Bärbel Kröger, Johanna Störiko, Jörg Wettlaufer

Workshop proposal **DH Benelux 2023 / KBR**

Title: Semantic Web and Linked Open Data in Historical Sciences

Short Description of the content and relevance for the digital humanities

The workshop introduces participants to the use of Semantic Web Technologies and Linked Open Data in Digital Historical Studies with a special focus on Wikidata[1]. Semantic Web and Linked Open Data are highly relevant for knowledge representation in the Digital Humanities and especially important for collaborative processes of analyzing and sharing data.

Since the publication of the first concept of the "Semantic Web" as an extension of the World Wide Web (Berners-Lee and Lassila Hendler 2001), humanities scholars have been discussing the possibilities and limits of modelling their data within “Semantic Web”. The Resource Description Framework (RDF) data model and its serialisation in Turtle or N-Triples has become the standard in modelling machine-readable semantic statements. The RDF data model today forms the basis of various knowledge bases (DBpedia, Wikidata) and the knowledge graphs that are currently emerging in a variety of contexts. For this reason, the Semantic Web and the linking of openly accessible data (Linked Open Data) continue to be of particular interest to the digital humanities (Beretta 2021, Beretta & Alamercery 2020, Hiltmann & Riechert 2020, Meroño-Peñuela 2017, Meroño-Peñuela et al. 2014, Pollin 2017, Wettlaufer 2018, Wettlaufer et al. 2015).

This full-day workshop offers an introduction to the topic of "Semantic Web and Linked Open Data" with a focus on historical studies. It aims at participants without prior knowledge in the field of Semantic Web/Linked Open Data. The workshop divides into four parts, with practical exercises taking up about two thirds of the time.

At the beginning of the workshop, the basics of the Semantic Web, the Resource Description Framework as well as associated WWW standards are taught in an introductory presentation. Special attention will be paid to Wikidata and SPARQL[2], both of which will play a prominent role in the subsequent exercises. The following topics are planned for this first, introductory part:

> The idea of the Semantic Web: brief presentation of the basic idea, which developed from the core problem Natural Language Processing. The Resource Description Framework (RDF) as a foundation for formalised statements. The importance of stable URIs for the functioning of the Semantic Web. The basics of the query language SPARQL: namespaces and their importance, especially in the Semantic Web. RDF Schema and Ontologies for formulating more complex statements. Linked Open Data, Knowledge Graphs and the LOD cloud. Wikidata (and DBpedia) as central nodes of the LOD cloud. Overview of resources in the Semantic Web and the LOD cloud for the historical sciences.

In the second part of the workshop, participants will learn to use the Linked Data platform Wikidata and the basics of the query language SPARQL in practical exercises. Wikidata is not only currently the largest freely available knowledge graph, it also offers data under free licences and, just like Wikipedia, allows free collaboration on building the knowledge base. This makes it one of the most relevant data sources for the Semantic Web (Jacobsen et al. 2018). In addition, Wikidata's Query Service [3] [3]provides a low-threshold introduction to SPARQL that does not require any local installations or technical knowledge. Everything needed for the exercise is an internet-enabled device (preferably a laptop) and a Wikidata account. The Wikidata graphical user interface is also well

[1] https://www.wikidata.org/

[2] https://www.w3.org/TR/sparql11-query/

[3] https://query.wikidata.org/

suited for teaching the theoretical concepts of the Semantic Web without requiring any knowledge of computer science.

The first tutorial section therefore explains the data structures of Wikidata. Using an example entry, the theoretical concepts from the introductory part of the workshop are shown in their application within Wikidata. The example entry represents an item that can be linked to other items via properties. Such a link creates a statement. This can be described in more detail by so-called qualifiers. Qualifiers are particularly relevant for the use of Wikidata in the historical sciences, as they make it possible to specify the origin of information as well as potential limits to its validity. This also enables the modelling of divergent research results. In addition to the structure of Wikidata entries. This part of the workshop also covers an introduction to various RDF namespaces relevant in Wikidata. Finally, a Wikidata item can be annotated with multiple labels that allow it to be described and named in different languages.

The handling of these data structures will then be practised using data sets from the research project Germania Sacra [4], which deals with the research of ecclesiastical institutions and persons of the Middle Ages and the early modern period. The data will be prepared for the workshop in such a way that the participants can enter it into Wikidata on their own. Entering the data manually deepens the understanding of the data structures, but is not feasible with larger amounts of data. As an outlook the tools "QuickStatements"[5] and OpenRefine[6] are introduced, which enable the serial import of larger amounts of data.

The next block of the tutorial deals with the basics of the query language SPARQL. The aim is for the participants to develop an understanding of how research questions in the humanities can be formulated as queries and implemented with SPARQL on Wikidata. To do this, they must first investigate how the concepts in the question are modelled in Wikidata. Then a suitable query is formulated. This always follows the same basic structure with SELECT, WHERE and, if necessary, OPTIONAL, which can be supplemented with further, more complex commands. (See figure 2). These basic building blocks of SPARQL are first practised with simple queries such as "Find all records for bishops with a WIAG identifier". The exercise then goes into more depth on concatenating query patterns and querying labels from Wikidata. These basics of the query language SPARQL will also be linked to the formal basics of the Semantic Web by referring back to the theoretical part of the workshop. During the tutorial, demonstrations of new concepts alternate with work on exercises that build on each other. During the exercises, participants are invited to exchange ideas with others, compare results and ask questions. In this way, they gradually work out how to query the data sets entered in the first part of the tutorial.

The technical framework in which Wikidata is embedded offers users freely available tools with which queried data can be evaluated and visualised. These include an interactive graph, but also a timeline, a map, an image gallery and many other visualisation options. The final part of the tutorial explains and explores these tools. Finally, the result of this hands-on tutorial part is a visualisation of the data that the participants entered into Wikidata at the beginning of the workshop and queried with the acquired SPARQL knowledge.

In the fourth and final part of the workshop, participants will also learn how historical data is modelled in Wikidata. Following on from this, the potential of Wikibase for the historical sciences will be discussed. This includes the advantages and disadvantages of collecting data through a

[4] http://www.germania-sacra.de

[5] https://quickstatements.toolforge.org

[6] https://openrefine.org

collaborative process. A critical look will be taken at the questions of data quality, data modelling and the completeness of the data.

Alternatives to Wikidata are also discussed in this part of the workshop. Wikibase[7], the technical framework underlying Wikidata, can also be used as a stand-alone instance independent of Wikidata. This solution has the potential to use the advantages of the system to link data and at the same time build an independent data collection curated by the researchers themselves. There are several examples of this in the Digital Humanities, which will be briefly presented.

Smaller pilot projects, which were realised in the context of courses at the University of Göttingen, serve as the basis for a subsequent practice-oriented examination of independent Wikibase instances. Not only Wikidata, but also the Wikibase instance FactGrid[8] operated by the Gotha Research Centre at the University of Erfurt was enriched with data by students of the course. The focus was on the bishops of the Old Empire, who processed their servitude payments to the papal curia with the help of the Florentine banking house of the Medici family. With the knowledge acquired in the workshop, the participants can query this data and gain an insight into the practical benefits of Linked Open Data in the historical sciences.

During the workshop, the participants will be provided with the exercises including a sample solution. They will also receive an overview of all commands used in the exercise. The didactic concept includes an introductory teaching of basic knowledge, interactive exercises, group work and hands-on examples. After conveying the basic collaborative idea of the Semantic Web and an introduction to the underlying technical standards (RDF, RDFS, SPARQL), the focus is on imparting skills when using SPARQL on the Wikidata platform. A look at already existing applications of LOD in the historical sciences completes the workshop.

Bibliography

Beretta, Francesco. 2021. “A challenge for historical research: making data FAIR using a collaborative ontology management environment (OntoME)”, Semantic Web 12: 2, Special issue on Semantic Web for Cultural Heritage. https://doi.org/10.3233/SW-200416

Beretta, Francesco and Vincent Alamercery. 2020. "Du projet symogih.org au consortium Data for History - La modélisation collaborative de l’information au service de la production de données géo-historiques et de l’interopérabilité dans le web sémantique." Revue ouverte d’ingénierie des systèmes d’information 1(3):1-15. https://doi.org/10.21494/ISTE.OP.2020.0532

Berners-Lee, Tim, James Hendler and Ora Lassila. 2001. The Semantic Web: a new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. Scientific American 284(5), 34-43.

Hiltmann, Torsten and Thomas Riechert. 2020. “Digital Heraldry. The State of the Art and New Approaches Based on Semantic Web Technologies.” In L’édition en ligne de documents d’archives médiévaux, ed. by Christelle Balouzat-Loubet, Turnhout, 102-125.

Jacobsen, Annika et al. 2018. “Wikidata as an intuitive resource towards semantic data modeling in data FAIRification.” In Semantic Web Applications and Tools for Health Care and Life Sciences. Proceedings of the 11th International Conference Semantic Web Applications and Tools for Life Sciences (SWAT4HCLS 2018). Ed. by Christoper J.O. Baker, CEUR workshop proceedings Vol. 2275. http://ceur-ws.org/Vol-2275/

[7] https://www.wikimedia.de/projects/wikibase

[8] https://database.factgrid.de

Meroño-Peñuela, Albert, Ashkan Ashkpour, Marieke van Erp, Kees Mandemakers, Leen Breure, Andrea Scharnhorst, Stefan Schlobach and Frank van Harmelen. 2015. "Semantic Technologies for Historical Research: A Survey." Semantic Web Journal 6: 539-564.

Meroño-Peñuela, Alberto. 2017. "Digital Humanities on the Semantic Web: Accessing Historical and Musical Linked Data," Journal of Catalan Intellectual History (JOCIH) 1(11): 144-149. DOI: 10.1515/jocih-2016-0013

Pollin, Christopher and Georg Vogeler. 2017. Semantically Enriched Historical Data. Drawing on the Example of the Digital Edition of the 'Urfehdebücher der Stadt Basel", in: A. Adamou, E. Daga and L. Isaksen (Hg.): 2nd Workshop on Humanities in the Semantic Web (WHiSe), 27-32.

Wettlaufer, Jörg. 2018. "Der nächste Schritt? Digitale Editionen und Semantic Web." In Zeitschrift für Digitale Geisteswissenschaften, Sonderheft "Digitale Metamorphosen", ed. by Roland S. Kamzelak and Timo Steyer (= Sonderband der Zeitschrift für digitale Geisteswissenschaften, 2). DOI: 10.17175/sb002_007

Wettlaufer, Jörg, Christopher Johnson, Martin Scholz, Marc Fichtner, Sree Ganesh Thotempudi. 2015. "Semantic Blumenbach: Exploration of Text-Object Relationships with Semantic Web Technology in the History of Science," Digital Scholarship in the Humanities (DSH), Special Issue 'Digital Humanities 2014'; ed. by Melissa Terras, Claire Clivaz, Deb Verhoeven and Frederic Kaplan, 30 Supplement 1: i187-i198 https://academic.oup.com/dsh/article/30/suppl_1/i187/364720/

Workshop Organizers:

Bärbel Kröger, Göttingen Academy of Sciences and Humanities, Geiststraße 10, 37073 Göttingen, Germany, bkroege@gwdg.de, +49-551-3921558

Bärbel Kröger works as a digital humanist for the research project 'Germania Sacra', which investigates the history of the Church of the Holy Roman Empire. She is responsible for designing databases on the history of the Middle Ages and the early modern period. A special focus is on prosopography. One of her special interests is the use of crowd-based technologies such as Wikibase for scholarly research.

Johanna Störiko, Georg-August-University Göttingen, Institute for Digital Humanities, Nikolausberger Weg 23, 37073 Göttingen, Germany, johanna.stoeriko@uni-goettingen.de

Johanna Störiko is a computer scientist who specialized in Digital Humanities. She also holds a Bachelor of Arts in Historical Studies. Her fields of interest include digitalization and multimodal analysis of historical sources, Semantic Web technologies and Software Engineering in the Digital Humanities.

Dr. Jörg Wettlaufer, Academy of Sciences and Humanities, Digital Academy, Theaterstr. 7, 37073 Göttingen, Germany, jwettla@gwdg.de, +49 551 39 37047, https://joergwettlaufer.de

Jörg Wettlaufer is head of the digital academy and digital historian. His research interests include information retrieval, semantic web applications in the humanities and all topics related to digital history. He advocates towards an intensive dialogue between the Humanities and the Digital Humanities.

Target group: Beginners, historians, humanities scholars. No prior knowledge of the topic is required.

Number of possible participants: 5-25

Required technical equipment: From prior experience (https://digigw.hypotheses.org/4000) we know that we need no special technical equipment apart from a projector and wifi. However, it

should be a space that allows active support of the participants in the hands-on part of the session. Participants must also bring along appropriate hardware for the practical exercises themselves.

Intended Length: full-day workshop

Tentative schedule of the workshop:
9:00 Introductory round and introduction to the event (30 min.)
9:30 Part 1: Introduction Basics of the Semantic Web and LOD (60 min.)
10:30 Coffee break
11:00 Part 2: Exercise with Wikidata using examples (90 min.)
12:30 Lunch break
13:30 Part 3: Exercise SPARQL on Wikidata (90 min.)
15:00 Break
15:30 Part 4: Examples for the use of Wikidata and LOD in the historical sciences (90 min.)
17:00 End