http://www.volkstellingen.nl`.

One Excel workbook may contain several tables (one per sheet), but given an Excel workbook it applies only to one year and one census type (population census, occupation census or housing census). Some years (1795-1879) only have the population census, but all three types of censuses are present from 1889 onwards. The last censuses of the 19th century contain a higher number of tables because the complexity of these censuses, according to experts (see Figure 1).

The dataset contains 33,283 annotations at the cell level, with a similar distribution to the tables (see Figure 1). Annotations contain considerations (e.g. *this cell includes 2 persons of unknown age*, *this number is unreadable*), interpretations (e.g. *this total should take foreigners into account*) and corrections (e.g. *48* instead of *43*) made by digitizers, archivists or social historians.

We observe structural heterogeneity while comparing these tables (see Figure 2): *row and column headers* do not follow any recognizable pattern with respect

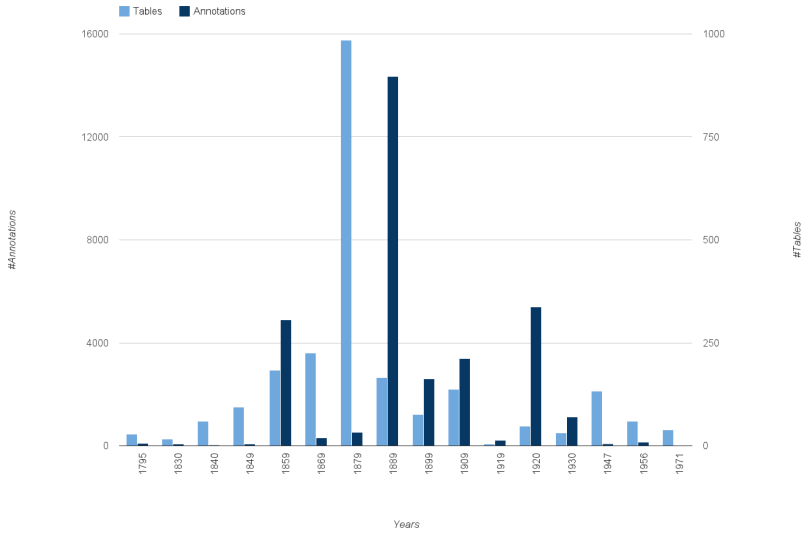

Fig. 1: Dutch historical census dataset counts for tables and annotations.

to their contents or sorting. 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' body make it even harder to identify a common structure. Our goal is to solve these differences, so all tables can be uniformly queried.

2.1 Harmonization semantics

We observe also semantic heterogeneity in these tables: meanings and values of variables (in the statistical sense) are not compatible in cross-year comparisons. We describe the main topics where semantic discordance plays a role, identifying specific data issues, and additional requirements.

Occupation is one of the most unwieldy variables in censuses. Comparative historical research on occupations is heavily hampered by misunderstandings with regards to occupational terminology across time and space [8]. Throughout the years occupational categories of censuses worldwide have evolved significantly, emerging and disappearing due to occupation evolution, and thus changing classification systems [11]. Accordingly, this evolution is reflected in the fact that the U.S. Census defined 287 separate occupations in 1950, compared to 441 in 1971. We have occupations such as *tile settlers* and *roof repairers* which were sometimes presented as one occupation and in other years separated.

Demographic information is at the core of the census. The composition of age distribution across different census years is a challenging aspect. Dissimilar age ranges are applied throughout different censuses making historical comparisons cumbersome.

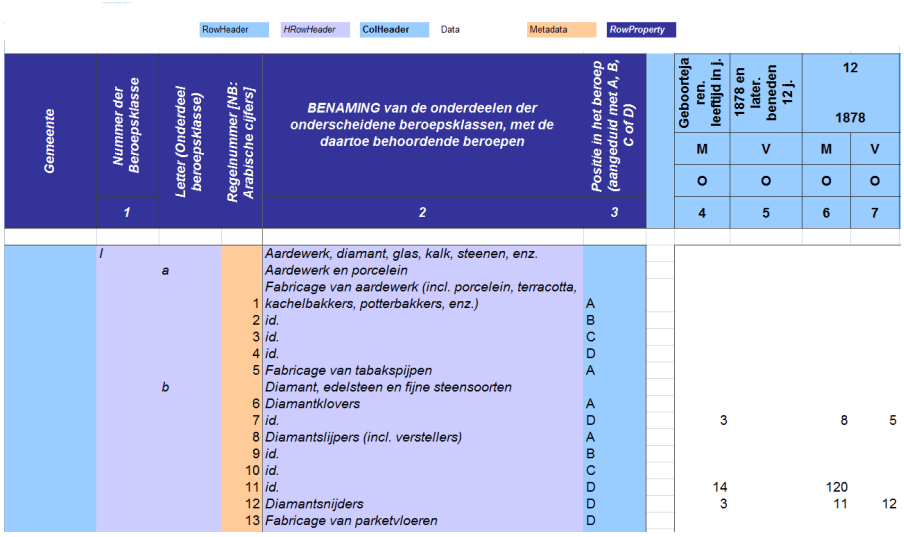

Fig. 2: One of the census tables. Legend illustrates common groups of cells identified.

The census contains rich *geographical* coverage of the Netherlands' state and spatial divisions. Giving geographical context to data enhances its meaning and makes allowances for their environment so as to enhance their comprehension [13]. However, comparative studies are severely hampered by changing boundaries. The composition of the municipality of Rotterdam, e.g., has changed significantly over time. Smaller municipalities may not exist anymore or have merged. This makes it difficult to trace these back in history without a historical classification system and GIS (geographic information system) referencing.

Errors have been introduced at many stages of the tables cycle: during its collection, during the digitization process, or in translating from images to Excel workbooks. Since we have to keep these workbooks faithful to originals, erroneous numbers can not simply be overwritten with corrected ones.

Provenance on activities that imply accessing, rewriting, reviewing or transforming the contents of the tables needs to be recorded. *Who did what, when, why?* are significant questions for researchers using these tables.

2.2 Current harmonization practice

We talked extensively with social historians to understand harmonization, the process of creating a unified, consistent data series from disparate census samples [2].

The current harmonization practice mainly uses external classification systems such as HISCO[5] to link and standardize the different occupations. By an

[5] The HISCO code is used to study historical patterns of work and social mobility in comparative perspective, and is built upon the 1000 most frequent male and female occupational titles (see `http://hisco.antenna.nl/`.)

extended process of trail, error and review a set of coding rules was developed allowing titles to be incorporated into the classification scheme. Compatibility with the International Labour Organisation's ISCO686 scheme allows the linkage of historical and contemporary data sets.

The current practice consists of a straightforward process of mapping occupation codes into an external classification system, such as HISCO. In most cases users can manually check the index for a given occupation title and find the corresponding HISCO code. In more difficult cases users have to choose themselves between different codes or in cases where the title cannot be found to proceed by using the logical structure of the HISCO code using the major, minor and unit structure. This also applies when harmonizing municipalities. As the internal classifications differ too much, several external classification systems (Amsterdam Code, CBS Code, Wageningen Code etc.) have been developed applying the principle of maximum and minimum varying codes [10].

The census applies disparate age distributions throughout the tables, making comparisons difficult. Harmonization uses statistical aggregation and imputation for solving this. For example, comparison of age ranges *14-18* and *19-20* with *14-15* and *16-20* can be achieved with the aggregate *14-20*, or by way of interpolation defining *14-15*, *16-17* and *18-20*.

3 A structured approach on harmonization

Since harmonization is the process social historians use to resolve data issues, it seems that the whole dataset should be harmonized before applying Linked Data principles. This approach has been followed by other initiatives, such as the United States 2000 census[6] and the 2001 Spanish census[7] have published census data as RDF, but in a microdata-based approach (microdata considers individual data at the muninicipality level, whilst census data consists of aggregated data). This seems the more straightforward approach, but there are several reasons to do the opposite: that is, to produce an RDF version of the dataset, and harmonize afterwards.

The first reason for this is that harmonization is not a standard process. Harmonization strongly relies on interpretation and specific goals. Researchers should be free to make their own interpretations and establish their own goals on how to harmonize census data. This is what historical research is all about.

Secondly, access to source data must always be guaranteed. Original and harmonized data must coexist in a single system to fulfill requirements of both source-oriented and goal-oriented historical research methods[8] [1].

Finally, as we show in section 4, Linked Data principles and semantic technologies can play an important role in solving and implementing harmonization.

[6] See `http://www.rdfabout.com/demo/census/`, only one year census, only microdata.

[7] See `http://gutenberg.dcs.fi.uva.es/~bhscmcyt/census/sparql_en.php`, only one year census, only microdata.

[8] In a source-oriented approach, historians have a priority on representing research data with a faithful schema with respect to structure of their sources. In a goal-oriented approach, the priority is to represent these data following a particular conceptualization of the domain that suits functional requirements.

We have defined and implemented a faithful conversion of the Dutch historical census data from Excel workbooks to RDF[9]. After defining appropriate data model and vocabularies, we developed a script producing RDF graph data, TabLinker, and a number of SPARQL queries to validate our requirements.

3.1 Generating census triples

We developed a specific tool, TabLinker[10], to produce RDF graphs according to this data model. Despite the existence of other tools translating tabular data into RDF (like `csv2rdf4lod` or `TopBraid Composer`), some of them used in the US census case, none met our input data (e.g. disparate table structure), output data (e.g. annotations, provenance support) and, most importantly, harmonization requirements. We organize all produced triples by TabLinker in a three layer architecture.

First, the *source* layer consists of a 1:1 copy of table contents. It represents tables in a flat, schema-agnostic way, providing direct access to original data. We use RDF to avoid an early, concrete conceptualization of the domain, according to principles of historical source-oriented approaches. Any necessary transformation due to functional requirements has to be done in a further stage (e.g. via `CONSTRUCT` SPARQL queries).

Since data contained in a census table is statistical data, we designed a per-cell data model according to the RDF Data Cube vocabulary[11] (see Figure 3). We use the term *observation* for all metadata describing a data cell and its context (`d2s:isObservation a qb:Observation`). A data cell is affected by a number of dimensions (`d2s:dimension a qb:DimensionProperty`) that correspond to the column and row headers that apply to that cell. The content of a data cell is always a number counting population (`d2s:populationSize a qb:MeasureProperty`).

We use the Open Annotation Core Data Model[12] and PROV[13] to add annotations and provenance to the cells. We attach an annotation to a cell using an attribute property (`d2s:attribute a qb:AttributeProperty`). The text in the body of an annotation can provide an updated number for the cell (`d2s:correctedPopulationSize a qb:MeasureProperty`). Each annotation has an author (`prov:wasGeneratedBy`).

Second, the *structure* layer consists of an interpretation of table layout. Tables have common sections (see legend in Figure 2). *Data* cells contain the census counts. *Headers* and *RowHeaders* contain hierarchical entities that describe the numbers in their corresponding *data* cells. *Properties* do the link between *data* and *RowHeaders*, and they are used as predicate names when linking a *data* cell with the *RowHeader* associated to it.

[9] Although the UK census case (`http://cdu.mimas.ac.uk/index.htm`) is very similar to ours, including harmonization and Linked Data principles, no data model, tool or guideline seems to be available yet.

[10] See `https://github.com/Data2Semantics/TabLinker`.

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

[12] See `http://www.openannotation.org/spec/core/`.

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

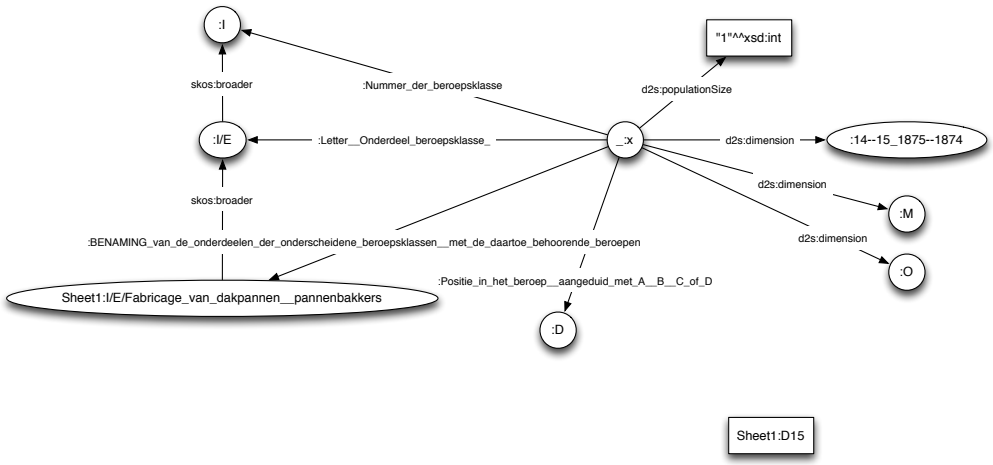

Fig. 3: Excerpt of one table cell as represented in the RDF graph (source layer).

To interpret table layout TabLinker needs tables to be marked, applying preset Excel styles to the main areas of the sheet (see Figure 2; different colors correspond to different styles). *Headers* and *RowHeaders* are read top-down and left-right, respectively. E.g., if A is a *Header* cell in top of the table subdivided in B, this is interpreted as `A skos:broader B`. Then, *data* cells are linked to their corresponding *Header* and *RowHeader* cells as Data Cube dimensions (`CellN28 d2s:dimension A, B`). One named graph is built per table.

Third, the *harmonization* layer lies on top and contains all mappings and classes necessary for internal linking (see Section 4).

3.2 Evaluation: querying the census

After successfully producing an RDF version of the census workbooks, RDF data was stored in a triplestore. The next step was to validate the RDF data, proving that meaningful queries for social historians can be issued against the RDF version of the census.

In order to have a measure of success, we considered previous studies on the Dutch censuses [1]. In these studies contents of the census are analyzed, aggregated and plotted in order to explain social historical phenomena. Thus, showing that a number of SPARQL queries can produce the same results as the studies can prove the faithfulness of the RDF representation of the census, as well as the correctness of the query. This was very important to our experts.

We produced a number of geo charts[14] representing results of two different queries that retrieve the same data on two different years. For several reasons (see Section 4) the same result cannot be achieved using one single query.

[14] See `http://cedar-project.nl/visualizing-sparql-query-results-on-the-census/`.

Comparisons show that, despite some alignment between the plots coming from SPARQL and the social studies, the standard Linked Data approach did not successfully produce satisfactory results with respect to our requirements.

4 Requirements for Linked Humanities Data

Following a typical implementation of Linked Data principles in the Dutch historical census dataset did not produce ready-to-publish RDF data. Some of our requirements were not successfully met. Our take is that there is a lack of Semantic Web tools to tackle some characteristics of the Dutch census dataset found in other Humanities datasets as well. This section describes a set of requirements for such tools.

4.1 A world of mappings

In Section 2.1 we have presented some census variables (like occupations and geographical coverage) that do not completely match in labeling or meaning. The standard procedure to solve these non-matching entities, so that all data can be uniformly queried, is to use direct mappings. Harmonization can be seen as an additional metadata layer mapping mismatching labels, e.g. stating `A owl:sameAs B` or `A skos:exactMatch B` if A and B equal or exactly match, respectively. This solution, though, arises many other problems that make it impracticable in our case.

Size and complexity TabLinker produced 2,288 highly disconnected named graphs from the same number of tables, because these tables are self-meaningful and non-dependant between them. Thus, TabLinker did not produce a single triple connecting them. Any mapping is an interpretation to be placed in the harmonization layer.

A possible solution to this disconnection is to produce explicit mappings between related entities. Listing 1.1 shows how to do this with SPARQL queries on the census dataset: lines between 8 and 13 contain the graph pattern to retrieve men population per municipality. Assuming that this graph pattern remains the same across the tables, there are 2,288 different queries to refer to *one* variable with all its possible labels. If T is the set of tables and V the set of variables encoded, there are $T \times V$ ways for referring to the variables. In the worse case, all variables and tables are targeted to be queried at once, which for the census means $2,288 \times 122 = 279,136$ mappings. The cost of connecting the graphs via SPARQL is clearly unaffordable: too many queries derive in complex queries and expensive joins.

A possible solution to manage this number of mappings is to materialize them in the triplestore. For example, if the source layer contains the labels `RoomsKatholiekeKerk`, `RomsKatholic` and `VaticanChristelijk` for the variable `Religie` (religion), all three can be harmonized creating a new resource `Catholic` and issuing three triples with `Catholic` as subject, `skos:altLabel` as predicate and the labels as objects. This way, multiple interpretations become multiple graphs, and SPARQL can refer to these graphs instead of explicitly mapping each possible literal. However, this has two drawbacks: for a per-cell access,

```
PREFIX d2s: <http://www.data2semantics.org/core/>
PREFIX d2sdata: <http://www.data2semantics.org/
                 data/VT_1879_10_H1_marked/Noordbrabant/>
PREFIX ns1: <http://www.data2semantics.org/core/Noordbrabant/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>

SELECT ?place ?size WHERE {
 ?cell d2s:isObservation [ d2s:dimension d2sdata:Totaal ;
                           d2s:dimension d2sdata:M ;
                           ns1:Kom__Buiten_de_kom___tijdelijk_
                           aanwezige__Schepen__Personen_zonder_
                           vaste_verblijfplaats ?place ;
                           d2s:populationSize ?size ] .
 ?place skos:prefLabel "Totaal voor de woningen in de gemeente"@nl .
}
ORDER BY DESC(?size)
```

Listing 1.1: SPARQL query for population of men per municipality, province of Noord-Brabant, year 1879.

queries still need to bind table dimensions to SPARQL variables (which produces expensive joins); and scalability is very complex, because domain experts have to link manually new literals with harmonizing classes.

One way of tackling with these interpretation graphs is using linksets. A linkset is a group of links that map two entities. Several linksets can be defined to describe different groups of mappings. For example, if author A_1 provides the mappings `X skos:exactmatch Y; owl:sameAs Y` and author A_2 does so with `X skos:related Y; skos:exactMatch Z`, a linkset engine can solve the mapping (i.e. replace all references to X as references to Y or Z) if the user chooses criteria from A_1 or A_2.

As we have seen, Humanities RDF datasets are constantly subject to interpretation, context and non-destructive updates at fine-grained level. These can result in an arbitrary increase of the graph size and complexity, making graphs unmanageable for query engines and reasoners.

Inconsistency Humanities datasets have specific requirements regarding reasoning. We have already introduced the problem of graph size when adding context, which certainly affects reasoning.

But mappings and classes in the harmonization layer do not need to be consistent at all. Author A_1, via annotations, can state `X owl:sameAs Y` and `Y owl:sameAs Z`, while author A_2 can state `X owl:differentFrom Z` at the same time. Author A_3 can define two disjoint classes U, V, while author A_4 can create a common subclass of both. Though being contradictory, these harmonization mappings are true: this is how authors interpret the dataset. Despite these inconsistencies, reasoning facilities are still needed for inferred knowledge discovery.

One possible solution is to contextualize inferencing, adapting the scope of the reasoner to make it consider only a subset of non-contradictory interpretations.

Concept similarity The described massive direct mapping scenario suggests that a more generic approach on concept similarity is necessary to improve management of concepts in the census and other Humanities datasets. We consider a theory of Concept Drift [14] as our definition for the *meaning of a concept*, its *label*, its *intension* and its *extension*.

Identity resolution [3] [6] plays a fundamental role as an harmonization operation in a Linked Data context. Very commonly, harmonizing values of a variable means grouping those that mean the same: different resources that have identical meaning have to be identified. In terms of Concept Drift theory, it is said that these concepts have identical *rigid intensions* (their fundamental properties are the same), while concept *labels* are different. Other scales of identity, like synonyms and homonyms, also fit here.

Many *harmonization concepts are based on other concepts*. Harmonization is also about taking two existing resources A and B and defining a new concept C in terms of A and B. For instance, a new concept `Christian` is introduced as a more abstract class holding both `Catholic` and `Protestant` as subclasses. But sometimes the hierarchy is specialized instead of generalized, or C only takes part of the *extension* of A and B (e.g. if A and B overlap and C takes the intersection).

Another case for harmonization occurs when *labels* remain the same over time. For instance, `CivilWorker` is used to refer to civil workers in 1879 as well as in 1920, but they are certainly not the same concept. While the *labels* are still the same, the meaning of the concept at its internal definition (its *intension*) has completely changed. A measure representing the *intension shift* of concepts is very valuable in this context [15].

Some *concepts are destroyed or created*, especially while trying to harmonize distant datasets in time. For instance, comparing the concept `ComputerScientist` in 1971 and 1791 makes no sense; likewise, `PublicSpeaker` does appear in a 1889 dataset, while it would rarely do in a 2012 one.

4.2 Beyond mappings

As shown, increase of size and complexity of census graphs due to direct mappings makes this strategy unsuitable for our case study: scalability of any system managing this knowledge base is severely compromised. This section exposes unmet requirements of the census (and other Humanities disciplines) to be considered as new challenges for Semantic Web technologies.

Dynamic ontologies Static, deterministic, agreed and formal definitions of meaning of concepts (the kind of definitions in which Semantic Web tools feel comfortable) are scarce in Humanities datasets. Formalization facilities provided e.g. by OWL do not fit at all with time-changing, non-deterministic, non-agreed and hardly formalizable concepts commonly found in the Humnanities [5].

Despite Ontology Versioning [4,7], current tools lack time-awareness, i.e. the ability of the formalism to define differently a class of objects depending on the

given moment in time, even if for that given moment no static definition has been explicitly attached. Many problems we encounter need variable, dynamic formalizations of concepts that change depending on a selected time frame. Having such definitions in the census case, the problem of having multiple, time-drifted labels for referring to a single occupation would no longer require direct mappings (neither big nor complex graphs), and concept similarity would be much easier to tackle.

Interpretation, often derived from contested concepts and hermeneutics [16], provides subjective definitions when there is no consensus about the meaning of a concept. Authors are free to define concepts as they consider, usually in an inconsistent or contradictory fashion. Agreement can not be imposed, and all definitions shall coexist in a system that takes into account this definition multiplicity. A broader scope on integrating interpretation to current Semantic Web formalisms can be also a solution for ontology inconsistency in the context of the Humanities.

Partitions and counting One peculiarity of the census tables is that they only contain aggregations, i.e. counts of population meeting certain conditions. Thus, a census table can be seen as a partition of the total population, assuming that variables are not overlapping (i.e. every person is counted once and only once). Given the case of merging two variables of different tables that only match partially, how their numbers (i.e. data cells with the population counts) should be added? To what extent is possible to infer an individual (instance) of a variable (class) considering only data available in the table?

Format round-tripping In our case, census data has been converted from Excel workbooks into RDF named graphs. Many other formats are of interest of several user groups, like relational databases, CSV files or XML documents. The choice of RDF as an intermediate representation also has to meet the requirement of round-tripping: it must be possible to reconstruct the original Excel workbooks from the generated named graphs. Moreover, regardless of the transformations done between formats and the expressivity of their languages, internal equivalence with the original data has to be kept so that it can be reconstructed.

5 Conclusions

In this paper we have presented a conversion of the Dutch historical census dataset into RDF, taking into account key requirements from social scientists and historians, like guaranteeing a permanent access to original data and providing a framework for implementing harmonization. We have shown that harmonization is necessary in order to align the highly disconnected named graphs that TabLinker produces. Once done, visualizations could be greatly improved, consistently built and easily generated with appropriate uniform queries.

We have shown that the extraordinary size and complexity of these graphs explodes when direct mappings between disconnected (but related) entities are made explicit. Even proposing a three-tier architecture for organizing this knowledge into source, structure and harmonization layers, SPARQL queries that grab homogeneous data across multiple census tables turn extraordinarily complex. A

standard Linked Data approach on the census also led us to inconsistency (due to contradictory harmonization statements) and to explore concept similarity.

Non-met requirements in the Dutch historical census case are our starting point to throw new challenges for Semantic Web technologies that could enormously help the Humanities solving their specific concerns on Linked Data. Awareness of time, interpretation and context as fundamental factors that greatly influence the definition of meaning of contested concepts can help creating a new version of the Semantic Web also suitable for Humanities data.

Acknowledgements The work on which this paper is based has been partly supported by the Computational Humanities Programme of the Royal Netherlands Academy of Arts and Sciences, under the auspices of the CEDAR project. For further information, see `http://ehumanities.nl`. This work has been supported as well by the Dutch national program COMMIT. Authors want to thank all contributing colleagues, especially Andrea Scharnhorst, Kees Mandemakers, Christophe Guéret and Frank van Harmelen.

References

1. Boonstra, O., et al.: Twee eeuwen Nederland geteld. Onderzoek met de digitale Volks-, Beroeps- en Woningtellingen 1795–2001. DANS en CBS (2007)
2. Esteve, A., Sobek, M.: Challenges and methods of international census harmonization. Historical Methods 36(2), 37–41 (2003)
3. Halpin, H., et al.: When owl: sameas isn't the same: An analysis of identity in linked data. In: International Semantic Web Conference (1). pp. 305–320 (2010)
4. Heflin, J., Hendler, J.A.: Dynamic ontologies on the web. In: AAAI/IAAI. pp. 443–449 (2000)
5. Ide, N., Woolner, D.: Historical Ontologies, chap. Words and Intelligence II: Essays in Honor of Yorick Wilks, pp. 137–152. Springer (2007)
6. Isele, R., et al.: Silk server - adding missing links while consuming linked data. 1st International Workshop on Consuming Linked Data, Shanghai (2010)
7. Klein, M., Fensel, D.: Ontology versioning on the semantic web. In: Stanford University. pp. 75–91 (2001)
8. van Leeuwen, M.H.D., Maas, I., Miles, A.: Creating a historical international standard classification of occupations. Historical Methods 37(4), 186–197 (2004)
9. Mandemakers, K., Dillon, L.: Best practices with large databases on historical populations. Historical Methods 37(1), 34–38 (2004)
10. van der Meer, A., Boonstra, O.: Repertorium van Nederlandse Gemeenten, 1812-2006, waaraan toegevoegd de Amsterdamse code. DANS, Den Haag (2006)
11. Meyer, P.B., Osborne, A.M.: Proposed category system for 1960-2000 census occupations. U.S. Bureau of Labor Statistics (2005)
12. Ruggles, S.: US-MN Historical Census Projects. Historical Methods 1(28), 6 (1995)
13. St-hilaire, M., Moldofsky, B., Richard, L., Beaudry, M.: Geocoding and Mapping Historical Census Data. Historical Methods 40(2), 76–91 (2007)
14. Wang, S., Schlobach, S., Klein, M.C.A.: What is concept drift and how to measure it? In: EKAW. pp. 241–256 (2010)
15. Wang, S., Schlobach, S., Klein, M.C.A.: Concept drift and how to identify it. J. Web Sem. 9(3), 247–265 (2011)
16. White, G.: Semantics, hermeneutics, statistics: some reflections on the semantic web. In: Proceedings of the 25th BCS Conference on HCI. pp. 24–28 (2011)