The dataLegend ecosystem for historical statistics

Hoekstra, Rinke; Meroño-Peñuela, Albert; Rijpma, Auke; Zijdeman, Richard; Ashkpour, Ashkan; Dentler, Kathrin; Zandhuis, Ivo; Rietveld, Laurens

published in
Journal of Web Semantics
2018

DOI (link to publisher)
10.1016/j.websem.2018.03.001

document version
Publisher's PDF, also known as Version of record

Web Semantics: Science, Services and Agents on the World Wide Web 50 (2018) 49–61

Contents lists available at ScienceDirect

Web Semantics: Science, Services and Agents on the World Wide Web

journal homepage: www.elsevier.com/locate/websem

The dataLegend ecosystem for historical statistics☆

Rinke Hoekstra [a,g], Albert Meroño-Peñuela [a,*], Auke Rijpma [b], Richard Zijdeman [c,h], Ashkan Ashkpour [c], Kathrin Dentler [d], Ivo Zandhuis [e], Laurens Rietveld [f]

[a] Department of Computer Science, Vrije Universiteit Amsterdam, Netherlands
[b] Utrecht University, Utrecht, Netherlands
[c] International Institute of Social History, KNAW, Amsterdam, Netherlands
[d] Academisch Medisch Centrum, Amsterdam, Netherlands
[e] Ivo Zandhuis Research & Consultancy, Haarlem, Netherlands
[f] Triply, Netherlands
[g] Faculty of Law, University of Amsterdam, Netherlands
[h] Faculty of Social Sciences, University of Stirling, United Kingdom

ARTICLE INFO

Article history:
Received 27 January 2017
Received in revised form 24 January 2018
Accepted 1 March 2018
Available online 10 March 2018

Keywords:
Digital humanities
Structured data
Linked data
QBer

ABSTRACT

The main promise of the digital humanities is the ability to perform scholarly studies at a much broader scale, and in a much more reusable fashion. The key enabler for such studies is the availability of sufficiently well described data. For the field of socio-economic history, data usually comes in a tabular form. Existing efforts to curate and publish datasets take a top-down approach and are focused on large collections, produce scarce metadata, require expertise for effective integration, provide poor user support while producing mappings, and present issues at data access. This paper presents the datalegend platform, which addresses the *long tail* of research data by catering for the needs of individual scholars. datalegend allows researchers to publish their (small) datasets, link them to existing vocabularies and other datasets, and thereby contribute to a growing collection of interlinked datasets. We present the architecture of datalegend; its core vocabularies and data; and QBer, an interactive, user supportive mapping generator and RDF converter. We evaluate our results by showing how our system facilitates use cases in socio-economic history.

1. Introduction

In a 2014 article in CACM, [1] describes digital humanities as a "movement and a push to apply the tools and methods of computing to the subject matter of the humanities". As the fuel of the computational method, the key enabler for digital humanities research is the availability of data in digital form. At the inauguration of the Center for Humanities and Technology (CHAT), José van Dijck, the president of the Royal Netherlands Academy of Arts and Sciences, characterizes progress in this field as the growing ability to tremendously increase the scale at which humanities research takes place, thereby allowing for much *broader* views on the subject matter [2]. Tackling this challenge for the digital humanities requires straightforward *transposition* of research queries from one humanities dataset to another, or even allow for direct *cross-dataset querying*. It is widely recognized that Linked Data technology is the most likely candidate to fill this gap [3–5].

However, current efforts to increase the availability and accessibility of these data – a requisite for such large scale humanities research – do not suffice. In particular, these solutions present pitfalls in five key aspects of data integration: the *scale and distribution* of data, which neglects the "long tail of research data" [6] of small datasets produced by individual researchers; data *publishing and archiving*, where limited provenance and metadata is provided along with the content; data *conversion to RDF*, which demands too much Semantic Web expert knowledge from users; schema and instance *mapping*, which offers poor user assistance; and data *access*, which does not ease the reusability of research queries. Hence, our aim is to address the limitations of current data publishing practice in the digital humanities, and socio-economic history in particular. While we acknowledge that, in isolation, these are Semantic Web problems in their own right, we consider this particular combination of challenges to best fit the digital humanities domain; and our proposed combination of solutions – the dataLegend ecosystem – the core novelty of our contribution.

☆ This is a significantly revised and extended version of a paper published as part of the ESWC 2016 post-conference proceedings, based on the WHISE 2016 workshop: Hoekstra et al. (2016).

* Corresponding author.
E-mail addresses: rinke.hoekstra@vu.nl (R. Hoekstra), albert.merono@vu.nl (A. Meroño-Peñuela), a.rijpma@uu.nl (A. Rijpma), richard.zijdeman@iisg.nl (R. Zijdeman), ashkan.ashkpour@iisg.nl (A. Ashkpour), k.dentler@amc.nl (K. Dentler), ivo@zandhuis.nl (I. Zandhuis), laurens@triply.cc (L. Rietveld).

https://doi.org/10.1016/j.websem.2018.03.001
1570-8268/

This paper presents the dataLegend platform and its key components, QBer[1] [7] and `grlc` [8]. dataLegend integrates a selection of large datasets from social history. QBer is a user-facing web application that allows *individual* researchers to upload, convert and link 'clean' data to existing datasets and vocabularies on the platform without compromising the detail and heterogeneity of the original data. Under the hood, we convert all data to RDF, following the principle of information hiding to not bother scholars with technical aspects. An inspector-view displays the result of the mappings – a growing network of interconnected datasets – in a visually appealing manner. The most important incentive is the ability to allow for transposing research queries across datasets, and the ability to perform cross-dataset querying. These features are enabled by `grlc`, a thin server that exposes Linked Data access methods (SPARQL, Linked Data fragments, dumps, RDFa, etc.) as RESTful APIs in a shareable and reusable manner. In general, the dataLegend ecosystem aims at addressing these issues by proposing a viable research ecosystem for Linked Humanities Data where all Linked Data remains under the hood, and combines:

- an automatic, highly-scalable, Linked Data-based data integration front-end, back-end, and pipeline for datasets in the long tail of research data;
- a profitable and indirect generation of provenance and metadata for better assessment and findability;
- COW,[2] a scalable, high-performance tabular-to-RDF conversion tool with no Semantic Web technical knowledge needed;
- QBer,[3] an interactive web application that allows non-technical scholars to interactively map, convert and publish their tabular data to RDF, and share their mappings;
- grlc,[4] an easy way of writing, sharing and reusing SPARQL queries as RESTful Web APIs

The rest of the paper is organized as follows. In Section 2 we survey the most important related work to the five core aspects addressed by our platform. In Section 3 we summarize the integrated datasets in dataLegend, together with novel concept schemes to describe *historical occupations* and *historical religions* in Linked Data. We describe the architecture of QBer, `grlc` and the other dataLegend systems in Section 4. In Section 5 we describe use cases that evaluate the ability of the dataLegend components to fulfill their requirements, and we conclude in Section 6.

2. Related work

The dataLegend ecosystem aims at addressing five key pitfalls of current data integration solutions for humanities scholars working with structured data. Concretely, these issues are found at the *scale and distribution* of data, data *publishing and archiving*, data *conversion to RDF*, schema and instance *mapping*, and data *access*.

Scale and distribution of data

Integration of tabular data sources is a key requirement in quantitative historical research [9]. Inherently structured, large in number, and scattered over the Web, tabular historical sources are the most promising type of data when it comes to using existing computational methods and tools. Examples are the North Atlantic Population Project (NAPP) [10], the Clio-Infra repository [11], and the Mosaic project.[5] However, data curation projects like these, which focus on collections of sufficient importance and size, are problematic in two ways. First, their scale is unsuited for the large volumes of important – but sometimes idiosyncratic – smaller datasets created by individual researchers: the long tail of research data [6]. Despite evidence that sharing research data results in higher citation rates [12], it is difficult and there is little incentive for researchers to make their data available with sufficiently rich, machine interpretable metadata [13]. And second, they enforce commitment to a shared standard of data harmonization that leads to loss of detail: the bigger a project is, the higher the cost of reconciling heterogeneity between large numbers of sources and their time, spatial, and demographic dimensions. In summary, data integration in the humanities typically focuses on *large, important collections*, and does not cater for the long tail of research data. An enforced standardization on a large scale means loss of flexibility, nuance and detail typically found in smaller, individually-maintained datasets. Similarly, data integration in Linked Data is also based on a large scale, *one-off effort* (e.g. in RISIS SMS [14],[6] DIVE[7] and WarSampo [15]), or is amongst pre-existing Linked Data sources (OpenPHACTS [16]).

Data publishing & archiving

Data publishing and archiving platforms such as EASY[8] (in the Netherlands), Dataverse[9] or commercial platforms such as Figshare,[10] Dryad[11] or Mendeley Data[12] aim to lower the threshold for data publishing, and cater for increasing institutional pressure to archive research data. However, the functionality of these platforms is limited with respect to the types of *provenance* and *content* metadata that can be associated with publications, and they do not offer the flexibility of the Linked Data paradigm [17]. This has a detrimental effect on both findability and reusability of research data, two of the key aspects of the FAIR guiding principles [18]. Large repositories that adhere to these principles, such as OpenPHACTS [16], the RISIS SMS [14] and – in the humanities – WarSampo [15], focus more on publishing Linked Data than on continuous ingestion of new, heterogeneous, non-linked data. Therefore, data archiving and publication platforms *do not provide support for generating rich metadata* or data conversion. And, apart from institutional pressure, there is little incentive to publish data in adherence to the FAIR principles.

From tabular data to RDF

In socio-economic history, a central challenge is to query data combined from multiple tabular sources: spreadsheets, databases and CSV files. The multiple benefits of Linked Data as a data integration method [3] encourage the representation of tabular sources as Linked Data.[13] CSV and HTML tables can be represented in RDF using CSV2RDF and DRETa [19,20]. For other tabular formats, like Microsoft Excel, Google Sheets, and tables encoded in JSON or XML, larger frameworks are needed, like OpenCube [21], Grafter [22], and the combination of OpenRefine and DERI's RDF plugin [23,24]. Several mapping languages exist that allow fully

1 A screencast of the system is available at https://vimeo.com/158153564.
2 See https://github.com/CLARIAH/COW.
3 See https://github.com/CLARIAH/qber.
4 See https://github.com/CLARIAH/grlc.
5 See https://www.clio-infra.eu and http://www.censusmosaic.org/.
6 See http://sms.risis.eu.
7 See http://diveproject.beeldengeluid.nl/.
8 See http://easy.dans.knaw.nl.
9 See http://dataverse.harvard.edu and http://dataverse.nl.
10 See http://figshare.com.
11 See http://datadryad.org.
12 See http://data.mendeley.com.
13 For a comprehensive list, see e.g. https://github.com/timrdf/csv2rdf4lod-automation/wiki and http://www.w3.org/wiki/ConverterToRdf.

automated conversion of tabular data to RDF. R2RML is the W3C specification for mapping a relational data model to RDF.[14] It can be used to present a database as virtual RDF graph, or convert it in its entirety and is supported by dedicated tools such as D2RQ[15] and triple stores such as Stardog.[16] RML [25] is a generalization of R2RML that extends it with the ability to define mappings for a variety of input formats (CSV, XML, JSON, HTML, ...).[17] R2RML and RML mappings are specified in RDF. The CSV on the Web (CSVW) standard[18] specifies schema files for CSV that use an extension of JSON-LD. The schema file allows for schema compliance checking, but it can also feed a straightforward conversion to RDF. Many of these converters operate under the assumption that one table row equals one observation (record). Datasets in social history, however, are often presented as multidimensional views that use other tabular layout features, such as hierarchical headers and spanning cells [26,9] (see Fig. 1). TabLinker [27] addresses these with a semi-automatic approach that represents multidimensional tables as RDF Data Cube using expert annotations [28]. The RMLEditor[19] [29] is an editor for the RML mapping language, and allows users to map legacy data to an RDF-based schema by bringing elements from a tabular data representation to the graph-based visualization of RDF. Similarly, the Karma tool for bringing structured data to the Semantic Web[20] combines a tabular representation of the source data with a graph-based schema view for mappings, assisting users with mapping suggestions. Crucially, both tools require users to have some familiarity with the graph data model of RDF, which is unnecessarily detrimental to the user experience of non-computer scientists. Hence, existing systems for mapping and converting data conversion are *targeted to tech-savvy users*. However, in our case, prospective users want to benefit from Linked Data but are unlikely to have any interest in the underlying technology.

Mappings between datasets

The systems discussed above map legacy data to standard RDF schemas. In socio-economic history, there are standardized code lists (HISCO for historical occupations [30], SDMX COG on sex, etc.) that play a crucial role in connecting datasets at the *instance* level. Work in ontology and vocabulary alignment, as in the OAEI,[21] aim to perform *automatic* alignments. Given the very specific (historic) meaning of terms in our datasets, these techniques are likely to be error-prone, hard to optimize due to the heterogeneity of socio-historical data, and unacceptable to scholars. Interactive alignment tools, such as Amalgame [31] are more promising, but treat the alignment task in isolation rather than as part of the data publishing process. Anzo for Excel[22] is an extension for Microsoft Excel for mapping spreadsheet data to ontologies. RightField[23] allows for selecting terms from an ontology from within Excel spreadsheets to annotate experiment results, but relies on a predefined template. TopBraid Composer[24] uses separate files for capturing mappings. Similarly, TabLinker [27] mappings are driven by Excel worksheets, but the information that is coded against has to be found and entered manually. In general, all these mapping tools focus on *generating* mappings; isolate the mapping task from the data; and do not provide users with sufficient support for selecting the right URI to map against.

Access to data

Most humanities data integration projects culminate in a website where *subsets* of the data can be *downloaded* or *visualized*, but cannot be programmatically accessed, isolating the data from efforts to cross-query over multiple datasets. In Linked Data, this is typically solved by letting clients query the data via SPARQL [32]. The problem with this approach is twofold: first, users without knowledge of SPARQL need to learn it or seek help from the community; and second, SPARQL queries can rarely be reused in more than one dataset. Current solutions like OpenPHACTS [33] and BASIL [34] propose to use RESTful APIs on top of SPARQL for better automation, but do not address ease of access nor reusability.

3. Datasets and concept schemes

3.1. Datasets

In Section 2, we made the case against starting a new, large scale harmonization effort. However, datalegend does include a number of core datasets from social, demographic, and economic history. There are two reasons for this: first, it *incentivizes* researchers to use the datalegend platform. It should be the place to go for researchers who want easy access to a well-documented and clean dataset. The second reason is that we want to provide a foundation for users to link their own datasets to. This approach distinguishes datalegend from other data integration efforts: individual users can contribute their own links through the system, leading to a growing, heterogeneous but interconnected web of linked statistical data.

For example, if a researcher has a historical census-dataset, she can link it up with other data on the same region or a similarly structured dataset in another region to facilitate comparisons. Especially at the start of this initiative, it is important that datasets are present to provide a full experience to users. Where datasets already contain (implicit) links we also convert these in a standardized fashion. We target data at three levels of observations:

macro data about countries or regions,
micro data about individuals or households, and
meso data that fall in between these categories, such as on worker unions.

Linking across these levels provides unprecedented opportunities for multilevel research designs. It can reveal how processes operating at the country level (say, income inequality) influence decisions made at the individual level (say, to enroll in school).

Secondly, we distinguish between *cross-sectional* and *longitudinal* data. Cross-sectional data concern one point in time for each subject (and individual, country, etc.), while longitudinal data contains multiple observations over time for each subject. Research designs that employ longitudinal data allow for far more sophisticated analyses. However, the survival of source material means that in many cases cross-sectional data is the only kind available. A third data type is *intergenerational* data: longitudinal micro-data that follows individuals from one generation to the next (i.e. parents and children).

We selected datasets from the fields of social, demographic, and economic history. First, there is the conversion of a large body of historical macro-data previously collected and harmonized in the Clio-Infra project. Its scope is to provide useful data for studying

14 See http://www.w3.org/TR/r2rml/.
15 See http://d2rq.org.
16 See http://stardog.com.
17 See http://rml.io.
18 See http://www.w3.org/TR/csvw/.
19 See http://rml.io.
20 See https://github.com/usc-isi-i2/Web-Karma.
21 The Ontology Alignment Evaluation Initiative, see oaei.ontologymatching.org/.
22 https://www.w3.org/2001/sw/wiki/Anzo.
23 https://www.sysmo-db.org/rightfield.
24 See http://www.topquadrant.com and https://www.w3.org/2001/sw/wiki/TopBraid.

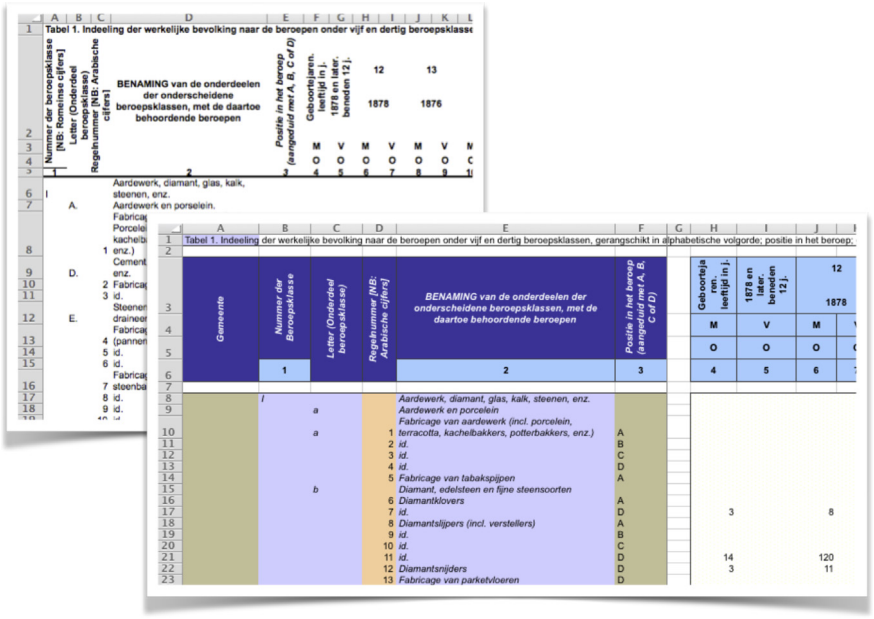

Fig. 1. TabLinker annotation of eccentric spreadsheet data in Excel. Experts can annotate different components of multidimensional tables, such as dimensions and values, with various styles for further processing.

global issues, such as inequality and development. Thus creating a new fruitful ground for testing hypotheses of new economic theories, as well as quantifying the past. The data currently available include 76 indicators that cover the fields of demography, institutions, human capital, production, agriculture, prices and wages, gender equality, labor relations, environment, finance, and national accounts. About 210 countries and territories are covered, with the origin of the data time span reaching as far back as 1500 for several series, and mid-19th century for most.

By including it we hope to facilitate the accessibility and dissemination of this dataset and make sure other data can be linked to it. A substantial effort is also made at curating and converting micro-datasets as these can be readily augmented by macro-datasets. In social and economic history, the past decades have seen a move towards data about individuals and households as this is where the decisions that interest economic and social historians are taken. Another focus is on data from the pre-industrial era. Knowledge about this period is more limited: datasets are rare and often not easy to use. We therefore estimate that the returns on investing in these datasets is high. Table 1 lists datasets that have been, or will be included as part of the datalegend platform.

Some of these datasets have specific licenses that constrain their (re)distribution, which directly affects dataLegend. We address this with three specific components of the platform: SPARQL access, Druid roles and authorization, and VoID metadata descriptions (see Section 4). In SPARQL access, we set two different levels of data access, *public* and *private*, which we set accordingly to the requirements of licenses. Only users with authorization and permissions over the original dataset can access Linked Data of e.g. datasets with no-redistribution policies. In Druid roles and authorization, we mimic GitHub's *roles* and *organizations*[25] to control permissions on users, specifically concerning licensing. Finally, the inclusion of VoID terminology in the nanopublication graph (see Section 4.2) contains triples conveying the licensing information to final users.

[25] See https://github.com/blog/674-introducing-organizations and https://help.github.com/articles/repository-permission-levels-for-an-organization/.

3.2. Essential concept schemes

Concept schemes and classification systems are key in dataLegend, since they allow researchers to semantically describe the contents of their datasets by reusing existing terminologies, thus enabling the discovery of common relations in the data of others. However, the effort to maintain well structured, harmonized, comparable, reusable, and provenance-aware classification systems for quantitative humanities datasets is not new, and has a long tradition in the social sciences. The Data Documentation Initiative [35] (DDI) is an international standard for describing statistical data files and designing codebooks of classifications. The Statistical Data and Metadata eXchange [36] (SDMX) aims at standardizing the mechanisms and processes for the exchange of statistical data and metadata among international organizations, through (among other standards) code lists in Content-Oriented Guidelines (COG). The Generic Statistical Information Model [37] (GSIM) follows a similar approach to organize statistical information objects. Following these, the Consortium of European Social Science Data Archives [38] (CESSDA) encourages and enforces the use of common classifications. These efforts excel in standardizing, structuring and preserving classification systems, although they face the remarkable challenge of achieving interoperability between them.

It is not the goal of dataLegend to compete with these standards. On the contrary, our proposed Linked Data-based approach rather supports an easy means for individual researchers to plug their datasets into their wealth of variables and codes, making it easy to reach broader classification systems and harmonize data over time and space. Moreover, dataLegend attempts to address the still manual process of aligning and mapping all these systems among themselves, by providing a platform where users can convert and map these systems in a Linked Data space. In this regard, some of these classification systems, like those provided by SDMX, have been already represented as Linked Data and are available in the system [28].

Adding to these, in this Section we describe two pioneering efforts by dataLegend for bringing historical vocabularies to the Semantic Web: the Historical International Standard Classification

Table 1
Datasets that are included, and will or could technically ([a]) be distributed in dataLegend.

Clio-Infra	https://www.clio-infra.eu
Thombos	http://www.sciencedirect.com/science/article/pii/S1081602X11000625
HSN	https://socialhistory.org/en/hsn/index
Campop	http://www.campop.geog.cam.ac.uk/datasets/
Bagdad to London	http://www.cgeh.nl/urbanisation-hub-clio-infra-database-urban-settlement-sizes-1500-2000, http://dx.doi.org/10.7910/DVN/24747, http://www.mitpressjournals.org/doi/abs/10.1162/REST_a_00284
HISCO	https://socialhistory.org/ru/projects/hisco-history-work
HISCAM	http://www.hiscam.org/
CEDAR	https://www.cedar-project.nl
Strikes	https://datasets.socialhistory.org/dataverse/Aggregate
Henry-Fleury	http://www.jstor.org/stable/20023818
NAPP[a]	https://www.nappdata.org/napp/
LINKS	https://socialhistory.org/en/hsn/linking-system-historical-family-reconstruction-links
CMPGD	http://www.icpsr.umich.edu/icpsrweb/ICPSR/series/00265
Opgaafrollen	http://dspace.library.uu.nl/handle/1874/257094
Mosaic	http://censusmosaic.org
Catasto	http://www.disc.wisc.edu/archive/catasto/index.html
Treechecker	http://www.treechecker.net

[a] For these datasets we have mimicked their current distribution model of a single-user license, but we are not redistributing them until given authorization.

of Occupations (HISCO) [30], and the Linked International Classification for Religions (LICR).[26] A key challenge of combining datasets with similar dimensions (variables) is to cope with ontological differences, especially over long periods of time. Historians, often confronted with this issue, solve this problem via knowledge driven data harmonization practices [39], usually resulting in aggregation of data to the smallest common denominator. In addition, historians can only take advantage of the possibilities provided by Linked Data if they are provided with domain specific knowledge in the form of coded dimensions.

HISCO deals with the representation of work from the past. Various historical datasets contain information on occupational titles, an important indicator for the study of social and economic inequality. However, occupational titles change meaning over time and the meaning may differ between different languages. HISCO resolves the issue of incomparability of occupations by looking at similarity in the activities undertaken in an occupation (e.g. lifting, planning, traveling). It harmonizes occupations between different languages and over time. To do so, it assigns to groups of very similar occupations a *unique HISCO code* from a *hierarchical tree*,[27] that conveys the semantics of that occupation while keeping the link with the original language-dependent string.[28] HISCO is linked to various standardized classifications about status and stratification, like HISCAM [40]. Users with data that can be linked to HISCO, automatically obtain status information.

Religion is a key variable in social research appearing in a large number of datasets, whether historical or contemporary. In order to stimulate more similar efforts within our community we have created a classification system for religious denominations, called the Linked International Classification for Religions (LICR). The LICR classification is built from a bottom-up and data driven approach, based on data found in NAPP [10], IPUMS (Integrated Public Use Microdata) and the HL7 (Health Level Seven) classifications. LICR links these classifications together, and provides more detail compared to the aforementioned systems separately. For example the IPUMS[29] classification provides 8 sub-denominations for Islamic religions but only one main group for Jewish religions, while the NAPP[30] classification provides great detail for Jewish religions but none for the Muslim denomination. We combine the level of detail from these systems and create new standards codes in LICR.[31] The inter-classification mapping allows users to interchangeably link their data between different systems. In addition, LICR provides rich textual descriptions of religious denominations by linking these to DBpedia. Doing so, LICR is the first Linked Data classification for religious denominations adhering to the principles of a five-star Linked Data Classification. Fig. 2 shows the LICR definition of 'Confucianism', linked to NAPP, IPUMS and HL7.

4. The dataLegend platform

The dataLegend platform serves as the central hub against which all user facing services are built. It consists of three layers (see Fig. 3): a data *ingestion* layer, a data *publication* layer, and a central data *management* layer. The data ingestion layer (see Section 4.1 and the three bottom boxes in Fig. 3) is responsible for taking data from its native form and making them ready for publication as Linked Data. This takes place in two parallel workflows that are tailored for different requirements: *COW*, and *QBer*. Once this is done, Linked Data representations of the original data sources are stored as datasets in the data management layer. The data management layer (see Section 4.2 and the central boxes in Fig. 3) is responsible for storing, updating the state, and granting access to the Linked Data datasets in dataLegend. It takes the output produced by the ingestion layer and uses it to perform a schema- or a mapping-based *conversion*. Then, it stores the produced RDF files in two systems: *GitLab*, which curates fine-grained versions of the datasets in their original and RDF formats; and *Druid*, which stores triples in a highly performant and scalable manner using HDT technology [41,42]. Finally, the Linked Data datasets are synced to a triplestore to facilitate their querying via SPARQL. The data publication layer (see Section 4.3 and the top orange boxes in Fig. 3) is responsible for enabling external access to data in dataLegend in three different ways: via a SPARQL endpoint, Linked Data APIs, and Linked Data Fragments. The additional Data Management API connects with the data management layer and

[26] See https://datasets.socialhistory.org/dataverse/LICR/ and http://data.datalegend.net/doc/resource/LICR/vocabulary.

[27] See http://historyofwork.iisg.nl/major.php.

[28] See e.g. https://goo.gl/eSo1pT for all occupational titles of HISCO code 54010, *domestic worker*.

[29] See https://international.ipums.org/international-action/variables/RELIGION.

[30] See https://www.nappdata.org/napp-action/variables/RELIGION.

[31] See e.g. http://data.datalegend.net/doc/resource/LICR/5374 for the denomination *Seventh Day Baptist*.

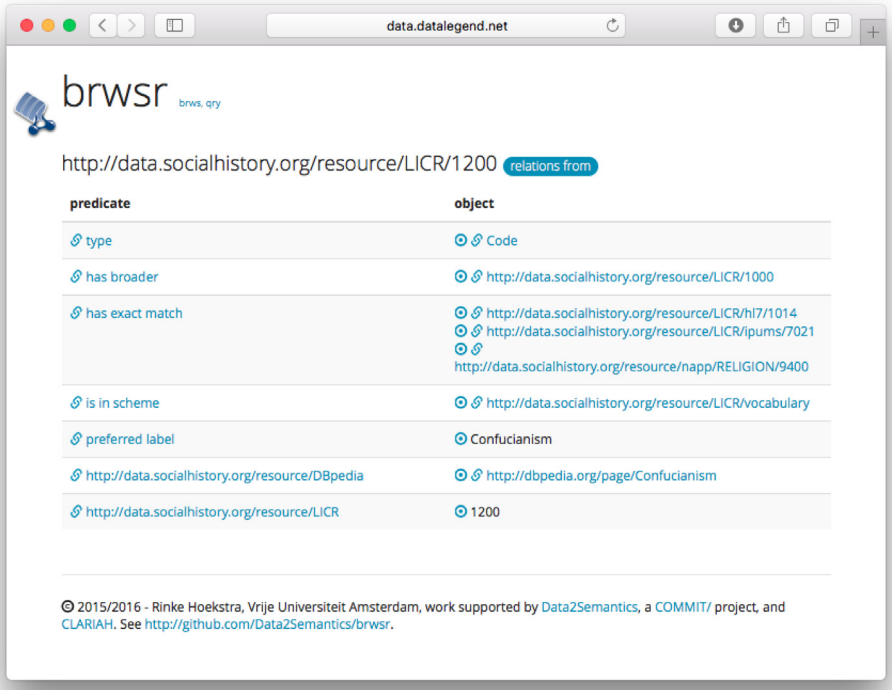

Fig. 2. A brwsr interface showing the LICR definition of 'Confucianism'.

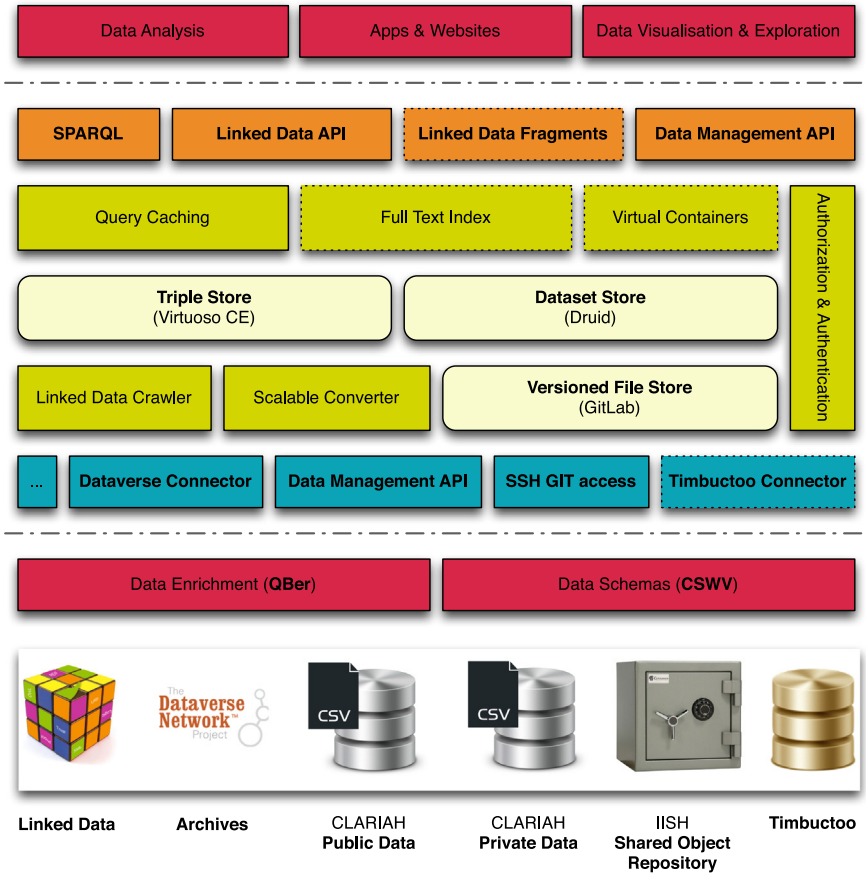

Fig. 3. Architecture of the dataLegend platform. From bottom to top: data sources; data ingestion layer (QBer, COW, and various connectors); data management layer (crawler, converter, file and triple stores, caches and indexes); and data publication layer (SPARQL, API, Linked Data Fragments).

allows other applications of the platform to upload, update, and remove datasets at their convenience. The rest of this section describes the components in these layers, the underlying design decisions, and their interaction.

4.1. Data ingestion

dataLegend has two different means for ingesting structured data: COW and QBer. COW is a batch converter and is intended for users that: (1) own datasets serialized as (sets of large) CSV files; (2) want to publish those following the CSVW W3C specification; (3) are familiar with executing command-line scripts and editing JSON files. In this scenario, users perform the following steps: create CSVW mappings and push the JSON mappings and CSV files into GitLab. GitLab webhooks then detect the newly pushed content and perform two actions: *(a)* execute the scalable converter, which uses the JSON mapping files and the original CSV files to create a CSVW RDF compliant conversion in NQuads notation; *(b)* push the NQuad files to the triplestore.

QBer, on the other hand, aims at users that: (1) own datasets serialized as CSV, TSV or Excel files; (2) want to publish those on the dataLegend platform without prior knowledge of Semantic Web technology; and (3) require a user interface to declare mappings between strings of the original files and Linked Data resources. This scenario is cared for by the QBer web application.[32] With QBer, users can use a graphical interface to browse their files, import them, indicate the existing resources they want to use for mapping (e.g. SDMX dimensions and SKOS concept schemes from harmonized sources or contributed by peers), and map original values to these LOD resources.

Using QBer consists of interacting with three main views: the *welcome screen*, the *mapping screen*, and the *inspector*. In the welcome screen, users first authenticate with OAuth compatible services (e.g. Google accounts), and then select a raw dataset to work with. Datasets can be selected directly from the dataLegend versioned file store, uploaded from Dropbox, or imported from a Dataverse collection by providing a DOI.

Once a dataset is loaded, QBer displays the mapping screen (Fig. 4). This screen is divided into the *variables sidebar* (left) and the *variable panel* (right). The sidebar allows the user to search and select a variable (i.e. column) from the dataset. Once the user clicks on one variable, the variable panel will show that variable's details: the *variable category*, the *variable metadata*, and the value *frequency table*. A next version of QBer will present the data in a more familiar table structure.

We distinguish between three *variable categories*: *coded*, *identifier* and *other*. Values for coded variables are mapped to corresponding concepts (skos:Concept) within a skos:ConceptScheme, which establishes all possible values the variable can take. If the variable is of type *identifier*, its values are mapped to dataset specific minted URIs. Finally, the values of variables of type *other* are mapped to literals instead of URIs. The 'Community' button gives access to all known predefined datacube dimensions. These come from LSD Dimensions, an index of dimensions used in Data Structure Definitions of RDF Data Cubes on the Web [43] and from datasets previously processed by QBer that now reside on the platform.

The *frequency table* panel has three purposes. First, it allows for quick inspection of the distribution of all values of the selected variable, by displaying their frequency. Second, if the variable type is "coded", it lets the user map the default minted URI for the chosen value to any skos:Concept within the selected skos:ConceptScheme in the variable metadata panel. QBer also has a batch mapping mode that prompts the user to map all values of the variable interactively. Third, if the variable type is "other", users can specify their own literal-to-literal transformations by providing their own transformation functions; this is useful e.g. if the original data needs to be expressed in different units of measure, or if strings need a systematic treatment. Finally, the panel shows the current mappings for values of the selected variable.

Mappings can be materialized in two ways. Users can click on *Save* in the navigation bar, which stores the current mapping status of all variables in their local cache. Clicking on *Submit* sends the mappings to the data management API, which converts the source file, and integrates them with other datasets in the hub.

4.2. Data management

Storage

The data management layer revolves around three core storage components. A GitLab instance[33] that provides a service layer around a Git-based versioned file store. We use GitLab to store the original datasets (in CSV, Excel or other formats) alongside a mapping file (see 4.1) and an NQuads serialization of their RDF representation. Secondly, a Virtuoso Open Source triple store[34] hosts all (latest) versions of RDF representations of the datasets present in GitLab.

This solution does not scale with large data volumes. For this reason, dataLegend's custom built dataset store Druid (see Fig. 3, *dataset store*) provides large scale storage and management of all datasets using HDT files [41,42].[35] Druid is able to spawn custom triple stores automatically, by selecting datasets from within a simple web interface and using Docker containers. In the long run, we expect Druid to replace the current Virtuoso triple store. The Druid backend is an evolution of the LOD Laundromat[36] storage layer [44] that allows us to store named graphs across multiple HDT files (HDT is restricted to triples). This is essential because dataLegend adopts the Nanopublication vocabulary [45] for separating publication information, provenance and the asserted data across multiple named graphs.

Conversion

The *scalable converter* component, aka COW,[37] takes care of the conversion of all non-RDF data provided to the system. It allows for parallel processing of very large CSV files by running a converter process for line-based chunks of the file. Conversion is driven by either a *mapping* or a *schema* file that instructs the converter how to interpret the tabular data.

Mapping-based conversion

The output of a QBer session is a custom JSON document that specifies the *mappings* used to construct Linked Data version of a source document. It specifies for each column what type of variable it represents (a dimension property, a coded property or a measure), and for specific values within those columns what URIs they should be mapped to (e.g. from standard vocabularies present in datalegend, see Section 3.2).

As mentioned before, the dataset is represented in RDF as a Nanopublication [45] with provenance metadata in PROV,[38] where the assertion-graph is an RDF Data Cube representation of the data [28].[39] This RDF representation is a verbatim conversion of

32 See http://qber.datalegend.net.

33 See http://gitlab.org.

34 See https://github.com/openlink/virtuoso-opensource.

35 See http://druid.datalegend.net.

36 See http://lodlaundromat.org.

37 See https://github.com/CLARIAH/COW.

38 See https://www.w3.org/TR/prov-o/.

39 See https://www.w3.org/TR/vocab-data-cube/.

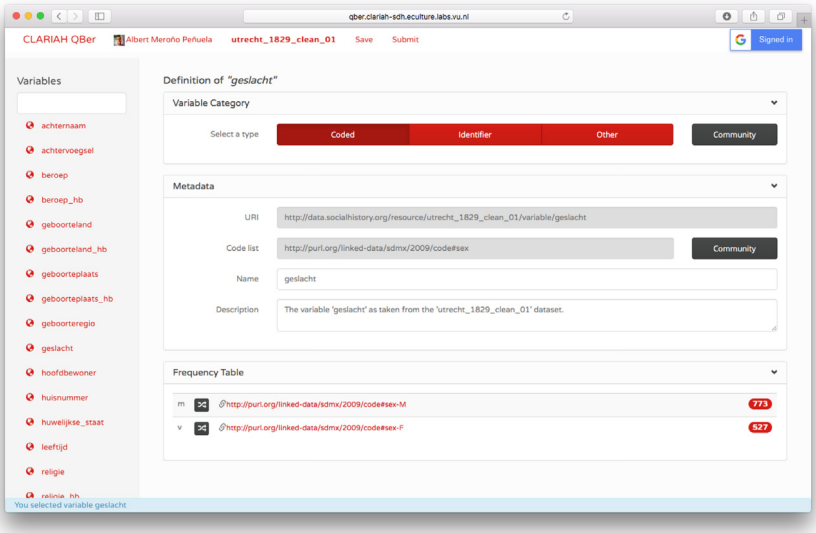

Fig. 4. Variable mapping screen of QBer with the variable 'geslacht' (sex) selected.

the original values of each cell; mappings between the original values and pre-existing vocabularies are explicitly represented using SKOS mapping relations. This scheme allows for the co-existence of alternative interpretations (by different scholars) of the data, thus overcoming the standardization-limitation alluded to in Section 2.

Schema-based conversion

The datasets and vocabularies discussed in Section 3 are typically provided as collections of CSV dumps from relational database tables. Unfortunately, statistical datasets often use specific codes to indicate e.g. unavailability of data, imprecise information, etc., typically captured in non machine-interpretable documents annexed to the data. In dataLegend, we capture this information as CSVW schema files that can be semi-automatically built. This allows to credit the original authors of the data, and drive conversion to Linked Data through the csv2rdf specification.[40] In order to allow more expressive formatting patterns (i.e. numeric codes ('01110') for standard terms represented as integers (1110); concatenations of values; multiple namespaces for values of one column; etc.) we extend the CSVW standard with the Jinja2 templating language,[41] of which CSVW is a subset. For instance, the pattern "This is my {{name|title}}" results in the {{name|title}} part to be replaced with the value of the cell in the 'name' column of the current row in title case. This means that without additional coding, we can accommodate a large portion of the limitations of CSVW.[42]

4.3. Data publication

The traditional means to access Linked Data is through a SPARQL endpoint. Writing SPARQL queries is widely recognized as a difficult task, especially among users that have no training in Linked Data or Semantic Web related technologies. However, this should not impede access to the integration-favorable dataspace of Linked Data for these users. To address this, and to automatize the process of building data APIs that use SPARQL queries, we generate a Linked Data API on the fly, using the `grlc` middleware [8].[43] `grlc` uses SPARQL queries stored in GitHub[44] and their logical repository-based organization to generate an OpenAPI[45] compliant RESTful API. This means that queries can be written (and further reused) in a collaborative manner. Fig. 5 shows a screenshot of the SwaggerUI website generated from an API spec based on queries stored in our GitHub repository. When users execute these simple API calls, `grlc` executes their equivalent SPARQL queries under the hood. This means that users do not need to deal with the writing, storage, execution, and result formatting of these queries; they only interact with a regular Web API, where each call name is identified by a URI. Moreover, users can exchange these URIs to share research questions over data, effectively making data actionable.

All Linked Data in dataLegend is published as dereferenceable Cool URIs[46] through the `brwsr`[47] utility. `brwsr` is a lightweight Linked Data browser similar to Pubby that implements content negotiation to serve representations of resources both as HTML, and as RDF/XML, Turtle and JSON-LD. It can connect to multiple SPARQL endpoints, ingest externally hosted Linked Data, browse across multiple namespaces, and optionally serves data from local files. The web interface of `brwsr` calls the http://preflabel.org service to retrieve preferred labels for known resources.

The *Inspector*, shown in Fig. 6, builds on top of the SPARQL endpoint and allows users to explore the contents of the dataLegend platform. The visualization shows a graph of nodes and edges, with different icons representing different node types for *users*, *datasets*, *data structure definitions*, and *dimensions*. A Data Structure Definition "defines the structure of one or more datasets. In particular, it defines the dimensions, attributes and measures used in the dataset along with qualifying information such as ordering of dimensions and whether attributes are required or optional" [28]. A dimension is essentially a variable used in a dataset. Users can interact with the inspector in several ways (hovering to show metadata; zooming and panning; clicking for browsing with `brwsr`). This gives publishers of datasets (i.e. humanities scholars) an intuitive feel of how the data is interconnected, and works as a reward mechanism for scholars who spent more effort in annotating their data.

40 See https://www.w3.org/TR/csv2rdf/.

41 See http://jinja.pocoo.org.

42 For further detail, see http://csvw-converter.readthedocs.io/en/latest/#the-schema and https://github.com/CLARIAH/iribaker.

43 See https://github.com/CLARIAH/grlc. A public instance of `grlc` is available at http://grlc.io.

44 See example queries at https://github.com/CLARIAH/wp4-queries/.

45 See https://www.openapis.org/.

46 See https://www.w3.org/TR/cooluris/.

47 See https://github.com/Data2Semantics/brwsr.

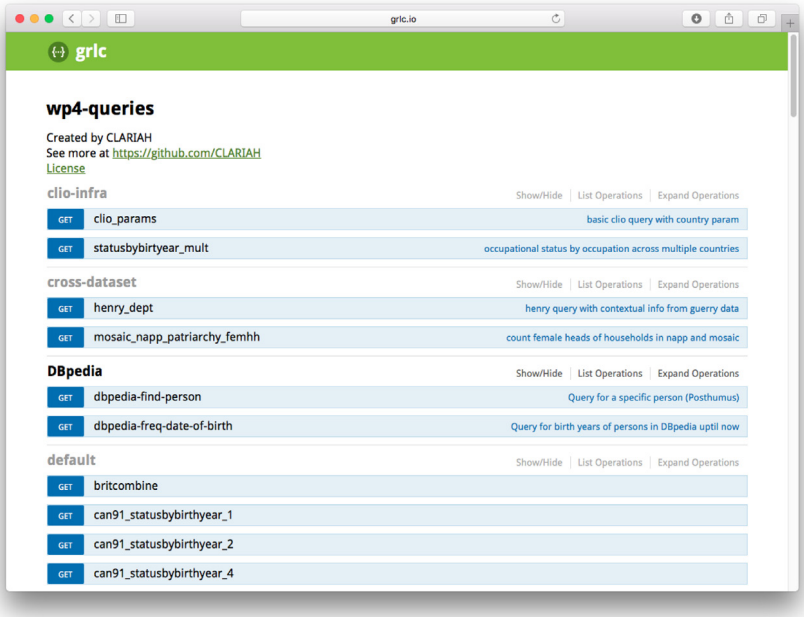

Fig. 5. A SwaggerUI website generated by `grlc` from OpenAPI specs based on SPARQL queries hosted in our GitHub repository at https://github.com/CLARIAH/wp4-queries/.

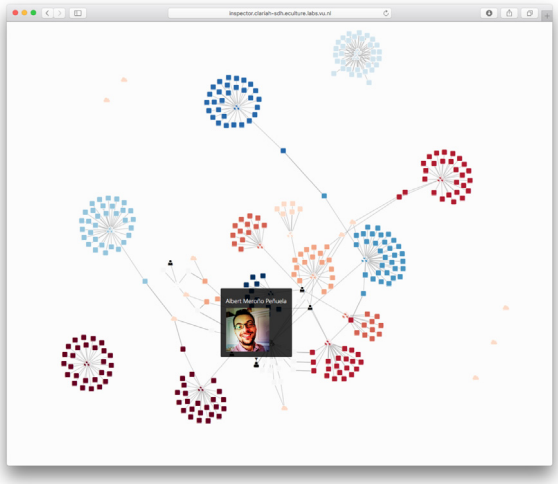

Fig. 6. The inspector view over the datasets currently in dataLegend.

5. Evaluation

In this section, we evaluate the dataLegend ecosystem by means of four use cases in socio-economic history research. The first two use cases investigate the question of whether dataLegend indeed allows research to be carried out on a broader scale, qualitatively comparing old and new workflows. The third use case measures quantitatively the speed up gain at answering domain-expert queries with respect to the absence of the ecosystem, highlighting added value – reusability and augmentation – that is exclusively due to Linked Data. In the last use case, we show how dataLegend enables query reusability by transposing queries that were originally built to answer a research question aimed at a different dataset.

5.1. Use Case 1: Early life conditions

Economic and social history takes questions and methods from the social sciences to the historical record. An important line of research in social and economic history focuses on the determinants of historical inequality. One hypothesis here is that prenatal [46] and early-life conditions [47] have a strong impact on socioeconomic and health outcomes later in life. A recent study on the United States found that people born in the years and states hit hardest during the Great Depression of the 1930s had lower incomes and higher work disability rates in 1970 and 1980 [48]. This study inspired this use case.

Most studies on the impact of early life conditions are case studies of single countries. Therefore, the extent to which results can be generalized – their external validity – is difficult to establish (e.g., differing impact of early life conditions in rich and poor countries). Moreover, historical data is often idiosyncratic. This means that dataset-specific characteristics such as sampling and variable coding schemes might influence the results (see Section 2).

In this use case, we explore the relation between economic conditions in individuals' birth year and occupational status in the historical census records of Canada and Sweden in 1891. In many cases it would be necessary to link the two census datasets so that they can be queried in the same way. Here, however, we use two harmonized datasets from the North Atlantic Population Project (NAPP, Canada 1891 [49] and Sweden 1890 [50]). We emphasize here that we use this data internally and for experimental purposes as this data is not meant for redistribution. The data is therefore only available to the researchers of this use case, using methods outlined in Sections 3.1 and 4.2. Economic conditions are measured using historical GDP per capita figures from the Clio-Infra repository [11]. Because our outcome is occupational status, we have to enrich the occupations in the census with occupational codes and a status scheme. Because the NAPP-project uses an occupational classification that provides no internationally comparable occupational status scores, we have to map their occupational codes to

the HISCO system, so that we can use the HISCAM cross-nationally comparable occupational status scheme [30,40].[48]

In general terms, the data requirements are typical of recent trends in large database usage in economic and social history:

(1) the primary unit of analysis is the individual (microdata);
(2) a large number of observations is analyzed;
(3) multiple micro-datasets are analyzed;
(4) microlevel observations are linked to macro-level data through the dimensions time and geographical area;
(5) qualitative data is encoded to extract more information from it.

5.1.1. Current workflow

The traditional workflow to do this would include the following steps. First, the researcher has to find and download the datasets from multiple repositories. The datasets, which come in various formats, then have to be opened, and, if necessary, the variables have to be renamed, cleaned, and re-encoded to be able to join them with other datasets. We can rely on previous cleaning and harmonization efforts of the NAPP project, but in many other situations the researcher would have to do this manually. Finally, the joined data has to be saved in a format that can be used by a statistical program.

5.1.2. New workflow

Using QBer and datalegend, the workflow is as follows. Linked-data tools are used to discover data on the platform. In our case, we used the Inspector, a linked data browser[49] and exploratory SPARQL queries. The Inspector provides a simple overview of all datasets in the CSHD.[50] Note that to discover datasets and especially linked datasets, it is necessary that someone uploaded the datasets and created the links in the first place, for example by linking datasets to a common vocabulary. While it is unavoidable that someone has to do this at some point, the idea behind datalegend is that if it is done once, the results can be re-used by other researchers.

The next step is to specify queries against the data, and store them on GitHub. The result sets that these queries produce against datalegend are then used to create the dataset that is to be analyzed. The web interface of `grlc` can be used to explore the parameters one can use for each query: grlc populates pull-down menus with potential bindings for each variable. The straightforward HTTP interface, combined with a CSV return format, allows for direct integration in statistical environments such as R.

5.2. Use Case 2: Railway strike

The second use case takes the form of a user study. It is about the "Dwarsliggers"[51] dataset by Ivo Zandhuis that collects data pertaining to a solidarity strike at the maintenance workshop of the Holland Railway Company (*Hollandsche IJzeren Spoorweg-Maatschappij*), in the Dutch city of Haarlem in 1903. From a sociological perspective, strikes are of interest for research on social cohesion as they deal both with the question of when and why people live peaceful together (even when in disagreement) and the question of how collective action is successfully organized, a prerequisite for a successful strike. The Dwarsliggers dataset is one of the few historical cases where data on strike behavior is available at the *individual* level.

The creation and use of this dataset is exemplary of the workflow of small to medium quantitative historical research projects in the sense that it relies on multiple data sources that need to be connected in order to answer the research questions. We briefly discuss this workflow, and then show the impact that QBer and datalegend have.

5.2.1. Current workflow

Zandhuis' current workflow is very similar to the one reported in the first use case. He first digitized the main dataset on the strike behavior of employees at the maintenance workshop of the railway company ($N = 1163$). Next, he gathered data from multiple sources in which these employees also appear, adding individual characteristics that explain strike behavior. For example, he derived family situations from the Dutch civil registers, and the economic position from tax registers, resulting in a separate dataset per source. Next, he inserted these datasets into a SQL database. In order to derive a concise subset to analyze his research questions, using e.g. QGIS, Gephi or R, he wrote SQL queries to extract the relevant information. These queries are usually added as an appendix to his research papers.

5.2.2. New workflow

In collaboration with Zandhuis, we revisited this workflow using QBer. Zandhuis, as most historians, uses spreadsheets to enter data, and uses a specific layout to enhance the speed and quality of data entry. The first step was to convert the data to a collection of .csv files. This is just a temporary limitation, as datalegend is not necessarily restricted to CSV files. It uses the Python Pandas library[52] for loading tabular files into a data frame.

The second step involves visiting each data file in turn, and linking the data to vocabularies and through them to other datasources. Data about the past often comes with a wide variety of potential values for a single variable. Religion, for example, can have dozens of different labels as new religions came about and old religions disappeared. As described in Section 4.1, QBer provides access to a large range of such classifications, basically all those available in the Linked Data cloud and datalegend. For example, QBer provides all occupation concepts from the HISCO classification used in the first use case [30].

Researchers can use occupational labels to get the correct codes from the latest version of this classification and, eventually, concepts linked to it. QBer however also shows the results of earlier coding efforts, so that historians can benefit from these (e.g. another dataset may have the same literal value already mapped to as HISCO code).

This step is new compared to Zandhuis' original workflow. The linking of occupational labels now enables him to combine an employee with his social status (HISCAM). This allows him to directly include a new, relevant, aspect in his study. Moreover, since QBer makes coding decisions explicit, they can be made subject to the same peer review procedure used to assess the quality of a research paper. In dataLegend, original values of the dataset and the mapped codings (potentially by different researchers) live side-by-side. Thus QBer adds to the ease of use in coding variables, increases flexibility by allowing for multiple interpretations, and allows for more rigorous evaluation of coding efforts. The inspector graph of Fig. 6 depicts the result of the new workflow.

The third step was then to query the datasets in order to retrieve the subset of data needed for analysis. As in the first use case, we design SPARQL queries that, when stored on GitHub, can be directly executed through the grlc API. This makes replication of research

[48] https://github.com/rlzijdeman/o-clack and http://www.camsis.stir.ac.uk/hiscam/.

[49] https://github.com/Data2Semantics/brwsr.

[50] Currently at http://inspector.datalegend.net/overview.

[51] In Dutch, a "dwarsligger" can mean both a railroad tie, and an obstructive person.

[52] See http://pandas.pydata.org.

much easier: rather than including the query as an appendix of a research paper, the query is now a first order citizen and can even be applied to other datasets that use the same mappings. Again, through the API, these queries can easily be accessed from within R, in order to perform statistical analysis. Indeed, the grlc API is convenient, but it is a lot to ask non-computer science researchers to design SPARQL queries. However, as we progress, we expect to be able to identify a collection of standard SPARQL query templates that we can expose in this manner (see also [51]).

To illustrate this, consider that since the Dwarsliggers collection contains multiple datasets on the same individuals at the same point in time, there are multiple observations of the same characteristics (e.g. age, gender, occupation, religion). However, the sources differ in accuracy. For example, measuring marital status is one of the key aims of the civil registry, while personnel files may contain information on marital status, but it is not of a key concern for a company to get this measurement right. By having all datasets mapped to vocabularies through QBer and having the queries stored in GitHub and executed by grlc, each query can readily be repeated using different sources on the same variables. This is useful as a robustness check of the analysis or even be used in what historians refer to as a 'source criticism' (a reflection of the quality and usefulness of a source). This, again, is similar to the first use case, but it emphasizes an additional role for the queries as so-called 'edit rules' [52]. dataLegend automatically generates data provenance and publishes it next to each dataset as nanopublications, which make source criticism more easily accountable.

To conclude, this use case shows that the QBer tool and related infrastructure provides detailed insight in how the data is organized, linked and analyzed. Furthermore, the data can be queried live. This ensures reusable research *activities*; not just reusable *data*.

5.3. Use Case 3: The Dutch historical censuses

We use the dataLegend ecosystem to curate and republish the CEDAR dataset, a Linked Data harmonized version of the Dutch historical censuses (1795–1971) [53]. Publishing the Dutch historical censuses as five-star Linked Open Data has had an fundamental impact in the efficiency and insight that historians and social scientists get from the study this dataset [26]. Due to the limitations of legacy spreadsheet formats, the dataset could not be utilized to its full potential, especially on research focused on specific comparable years [54]. To address this, researchers have identified data *harmonization* as a key aspect. Previously, if researchers wanted to know, for instance, the number of houses under construction in the Netherlands per municipality between 1859 and 1920,[53] they had to do a number of tasks. First, they had to extract data from 47 different Excel tables. Second, they had to *transform* these data in order to harmonize values. And third, they had to *replace* string values with some standard code. But above all, the results of these tasks where *hardly reusable*, meaning that the next query would take a similar completion time. In this rather inefficient scenario, query time was measured in number of days.

By using explicit harmonization rules and links to standard schemas in dataLegend for occupations, municipalities, religions and house types (see Section 3.2), 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 2 shows the number of tables that users had to open and the number of cells they had to manipulate to answer a set of prototypical queries in social history [54]. Hence, major advantages dataLegend provides for scholars are (a) a speed-up of query answering; (b) a reusable core of schemas for harmonization; and (c) an augmentation of their data by means of links to shared codes in these schemas with external datasets. Importantly, while the scalability mentioned in (a) would be certainly possible in other solutions (e.g. relational systems), the reusability and augmentation of (b) and (c) are characteristic of Semantic Web systems. For example, links from this dataset to `gemeentegeschiedenis.nl` [55] (Dutch historical municipality names as Linked Data) and DBpedia allow to instantly compare the current population of Dutch municipalities with their historical figures by using SPARQL federation.

5.4. Use Case 4: An ecosystem of reusable research

The researchers in the two first use cases correspond to two roles. The inequality use case illustrates a user who is primarily interested in data for the purpose of comparative and cross-dataset research. The railway strike shows a data owner who wants to publish and analyze his data and benefits from pre-existing data in dataLegend. Although both use cases show benefit for both roles, they reflect fairly traditional data driven processes. While more sophisticated models are required to disentangle cohort, period and age effect [56], the results suggest that in Canada in 1891 the expected effects of early life-conditions are found: higher GDP per capita in a person's birth year was associated with higher occupational status at the time of the census. However, in Sweden, the opposite was the case (see Fig. 7).

This last use case only emerged after one of the authors of this paper decided to have a closer look at the results of Use Case 1. In just under 15 min, he was able to reproduce the analysis for Canada and Sweden, and show that by adding an additional correction for age, the respective positive and negative correlation apparent in respectively Canada and Sweden are not only both negative, but also not significant. Essential in this reproduction of research is that the queries used in Use Case 2 were available on GitHub,[54] and exposed as a RESTful API through grlc [8].[55] This API enables data consumers to transpose and re-execute the same queries in different datasets, by minimally modifying *parameter values* — which `grlc` replaces in specifically designed variables in the SPARQL queries as templates[56] – and *endpoint URIs* – which can be defined both in a per-query or per-API basis. As a result, queries become reusable *actionable links* that users can invoke in their analyses by simply making HTTP requests.

Such reusable actionable links allow the research results of this use case, as shown in the R analytical scripts at https://github.com/CLARIAH/wp4-queries-censusmicro/blob/master/can91_statbybirthyear_grlc.r. This highlights the third role: a user who wants to build on earlier existing queries (not just data), and thus has the most to gain from our approach. The dataLegend platform plays an essential part in making sure that these different users meet, and collectively increase both the speed and quality of research. This shows the relative ease at which the platform facilitates reusable research questions by means of query transposition.

6. Conclusion

The preceding sections presented the dataLegend platform for linked statistical data. It aims to address the unique combination of challenges in data curation for digital humanities, by facilitating: (1) high-scale access to the *long tail* of research data; (2) Linked

53 Additional example queries at http://lod.cedar-project.nl/cedar/data.html.

54 See queries at https://github.com/CLARIAH/wp4-queries-censusmicro/blob/master/swe90_statusbybirthyear.rq and https://github.com/CLARIAH/wp4-queries-censusmicro/blob/master/can91_statusbybirthyear_5.rq.

55 See equivalent API at http://grlc.io/api/CLARIAH/wp4-queries-censusmicro.

56 According to the BASIL parameter mapping specification; see https://github.com/the-open-university/basil/wiki/SPARQL-variable-name-convention-for-WEB-API-parameters-mapping.

Table 2
Example queries automated by the integration process of the dataset in dataLegend. 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.

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

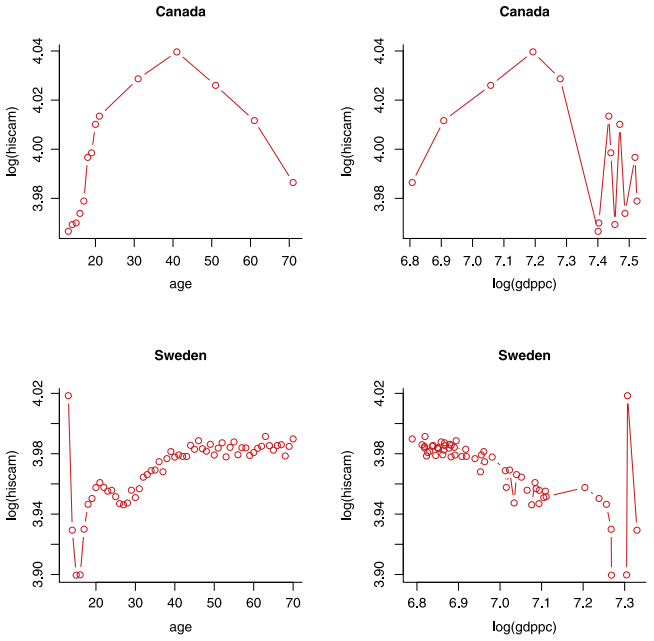

Fig. 7. HISCAM scores versus log(GDP per capita) in Canada (1891) and Sweden (1891).

Data mapping and conversion for non-Linked Data experts with QBer and COW, allowing the linkage of legacy datasets and their use in more sophisticated analyses; (3) a profitable and indirect generation of provenance and metadata at publication time; and (4) cross-dataset querying and reuse of queries with `grlc`. While some of these challenges – like high scalability – find some solutions in other technological parcels (e.g. relational systems), some others – like reuse of schemas, augmentation of datasets, and transposition of queries – are characteristic of Semantic Web systems in general, and of dataLegend in particular.

dataLegend enables individual scholars to publish and use their data in a flexible manner. QBer allows researchers to publish their (small) datasets, link them to existing vocabularies and other datasets, and thereby contribute to a growing collection of interlinked datasets hosted by the datalegend. The dataLegend platform offers services for inspecting data, and its use of the grlc API gateway ensures reusable querying across multiple datasets. We illustrated these features by means of four use cases. The first shows the ability of Linked Data to significantly lower the effort needed to do comparative research (even when the data was published as part of the same larger standardization effort). The second use case shows how publishing data through QBer allows individual researchers to have more grip on their data, to be more explicit regarding data interpretation (coding) and, via the platform, to be able to answer more questions for free (e.g. the mapping through HISCO to HISCAM). The third use case proves the order-of-magnitude decrease in effort at querying poorly linked legacy datasets, the work that reusable standard taxonomies (like HISCO and LICR) can save, and the data augmentation that links to these standard taxonomies can bring. The fourth use case shows how the hard work of other scholars, both in data curation and in the formulation of queries over the data, can be readily reproduced and used to further the field thanks to the notion of *actionable links*.

Of course, there still is room for expansion. To ensure uniqueness of identifiers, historical 'codes' need to be mapped to URIs. This is technically trivial, but historians are not used to these lengthy identifiers in their statistical analyses. Secondly, formulating research questions as queries requires an understanding of the structure of the data. Given the large numbers of triples involved, this can be difficult. As said above, standard APIs based on SPARQL query templates should solve some of this problem, but offering a user-friendly data inspection tool is high on our list. SPARQL templates allow us to solve another issue: allowing for free-form querying can have a detrimental effect on performance. The use of templates enables more efficient use of caching strategies. A building block for this approach is the grlc service, which serves SPARQL queries as Linked Data APIs using Swagger, an API specification format and user interface also for non API experts. Similarly Druid provides highly scalable HDT-based storage of the RDF files, and will in the future allow users to deploy and populate a custom triple store from a simple web console.

But even without such improvements, we believe that the use cases show that dataLegend already broadens the scope of supported workflows and data in our ecosystem, and brings the benefits of Linked Data and the Semantic Web at the fingertips of humanities scholars; an important step towards FAIR data management.

Acknowledgments

This work was funded by the NWO (Dutch Science Foundation) funded CLARIAH project and the Data2Semantics (COMMIT P23).

References

[1] T. Haigh, We have never been digital, Commun. ACM 57 (9) (2014) 24–28.
[2] E. Renckens, Digital humanities verfrissen onze blik op bestaande data, E-Data & Res. 10 (2016).
[3] T. Heath, C. Bizer, Linked Data: Evolving the Web into a Global Data Space, first ed., Morgan and Claypool, 2011, pp. 1–136.

[4] A. Meroño-Peñuela, Semantic web for the humanities, in: P. Cimiano, O. Corcho, V. Presutti, L. Hollink, S. Rudolph (Eds.), the Semantic Web: Semantics and Big Data. 10th International Conference, ESWC 2013, Proceedings, in: LNCS, vol. 7882, Springer-Verlag, Berlin, Heidelberg, 2013, pp. 645–649.

[5] A. Meroño-Peñuela, Refining Statistical Data on the Web (Ph.D. thesis), Vrije Universiteit Amsterdam, 2016.

[6] A.R. Ferguson, J.L. Nielson, M.H. Cragin, A.E. Bandrowski, M.E. Martone, Big data from small data: data-sharing in the'long tail'of neuroscience, Nature Neurosci. 17 (11) (2014) 1442–1447.

[7] R. Hoekstra, A. Meroño-Peñuela, K. Dentler, A. Rijpma, R. Zijdeman, I. Zandhuis, An ecosystem for linked humanities data, in: H. Sack, G. Rizzo, N. Steinmetz, D. Mladenić, S. Auer, C. Lange (Eds.), The Semantic Web: ESWC 2016 Satellite Events, Heraklion, Crete, Greece, May 29 –June 2, 2016, Revised Selected Papers, Springer International Publishing, Cham, 2016, pp. 425–440. http://dx.doi.org/10.1007/978-3-319-47602-5_54.

[8] A. Meroño-Peñuela, R. Hoekstra, grlc makes GitHub taste like linked data APIs, in: The Semantic Web: ESWC 2016 Satellite Events, Heraklion, Crete, Greece, May 29 –June 2, 2016, Revised Selected Papers, Springer, 2016, pp. 342–353. http://dx.doi.org/10.1007/978-3-319-47602-5_48.

[9] A. Meroño-Peñuela, A. Ashkpour, M. van Erp, K. Mandemakers, L. Breure, A. Scharnhorst, S. Schlobach, F. van Harmelen, Semantic technologies for historical research: A survey, Semant. Web –Interopera. Usability Appl. 6 (6) (2015) 539–564.

[10] S. Ruggles, E. Roberts, S. Sarkar, M. Sobek, The North Atlantic population project: Progress and prospects, Historical Methods: J. Quant. Interdiscip. History 44 (1) (2011) 1–6.

[11] J. Bolt, M. Timmer, J.L. van Zanden, GDP per capita since 1820, in: How was Life? Global Well-Being Since 1820, Organisation for Economic Co-operation and Development, 2014, pp. 57–72.

[12] H.A. Piwowar, R.S. Day, D.B. Fridsma, Sharing detailed research data is associated with increased citation rate, PLoS One 2 (3) (2007) e308.

[13] C. Tenopir, S. Allard, K. Douglass, A.U. Aydinoglu, L. Wu, E. Read, M. Manoff, M. Frame, Data sharing by scientists: Practices and perceptions, PLoS One 6 (6) (2011) e21101.

[14] P. van den Besselaar, A. Khalili, K.A. de Graaf, A. Idrissou, A. Loizou, S. Schlobach, F. van Harmelen, Towards an Open Infrastructure for Science, Technology and Innovation data. Technical Report, OECD. URL https://www.oecd.org/sti/186%20-%20VanDenBesselaar%20et%20al_RISIS.pdf.

[15] E. Hyvönen, E. Heino, P. Leskinen, E. Ikkala, M. Koho, M. Tamper, J. Tuominen, E. Mäkelä, Warsampo data service and semantic portal for publishing linked open data about the second world war history, in: H. Sack, E. Blomqvist, M. d'Aquin, C. Ghidini, S.P. Ponzetto, C. Lange (Eds.), The Semantic Web: Latest Advances and New Domains (ESWC 2016), Springer-Verlag, 2016.

[16] A.J.G. Gray, P.T. Groth, A. Loizou, S. Askjaer, C.Y.A. Brenninkmeijer, K. Burger, C. Chichester, C.T.A. Evelo, C.A. Goble, L. Harland, S. Pettifer, M. Thompson, A. Waagmeester, A.J. Williams, Applying linked data approaches to pharmacology: architectural decisions and implementation, Semant. Web 5 (2) (2014) 101–113.

[17] R. Hoekstra, P. Groth, Linkitup: Link Discovery for Research Data. AAAI Fall Symposium Series Technical Reports, 2013, pp. 28–35.

[18] M. Wilkinson, et al., The FAIR guiding principles for scientific data management and stewardship, Sci. Data (2016).

[19] T. Lebo, J. McCusker, csv2rdf4lod, Technical Report, Tetherless World, RPI, 2012 https://github.com/timrdf/csv2rdf4lod-automation/wiki.

[20] E. Muñoz, A. Hogan, A. Mileo, DRETa: Extracting RDF from wikitables, in: Int. Semantic Web Conference, Posters and Demos, CEUR-WS, 2013 98–92.

[21] E. Kalampokis, A. Nikolov, et al., Exploiting linked data cubes with opencube toolkit, in: Posters and Demos Track, 13th International Semantic Web Conference, ISWC2014, vol. 1272, CEUR-WS, Riva del Garda, Italy, 2014. http://ceur-ws.org/Vol-1272/paper_109.pdf.

[22] D. Roman, N. Nikolov, et al., DataGraft: One-stop-shop for open data management, Semantic Web – Interoperability, Usability, Applicability (2016) to appear, http://www.semantic-web-journal.net/content/datagraft-one-stop-shop-open-data-management.

[23] DERI, RDF Refine - a Google Refine extension for exporting RDF, Technical Report, Digital Enterprise Research Institute, 2015. http://refine.deri.ie/.

[24] T. Morris, T. Guidry, M. Magdinie, OpenRefine: A Free, Open Source, Powerful Tool for Working with Messy Data, Technical Report, The OpenRefine Development Team, 2015. http://openrefine.org/.

[25] A. Dimou, M. Vander Sande, P. Colpaert, R. Verborgh, E. Mannens, R. Van de Walle, Rml: a generic language for integrated rdf mappings of heterogeneous data, in: Proceedings of the 7th Workshop on Linked Data on the Web, LDOW14 - WWW14, Seoul, South Korea, 2014.

[26] A. Ashkpour, A. Meroño-Peñuela, K. Mandemakers, The dutch historical censuses: Harmonization and RDF, Historical Methods: J. Quant. Interdiscip. History 48 (2015).

[27] A. Meroño-Peñuela, A. Ashkpour, L. Rietveld, R. Hoekstra, S. Schlobach, Linked humanities data: The next frontier?, in: 2nd International Workshop on Linked Science (LISC2012), ISWC, vol. 951, CEUR-WS, 2012. http://ceur-ws.org/Vol-951/.

[28] R. Cyganiak, D. Reynolds, J. Tennison, The RDF Data Cube Vocabulary, Technical Report, W3C, 2013. http://www.w3.org/TR/vocab-data-cube/.

[29] P. Heyvaert, A. Dimou, A.-L. Herregodts, R. Verborgh, D. Schuurman, E. Mannens, R. Van de Walle, RMLEditor: A Graph-based Mapping Editor for Linked Data Mappings, in: H. Sack, E. Blomqvist, M. d'Aquin, C. Ghidini, P.S. Ponzetto, C. Lange (Eds.), The Semantic Web –Latest Advances and New Domains, ESWC 2016, in: Lecture Notes in Computer Science, vol. 9678, Springer, 2016, pp. 709–723. http://dx.doi.org/10.1007/978-3-319-34129-3_43.

[30] M. van Leeuwen, I. Maas, A. Miles, HISCO: Historical International Standard Classification of Occupations, Leuven University Press, 2002.

[31] J. van Ossenbruggen, M. Hildebrand, V. de Boer, Interactive vocabulary alignment, in: Research and Advanced Technology for Digital Libraries, TPDL 2011, in: LNCS, vol. 6966, Springer-Verlag, Berlin, Heidelberg, 2011, pp. 296–307.

[32] The World Wide Web Consortium (W3C), SPARQL Query Language for RDF, http://www.w3.org/TR/rdf-sparql-query/.

[33] P. Groth, A. Loizou, A.J. Gray, C. Goble, L. Harland, S. Pettifer, API-centric Linked Data integration: The Open PHACTS Discovery Platform case study, Web Semant.: Sci. Serv. Agents World Wide Web 29 (2014) 12–18. Life Science and e-Science.

[34] E. Daga, L. Panziera, C. Pedrinaci, A BASIlar approach for building web APIs on top of SPARQL endpoints, in: Services and Applications over Linked APIs and Data SALAD2015, ISWC 2015, vol. 1359, CEUR Workshop Proceedings, 2015. http://ceur-ws.org/Vol-1359/.

[35] Data Documentation Initiative (DDI), 1996, http://www.ddialliance.org/.

[36] Statistical Data and Metadata eXchange (SDMX), 2013, http://sdmx.org/.

[37] Generic Statistical Information Model (GSIM).

[38] Consortium of European Social Science Data Archives (CESSDA).

[39] A. Ashkpour, K. Mandemakers, O. Boonstra, Source oriented harmonization of aggregate historical census data: A flexible and accountable approach in RDF, Soc. Hist. Res. 41 (4) (2016) 291–321.

[40] P.S. Lambert, R.L. Zijdeman, M.H. Van Leeuwen, I. Maas, K. Prandy, The construction of HISCAM: A stratification scale based on social interactions for historical comparative research, Historical Methods: J. Quant. Interdiscip. History 46 (2) (2013) 77–89.

[41] M.A. Martnez-Prieto, M. Arias, J.D. Fernndez, Exchange and consumption of huge RDF data, in: The Semantic Web: Research and Applications, Springer, 2012, pp. 437–452.

[42] J.D. Fernndez, M.A. Martnez-Prieto, C. Gutirrez, A. Polleres, M. Arias, Binary RDF representation for publication and exchange (HDT), Web Semant. Sci. Serv. Agents World Wide Web 19 (2013) 22–41.

[43] A. Meroño-Peñuela, LSD dimensions: Use and reuse of linked statistical data, in: Knowledge Engineering and Knowledge Management, EKAW 2014, in: LNCS, vol. 8982, 2014, pp. 159–163.

[44] W. Beek, L. Rietveld, H.R. Bazoobandi, J. Wielemaker, S. Schlobach, LOD Laundromat: A uniform way of publishing other peoples dirty Data, in: The Semantic Web –ISWC 2014, 2014.

[45] P. Groth, A. Gibson, J. Velterop, The anatomy of a nanopublication, Inf. Serv. Use 30 (1–2) (2010) 51–56.

[46] D.J. Barker, The fetal and infant origins of adult disease, BMJ: Br. Med. J. 301 (6761) (1990) 1111.

[47] J.J. Heckman, Skill formation and the economics of investing in disadvantaged children, Science 312 (5782) (2006) 1900–1902.

[48] M.A. Thomasson, P.V. Fishback, Hard times in the land of plenty: The effect on income and disability later in life for people born during the great depression, Explor. Econ. History 54 (2014) 64–78.

[49] K. Inwood, C. Jack, National Sample of the 1891 Census of Canada, University of Guelph, Guelph, Canada, 2011.

[50] National Sample of the 1890 Census of Sweden, Version 1.0, The Swedish National Archives and Umeå University, and the Minnesota Population Center, Minneapolis, MN, Minnesota Population Center [distributor], 2011.

[51] A. Meroño-Peñuela, R. Hoekstra, grlc Makes GitHub Taste Like Linked Data APIs, The Semantic Web - ESWC 2016 Satellite Events, Heraklion, Crete, Greece, May 29–June 2, 2016, Revised Selected Papers, in: LNCS, vol. 9989, Springer, 2016, pp. 342–353.

[52] A. Meroño-Peñuela, C. Guéret, S. Schlobach, Linked edit rules: A web friendly way of checking quality of RDF data cubes, in: 3rd International Workshop on Semantic Statistics (SemStats 2015), ISWC, CEUR, 2015.

[53] A. Meroño-Peñuela, C. Guéret, A. Ashkpour, S. Schlobach, CEDAR: The Dutch Historical Censuses as Linked Open Data, Semant. Web –Interoper. Usability Appl. 8 (2) (2017) 297–310.

[54] O. Boonstra, P. Doorn, R. van Horik, J. van Maarseveen, K. Oudhof, Twee Eeuwen Nederland Geteld. Onderzoek met de digitale Volks-, Beroeps- en Woningtellingen 1795-2001, DANS en CBS, The Hague, 2007.

[55] I. Zandhuis, M. den Engelse, E.M. Gillavry, Dutchhistorical toponyms in the Semantic Web, in: Population Reconstruction, Springer, 2015, pp. 23–41.

[56] L.M. Bartels, S. Jackman, A generational model of political learning, Elect. Stud. 33 (2014) 7–18.