Historical Quantitative Reasoning on the Web

Albert Meroño-Peñuela[1,2] and Ashkan Ashkpour[3]

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

Abstract. The Semantic Web is an extension of the Web through standards by the World Wide Web Consortium (W3C) [4]. These standards promote common data formats and exchange protocols on the Web, most fundamentally the Resource Description Framework (RDF). Its ultimate goal is to make the Web a suitable data space for humans (using the paradigm of Linked Documents) as well as for machines (using the paradigm of Linked Data). The latter has experimented an enormous growth in the last years, mostly due to the adoption of Linked Data publishing practices by institutions, governments and users, giving birth to the Linked Open Data cloud: a Web-graph of 100 billion interconnected facts and schemas (also called ontologies) [3]. Many domains have published their datasets as Linked Open Data, including History [6,12], converting them to RDF and linking them to related historical datasets and concepts on the Web. One of the fundamental problems of historical research are so-called gaps in historical evidence: the non-existence of relevant historical data for a particular matter. The very exercise of historical research has a primary focus on filling these gaps, generating the missing knowledge (to a level of certainty) using diverse methods. When raw historical datasets are published on the Semantic Web as Linked Data, these gaps still exist [2]. In this paper, we investigate whether existing Semantic Web technology is useful to fill these gaps of historical knowledge. We study the way in which social historians derive new knowledge from historical quantitative sources to fill missing gaps [2,13], and we mimic their behavior in a Semantic Web setting, by adapting existing technologies to, first, identify these gaps [14] and, second, to fill them [17] We explore whether these mechanisms can be generalized to be applied to other domains [18].

Keywords: Reasoning, Semantic Web, Statistics, Social and Economic History

1 Introduction

The Semantic Web is an evolution of the traditional Web, which is made of interlinked HTML pages that humans can read, into a Semantic Web, made of interlinked data that machines can process meaningfully [4]. This view is being today implemented through an active community and a set of standards that

pursue the realization of this vision. The Resource Description Framework (RDF), published and maintained by the World Wide Web Consortium[4] (W3C), is the basic layer on which the Semantic Web is built. RDF is language designed as a metadata data model. It can be used as a conceptual description method: entities of the world are represented with nodes (e.g. *Dante Alighieri*, *The Divine Comedy*), and the relationships between them are represented with edges connecting them (e.g. Dante Alighieri *wrote* The Divine Comedy). Linked Data[5] is the method of publishing and interlinking structured data on the Web using RDF and standard vocabularies[6]. The Linked Open Data (LOD) cloud [7] contains all linked datasets as Linked Data on the Web, and contains about 100 billion RDF statements [3].

Many historical datasets can be found in the LOD cloud, including relevant datasets for Social History [12]. For example, the 2000 U.S Census[8] is published as Linked Data in RDF, providing population statistics on various geographic levels. The Canadian health census uses LOD principles to provide greater access to the data and to promote greater interoperability, impossible to achieve with conventional data formats [5]. Holding a deeper historical dimension, the CEDAR project has published the Dutch historical censuses as 5-star Linked Open Data [13]. Among other goals, the CLARIAH project[9] aims at publishing a Structured Data Hub of Humanities Data in RDF, allowing multiple humanities datasets (among them socio-historical datasets) to be interrogated in an combined manner [10]. Semantic Web technology and the Linked Data methodology have proven to be effective in the integration of multiple data sources in these and other domains. Current efforts focus on improving access to these integrated, Linked Data through more widely adopted technologies, like Web APIs [8].

However, data problems that were present in the original datasets are often carried over to their Linked Data representation. Consequently, there is a growing concern about quality of data on the Semantic Web, which is currently being addressed by multiple methods for measuring the quality of Linked Data [1,19]. In historical quantitative datasets, one such a data quality problem is related to the presence of significant *data gaps*. Data gaps are missing values of observations that, according to the valid dimensions of the dataset, should be there for completeness. An example would be the non-existence of the population count of butchers in the Dutch municipality of Achtkarspelen in 1889, while that count is documented for both previous (1879 and before) and subsequent (1899 and after) censuses. The challenge of these data gaps are twofold: first they need to be *identified* and, second, they need to be *corrected*. None of both are trivial, and hamper further analysis of socio-historical datasets represented as Linked Data. Moreover, work addressing these two issues do not consider the particularities of Linked Data and Web standards.

---

[4] See `http://www.w3.org/`

[5] See also `http://www.w3.org/DesignIssues/LinkedData.html`

[6] A summary of these vocabularies can be found at `http://lov.okfn.org/dataset/lov/`

[7] A summary of contained datasets can be found at `http://linkeddata.org/`

[8] See `http://datahub.io/dataset/2000-us-census-rdf`

[9] `http://www.clariah.nl/`

In this paper, we first summarize our methodology of converting socio-historical datasets to Linked Data, with a use case on the Dutch historical censuses[10] (1795–1971) (Section 2). Further, we describe two Web-friendly techniques we developed to address the two sides of the problem of *data gaps*. First, to identify them we propose Linked Edit Rules, a methodology to express constraint checks of National Statistical Offices (NSOs) or "edit rules" as Linked Data (Section 3). Second, to correct them we propose SCRY, an easy and standards-compliant way of transforming Linked Data via the RDF query language, SPARQL 4. To conclude, we discuss the extensibility of these methods to other data curation problems in History, their applicability to other domains, and future work (Section 5).

2 Social History as Linked Data

Historical quantitative datasets[11] are spread and scattered over the Web. For instance, the North Atlantic Population Project (NAPP) [16] publishes an harmonized time series of historical census micro-data from multiple countries. But more often, socio-historical datasets are not standardized nor harmonized. Multiple historical prices and wages datasets can be found in the form of Excel Spreadsheets[12]. The Dutch historical censuses is a collection of 2,288 disconnected tables with demographic, labour and housing information, which is key to understand the 1795–1971 period of Dutch history. All these disconnected data sources need to be interrogated in a combined fashion in order to answer scholars' research questions.

Integrating tabular data on the Web is not a trivial task, but some recent work is contributing on paving the way. A promising approach is to convert these datasets to Linked Data, the Semantic Web data publication method, which is known to be effective to solve important data integration problems [8]. Consequently, our proposal is to represent socio-historical datasets as Linked Data, in a way they can be linked to related datasets and concepts on the Web.

We propose multiple tools to address this. Our recent work on QBer [10] allows humanities scholars to self-publish their own CSV or SAV files that contain statistical data, as Linked Data compliant with the RDF Data Cube and CSV on the Web standards [7]. To enhance this process, we offer scholars the possibility to reuse more than 2,000 existing dataset dimensions[13], which are currently being used in the Web of Linked Data, to describe data of their own [11]. URI identifiers for common variables (such as `sdmx-dimension:sex`, `sdmx-dimension:refPeriod` or `sdmx-dimension:refArea`) are already there.

[10] See `http://www.volkstellingen.nl/`

[11] Our research focuses primarily on quantitative historical datasets, a term we use to describe datasets containing structured or semi-structured data, in contrast to unstructured datasets like e.g. textual corpora.

[12] See e.g. `http://www.iisg.nl/hpw/link.php`

[13] With *dimensions* we refer to the variables in a dataset, typically the column headers in a CSV file.

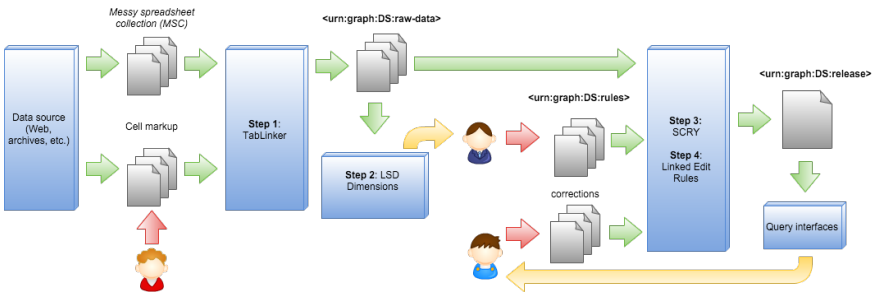

Fig. 1: Proposed workflow for integrating "messy" spreadsheet collections on the Web. Red arrows indicate expert intervention; yellow arrows indicate semi-automatic steps; and green arrows indicate totally automatic stages.

For socio-historical datasets that lack the curation of CSV files, like layout-intensive spreadsheets, we propose the *Integrator*[14] [13]. The Integrator is a full pipeline to convert so-called "messy" spreadsheet collections[15] to Linked Data. It was originally developed to convert the Dutch historical censuses to RDF. Currently it covers a wide variety of datasets from multiple domains. The workflow of the Integrator is shown in Figure 1. In summary, the Integrator is a semi-automatic method, where expert intervention is required in two stages: *(1)* to mark-up source data files, indicating where data, headers and other important spreadsheet regions are; and *(2)* to provide harmonization mappings that will convert strings to linked URIs in a graph. This approach presents three significant advantages. First, it keeps a source-oriented representation of the *raw data* as RDF, allowing users to query for the original values of the spreadsheets. Second, it links elements of this raw data to the *harmonization rules*, i.e. to the transformations that need to be applied to these source data in order to be cleaned and standardized. Third, it provides the *released* dataset as a set of observations that are linked to both the raw data and the harmonization rules via PROV [9], the Semantic Web standard for provenance representation. This way, users of the data can easily follow and reproduce all transformations applied to all data points of the source data to harmonize them.

The improvement on access to socio-historical data published on the Web as Linked Data can be appreciated by the ease with which new tools and applications can be developed on top. For example, historical maps are very easily drawn making use of the Dutch historical censuses as Linked Data via SPARQL (see Figure 2). Queries over the harmonized dataset can finally be made with no data munging efforts[16].

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

[15] For an example of such collection, see `https://github.com/CEDAR-project/DataDump/tree/master/xls`

[16] See e.g. `http://lod.cedar-project.nl/cedar/data.html` and `http://grlc.clariah-sdh.eculture.labs.vu.nl/CEDAR-project/Queries/api-docs`.

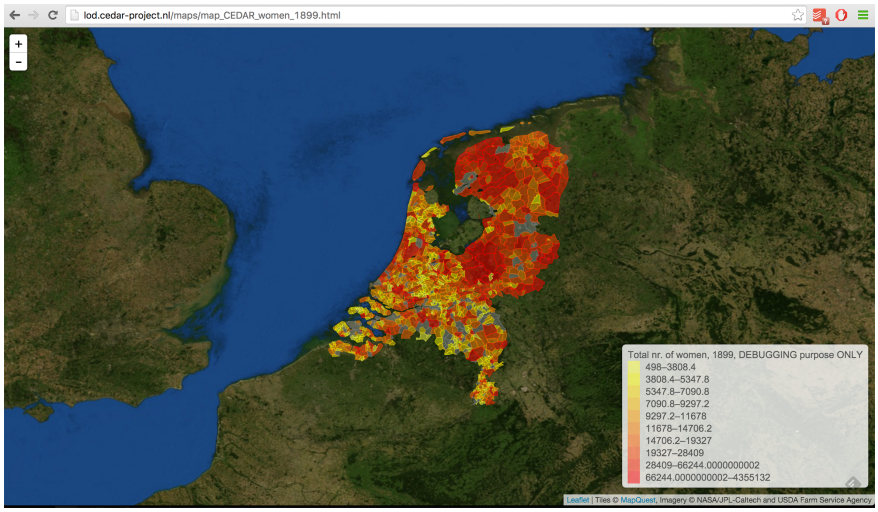

Fig. 2: Screenshot of a client map-drawing interface. Population data is retrieved from the SPARQL endpoint of the Dutch historical censuses, and merged with externally fetched temporal Dutch municipality shapes.

3 Identifying Data Gaps with Edit Rules

Missing data is common in historical datasets. The map of Figure 2 already reveals some areas of the Dutch geography for which data is not available. However, the identification of such missing data points beyond human visual exploration is of central interest for the development of automatic methods of data quality assessment.

To identify data gaps, and other data flaws and obvious inconsistencies in their datasets, NSOs use so-called "edit rules". identifying so-called *obvious inconsistencies*. An obvious inconsistency occurs when the datasets contains a value or combination of values that cannot correspond to a real-world situation. For example, a person's age cannot be negative, a man cannot be pregnant and an underage person cannot possess a driving license. In order to validate data cubes against obvious inconsistencies, statisticians express this knowledge as *edit rules*. Edit rules are used to automatically detect inconsistent data points in statistical datasets. Examples are shown in Listing 1.1. Edit rules can be *micro-* or *macro-*, depending on whether they constrain the data in a per-record or inter-record manner, respectively.

To facilitate the exchange, publication, share and reuse of these edit rules, we propose to express them as Linked Data, in so-called *Linked Edit Rules* (LER) [14]. With LER, it is easy to identify data gaps, outliers, and other data flaws that hamper the soundness and completeness of socio-historical datasets represented as Linked Data[17].

[17] See examples at `http://linkededitrules.org/`

```
dat1 : ageGroup %in% c('adult', 'child', 'elderly')
dat7 : maritalStatus %in% c('married', 'single', 'widowed')
num1 : 0 <= age
num2 : 0 < height
num3 : age <= 150
num4 : yearsMarried < age
cat5 : if(ageGroup == 'child') maritalStatus != 'married'
mix6 : if(age < yearsMarried + 17) !(maritalStatus %in% c('married','widowed'))
mix7 : if(ageGroup %in% c('adult', 'elderly') age >= 18
mix8 : if(ageGroup %in% c('child', 'elderly') & 18 <= age) age >= 65
mix9 : if(ageGroup %in% c('adult', 'child')) 65 > age
```

Listing 1.1: Examples of micro-edits in the R `editrules` package.

4 Filling the Gaps via Data Transformations

Once identified, the final step on addressing data gaps is to *transform* them into actual data. Multiple statistical methods exist to address the problem of `N/A` data, such as imputation. In statistics, imputation is the process of replacing missing data with substituted values.

Imputation of statistical data in Linked Data has advantages, but also pitfalls. On the one hand, it is very easy to leverage the expressivity of vocabularies RDF Data Cube to express the fact that an observation has an imputed value, instead of an observed one. This can be easily achieved by using component properties of the vocabulary such as `qb:AttributeProperty`. On the other hand, Linked Data technology such as SPARQL has scarce support for statistical functionality, or any non-standard SPARQL provided functionality for that matter.

To address the later, and easily impute missing Linked Data using SPARQL in a standards compliant way, we propose to use SCRY [18]. SCRY is a lightweight SPARQL endpoint that leverages query federation to execute non-standard functionality in a remote server. For instance, the following is a valid SPARQL query that uses an external SCRY endpoint to impute missing values in a data cube:

```
PREFIX :       <http://scry.rocks/example/>
PREFIX scry:   <http://scry.rocks/>
PREFIX impute: <http://scry.rocks/math/impute?>
PREFIX mean:   <http://scry.rocks/math/mean?>
PREFIX sd:     <http://scry.rocks/math/stdev?>

SELECT ?obs ?dim ?imputed_val WHERE {
?obs a qb:Observation .
?dim a qb:DimensionProperty|qb:MeasureProperty .
FILTER NOT EXISTS { ?obs ?dim ?val .}
?other_obs ?dim ?other_val .

SERVICE <http://sparql.scry.rocks/> {
```

```
    SELECT ?imputed_val {
      GRAPH ?g1 {impute:v scry:input  ?other_val ;
                          scry:output ?imputed_val .}
    }
}
}
```

The specific method used in the implementation of the remote procedure `impute:v` is left to the user, which can select any imputation method of her choice. A simple example would be to return an average of the values of all other observations which do not miss the value. The specific method chosen is independent of the primary endpoint that receives the query, and thus users are free to implement any functionality that SCRY may call from Python code.

5 Discussion and Future Work

In this paper, we have surveyed current methodologies on representing quantitative historical datasets, concretely socio-historical ones, as Linked Data. We have also shown how state of the art techniques in constraint checking and data transformation in Linked Data can be used to identify and correct important data gaps in socio-historical datasets, a key data-problem in the research workflows of History scholars.

We argue that a greater degree of genericity and extensibility of these approaches should be possible at two different levels. First, while we recognize the importance of data gaps in historical datasets, we see that the methodologies proposed can hold their validity in detecting, and correcting, other types of data flaws besides missing data. Other data problems such as outlier detection, data normalization, data standardization, data quality checking, and data quality improvement are important data curation issues that could be, to a certain extent, addressed with the proposed combined methods of the Integrator (Section 2), LER (Section 3) and SCRY (Section 4). Second, none of the proposed solutions depend on being applied to Linked Datasets of socio-historical data specifically. Moreover, the Integrator, LER and SCRY are domain-agnostic, which means that as far as the data is encoded in RDF and published under the Linked Data principles[18], they work with independence from the domain the data belongs. Furthermore, current experiments to prove this genericity are on the way[19].

We plan to extend this work in multiple directions. First, we plan on achieving the aforementioned extensibility and genericity by executing our proposed methods in more datasets, domains, and over different types of data flaws. Second, we will implement user interfaces that enhance the usage of these methods by non-Linked Data savvy scholars, hence improving their ease of use. Finally, we will increase the accessibility of these methods by leveraging more wide-spread

[18] See `https://www.w3.org/DesignIssues/LinkedData.html`

[19] See `https://github.com/albertmeronyo/DataDump-uk-messy` and `https://github.com/albertmeronyo/DataDump-prices-wages`

Web technologies to bridge the gap with Linked Data languages such as RDF and SPARQL. To this end, we will investigate the use of Linked Data APIs around SPARQL endpoints through `grlc`[20] [15], and in particular on how to integrate quality detection mechanisms like LER, and data transformation methods like SCRY, with automatically-generated Web APIs.

References

1. Albertoni, R., Guéret, C., Isaac, A.: Data Quality Vocabulary (First Public Working Draft). Tech. rep., World Wide Web Consortium (2015), http://www.w3.org/TR/2015/WD-vocab-dqv-20150625/
2. Ashkpour, A., Meroño-Peñuela, A., Mandemakers, K.: The Dutch Historical Censuses: Harmonization and RDF. Historical Methods: A Journal of Quantitative and Interdisciplinary History 48, 230–245 (2015)
3. Beek, W., Rietveld, L., Bazoobandi, H.R., Wielemaker, J., Schlobach, S.: LOD Laundromat: A Uniform Way of Publishing Other People's Dirty Data. In: The Semantic Web – ISWC 2014 (2014)
4. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. Scientific American 284(5), 34–43 (2001)
5. Bukhari, A.C., Baker, C.J.: The Canadian health census as Linked Open Data: towards policy making in public health. In: 9th International Conference on Data Integration in the Life Sciences (2013)
6. Burrows, T.: A data-centred 'virtual laboratory' for the humanities: Designing the Australian Humanities Networked Infrastructure (HuNI) service. Literary and Linguistic Computing: The Journal of Digital Scholarship in the Humanities 28(4), 576–581
7. Cyganiak, R., Reynolds, D., Tennison, J.: The RDF Data Cube Vocabulary. Tech. rep., World Wide Web Consortium (2013), http://www.w3.org/TR/vocab-data-cube/
8. Groth, P., Loizou, A., Gray, A.J., Goble, C., Harland, L., Pettifer, S.: API-centric Linked Data integration: The Open PHACTS Discovery Platform case study. Web Semantics: Science, Services and Agents on the World Wide Web 29(0), 12 – 18 (2014), http://www.sciencedirect.com/science/article/pii/S1570826814000195, life Science and e-Science
9. Groth, P., Moreau, L.: PROV-Overview. An Overview of the PROV Family of Documents. Tech. rep., World Wide Web Consortium (W3C) (2013), http://www.w3.org/TR/prov-overview/
10. Hoekstra, R., Meroño-Peñuela, A., Dentler, K., Rijpma, A., Zijdeman, R., Zandhuis, I.: An Ecosystem for Linked Humanities Data. In: Proceedings of the 1st Workshop on Humanities in the Semantic Web (WHiSe 2016), ESWC 2016 (2016), under review
11. Meroño-Peñuela, A.: LSD Dimensions: Use and Reuse of Linked Statistical Data. In: Knowledge Engineering and Knowledge Management (EKAW 2014). LNCS, vol. 8982, pp. 159–163 (2014)
12. 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. Semantic Web – Interoperability, Usability, Applicability 6(6), 539–564 (2015)

[20] See https://github.com/CLARIAH/grlc

13. Meroño-Peñuela, A., Ashkpour, A., Guéret, C., Schlobach, S.: CEDAR: The Dutch Historical Censuses as Linked Open Data. Semantic Web – Interoperability, Usability, Applicability (2015), In press
14. 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)
15. Meroño-Peñuela, A., Hoekstra, R.: grlc Makes GitHub Taste Like Linked Data APIs. In: Proceedings of the Services and Applications over Linked APIs and Data workshop, ESWC 2016 (2016), under review
16. Ruggles, S., Roberts, E., Sarkar, S., Sobek, M.: The North Atlantic Population Project: Progress and Prospects. Historical Methods: A Journal of Quantitative and Interdisciplinary History 44(1), 1–6 (Jan 2011)
17. Stringer, B., Meroño-Peñuela, A., Loizou, A., Abeln, S., Heringa, J.: SCRY: Enabling quantitative reasoning in SPARQL queries. In: Semantic Web applications and tools for life sciences (SWAT4LS 2015), December 7-10 Cambridge, England (2015) (2015)
18. Stringer, B., Meroño-Peñuela, A., Loizou, A., Abeln, S., Heringa, J.: To SCRY Linked Data: Extending SPARQL the Easy Way. In: Proceedings of the Diversity++ workshop. International Semantic Web Conference (ISWC 2015) (2015)
19. Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S.: Quality Assessment for Linked Data: A Survey. Semantic Web – Interoperability, Usability, Applicability (2014), http://www.semantic-web-journal.net/system/files/swj773.pdf