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 data integration and 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). After 1971, censuses stopped from being collected from the entire population, mostly due to social opposition, and authorities switched to use municipal registers and sampling. Three facts make the 1795-1971 dataset self-contained and of special interest to historians and social scientists: it is based on counting the whole Dutch population, instead of sampling; it provides an unprecedented level of detail, hardly comparable to modern censuses due to privacy regulations; and the survey microdata from which the aggregations were originally built is almost entirely lost.

The census aggregations were written down in tables and published in books, archived by the Central

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

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 the historical censuses books 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 Networked 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 a collection of 507 Excel spreadsheets, containing 2,288 census tables.

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 spreadsheets 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 pre-requisite in order to enable greater use of the census by researchers and citizens.

Related work. Previous work addresses Linked Historical Data as the general issue of enriching Semantic Web graphs with temporal information. For instance, authors of [7] expose a knwoledge graph that automatically integrates a spatio-temporal dimension from Wikipedia, GeoNames and WordNet data. Similarly, [13] proposes a generic approach for inserting temporal information to RDF data by gathering time intervals from the Web and knowledge bases. Authors in [5] focus on using the temporal aspect of Linked Data snapshots to keep track of the evolution of data over time. Besides addressing this temporal dimension, *historical* can also refer to the data and methods used by historians; the adoption of semantic technologies by these data and methods has been previously surveyed [11].

Contributions. The goal of CEDAR[7] is to integrate the Dutch historical censuses in these spreadsheets using Web technologies and standards; to publish the result of this integration as five-star Linked Open Data; and to investigate how semantic technologies can improve the research workflow of historians. Concretelly, the main contributions of the dataset are:

- It is the first historical census data made available as LOD, integrated and Web-enabled from heterogeneous sources;
- it is released together with auxiliary resources, like historical classification schemes and integration mappings;
- it is linked to other datasets in the LOD cloud to improve its exposure and richness.

Additionally, the Dutch historical censuses Linked Open Data comes with the following features:

- Historical statistics on two centuries of Dutch history, fully compliant with RDF Data Cube [4];
- Standardization and harmonization procedures encoded using Open Annotations [14];
- Full tracking of provenance in all 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] See http://www.cbs.nl/
[3] See http://www.iisg.nl/
[4] See http://www.knaw.nl/
[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://lod.cedar-project.nl:8888/cedar/page/harmonised-data-dsd
[9] See http://lod.cedar-project.nl/cedar/
[10] See http://lod.cedar-project.nl/cedar/stats.html

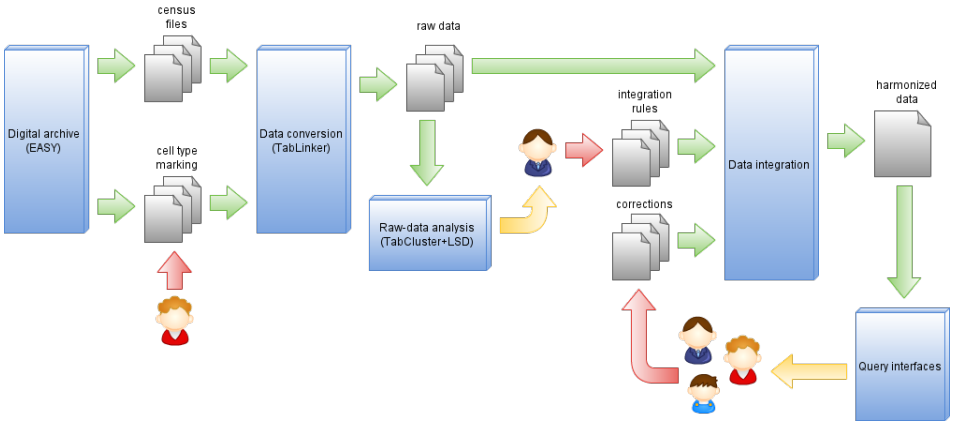

Fig. 1. Integration pipeline for the CEDAR data. The workflow starts at the archiving system, where the original Excel files are stored and retrieved using its API. Raw data is produced after interpreting complex table layout. These raw data are later transformed into harmonized data by applying integration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

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[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). We define 6 markup styles that allow us to distinguish different cell roles (row headers, hierarchical row headers, column headers, data cells, metadata cells and row properties) within spreadsheets. 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 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. We choose RDF Data Cube (QB) as our goal data model to express the census data in RDF, since 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" [4]. In QB, data points are called *observations*, primarily composed of a *measure* (*e.g.* "3 inhabitants") and a set of *dimensions* qualifying that measure (*e.g.* "males", "unemployed", "in Amsterdam"). Dimensions can be arbitrarily combined to refer to unique observations in the cube.

However, the source tables lack critical information needed to generate a complete and sound QB dataset.

[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`

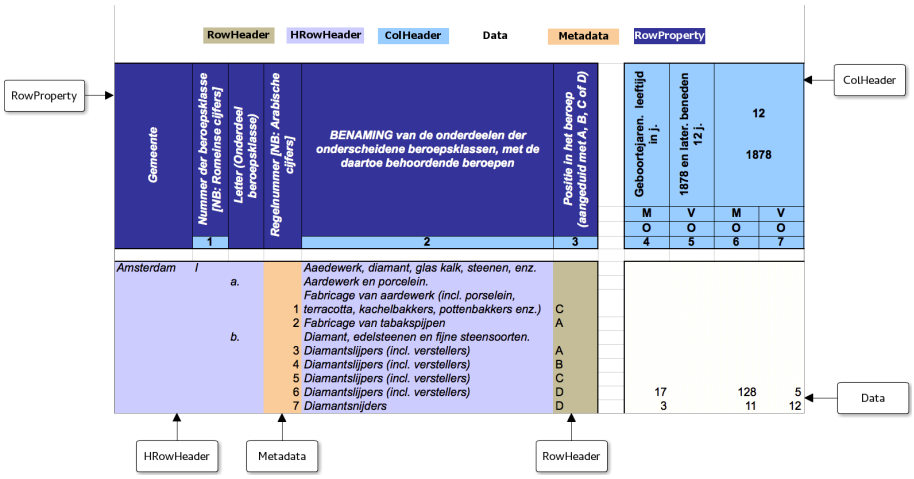

Fig. 2. One of the census tables of the dataset (occupation census of 1889, province of Noord-Holland). Colour markup is manually added and does not belong to the original data.

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.

After a 2 hour technical training, two people styled the 2,288 sheets of the dataset in 25 hours with the markup discussed in Section 2.1. Using such styles, 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 [14] 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.* mapping "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]

```
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.

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

```
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.

```
cedar:BRT_1889_02_T1-S0-K17-h a qb:Observation ;
    maritalstatus:maritalStatus maritalstatus:single ;
    cedar:occupationPosition cedar: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.

(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). We develop two tools to help experts on this process: LSD Dimensions, and TabCluster. LSD Dimensions [9] is an observatory of RDF Data Cube dimensions, codes, concept schemes and data structure definitions available now in the Linked Open Data cloud. It allows the reuse of these statistical resources by data owners and publishers. In case a specific concept scheme of interest is not available yet, we propose TabCluster. TabCluster [10] is a concept scheme generator that leverages syntactic and semantic properties of non-standardized data cubes to assist data modelers on building concept schemes.

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. 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 [12].

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`.

2.6. Named Graphs and URI Policy

To organise the generated census triples we make them available in three different named graphs[18]:

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

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

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

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

– 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`.

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.[19]. 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.

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.

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.

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*[20].

In some variables this problem is 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**, **occupation position** (*i.e.* rank class that a worker was assigned), **housing type situation** and **residence status**. 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[21]. We mint our own URIs for dimensions *marital status* (`maritalstatus:maritalStatus`) and *occupation position* (`cedar:occupationPosition`). *Marital status* can get

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

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

[21] 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#`

Dimension	New	Value / code in scheme	#Refs
cedar:houseType	✓	cedar:house-BewoondeHuizen	88,737
		cedar:house-BewoondeSchepen	28,573
		cedar:house-BewoondeWagens	4,221
		cedar:house-HuizenAanbouw	14,323
		cedar:house-OnbewoondeHuizen	51,599
		cedar:house-OverigeGebouwen	23,344
cedar:isTotal	✓	"0" or "1"	205,606
cedar:population	✓	xsd:integer	710,462
cedar:residenceStatus	✓	resStatus:AltijdAanwezig	7,640
		resStatus:FeitelijkeAanwezig	81,625
		resStatus:JuridischAanwezig	220,293
		resStatus:TijdelijkAanwezig	119,373
		resStatus:TijdelijkAfwezig	55,733
		resStatus:WerkelijkTotaal	21,403
sdmx-dimension:refArea	×	From gg:10002 to gg:11447	692,491
sdmx-dimension:sex	×	sdmx-code:sex-M	220,661
		sdmx-code:sex-F	213,991

Table 3

Dimensions attached to released observations, with their intended values (codes forming concept schemes), whether they are new or reused, and their frequency in the dataset. Prefixes are described in Table 4. References to codes in the concept schemes are expanded from a much smaller number of mapping rules, as shown in Table 5.

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 cedar:job-D (ordinary workers of the lowest rank, usually assigned to youth), cedar:job-C (ordinary workers with other lower-rank workers under their responsibility), cedar:job-B (foremen and other workers with many labour below their hierarchy) or cedar:job-A (directors or owners of businesses). The dimension *housing type situation* indicates the type of house inhabitants were counted in (occupied/empty houses, occuped/empty living ships, houses in construction), and *residence status* qualifies the status of the counted residents (present, legally registered and present, temporarily present, temporarily absent) (see Table 3).

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[22] in a data-driven way, assisted by domain experts in social history.[23] We use the dimension cedar: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.[24] 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 (see Table 5).

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 (see Table 5).

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 [10]). In any case, we use mappings to these classifications to standardize the census values (see Table 5). We use such mappings to issue 256,952 links to historical religious denominations.

Table 3 shows a summary of the different dimensions mapped into observations so far, together with the codes associated to them, the number of times they are referenced, and whether they are created or reused from existing vocabularies. We also make available the RDF describing the created vocabularies[25], and we foresee a future reuse of these vocabularies by publishers of historical aggregated censuses of other countries. Table 4 lists all prefixes used. To standardize dimensions and their values, we create mapping rules and scripts; a summary of these is shown in Table 5

[22] See concept scheme at https://goo.gl/mt1dsn

[23] See [10] for an approach to build such taxonomies automatically

[24] See http://www.gemeentegeschiedenis.nl/

[25] See https://github.com/CEDAR-project/Vocab

Prefix	URI	Content
oa	http://www.w3.org/ns/openannotation/core/	Open Annotations vocabulary
prov	http://www.w3.org/ns/prov#	PROV ontology
dcat	http://www.w3.org/ns/dcat#	Data Catalog vocabulary
qb	http://purl.org/linked-data/cube#	RDF Data Cube (QB) vocabulary
sdmx-dimension	http://purl.org/linked-data/sdmx/2009/dimension#	QB dimensions
sdmx-code	http://purl.org/linked-data/sdmx/2009/code#	QB codes (dimension values)
sdmx-attribute	http://purl.org/linked-data/sdmx/2009/attribute#	QB attributes
cedar	http://bit.ly/cedar#	CEDAR terms: population, totals, variable descriptions
cedar-data	http://lod.cedar-project.nl:8888/prices-wages/resource/	CEDAR data points
resStatus	http://bit.ly/cedar-residenceStatus#	Residence status codes
maritalstatus	http://bit.ly/cedar-maritalstatus#	Marital status codes
gg	http://www.gemeentegeschiedenis.nl/amco/	Dutch historical municipalities (AMCO codes)
tablink	http://bit.ly/cedar-tablink#	TabLinker spreadsheet cell types
hisco	http://bit.ly/cedar-hisco#	Historical occupations (HISCO codes)

Table 4

Prefixes used in the dataset.

Variable	Mapping file	Generation	#Mapping rules
Age	https://goo.gl/5NIIqE	Expert-based; SPARQL	16,398
Belief	https://goo.gl/i1H2j4	Expert-based	582
City	https://goo.gl/poFcxo	Expert-based; string similarity script	42,294
Housing type	https://goo.gl/fdc0s8	Expert-based	3,484
Marital status	https://goo.gl/2rYLyu	Expert-based	10
Occupation	https://goo.gl/CUVSGc	Expert-based	21,851
Occupation position	https://goo.gl/y7NoYw	Expert-based	4
Province	https://goo.gl/yShX7w	Expert-based	18
Sex	https://goo.gl/ZtVS3z	Expert-based	10
Total	https://goo.gl/978YSy	Expert-based; SPARQL	38
Housing type situation	https://goo.gl/IEWfBf	Expert-based	22
Residence status	https://goo.gl/TRra0U	Expert-based	40

Table 5

Type and number of mapping rules created per variable type. These mappings expanded into a much greater number of references to codes in concept schemes, as shown in Table 3.

(string similarity scripts for cities can be found online[26]).

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 traditionally 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 know *e.g.* the number of houses under construction in the Netherlands per municipality between 1859 and 1920[27], they had to consult 47 different Excel tables and run into laborious data transformations. Moreover, keeping track of provenance of all performed operations was cumbersome and relied on data munging 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. Table 6 shows the number of tables that users had to open and the number of cells they had to manipulate to answer a set of example queries. Most of these queries have been already manually investigated by social historians [3]. Hence, major milestones the

[26] See https://github.com/CEDAR-project/1909-exception-maker and https://github.com/CEDAR-project/CityVariantMapper.

[27] Additional example queries at http://lod.cedar-project.nl/cedar/data.html

dataset provides for History scholars are (a) speed-up of query answering; and (b) full provenance tracks of every data point down to the historical sources. Using the SPARQL endpoint, social scientists can retrieve data that gives support to hypotheses that previously could only be assumed. In addition, links to external datasets facilitate answering queries that users hardly could perform otherwise; for instance, links to `gemeentegeschiedenis.nl` and DBpedia allow to instantly compare nowaday's population of Dutch municipalities with their historical figures, via SPARQL 1.1 federation. Moreover, dimension standardization enables new query solutions that were only possible through extensive manual work and expert knowledge.

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 [15], is a fundamental topic to explore.

A set of tools built on top of the dataset is already available. For instance, social historians of the NLGIS project[28] query the endpoint to get historical census data and plot it in a map. Computational musicologists do research by linking the CEDAR dataset with their own historical singers database [8].

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 international 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.

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` and `http://lod.cedar-project.nl/cedar-mini/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 bulk downloads at `https://github.com/CEDAR-project/DataDump`.

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.

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

[28] See `http://www.nlgis.nl/`

[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] See `http://goo.gl/yfvUTl`

[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/`

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

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

Query	#Tables	#Cells
Inhabited houses in Zuid-Scharwoude in 1899	1	1
Occupied houses and living ships per municipality	59	80,032
Legally registered and present inhabitants per municipality	34	23,086
Houses under construction	47	4,478
Empty houses	59	34,834
Temporarily present inhabitants in ships	35	4,255
Temporarily present inhabitants per municipality	47	74,462
Temporarily absent inhabitants per municipality	34	37,044
Temporarily present inhabitants in wagons	13	426
Number of houses according to their type, from 1859 until 1920	59	136,768
Average	38.8	39,538.6

Table 6

Example queries automated by the integration process of the dataset. For each query, we detail the number of tables that users had to open and the number of cells they had to manipulate in order to reach a query answer. Unless stated, reference periods cover from 1859 until 1920. SPARQL translations of these queries can be found at `http://lod.cedar-project.nl/cedar/data.html`.

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

[5] Valeria Fionda and Giovanni Grasso. Linking Historical Data on the Web. In *Poster session, 13th International Semantic Web Conference (ISWC2014)*, 2014.

[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] Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, and Gerhard Weikum. YAGO2: A Spatially and Temporally Enhanced Knowledge Base from Wikipedia. *Artificial Intelligence*, 194(28):3161–3165, 2013.

[8] Berit Janssen, Albert Meroño-Peñuela, Ashkan Ashkpour, and Christophe Guéret. Tracking Down the Habitat of Folk Songs. *eHumanities eMagazine*, (4), 2015.

[9] Albert Meroño-Peñuela. LSD Dimensions: Use and Reuse of Linked Statistical Data. In *Proceedings of the 19th International Conference on Knowledge Engineering and Knowledge Management, EKAW 2014*, 2014.

[10] 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.

[11] Albert Meroño-Peñuela, Ashkan Ashkpour, Marieke van Erp, Kees Mandemakers, Leen Breure, Andrea Scharnhorst, Stefan Schlobach, and Frank van Harmelen. Semantic Technologies

[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.

for Historical Research: A Survey. *Semantic Web – Interoperability, Usability, Applicability*, 2014. on press.

[12] 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.

[13] Anisa Rula, Matteo Palmonari, Axel-Cyrille Ngonga Ngomo, Daniel Gerber, Jens Lehmann, and Lorenz Bühmann. Hybrid Acquisition of Temporal Scopes for RDF Data. In *Proc. of the Extended Semantic Web Conference 2014*, 2014.

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

[15] 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.