DE GRUYTER OPEN

Journal of Catalan Intellectual History (JOCIH)
Online ISSN 2014-1564 DOI: 10.1515/jocih-2016-0013
https://www.degruyter.com/view/j/jocih

Received 11/01/2017 / Accepted 12/04/2017

Digital Humanities on the Semantic Web: Accessing Historical and Musical Linked Data

Albert Meroño-Peñuela* (Knowledge Representation & Reasoning, Department of Computer Science, Vrije Universiteit, Amsterdam)

albert.merono@vu.nl http:// orcid.org/0000-0003-4646-5842

Abstract. Key fields in the humanities, such as history, art and language, are central to a major transformation that is changing scholarly practice in these fields: the so-called Digital Humanities (DH). A fundamental question in DH is how humanities datasets can be represented digitally, in such a way that machines can process them, understand their meaning, facilitate their inquiry, and exchange them on the Web. In this paper, we survey current efforts within the Semantic Web and Linked Data, a family of Web-compatible knowledge representation formalisms and standards, to represent DH objects in quantitative history and symbolic music. We also argue that the technological gap between the Semantic Web and Linked Data, and DH data owners is currently too wide for effective access and consumption of these semantically enabled humanities data. To this end, we propose *grlc*, a thin middleware that leverages currently existing queries on the Web (expressed in, e.g., SPARQL) to transparently build standard Web APIs that facilitate access to any Linked Data.

Keywords: Digital Humanities, History, Music, Semantic Web, Statistical Data, APIs

1 Introduction

The traditional disciplines of the humanities (history, languages, law, art, philosophy and religion) are nowadays undergoing a major transformation under the label of *Digital Humanities* (DH) (Schreibman et al. 2004). The DH encompass two major, often simultaneous, ways of understanding the intersection between computer science and the humanities: (1) the employment of technology in the pursuit of humanities research; and (2) the subjection of technology to humanistic questioning and interrogation (Schreibman et al. 2004). How to represent prototypical humanities data (e.g., historical texts, census tables, holy scriptures, paintings, music, poetry, etc.) in digital form is, therefore, a natural question from the DH perspective and a necessary step to enabling computers to read and process, efficiently and meaningfully, the content of humanities datasets.

However, humanities datasets are inherently difficult to express digitally. Their scattered distribution on the Web, and their diversity in syntactically and semantically heterogeneous languages, hamper their use, integration, and potential (Meroño-Peñuela et al. 2013). Moreover, the lack of explicitly and semantically meaningful *links* between these datasets – which very often share common resources and concepts – prevents an

* Albert Meroño-Peñuela is a postdoc at the Vrije Universiteit Amsterdam and the International Institute of Social History (IISG). He is currently working in WP4 of CLARIAH, which aims at facilitating the integration of socio-historical datasets using Web technology. He obtained his PhD at Vrije Universiteit, also holds a bachelor in Informatics Engineering from Universitat Politecnica de Catalunya (FIB-UPC), and has previously worked at the Institute of Law and Technology (IDT-UAB).

automatic and intelligent retrieval and use by applications that consume data. *The Semantic Web* (Berners-Lee et al. 2001) aims at providing the necessary building blocks to support a machine-processable *Web of data* – extending the most widespread human-readable *Web of documents*. In such a Web of data, information is expressed in the form of statements (or triples) using the Resource Description Framework (RDF), which connects arbitrary things identified by global Web identifiers (URIs). Due to the noticeable overlap between the mission of the Semantic Web, and the need for data representation formalisms for the DH, recent approaches investigate the use of the former in order to address the latter, primarily in quantitative history (Meroño-Peñuela, 2016). The application of Web-enabled knowledge representation methods (Linked Data, ontologies) and data science (data integration, data preparation, provenance) has lead to a Semantic Web that is also rich in interlinked historical and cultural heritage knowledge (Meroño-Peñuela, 2016; Schreiber et al. 2008). Furthermore, recent work proposes to represent fine-grained symbolic music using the same Semantic Web building blocks, effectively interconnecting not just music metadata, but music itself (i.e., notes) on the Web (Meroño-Peñuela and Hoekstra, 2016). In summary, users and the applications that consume data can query today an immensely rich and interconnected knowledge graph (the so-called Linked Open Data cloud, or LOD cloud) of more than 100 billion statements (Heath and Bizer, 2011), many of them of a DH nature (historical records, cultural heritage, museum works, government archives, music, etc.).

Nonetheless, the benefits that come with the LOD cloud and the Semantic Web are very often distanced from final users and the data consuming applications. The problem is the inherently steep learning curve that Semantic Web technologies such as SPARQL and RDF have. This is very often a challenge for non-technology savvy users, and amongst them many (digital) humanities scholars. Ease of access to semantically integrated data in the LOD cloud is a challenge not only for human users, but also for Web data consuming applications. These applications need to implement specific technology-dependent access methods (e.g., libraries) that can deal with semantic queries in SPARQL and RDF, which creates additional coupling, and increases the software complexity and maintenance costs. Obviously, accessing Linked Open Data in a standard way is a problem that is shared amongst both DH and non-DH data publishers. To address this issue, we propose the use of *grlc*, a thin middleware layer that serves as an interface between applications consuming Web data and Linked Data publishers, and automatically builds universal and standard Web APIs by reusing distributed queries (Meroño-Peñuela and Hoekstra, 2016). These resulting Web APIs are understandable and usable by most Web client applications without any Semantic Web technology-specific requirements.

In this paper, we describe the use of Semantic Web technology to (a) publish, refine and semantically connect DH datasets on the Web; and (b) make these datasets more accessible to the wider spectrum of Web data consuming applications. First, in Section 2 we survey our current work on publishing prototypical DH datasets on the Semantic Web, with concrete use cases in quantitative history (Section 2.1) and symbolic music (Section 2.2). Second, in Section 3 we describe *grlc*, a universal method for accessing Linked Open Data via methodically curated and automatically created Web APIs.

2 Digital Humanities in the Semantic Web

Semantic Web technologies, such as RDF, Linked Data, SPARQL and OWL, offer excellent opportunities to represent DH datasets digitally, in a way that computers can process their content and follow meaningful links to other related resources and concepts on the Web. In this section we summarise work in two different application areas of Semantic Web technologies for DH objects: quantitative history and symbolic music.

2.1 Quantitative History

Quantitative history deals with the acquisition of knowledge about our past using statistical data and registries. To obtain such knowledge, the statistical data are subject to analysis by statistical methods and tools. However, analysis can only be performed once the statistical data have been adequately prepared; empirical studies show that this preparation takes at least 60% of the total time spent. This is due to a number of typical *data*

preparation problems. First, non-standard legacy formats are subject to decaying support over time that negatively affects the accessibility of the data. Second, data errors, typos and other flaws are hard to detect and correct, and they affect how meaningful the results will be in the analysis. Third, data curation procedures are often hard-coded in implementations or hidden in closed-source systems, which hinders their reusability. Moreover, if these datasets also include a *historical* dimension, two additional problems occur. First, operational sources of historical statistics have often been lost over time, leaving partially analytical views as the only representation preserved in archives. Second, time series are usually poorly harmonised, due to the incompatibility of changing classification systems. Data scientists try to resolve all these data preparation issues by resorting to painful data munging, which results in results in labour intensive and computationally expensive operations, as previously noted (Meroño-Peñuela, 2016).

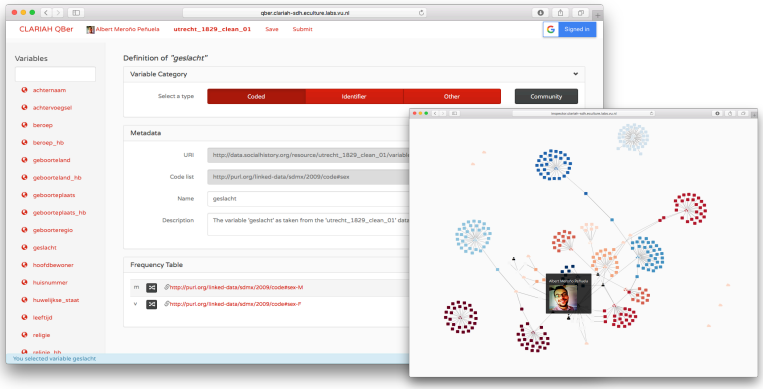

Fig. 1. *Variable mapping screen of QBer with the variable 'geslacht' (sex) selected. The inspector can be seen in the bottom right.*

In order to overcome these data preparation issues, in previous work, we have proposed QBer within the CSDH structured data hub (Hoekstra et al. 2016), an interface for transforming and publishing historical datasets on the Web as Linked Data – without any previous knowledge required for Semantic Web technologies. It should be *easy* and *profitable* for individual researchers to enrich and publish their historical data. To achieve this, we developed QBer (see Figure 1), 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. Through a sequence of screens, users can map column names to existing variable URIs on the Web with the same meaning as theirs. Similarly, they can map the values of these columns to different *skos:Concept* instances within a previously selected *skos:ConceptScheme*. This way, both column names and column values are mapped to Linked Data URIs, forming an RDF graph of historical statistical information. Then, the data is converted to a Nanopublication with provenance metadata in PROV, where the assertion-graph is an RDF Data Cube representation of the data. 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. QBer is open source[1], and a public instance can be found online at http://qber.clariah-sdh.eculture.labs.vu.nl/

2.2 Symbolic Music

Many prototypical DH objects (texts, images, video, audio), for which some *metadata* already exists in the LOD cloud in the form of RDF statements, are not (yet) machine interpretable from a semantic perspective, or require

1 See https://github.com/clariah/qber

the use of legacy tooling to do so. Our emphasis here is that, despite the resources that describe such DH objects, and promote them to first-class citizens of the Web as Linked Data (e.g., the DBpedia page about "Hey Jude" by the Beatles) – the objects themselves (i.e., the *song* "Hey Jude" by the Beatles) remain, to a large extent, non-interoperable.

Music is one of the most predominant of such DH objects. Music metadata has received considerable attention by the Linked Data community: DBTune.org, MusicBrainz, MySpace and BBC Music link several music-related data sources on the Semantic Web. Although the publication of music metadata about artists, songs, albums, and musical events is an obvious contribution to the variety of Linked Data, music itself is currently only composed, published, and exchanged offline, or using monolithic systems.

In previous work (Meroño-Peñuela and Hoekstra, 2016), we have shown that it is possible to convert any piece of digital music in MIDI format as Linked Data, using the *midi2rdf*[2] converter suite. This representation expresses MIDI files as RDF graphs, making digital music not only interoperable amongst Web agents, but also better readable by humans. Here, we want to emphasise the potential for expressing symbolic music as Linked Data. First, the linkage of the very fine-grained parts of musical pieces (e.g., notes, silences, etc.,) with their metadata can provide musicologists with new tools to gain new knowledge about music. Second, shared resources amongst different musical pieces (i.e., chords, notes, instruments, keys) might unravel new ways of comparing and clustering music. Third, the ability to point to global, de-referenceable parts of music allows for annotating it to an unprecedented level of detail, relying only on standard Web technologies and ensuring that the musical content remains machine-processable. Finally, MIDI music as Linked Data can be linked to any other DH resources also expressed as Linked Data, thus enabling the possibility of traversing RDF graphs of a particular piece of (classical) music, its scores, its metadata, the books that have been written about it, and to any other related resource available on the Web.

3 A More Accessible Semantic Web with on-the-Fly APIs

Using (DH) data that has been published as Linked Data on the Semantic Web is *difficult*, despite the inherent advantages. Linked Data requires users to be familiar with concepts and languages like URIs, RDF, triples, and SPARQL; and to be able to use these adequately. Such a requirement is often too onerous even for simply accessing and consuming the Linked Data. On the other hand, these technologies are necessary to link, publish and query Linked Data.

In previous work (Meroño-Peñuela and Hoekstra, 2016), we presented a compromise solution, *grlc*, which only requires a one-time effort with SPARQL to build an API that provides universal access to Linked Data. Assuming that the SPARQL queries that retrieve some Linked Data of interest are hosted publicly in GitHub[3], *grlc* retrieves them to build an OpenAPI specification, and executes them transparently via a RESTful API. This means that neither the final users or applications need to deal with SPARQL anymore; opening a URL or clicking a button is enough to get the same results directly from the LOD cloud.

grlc provides three basic operations: (1) it generates the OpenAPI specification of a given GitHub repository; (2) it generates the Swagger UI (see Figure 2) to provide an interactive user-facing front end of the API contents; and (3) it translates requests to call the operations of the API against a SPARQL endpoint with several parameters. If the GitHub repository at *https://github.com/:owner/:repo* contains SPARQL queries, *grlc* uses these, together with organisational repository information from the GitHub API, to build the API interface automatically. Assuming that *grlc* is running at *:host*, these operations are available at the following routes:

- *http://:host/:owner/:repo/spec* : JSON OpenAPI-compliant specification

2 See https://github.com/albertmeronyo/midi2rdf/
3 See https://github.com/

- *http://:host/:owner/:repo/api-docs* : Swagger-UI, rendered using mappings to the SPARQL queries and the GitHub repository information, as shown in Figure 2.
- *http://:host/:owner/:repo/:operation?p_1=v_1...p_n=v_n* : request to *:operation* with parameters $p_1,...,p_n$ taking values $v_1,...,v_n$.

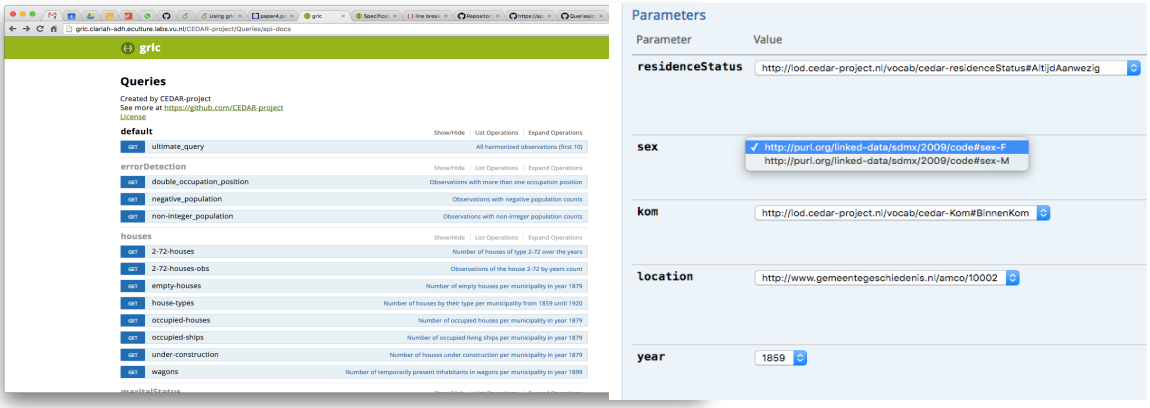

Fig. 2. *(Left) Screenshot of the API user interface generated with* grlc. *(Right) Screenshot of the parameter enumeration feature, making it easier for users to explore possible parameter values.*

A public instance of *grlc* is available at http://grlc.io. This instance can be freely used by users and applications to access any Linked Data on the Web via neatly organised, on-the-fly APIs. We particularly encourage users owning, or willing to access, DH datasets, to use it to ease their consumption and enhance their linkage with related resources and concepts on the Web.

4 Conclusions

In this paper, we have proposed a solution for representing prototypical digital humanities (history, music, philosophy, etc.,) datasets digitally, using the tools, formalisms and languages provided by Linked Data and the Semantic Web. We have shown the use of these technologies for two different kinds of DH objects: historical statistical registries (Section 2.1) and symbolic music (Section 2.2). We have also argued that, despite the benefits that Linked Data provides for representing such DH objects as semantically enabled Web resources, using these resources is difficult for users and applications that are not familiar with these technologies. For this, we have proposed two solutions: (1) QBer, which aids users at converting and publishing their table-like DH datasets as Linked Data; and (2) *grlc*, which enables API-based access to any Linked Data and saves users and applications from learning new query languages, like SPARQL.

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

5 References

Meroño-Peñuela, A., Schlobach, S., van Harmelen, F.: *Semantic Web for the Humanities*, In: Proceedings of the 10th Extended Semantic Web Conference, ESWC 2013, Montpellier, France, May 28–30, 2013. Philipp Cimiano et al. (Eds.), Lecture Notes in Computer Science 7882 Springer, 2013, pp. 645–649.

Meroño-Peñuela, A.: *Refining Statistical Data on the Web*. Vrije Universiteit Amsterdam, 2016.

Berners-Lee, T., Hendler, J., Lassila, O.: *The Semantic Web*. In: Scientific American 284 (5) (2001):34–43.

Schreiber, G., Amin, A., Aroyo, L., van Assem, M., de Boer, V., Hardman, L., Hildebrand, M., Omelayenko, B., van Ossenbruggen, J., Tordai, A., Wielemaker, J., Wielinga, B.: *Semantic annotation and search of cultural-heritage collections: The MultimediaN E-Culture demonstrator*. Web Semantics: Science, Services and Agents on the World Wide Web, 6(4) (2008): 243–24.

Heath, T., Bizer, C.: *Linked Data: Evolving the Web into a Global Data Space* (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1–136. Morgan & Claypool, 2011.

Meroño-Peñuela, A., Hoekstra, R.: *The Song Remains the Same: Lossless Conversion and Streaming of MIDI to RDF and Back*. In: 13th Extended Semantic Web Conference (ESWC 2016), posters and demos track. May 29th — June 2nd, Heraklion, Crete, Greece, 2016.

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) (2015): 539–564. IOS Press.

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.) The Semantic Web. ESWC 2016. Lecture Notes in Computer Science, vol. 9989, pp. 342–353. Spring, 2016.

Schreibman, S., Siemens, R., Unsworth, J. (Eds.): *A Companion to Digital Humanities*. Oxford: Blackwell, 2004.

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, May 29th, Heraklion, Crete, Greece, 2016.