An Ecosystem for Linked Humanities Data

Rinke Hoekstra[1,5(✉)], Albert Meroño-Peñuela[1,4], Kathrin Dentler[1],
Auke Rijpma[2,7], Richard Zijdeman[2,6], and Ivo Zandhuis[3]

[1] Department of Computer Science,
Vrije Universiteit Amsterdam, Amsterdam, Netherlands
{rinke.hoekstra,albert.merono,k.dentler}@vu.nl
[2] International Institute of Social History, KNAW, Amsterdam, Netherlands
{auke.rijpma,richard.zijdeman}@iisg.nl
[3] Ivo Zandhuis Research and Consultancy, Haarlem, Netherlands
ivo@zandhuis.nl
[4] Data Archiving and Networked Services, KNAW, The Hague, Netherlands
[5] Faculty of Law, University of Amsterdam, Amsterdam, Netherlands
[6] University of Stirling, Stirling, UK
[7] Utrecht University, Utrecht, Netherlands

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. This paper presents QBer and the underlying structured data hub, which address the *long tail* of research data by catering for the needs of individual scholars. 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. We present QBer, and evaluate our first results by showing how our system facilitates three use cases in socio-economic history.

Keywords: Digital humanities · Structured data · Linked Data · QBer

1 Introduction

In a 2014 article in CACM, [10] 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 Dutch Royal Academy of 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 [29]. Tackling this important challenge for the digital

H. Sack et al. (Eds.): ESWC 2016 Satellite Events, LNCS 9989, pp. 425–440, 2016.
DOI: 10.1007/978-3-319-47602-5_54

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. We argue that current efforts to increase the availability and accessibility of this data do not suffice. They do not cater for the "long tail of research data" [8], the large volumes of small datasets produced by *individual* researchers; and existing Linked Data tooling is too technology-oriented to be suitable for humanities researchers at large.

This paper presents QBer and the underlying CLARIAH Structured Data Hub (CSDH),[1] whose aim is to address the limitations of current data-publishing practice in the digital humanities, and socio-economic history in particular. The CSDH integrates a selection of large datasets from this domain, while QBer is a user-facing web application that allows *individual* researchers to upload, convert and link 'clean' data to existing datasets and vocabularies in the hub without compromising the detail and heterogeneity of the original data (see Sect. 2). Under the hood, we convert all data to RDF, but QBer does 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 (See Fig. 2). The visualization is just one of the incentives we are developing. The most important incentive will be the ability to allow for transposing research queries across datasets, and the ability to perform cross-dataset querying. According to Wikipedia, an ecosystem is "a community of living organisms in conjunction with the nonliving components of their environment (things like air, water and mineral soil), interacting as a system".[2] Translating that to our situation, it means that QBer and the CSDH should interact with users that have different roles, that mutually benefit from each other, and that collectively – as a system – exhibit behavior that would not otherwise emerge.

Section 4 describes two use-cases that evaluate the ability of QBer and the CSDH to fulfill that promise. We first discuss related work in Sect. 2 and describe the QBer and CSDH systems in Sect. 3.

2 Related Work

Historical data comprises text, audiovisual content or – in our case – data in the more traditional sense: structured data in tabular form. Preparing historical data for computational analysis takes considerable expertise and effort. As a result, digital data curation efforts are organized (and funded) in a top-down fashion, and focus on the enrichment of individual datasets and collections of sufficient importance and size. Examples are the North Atlantic Population Project (NAPP) [31], the Clio-Infra repository [5], and the Mosaic project.[3] Such projects face three important issues:

[1] A screencast of the system is available at https://vimeo.com/158153564.

[2] See http://en.wikipedia.org/wiki/Ecosystem.

[3] See https://www.clio-infra.eu and http://www.censusmosaic.org/.

1. They often culminate in a website where *subsets* of the data can be *downloaded*, but cannot be programmatically accessed, isolating the data from efforts to cross-query over multiple datasets.
2. They enforce commitment to a shared standard: standardization leads to loss of detail, and thus information. The bigger a project is, the higher the cost of reconciling heterogeneity – in time, region, coding etc. – between the large number of sources involved.
3. 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 [8].

For this last reason, it is difficult for individual researchers to make their data available in a sustainable way [33]. Despite evidence that sharing research data results in higher citation rates [28], researchers perceive little incentive to publish their data with sufficiently rich, machine interpretable metadata. Data publishing and archiving platforms such as EASY[4] (in the Netherlands), Dataverse[5] or commercial platforms such as Figshare[6] and Dryad[7] aim to lower the threshold for data publishing, and cater for increasing institutional pressure to archive research data. However, as argued in [14], the functionality of these platforms – data upload, data landing page, citable references, default licensing, long term preservation – 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. This has a detrimental effect on both findability and reusability of research data.

The FAIR guiding principles for data management and stewardship [35], designed and endorsed by "a diverse set of stakeholders – representing academia, industry, funding agencies, and scholarly publishers", corroborate this view. Data should be *findable*, *accessible*, *interoperable* and *reusable*. These goals can only be achieved when, among others, data is (globally) uniquely identifiable using URIs, accessible at these identifiers, described using rich metadata in a formal language (RDF), using vocabularies and attributes that use FAIR principles, and published with a clear data usage license and provenance [35, Box 2.].

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 [11] encourage the representation of tabular sources as Linked Data.[8] CSV and HTML tables can be represented in RDF using CSV2RDF and DRETa [18,26]. For other tabular formats, like Microsoft Excel, Google Sheets, and tables encoded in JSON or XML, larger frameworks are needed, like Opencube [15], Grafter [30], and the

[4] See http://easy.dans.knaw.nl.

[5] See http://dataverse.harvard.edu and http://dataverse.nl.

[6] See http://figshare.com.

[7] See http://datadryad.org.

[8] For a comprehensive list, see e.g. https://github.com/timrdf/csv2rdf4lod-automation/wiki and http://www.w3.org/wiki/ConverterToRdf.

combination of OpenRefine and DERI's RDF plugin [7,25]. The RMLEditor[9] [13] is an editor for the R2RML mapping language,[10] and allows users to map legacy data to an RDF-based schema by bringing elements from a tabular data representation to the familiar graph-based visualization of RDF.

Perhaps closest to our work is the Karma [16,32, a.o.] tool for bringing structured data in JSON, CSV, XML and other formats to the Semantic Web.[11] Karma's user interface combines a tabular representation of the source data with a graph-based schema view for mapping the schema of legacy data files to an ontology. This mapping can then be used to transform data to RDF or to produce a R2RML mapping file. To assist users, it can automatically provide suggestions for mappings and rewrites, both at a schema level and at the data level.

The above tools cater for tabular data that represents one observation (record) per row. Datasets in social history, however, are often presented as multidimensional views that use other tabular layout features, such as hierarchical headers, column and row spanning cells, and arbitrary data locations [1,21] TabLinker [22] uses a semi-automatic approach to represent multidimensional tables as RDF Data Cube [6]. As in TopBraid Composer,[12] TabLinker uses external mapping files rather than an interactive interface.

An important question is: how can mappings be created? Work in ontology and vocabulary alignment, as in the OAEI,[13] or identity reconciliation, aim to perform *automatic* alignments. Given the often very specific (historic) meaning of terms in our datasets, these techniques are likely to be error-prone, hard to optimize (given the heterogeneity of our data) and unacceptable to scholars. Interactive alignment tools, such as Amalgame [27], developed for the cultural heritage and digital humanities domains, are more promising, but treat the alignment task in isolation rather than as part of the data publishing process. Anzo for Excel[14] is an extension for Microsoft Excel for mapping spreadsheet data to ontologies. Similarly, [2] and RightField[15] allow for selecting terms from an ontology from within Excel spreadsheets, but these require the data to conform to a pre-defined template. TabLinker processes mappings expressed in Excel worksheets.

To summarize, existing tools are not suitable for two reasons:

1. they are targeted to relatively tech-savvy users; data engineers that understand the RDF data model, the use of RDF Schema, and for whom the conversion to RDF is a goal in itself. In our case, prospective users will benefit from interlinked data, but are unlikely to have any interest in the underlying technology.

[9] See http://rml.io.

[10] See http://www.w3.org/TR/r2rml/.

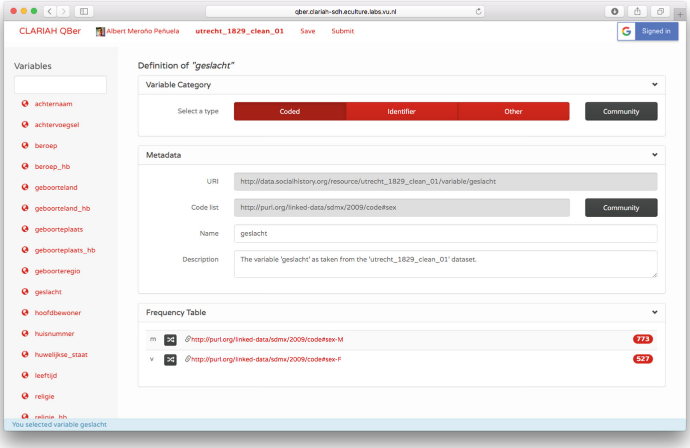
[11] See https://github.com/usc-isi-i2/Web-Karma.

[12] https://www.w3.org/2001/sw/wiki/TopBraid.

[13] The Ontology Alignment Evaluation Initiative, see oaei.ontologymatching.org/.

[14] https://www.w3.org/2001/sw/wiki/Anzo.

[15] https://www.sysmo-db.org/rightfield.



Definition of *"geslacht"*

URI: http://data.socialhistory.org/resource/utrecht_1829_clean_01/variable/geslacht

Code list: http://purl.org/linked-data/sdmx/2009/code#sex

Name: geslacht

Description: The variable 'geslacht' as taken from the 'utrecht_1829_clean_01' dataset.

m	http://purl.org/linked-data/sdmx/2009/code#sex-M	773
v	http://purl.org/linked-data/sdmx/2009/code#sex-F	527

You selected variable geslacht

Fig. 1. Variable mapping screen of QBer with the variable 'geslacht' (sex) selected.

2. they focus on *generating* mappings, isolate the mapping task from the data, or do not provide users with sufficient support for selecting the right URI to map against.

3 QBer and the Structured Data Hub

To create a viable research ecosystem for Linked Humanities Data of all sizes, we need to combine expert knowledge with automated Linked Data generation. It should be *easy* and *profitable* for individual researchers to enrich and publish their data via our platform.

To achieve the first goal, we developed QBer;[16] an interactive tool that allows non-technical scholars to convert their data to RDF, to map the 'variables' (column names) and values in tabular files to Linked Data concept schemes, and to publish their data on the structured data hub. What sets QBer apart is that all Linked Data remains under the hood.

To achieve the second goal, we build in direct feedback (reuse of existing content, visualizations, etc.) on top of the CSDH and demonstrate the research benefits of contributing data to it (see Sect. 4). We illustrate QBer by means of a walk-through of the typical usage of the tool, and then summarize its connection with the CSDH.

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 CSDH

[16] See https://github.com/CLARIAH/qber.

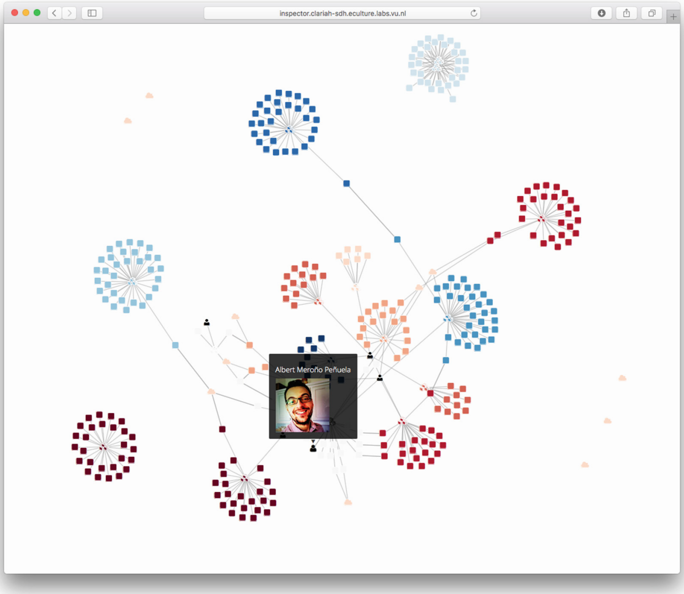


Fig. 2. The inspector view over the datasets currently in the CSDH

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. 1). 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*.

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 [20] and from datasets previously processed by QBer that now reside on the CSDH.

The *variable metadata* panel can be used to change the label and the description of the variable. If the variable has been specified to be "coded" in the previous pane, it can be linked to existing code lists curated by the Linked Open Data community. QBer provides access to all concept schemes from the Linked

Data cache[17] and the CSDH itself. If the variable is of type "other", this panel lets users define their own transformation function for literals.

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 CSDH API, which integrates them with other datasets in the hub. Under the hood, the data is converted to a Nanopublication [9] with provenance metadata in PROV, where the assertion-graph is an RDF Data Cube representation of the data [6]. The RDF representation is a verbatim conversion of the data; 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 Sect. 2.

The *Inspector*, shown in Fig. 2, allows users to explore the contents of the CSDH. The visualization shows a graph of nodes and edges, with different icons representing different node types. *User* nodes represent users that have submitted data to the hub, according to their provided OAuth identities in the welcome screen. *Dataset nodes* are shown as stacked boxes, and represent Data Structure Definitions[18] (DSD) submitted by these users. *Dimension nodes*, shown as rounded squares, represent dimensions (i.e. variables, columns in the raw data) within those DSD. Dimensions that are externally defined (e.g. by SDMX or some other external party) and are thus not directly used in datasets, are represented as cloud-icons. The color of nodes is based on the source dataset from which the information was gleaned. Users can interact with the inspector in several ways:

– hovering over a node displays a summary of its metadata, e.g. for person nodes this includes a depiction of the author;
– dragging nodes around redraws the force layout

[17] See http://lod.openlinksw.com, we aim to extend this with schemes from the LOD-Laundromat, http://www.lodlaundromat.org.

[18] According to [6], 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".

- scrolling zooms in and out
- clicking a node opens up a new browser tab with a tabular representation of the resource (using brwsr[19]).

QBer and the Inspector work on top of the CSDH API,[20] which carries out the backend functionality. This includes converting, storing, and managing `POST` and `GET` requests over datasets. The CSDH API functionality will be extended to cover standard data search, browsing and querying tasks. All data is currently stored in an OpenLink Virtuoso CE triplestore,[21] but since CSDH communicates through the standard SPARQL protocols it is not tied to one triple store vendor.

4 Evaluation

In this section, we evaluate the results of our approach by means of three use cases in socio-economic history research. The first use case investigates the question as to whether the CSDH indeed allows research to be carried out on a broader scale. In this case, we transpose a query that was built to answer a research question aimed at a dataset of one country to a dataset that describes another country. The second use case investigates the question as to whether our system facilitates the workflow of a typical individual researcher. A third use case shows how the CSDH can speed up the research cycle.

4.1 Use Case 1: Born Under a Bad Sign

Economic and social history takes questions and methods from the social sciences to the historical record. An important line of research focuses on the determinants of historical inequality. One hypothesis here is that prenatal [3] and early-life conditions [12] have a strong impact on socioeconomic and health outcomes later in life. For example, 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 [34]. 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 Sect. 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,

[19] See http://github.com/Data2Semantics/brwsr.

[20] https://github.com/CLARIAH/wp4-csdh-api.

[21] See http://www.openlinksw.com.

we use two harmonized datasets from the North Atlantic Population Project (NAPP) [31]. Economic conditions are measured using historical GDP per capita figures from the Clio-Infra repository [5]. 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 cannot provide internationally comparable occupational status scores, we have to map their occupational codes to the HISCO system,[22] so that we can use the HISCAM cross-nationally comparable occupational status scheme [17,19].[23]

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.

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.

New Workflow. Using QBer and the CSDH, the workflow is as follows. Linked-data tools are used to discover data on the hub. In our case, we used the CSDH Inspector, a linked data browser[24] and exploratory SPARQL queries. The Inspector provides a simple overview of all datasets in the CSHD.[25] Note that to discover datasets and especially linked datasets on the CSDH, 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 the hub is that if it is done once, the results can be re-used by other researchers.

Next, queries are built and stored on GitHub. The result sets that these queries produce against the data hub are then used to create the dataset that is to be analyzed. `grlc`, a tool we developed for creating Linked Data APIs using

[22] HISCO: Historical International Standard Classification of Occupations.

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

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

[25] Currently at http://inspector.clariah-sdh.eculture.labs.vu.nl/overview, but this may change.

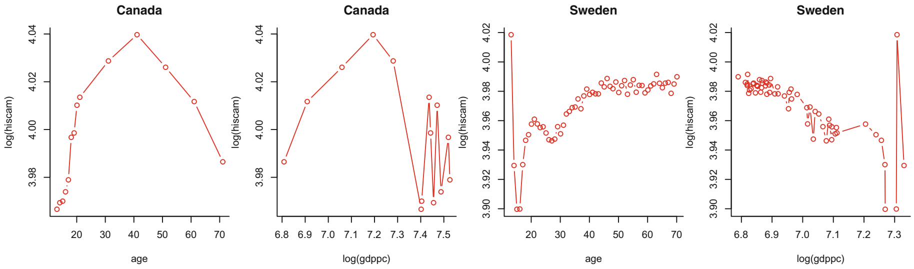


Fig. 3. HISCAM scores versus log (GDP per capita) in Canada (1891) and Sweden (1891)

SPARQL queries in GitHub repositories, was helpful in exploring the data on the hub and executing the eventual query [24].[26] Among others, it populates the API specification with allowed values for request parameters based on data available in the underlying repository. This tool can also be used to download the data directly into a statistical environment like `R` via `HTTP` requests, for example using `curl`. Alternatively, the CSDH can be queried directly from a statistical environment using SPARQL libraries.

Observations. While more sophisticated models are required to disentangle cohort, period and age effect [4], 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. 3). This shows the relative ease at which the CSDH facilitates reusable research questions by means of query transposition.

4.2 Use Case 2: Railway Strike

The second use case takes the form of a user study. It is about the "Dwarsliggers"[27] 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 it deals 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

[26] See http://grlc.io/ and https://github.com/CLARIAH/grlc.

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

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 the CSDH have.

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 behaviour 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 behaviour. 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.

New Workflow. In collaboration with Zandhuis, we revisited this workflow using QBer. Zandhuis, as many historians do, 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 the CSDH is not necessarily restricted to CSV files. It uses the Python Pandas library[28] 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 Sect. 3, QBer provides access to a large range of such classifications, basically all those available in the Linked Data cloud and the CSDH. For example, QBer provides all occupation concepts from the HISCO classification used in the first use case [19]. 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 the CSDH, 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. 2 depicts the result of the new workflow.

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

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 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 [24]).

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' [23].

Observations. 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*.

4.3 Use Case 3: An Ecosystem?

The researchers in the two use cases correspond to two roles. The railway strike shows a data owner who wants to publish and analyze his data and benefits from pre-existing data in the CSDH. The inequality use case illustrates a user who is primarily interested in CSDH data for the purpose of comparative and cross-dataset research. Although both use cases show benefit for both roles, they reflect fairly traditional data driven processes.

Our third use case only emerged after one of the authors of this paper decided to have a closer look at the results of Use Case 2. 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, and exposed as a RESTful API through `grlc`. This highlights the third role: a user who wants to build on earlier work (not just data), and thus has the most to gain from our approach. The CSDH plays an essential part in making sure that these different users meet, and collectively increase both the speed and quality of research.

5 Conclusion

The preceding sections presented QBer and the CSDH to address the limitations of existing digital humanities data curation projects in facilitating (1) the *long tail* of research data and (2) research at a broader scale, enabling cross-dataset querying and reuse of queries. We argued that existing Linked Data publishing and mapping tools do not meet the needs of scholars that are not technologically versed (or interested).

QBer and the CSDH enable 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 CSDH. The CSDH offers services for inspecting data, and (in combination with `grlc`) reusable querying across multiple datasets. We illustrated these features by means of two use cases. The first shows the ability of the Linked Data paradigm used in the CSDH 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 CSDH, to be able to answer more questions for free (e.g. the mapping through HISCO to HISCAM). The third 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.

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

Acknowledgments. This work was funded by the CLARIAH project of the Dutch Science Foundation (NWO) and the Dutch national programme COMMIT.

References

1. Ashkpour, A., Meroño-Peñuela, A., Mandemakers, K.: The Dutch historical censuses: harmonization and RDF. Hist. Methods J. Quant. Interdisc. Hist. **48** (2015)
2. van Assem, M., Rijgersberg, H., Wigham, M., Top, J.: Converting and annotating quantitative data tables. In: Patel-Schneider, P.F., Pan, Y., Hitzler, P., Mika, P., Zhang, L., Pan, J.Z., Horrocks, I., Glimm, B. (eds.) ISWC 2010, Part I. LNCS, vol. 6496, pp. 16–31. Springer, Heidelberg (2010). doi:10.1007/978-3-642-17746-0_2
3. Barker, D.J.: The fetal and infant origins of adult disease. BMJ Br. Med. J. **301**(6761), 1111 (1990)
4. Bartels, L.M., Jackman, S.: A generational model of political learning. Electoral. Stud. **33**, 7–18 (2014)
5. Bolt, J., Timmer, M., van Zanden, J.L.: GDP per capita since 1820. In: How Was Life? Global well-being since 1820, pp. 57–72. Organisation for Economic Co-operation and Development, October 2014
6. Cyganiak, R., Reynolds, D., Tennison, J.: The RDF data cube vocabulary. Technical report, W3C (2013). http://www.w3.org/TR/vocab-data-cube/
7. DERI: RDF Refine - a Google Refine extension for exporting RDF. Technical report, Digital Enterprise Research Institute (2015). http://refine.deri.ie/
8. Ferguson, A.R., Nielson, J.L., Cragin, M.H., Bandrowski, A.E., Martone, M.E.: Big data from small data: data-sharing in the 'long tail' of neuroscience. Nat. Neurosc. **17**(11), 1442–1447 (2014)
9. Groth, P., Gibson, A., Velterop, J.: The anatomy of a nanopublication. Inf. Serv. Use **30**(1–2), 51–56 (2010)
10. Haigh, T.: We have never been digital. Commun. ACM **57**(9), 24–28 (2014)
11. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data, 1st edn. Morgan and Claypool, Palo Alto (2011)
12. Heckman, J.J.: Skill formation and the economics of investing in disadvantaged children. Science **312**(5782), 1900–1902 (2006). http://www.sciencemag.org/content/312/5782/1900
13. Heyvaert, P., Dimou, A., Herregodts, A.-L., Verborgh, R., Schuurman, D., Mannens, E., Van de Walle, R.: RMLEditor: a graph-based mapping editor for linked data mappings. In: Sack, H., Blomqvist, E., d'Aquin, M., Ghidini, C., Ponzetto, S.P., Lange, C. (eds.) ESWC 2016. LNCS, vol. 9678, pp. 709–723. Springer, Heidelberg (2016). doi:10.1007/978-3-319-34129-3_43
14. Hoekstra, R., Groth, P.: Linkitup: link discovery for research data. In: AAAI Fall Symposium Series Technical Reports (FS-13-01), pp. 28–35 (2013)
15. Kalampokis, E., Nikolov, A., 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
16. Knoblock, C.A., et al.: Semi-automatically mapping structured sources into the semantic web. In: Simperl, E., Cimiano, P., Polleres, A., Corcho, O., Presutti, V. (eds.) ESWC 2012. LNCS, vol. 7295, pp. 375–390. Springer, Heidelberg (2012). doi:10.1007/978-3-642-30284-8_32

17. Lambert, P.S., Zijdeman, R.L., Van Leeuwen, M.H., Maas, I., Prandy, K.: The construction of HISCAM: a stratification scale based on social interactions for historical comparative research. Hist. Methods J. Quant. Interdisc. Hist. **46**(2), 77–89 (2013)
18. Lebo, T., McCusker, J.: csv2rdf4lod. Technical report, Tetherless World, RPI (2012). https://github.com/timrdf/csv2rdf4lod-automation/wiki
19. van Leeuwen, M., Maas, I., Miles, A.: HISCO: Historical International Standard Classification of Occupations. Leuven University Press, Leuven (2002)
20. Meroño-Peñuela, A.: LSD dimensions: use and reuse of linked statistical data. In: Lambrix, P., Hyvönen, E., Blomqvist, E., Presutti, V., Qi, G., Sattler, U., Ding, Y., Ghidini, C. (eds.) EKWA 2014 Satellite Events. LNCS, vol. 8982, pp. 159–163. Springer, Heidelberg (2015). doi:10.1007/978-3-319-17966-7_22
21. Meroño-Peñuela, A., Ashkpour, A., van Erp, M., Mandemakers, K., Breure, L., Scharnhorst, A., Schlobach, S., van Harmelen, F.: Semantic technologies for historical research: a survey. Seman. Web Interoperability Usability Applicability **6**(6), 539–564 (2015)
22. Meroño-Peñuela, A., Ashkpour, A., Rietveld, L., Hoekstra, R., Schlobach, S.: 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/
23. Meroño-Peñuela, A., Guéret, C., Schlobach, S.: 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)
24. Meroño-Peñuela, A., Hoekstra, R.: grlc makes GitHub taste like linked data APIs. In: Sack, H., Rizzo, G., Steinmetz, N., Mladenić, D., Auer, S., Lange, C. (eds.) ESWC 2016 Satellite Events. LNCS, vol. 9989, pp. 342–353. Springer, Heidelberg (2016)
25. Morris, T., Guidry, T., Magdinie, M.: OpenRefine: a free, open source, powerful tool for working with messy data. Technical report, The OpenRefine Development Team (2015). http://openrefine.org/
26. Muñoz, E., Hogan, A., Mileo, A.: DRETa: extracting RDF from Wikitables. In: International Semantic Web Conference, Posters and Demos, pp. 92–98. CEUR-WS (2013)
27. van Ossenbruggen, J., Hildebrand, M., de Boer, V.: Interactive vocabulary alignment. In: Gradmann, S., Borri, F., Meghini, C., Schuldt, H. (eds.) TPDL 2011. LNCS, vol. 6966, pp. 296–307. Springer, Heidelberg (2011). doi:10.1007/978-3-642-24469-8_31
28. Piwowar, H.A., Day, R.S., Fridsma, D.B.: Sharing detailed research data is associated with increased citation rate. PloS one **2**(3), e308 (2007). http://dx.plos.org/10.1371/journal.pone.0000308
29. Renckens, E.: Digital humanities verfrissen onze blik op bestaande data. E-Data Res. **10** (2016)
30. Roman, D., Nikolov, N., et al.: DataGraft: one-stop-shop for open data management. Sem. Web Interoperability Usability Applicability (2016, under review). http://www.semantic-web-journal.net/content/datagraft-one-stop-shop-open-data-management
31. Ruggles, S., Roberts, E., Sarkar, S., Sobek, M.: The North Atlantic population project: progress and prospects. Hist. Methods J. Quant. Interdisc. Hist. **44**(1), 1–6 (2011)

32. Szekely, P., Knoblock, C.A., Yang, F., Zhu, X., Fink, E.E., Allen, R., Goodlander, G.: Connecting the Smithsonian American art museum to the linked data cloud. In: Cimiano, P., Corcho, O., Presutti, V., Hollink, L., Rudolph, S. (eds.) ESWC 2013. LNCS, vol. 7882, pp. 593–607. Springer, Heidelberg (2013). doi:10.1007/978-3-642-38288-8_40
33. Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A.U., Wu, L., Read, E., Manoff, M., Frame, M.: Data sharing by scientists: practices and perceptions. PLoS ONE **6**(6), e21101 (2011). http://dx.doi.org/10.1371/journal.pone.0021101
34. Thomasson, M.A., Fishback, P.V.: 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. Hist. **54**, 64–78 (2014)
35. Wilkinson, M., et al.: The fair guiding principles for scientific data management and stewardship. Sci. Data (160018) (2016). http://www.nature.com/articles/sdata201618