Undefined 1 (2009) 1–5
IOS Press

CEDAR: The Dutch Historical Censuses as Linked Open Data

Editor(s): Name Surname, University, Country
Solicited review(s): Name Surname, University, Country

Albert Meroño-Peñuela [a,b], Christophe Guéret [a,b], Ashkan Ashkpour [c], and Stefan Schlobach [a]
[a] *Department of Computer Science, VU University Amsterdam, De Boelelaan 1081a, 1081HV Amsterdam, NL*
E-mail: {albert.merono, c.d.m.gueret, k.s.schlobach}@vu.nl
[b] *Data Archiving and Networked Services, Anna van Saksenlaan 10, 2593HT Den Haag, NL*
E-mail: {albert.merono, christophe.gueret}@dans.knaw.nl
[c] *International Institute of Social History, Cruquiusweg 31, 1019AT Amsterdam, NL*
E-mail: ashkan.ashkpour@iisg.nl

Abstract. In this document we describe the CEDAR dataset, a five-star Linked Open Data representation of the Dutch historical censuses, conducted in the Netherlands once every 10 years from 1795 to 1971. We produce a linked dataset from a digitized sample of 2,300 tables. The dataset contains more than 6.8 million statistical observations about the demography, labour and housing of the Dutch society in the 18th, 19th and 20th centuries. The dataset is modeled using the RDF Data Cube vocabulary for multidimensional data, uses Open Annotation to express rules of data harmonization, and keeps track of the provenance of every single data point and its transformations using PROV. We link these observations to well known standard classification systems in social history, such as the Historical International Standard Classification of Occupations (HISCO) and the Amsterdamse Code (AC), which in turn link to DBpedia and GeoNames. The two main contributions of the dataset are the improvement of straightforward data access for historical research, and the emergence of new historical data hubs, like classifications of historical religions and historical house types, in the Linked Open Data cloud.

Keywords: Social History, Census data, Linked Open Data, RDF Data Cube

1. Introduction

In this document we describe the CEDAR dataset, a five-star Linked Open Data conversion of the Dutch historical censuses dataset[1].

The Dutch historical censuses were collected from 1795 until 1971, in 17 different editions, once every 10 years. The government counted the entire population of the Netherlands, door by door, and aggregated the results in three different census types: demographic (age, gender, marital status, location, belief), occupational (occupation, occupation segment, position within the occupation), and housing (ships, private houses, government buildings, occupied status). These aggregations were written down in tables and published in books. These books are archived by the Central Bureau of Statistics[2] (CBS) and the International Institute of Social History[3] (IISH). In an effort to improve their systematic access, part of the tables in these books have been digitized as images in several projects between the CBS, the IISH and several institutes of the Royal Netherlands Academy of Arts and Sciences[4] (KNAW), such as Data Archiving and Net-

[1] See http://www.volkstellingen.nl/

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

[3] See http://www.iisg.nl/

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

worked Services[5] (DANS) and the Netherlands Interdisciplinary Demographic Institute[6] (NIDI). Beyond digitisation, these projects have translated part of this dataset, by manual input, into more structured formats. As a result, a subset of the dataset is available as Excel spreadsheets.

Challenges. The historical Dutch censuses have been collected for almost two centuries with different information needs at given times [2]. Census bureaus are notorious for changing the structure, classifications, variables and questions of the census in order to meet the information needs of a society. Not only do variables change in their semantics over time, but the classification systems in which they are organized also change significantly, making it extremely cumbersome to use the historical censuses for longitudinal analysis. The structures of the tables and changing characteristics of the census currently do not allow comparisons over time without extensive manual input of a domain expert. Even when converted into Web structured data, the need for harmonization across all years is a prerequisite in order to enable greater use of the census by researchers and citizens.

Contributions. The goal of CEDAR[7] is to make these spreadsheets available as five-star Linked Open Data, to make them comparable over time, and to investigate how semantic technologies can improve the workflow of research performed by historians. This paper provides details about the Dutch historical censuses Linked Open Data published so far. The dataset comes with the following features:

- Historical statistics on two centuries of Dutch history, fully compliant with RDF Data Cube [5];
- Standardization and harmonization procedures encoded using Open Annotation [12] at the header cell level;
- Full tracking of provenance in all our activities and consumed/produced entities as of PROV [6];
- Dereferenceable URIs[8];
- A human browseable web front-end[9];
- Dataset live statistics[10].

The rest of the paper is organized as follows. In Section 2 we describe our conversion pipeline. In Section 3 we provide a full description of the data model and the use of established vocabularies, along with the quantity, quality and purpose of links to other datasets. In Section 4 we argue the importance of the dataset and its availability, including plans for long term preservation of the produced Linked Open Data. We discuss the five-star conformance of the dataset and its known shortcomings in Section 5.

2. Data Conversion and Modeling

Our data conversion pipeline follows the diagram shown in Figure 1. In the following sections we describe this pipeline in more detail.

2.1. Data Conversion

In this section we describe the conversion process of the census tables from their original format to RDF. Further details on this conversion process can be found at [9].[11] The dataset consists of 2,288 tables represented as spreadsheets in 507 Excel files. Each Excel file may contain one or several spreadsheets, but one spreadsheet always contains one single census table. An example of such a table is shown in Figure 2. A specific interpretation of the eccentric layout of these tables is necessary to generate RDF triples expressing exactly the same information. For instance, the bottom right figure in Figure 2 should be read: there were 12 unmarried (*O* column) women (*V* column), 12 years old and born in 1878 (*12 1878* column) working as ordinary (row *D* in column *Positie in het beroep*, position in the occupation) diamond cutters (*Diamantsnijders* row) in the municipality of Amsterdam (column *Gemeente*, municipality). Consequently, this interpretation hampers a straightforward conversion of these tables, *e.g.* using well known generic community tools, to RDF. To this end, we developed TabLinker[12], a supervised Excel-to-RDF converter that relies on human markup on critical areas of these tables (see colors in Figure 2). With such markup, TabLinker can follow the same interpretation and generate meaningful RDF graphs across Excel files. The Integration pipeline shown in Figure 1 uses a fork of TabLinker, called TabLink[13], which generates raw data according

[5] See http://www.dans.knaw.nl/

[6] See http://www.nidi.knaw.nl/en/

[7] See http://cedar-project.nl/ and http://www.ehumanities.nl/

[8] See http://bit.ly/cedar-data#

[9] See http://lod.cedar-project.nl/cedar/

[10] See http://lod.cedar-project.nl/cedar/stats.html

[11] All conversion source code is available at https://github.com/CEDAR-project/Integrator/

[12] See https://github.com/Data2Semantics/TabLinker/

[13] See https://github.com/CEDAR-project/Integrator/blob/master/src/tablink.py

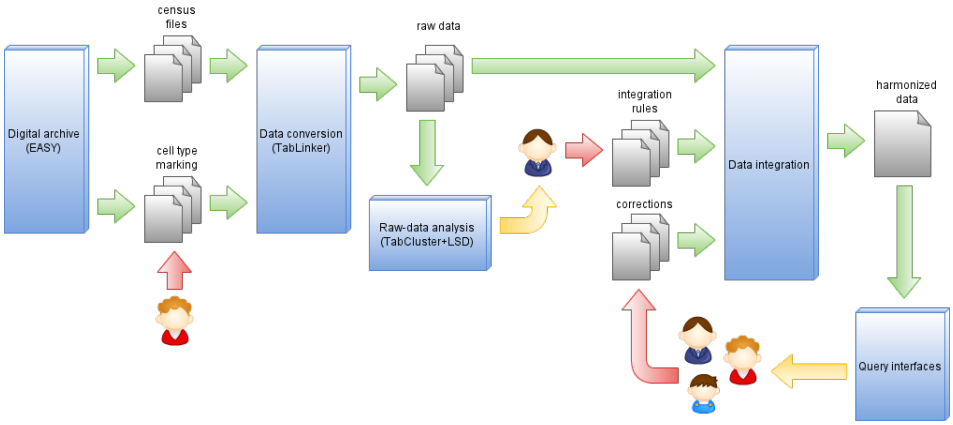

Fig. 1. Integration pipeline for the CEDAR data. The workflow starts at the archiving system, where the original Excel files are stored and retrieved using its API. Raw data is produced after interpreting complex table layout. These raw data are later transformed into harmonized data by applying integration rules encoded as Open Annotations. Red arrows indicate that manual input is required.

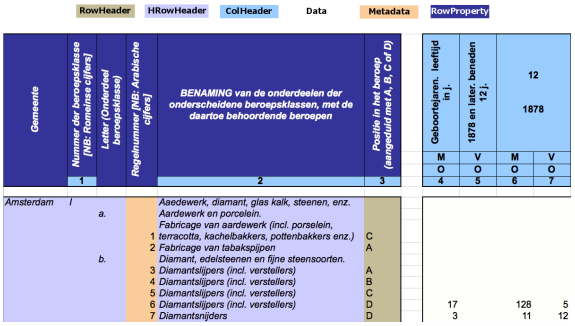

Fig. 2. One of the census tables of the dataset (occupation census of 1889, province of Noord-Holland). Colour markup is manually added and does not belong to the original data.

to our own table layout model instead of RDF Data Cube.

2.2. Raw Data

The Dutch historical censuses are multidimensional data covering a wide spectrum of statistics in population demography, labour force and housing situation. Since RDF Data Cube (QB) provides a means "to publish multi-dimensional data, such as statistics, on the web in such a way that they can be linked to related data sets and concepts" [5], we choose QB as our goal data model to express the census data as RDF.

However, the source tables lack critical information needed to generate a complete and sound QB dataset. Concretely, we miss mappings between dimensions with their corresponding values (*e.g.* it is said nowhere that column header *M* means *male* and relates to dimension *gender*, or that *O* means *unmarried* and relates to *marital status*). For this reason, we generate an agnostic RDF table layout representation as a first step, postponing the generation of proper RDF Data Cube.

Using the markup discussed in Section 2.1, TabLink first generates one `tablink:DataCell` for each data cell (*i.e.* cells marked as *Data* in Figure 2), attaching its value (the actual population count) and the `tablink:sheet` the observation belongs to (a legacy table identifier, *e.g.* `BRT_1889_02_T1-S0`). Secondly, the observation is linked with all its corresponding column and row headers (*i.e.* cells marked as *RowHeader*, *HRowHeader*, and *ColHeader* in Figure 2). An example is shown in Listing 1. Additionally, we create resources that describe the column and row headers, their types, labels, cell positions in the spreadsheets and hierarchical parent/child relationships with other headers (if any).

Because the result of this conversion stage is incomplete, due to the lack of further description of some dimensions and their mappings to standard values, codes and concept schemes, we call this the *raw* dataset conversion of the original Excel tables.

2.3. Integration Rules as Open Annotations

To solve the missing dimension-value mappings shown in Listing 1, we annotate header cells using Open Annotation [12] with *harmonization rules* (see Listing 2). This is a manual process performed by experts. With such rules we can explicitly indicate the dimension to which a specific value belongs. Moreover, we can extend the description of such value (*e.g.* map-

```
cedar:BRT_1889_08_T1-S0-K17 a tablink:DataCell ;
    rdfs:label "K17";
    tablink:dimension cedar:BRT_1889_08_T1-S0-A8 ;
    tablink:dimension cedar:BRT_1889_08_T1-S0-K6 ;
    tablink:dimension cedar:BRT_1889_08_T1-S0-J3 ;
    tablink:dimension cedar:BRT_1889_08_T1-S0-K4 ;
    tablink:dimension cedar:BRT_1889_08_T1-S0-K5 ;
    tablink:dimension cedar:BRT_1889_08_T1-S0-B8 ;
    tablink:dimension cedar:BRT_1889_08_T1-S0-C12 ;
    tablink:dimension cedar:BRT_1889_08_T1-S0-E17 ;
    tablink:dimension cedar:BRT_1889_08_T1-S0-F17 ;
    tablink:value "12.0" ;
    tablink:sheet cedar:BRT_1889_08_T1-S0 .
```

Listing 1: Raw RDF extracted for the cell K17 of the occupation census table of 1889, province of Noord-Holland. The `cedar` namespace maps to `http://lod.cedar-project.nl:8888/cedar/resource/`

```
cedar:BRT_1889_08_T1-S0-K4-mapping a oa:Annotation ;
    oa:hasBody cedar:BRT_1889_08_T1-S0-K4-mapping-body ;
    oa:hasTarget cedar:BRT_1889_08_T1-S0-K4 ;
    oa:serializedAt "2014-09-24"^^xsd:date ;
    oa:serializedBy
        <https://github.com/CEDAR-project/Integrator> ;
    prov:wasGeneratedBy
        cedar:BRT_1889_08_T1-S0-mapping-activity .

cedar:BRT_1889_08_T1-S0-K4-mapping-body a rdfs:Resource ;
    sdmx-dimension:sex sdmx-code:sex-F .
```

Listing 2: Mapping rules defined for *one* of the header cells associated to a data cell, in its corresponding annotation.

ping "O" with "unmarried" and "V" with "female") or map these values to dimensions that where not explicitly present in the original tables.

Some of these rules map the values extracted from the tables into standard *classification systems*. For instance, in order to query occupations consistently across the whole dataset, we map occupation dimension values (which are table dependent) to HISCO codes[14] (Historical International Standard Classification of Occupations). We proceed similarly with other dimensions like historical religions, house types and historical municipalities in the Netherlands, using scripts and mappings done manually by experts (see Sections 3.1 and 3.2).

[14] See `http://historyofwork.iisg.nl/`

```
cedar:BRT_1889_02_T1-S0-K17-h a qb:Observation ;
    maritalstatus:maritalStatus maritalstatus:single ;
    cedarterms:occupationPosition cedarterms:job-D ;
    cedar:population "12"^^xml:decimal ;
    sdmx-dimension:sex sdmx-code:sex-F ;
    prov:wasDerivedFrom cedar:BRT_1889_08_T1-S0-K17 ;
    prov:wasGeneratedBy
        cedar:BRT_1889_08_T1-S0-K17-activity .
```

Listing 3: Refined RDF Data Cube after applying harmonization rules in observation-attached OA annotations.

2.4. Harmonized RDF Data Cube

Using `CONSTRUCT` SPARQL queries, we process all the raw data produced by TabLink and apply all harmonization rules conveniently. As a result, we obtain refined, harmonized RDF Data Cube like shown in Listing 3. We generate a `qb:Observation` for each `tablink:DataCell`, and we link that observation to all its corresponding PROV triples.[15]

We also produce a `qb:DataStructureDefinition` (DSD) with all dimensions, attributes and measures used, and introduce several `qb:Slice` that group the observations by census type (VT, demography; BRT, occupations; and WT, housing) and year (from 1795 to 1971). The DSD can be browsed online[16], as well as the slices[17] and therefore all the observations.

2.5. Provenance

We implement provenance tracking with PROV [6] at all stages as shown in Figure 3[18]. We do this for a number of reasons. First, provenance allows us to ensure reproducibility of our conversion workflow. Second, it facilitates the debugging of all integration rules, since we can trace back all mappings, activities and entities involved in the generation of each `qb:Observation`. And third, we use it to meet the strong requirement of historians of being able to explain how every single harmonized value of the dataset is produced, back to the archived sources. For historians, ensuring independence and reliability of primary sources is fundamental, also in the Semantic Web [11].

[15] See `https://github.com/CEDAR-project/Integrator/blob/master/src/cubes.py#L226`

[16] `http://lod.cedar-project.nl:8888/cedar/harmonized-data-dsd`

[17] `http://lod.cedar-project.nl:8888/cedar/harmonized-data-sliced-by-type-and-year`

[18] Generated with `https://github.com/Data2Semantics/provoviz`

Fig. 3. Visualization of PROV entities and their transformations through our conversion workflow.

For the TabLink generation of raw data cubes, we log a specific `prov:Activity`, recording task timestamps (`prov:startedAtTime`, `prov:endedAtTime`), its `prov:Agent` (`prov:wasAssociatedWith`) and the specific markup used via `prov:used`.

Similarly, during the execution of the mappings described as OA annotations we record an additional `prov:Activity`, making explicit the use of each specific mapping in the harmonization rules via `prov:used`. The result is shown in Figure 3.

2.6. Named Graphs and URI Policy

To organise the generated census triples we make them available in three different named graphs[19]:

- The *raw* data triples, as extracted from the original tables, are in `<urn:graph:cedar:raw-data>`.
- All annotation mapping rules are contained in `<urn:graph:cedar:rules>`.
- The refined RDF Data Cube, produced after applying the mapping rules to the raw data, is located at `<urn:graph:cedar:release>`.

The resource URI naming policy is as follows: raw data cells are named following the schema `<cedar:(FILE-ID)-(SHEET-ID)-(CELL-ID)>` , like `<cedar:BRT_1889_08_T1-S0-K17>` (see Listing 1), where:

- `(FILE-ID)` is a legacy ID for the original Excel file, with the format `(TYPE)_(YEAR)_(PART)_(VOLUME)`, *e.g.* `BRT_1889_08_T1` refers to the occupation census (*BRT*) conducted in 1889, part 8, volume T1.
- `(SHEET-ID)` is an identifier of the sheet within a file, *e.g.* `S0` for the first sheet, `S1` for the second, etc.
- `(CELL-ID)` is an identifier of the cell within a sheet, *e.g.* `K17` for the cell in column K, row 17.

The annotations containing the mapping rules associated to each header cell that affects a data cell follow exactly the same encoding, but adding the suffix *"-mapping"* to the resource. For example, `<cedar:BRT_1889_08_T1-S0-K4-mapping>` identifies the annotation containing the mapping rules for the header cell `<cedar:BRT_1889_08_T1-S0-K4>`

Similarly, we identify the refined RDF Data Cube observations adding to the raw data URIs the suffix *"-h"*. For example, `<cedar:BRT_1889_08_T1-S0-K17-h>` identifies the `qb:Observation` we generate using the data cell `<cedar:BRT_1889_08_T1-S0-K17>` as a basis and applying the mapping rules defined at the annotation `<cedar:BRT_1889_08_T1-S0-K17-mapping>`.

[19] since we do not need them to be de-referenceable, we currently use URNs instead of URIs

Description	Count
Number of datasets processed	1,358
Expected number of datasets	2,308
Total number of observations	6,800,175

Table 1

Datasets processed, expected, and generated observations.

3. Linked Dataset Description

In this section we describe the CEDAR dataset in more detail. Table 1 shows some dataset statistics through its Data Structure Definition (DSD). Our conversion workflow is an ongoing process, since mapping rules in the observation annotations need to be manually curated. For this reason, we update these statistics every time we run the conversion workflow.[20]. This allows us to keep track of what is left to map. Currently 6,800,175 observations are generated and linked to one `qb:measureProperty` (population), one `qb:attributeProperty` (unit of measure, number of persons), and nine `qb:dimensionProperty`: year of birth, sex, occupation position, belief, occupation, reference area, marital status, reference period, and census type.

Table 2 shows a summary of the different dimensions correctly mapped with standard codes into observations so far.

3.1. Internal Links

The census tables often refer to variables and values with multiple synonyms: *e.g.* the value *female* for variable *sex* can be arbitrarily referred by *v*, *vrouw*, *vrouwen*, *vrouwelijk* or *vrouwelijk geslacht*[21].

[20] Full and regularly updated statistics can be found at `http://lod.cedar-project.nl/cedar/stats.html`

[21] *Vrouw* stands for *woman* in Dutch.

Dimension label	Occurrences	%
belief	253480	1.24%
censusType	4642360	22.75%
municipality	153248	0.75%
maritalStatus	1886415	9.25%
occupation	328790	1.61%
occupationPosition	8120	0.04%
province	43946	0.22%
refPeriod	4642360	22.75%
sex	3801431	18.63%

Table 2

Dimensions and their frequency over the dataset.

In some variables this problem is quite straightforward to solve via the mappings we define as annotations, and we manually code mappings that cover all possible synonyms. This is the case for the variables **sex**, **marital status** and **occupation position** (*i.e.* rank class that a worker was assigned). The dimension *sex* is coded with `sdmx-dimension:sex`, and the codes `sdmx-code:sex-F` (female) and `sdmx-code:sex-M` (male) as values[22]. We mint our own URIs for dimensions *marital status* (`maritalstatus:maritalStatus`) and *occupation position* (`cedarterms:occupationPosition`). *Marital status* can get as value one of the codes `maritalstatus:single` (denoting single individuals), `maritalstatus:married` (married) or `maritalstatus:widow` (widows). Likewise, *Occupation position* can get as value one of the codes `cedarterms:job-D` (ordinary workers of the lowest rank, usually assigned to youth), `cedarterms:job-C` (ordinary workers with other lower-rank workers under their responsibility), `cedarterms:job-B` (foremen and other workers with many labour below their hierarchy) or `cedarters:job-A` (directors or owners of businesses).

Other variables require a more complex schema of their possible values: for these QB suggests the use of concept schemes (also called *classification systems* in social history). The variable **house type**, which distinguishes military, civil, public and private buildings that were counted during the censuses, encodes building types in a taxonomic fashion. We manually build up this concept scheme[23] in a data-driven way, assisted by domain experts in social history.[24] We use the dimension `cedarterms:houseType` and an associated code list for this variable.

3.2. External Links

Other variables, like **province**, **municipality**, **occupation** and **belief**, also need complex schemas or taxonomies to encode their values (see Section 3.1). We link to external datasets to standardize these variables.

Province and *municipality* contain codes of Dutch provinces and municipalities from the past and are assigned as objects of predicates `sdmx-dimension:refArea`. Linking to GeoNames or DBPedia seems appropriate. However, Dutch provinces and municipalities suffered major changes during the historical censuses period. To address this, we issue links to `gemeentegeschiedenis.nl`.[25] `gemeentegeschiedenis.nl` is a portal that exposes standardized Dutch historical province and municipality names as Linked Open Data, based on the work done in the Amsterdamse Code (AC) [1]. 2,658,483 links are issued to provinces and municipalities in this dataset, based on previously existing manually curated mappings[26].

We follow a similar procedure to link values of the variable *occupation*. In this case, we rely on HISCO, which offers 1,675 standard codes for historical occupations. We issue 354,211 links to human-readable occupation description pages, also relying on existing manual mappings[27].

Other variables, like *belief* (religion), also need to be standardized by linking to standard classification systems. However, for these no proper historical classifications are available. In such cases, we create these classifications, either manually (relying on expert knowledge) or automatically (leveraging lexical and semantic properties [8]). In any case, we use mappings to these classifications to standardize the census values[28]. We use such mappings to issue 256,952 links to historical religious denominations.

4. Impact and Availability

4.1. Impact

Publishing the Dutch historical censuses as five-star Linked Open Data has a deep impact in the methodology that historians and social scientists have tradi-

[22] Some SDMX COG dimensions and codes are available in RDF at `http://purl.org/linked-data/sdmx/2009/dimension#` and `http://purl.org/linked-data/sdmx/2009/code#`

[23] See concept scheme at `goo.gl/mt1dsn`

[24] See [8] for an approach to build such taxonomies automatically

[25] See `http://www.gemeentegeschiedenis.nl/`

[26] Municipality-AC mappings at `http://goo.gl/yNf3LX`

[27] Occupation-HISCO mappings at `http://goo.gl/oiBTQc`

[28] See mapping files at `https://github.com/CEDAR-project/Integrator/tree/master/extra`

tionally followed to study this dataset [2]. Due to the limitations of the old formats, the dataset could not be utilized to its full potential. To the date, most of the research based on the historical Dutch censuses focused on specific comparable years [3]. To utilize the full potential of the historical censuses researchers have identified harmonization of the data as a key aspect, which we implement as rules in `oa:Annotation` annotations. Previously, if researchers wanted to map *e.g.* the evolution of the total number of inhabitants of a specific gender, in a specific municipality, and for a specific occupation, they had to consult more than 25 different Excel tables and run into manual data transformations. Moreover, keeping track of provenance of all performed operations was cumbersome and relied on good craftsmanship and delicate assumptions. By using explicit harmonization rules and links to standard classifications for occupations, municipalities, religions and house types, researchers can get answers to their queries in a blink of a time compared to the manual way of digging into disparate Excel tables. Hence, (a) speed of query answering and (b) a full provenance track of every single data point down to the historical sources are major milestones for History scholars. Using the SPARQL endpoint, social scientists can retrieve data that gives support to hypotheses that previously could only be assumed.

As five-star Linked Open Data, the census dataset is open for longitudinal analysis, especially for a study of change. Being a major interest for historical research, the change in structures of classifications, meaning of variables and semantics of concepts over time, known as concept drift [7,13], is a fundamental topic to explore. One possibility is to relate the change of meaning of census concepts over time with statistical variances of the distribution of individuals belonging to such concepts [10]. Another related question is the relationship that may hold between the correlation of statistical concepts and their semantic similarity [4].

The dataset sums to other initiatives on publishing census data on the Web as RDF Data Cube[29]. To the best of our knowledge, our is the first effort on publishing censuses with historical characteristics.

We have collected a number of SPARQL queries that we consider relevant for interested users. These are available in the CEDAR dataset front-end[30].

The CEDAR dataset was used in the hackathon held during the 2014 CEDAR symposium[31] with 11 attendees, and also in the 1st Digital History Datathon held at the VU University Amsterdam[32] with 13 attendees. The CEDAR dataset is listed as one of the datasets in the Challenge of the 2nd International Workshop on Semantic Statistics[33] (SemStats 2014), International Semantic Web Conference (ISWC 2014).

In addition, we log the usage of the dataset via any dereferenced URI or fired SPARQL query. These dataset usage data will be soon made available in the CEDAR website.

4.2. Availability

The CEDAR dataset, consisting of the raw Excel file conversions, the annotation mapping rules, and the harmonized RDF Data Cube, is served as Linked Open Data at `http://lod.cedar-project.nl/cedar/`. All URIs dereference via a Pubby installation on this server, which returns data formatted according to the requested format in the response header of HTTP requests. The SPARQL endpoint of the dataset can be found at `http://lod.cedar-project.nl/cedar/sparql`. All versions of the dataset, including the original Excel files (with and without markup), mappings, and the converted RDF data can also be retrieved as Turtle dumps at `https://github.com/CEDAR-project/Integrator/tree/master/data/input`.

The creation and update of the dataset is done through a software package, the CEDAR Integrator[34], developed for that purpose at the VU University Amsterdam and DANS under the LGPL v3.0 license[35]. The dataset is regularly dumped to a GitHub repository[36]. Updates are performed in order to correct errors and incomplete mappings our experts detect when supervising statistical analyses[37] that we automatically generate during the conversion process (see Section 2.1). For long term preservation, the dataset is (and will continue being) deposited into DANS EASY[38], a trusted digital archive for research data.

[29] See cases for Italy, France, Australia and Ireland at `http://www.istat.it/it/archivio/104317#variabili_censuarie`, `http://goo.gl/hiGZF9`, `http://stat.abs.gov.au/` and `http://data.cso.ie/`

[30] See `http://lod.cedar-project.nl/cedar/data.html`

[31] `http://cedar-project.nl/cedar-minisymposium-march-31st-april-1st-2014`

[32] See `http://cedar-project.nl/linkathon-at-the-vu/`

[33] See `http://semstats2014.wordpress.com/`

[34] See `https://github.com/CEDAR-project/Integrator`

[35] See `http://www.gnu.org/licenses/lgpl.html`

[36] See `https://github.com/CEDAR-project/DataDump/`

[37] See `http://lod.cedar-project.nl/cedar/stats.html`

[38] See `https://easy.dans.knaw.nl/`

5. Discussion

In this paper we present the steps followed and the results achieved by CEDAR to transform a two-star (Excel conversions of scanned census tables) representation of the Dutch historical censuses into five-star Linked Open Data (harmonized census resources using URIs and linked to external concept schemes) as part of the Computational Humanities Programme[39] of the Netherlands Royal Academy of Arts and Sciences[40].

We acknowledge a number of shortcomings in the dataset. Importantly, we are aware that the conversion is not complete. Not all observations reach the end of the pipeline, and the ones that do might not get linked to all the original dimensions of the tables. Moreover, our mappings can be incomplete (e.g they can leave out possible raw data values). To address this, we developed full statistical analyses on the conversion process[41]. With such analyses, we can quantify how far we are from completion and the work that still needs to be done on standardization. In addition, we are aware that an important demographic variable, **age**, has no mappings defined yet. Age ranges are aggregated differently in each census edition, and mappings need to define additional interpolation rules in order to generate comparable data. During the data generation we have issued temporal vocabularies (*e.g.* `cedarterms`) for some variables that we will modularize in separate data-hubs. For instance, *belief* and *houseType* deserve their own Web spaces to allow other historical datasets to link to them. Linking the census observations to other datasets is another challenge[42]. Finally, the census tables contain a number of subtotals, totals and partial results at different levels of aggregation. We plan on checking the consistency of these aggregation levels automatically, spotting possible source errors.

References

[1] Ad van der Meer and Onno Boonstra. *Repertorium van Nederlandse Gemeenten, 1812-2006, waaraan toegevoegd de Amsterdamse code*. DANS Data Guide 2, The Hague, 2006.

[2] Ashkan Ashkpour, Albert Meroño-Peñuela, and Kees Mandemakers. The Dutch Historical Censuses: Harmonization and RDF. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 2014. (to appear).

[3] O.W.A. Boonstra, P.K. Doorn, M.P.M. van Horik, J.G.S.J. van Maarseveen, and J. Oudhof. *Twee Eeuwen Nederland Geteld. Onderzoek met de digitale Volks-, Beroeps- en Woningtellingen 1795-2001*. DANS en CBS, The Hague, 2007. `https://www.knaw.nl/nl/actueel/publicaties/twee-eeuwen-nederland-geteld/@@download/pdf_file/Volkstelling_geheel_WEB_verkleind.pdf`.

[4] Sarven Capadisli, Albert Meroño-Peñuela, Sören Auer, and Reinhard Riedl. Semantic Similarity and Correlation of Linked Statistical Data Analysis. In *Proceedings of the 2nd International Workshop on Semantic Statistics (SemStats 2014), ISWC*. CEUR Workshop Proceedings, 2014.

[5] Richard Cyganiak, Dave Reynolds, and Jeni Tennison. The RDF Data Cube Vocabulary. Technical report, W3C, 2014. `http://www.w3.org/TR/vocab-data-cube/`.

[6] Paul Groth and Luc Moreau. PROV-Overview. An Overview of the PROV Family of Documents. Technical report, World Wide Web Consortium, 2013. `http://www.w3.org/TR/prov-overview/`.

[7] Albert Meroño-Peñuela. Semantic Web for the Humanities. In P. Cimiano, O. Corcho, V. Presutti, L. Hollink, and S. Rudolph, editors, *The Semantic Web: Semantics and Big Data. 10th International Conference, ESWC 2013, Proceedings*, volume 7882 of *LNCS*, pages 645–649, Berlin, Heidelberg, 2013. Springer-Verlag.

[8] Albert Meroño-Peñuela, Ashkan Ashkpour, and Christophe Guéret. From Flat Lists to Taxonomies: Bottom-up Concept Scheme Generation in Linked Statistical Data. In *Proceedings of the 2nd International Workshop on Semantic Statistics (SemStats 2014). International Semantic Web Conference (ISWC)*. CEUR Workshop Proceedings, 2014.

[9] Albert Meroño-Peñuela, Ashkan Ashkpour, Laurens Rietveld, Rinke Hoekstra, and Stefan Schlobach. Linked Humanities Data: The Next Frontier? A Case-study in Historical Census Data. In *Proceedings of the 2nd International Workshop on Linked Science (LISC2012). International Semantic Web Conference (ISWC)*, volume 951. CEUR-WS, 2012.

[10] Albert Meroño-Peñuela, Christophe Guéret, Rinke Hoekstra, and Stefan Schlobach. Detecting and Reporting Extensional Concept Drift in Statistical Linked Data. In *Proceedings of the 1st International Workshop on Semantic Statistics (SemStats 2013), ISWC*. CEUR Workshop Proceedings, 2013.

[11] Albert Meroño-Peñuela and Rinke Hoekstra. What Is Linked Historical Data? In *19th International Conference on Knowledge Engineering and Knowledge Management, EKAW 2014, Proceedings*, LNCS, Berlin, Heidelberg, 2014. Springer-Verlag. To appear.

[12] Robert Sanderson, Paolo Ciccarese, and Herbert Van de Sompel. Open Annotation Data Model. Technical report, W3C, 2013. `http://www.openannotation.org/spec/core/`.

[13] Shenghui Wang, Stefan Schlobach, and Michel C. A. Klein. Concept drift and how to identify it. *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, 9(3):247–265, 2011.

[39] See `http://www.ehumanities.nl/`

[40] See `http://www.knaw.nl/`

[41] See `http://lod.cedar-project.nl/cedar/stats.html`

[42] See already issued links at `http://cedar-project.nl/linkathon-at-the-vu/`. Historical newspapers at `http://kranten.delpher.nl/` are other interesting data to link.