

# SPARQL for (Digital) Humanists – Querying Wikidata and the MiMoTextBase

**Röttgermann, Julia**

roettger@uni-trier.de  
Trier Center for Digital Humanities, Trier University, Germany

**Duan, Tinghui**

duan@uni-trier.de  
Trier Center for Digital Humanities, Trier University, Germany

**Hinzmann, Maria**

hinzmannm@uni-trier.de  
Trier Center for Digital Humanities, Trier University, Germany

**Klee, Anne**

klee@uni-trier.de  
Trier Center for Digital Humanities, Trier University, Germany

**Konstanciak, Johanna**

konstanciak@uni-trier.de  
Trier Center for Digital Humanities, Trier University, Germany

**Schöch, Christof**

schoech@uni-trier.de  
Trier Center for Digital Humanities, Trier University, Germany

**Steffes, Moritz**

steffesm@uni-trier.de  
Trier Center for Digital Humanities, Trier University, Germany

## Introduction

Not only in cultural and memory institutions, but also in DH projects, an increasing uptake of the Linked Open Data paradigm is currently visible. How can data be presented in a way that is open, easily accessible, interoperably linked, machine-readable and available in the long-term? In the project "Mining and Modeling Text", we have chosen the open and free software Wikibase, which includes its own SPARQL endpoint and is used by a growing number of research projects besides Wikidata.<sup>1</sup>

The workshop aims to share theoretical and practical knowledge about modeling data in the humanities and especially literary history in the paradigm of Linked Open Data, to introduce the syntax of the query language SPARQL, and to demonstrate the advantages of modeling and providing data as knowledge graphs.

<sup>2</sup> The focus is on teaching SPARQL in theoretical and practical sessions. Participants should gain the competence to understand

the structure of SPARQL and to write both simple and moderately complex queries on their own.

```
1 # Which thematic concepts are represented in the corpus?
2 #defaultView:BubbleChart
3 prefix wd:<http://data.mimotext.uni-trier.de/entity/>
4 prefix wdt:<http://data.mimotext.uni-trier.de/prop/direct/>
5 SELECT ?topLabel (count(*) as ?count)
6 WHERE {
7   ?item wdt:P36 ?top . # P36 = thematic concept of the literary work
8   ?top rdfs:label ?topLabel .
9   filter(lang(?topLabel) = "en")
10 }
11 GROUP BY ?topLabel
12 ORDER BY desc(?count)
13
```

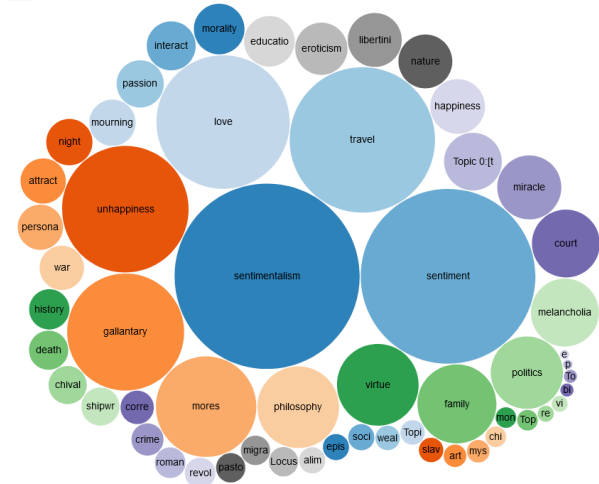


Fig. 1: Which thematic concepts are represented in the French novel 1751-1800? Query: <https://tinyurl.com/2fgqotp7>.

## Linked Open Data for the Humanities

We can observe that there is an increasing interest in the DH community to publish data in the form of Linked Open Data and to link it with the Semantic Web (Hogan et al. 2021; Nešić et al. 2021; Thornton et al. 2021; Alves 2022; Dörpinghaus 2022; Ohmukai / Yamada 2022; Zhao 2022). The project "Mining and Modeling Text" also aims at aggregating data from different information sources and linking them to further resources in the sense of the Linked Open Data paradigm (Schöch et al. 2022). The advantages of the complex process of mining and modeling the data only become clear in the diverse and flexible query options and can therefore not be separated from SPARQL.

SPARQL (SPARQL Protocol and RDF Query Language) is a graph-based query language for RDF (Resource Description Framework) published by the W3C in 2008. RDF is a data model used to represent resources on the World Wide Web. It is the central standard of the W3C, representing semantic data in the characteristic triple structure composed of subject-predicate-object. Starting from a single triple, the structure of a knowledge graph will be explained in the workshop and the "translation" of research questions in natural language into the SPARQL syntax will be practiced.

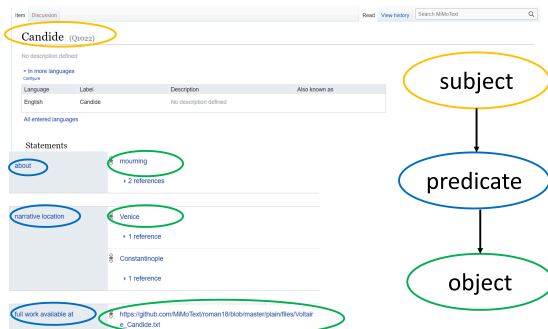


Fig. 2: Triple structure consisting of subject-predicate-object, here an example for a literary work ("Candide") from the MiMoTextBase: <http://data.mimotext.uni-trier.de/wiki/Item:Q1022>.

The SPARQL query language is composed of several building blocks: *pattern matching* (filtering of the dataset), *solution modifier* (processing of intermediate results) & *output* (output as table or graph; Arenas et al. 2010: 282). SPARQL allows users to identify new patterns in the data by recombining data sets and to formulate hypothesis-driven queries. The strengths of this query language, along with the structuring of knowledge using RDF triples, will be demonstrated through application examples during the workshop.

SPARQL queries are often issued within a single knowledge graph. However, it is also possible to query across multiple knowledge graphs, so-called federated queries (Prud'hommeaux / Buil-Aranda 2013). Here, the full potential of Linked Open Data becomes evident, as multiple datasets can be intelligently combined and insights can be gained from the combination of multiple graphs. Furthermore, redundancies between the different graphs can be avoided through these links, because in the sense of Linked Open Data, the collected knowledge of other graphs can be accessed. In the workshop, federated queries with Wikidata will be introduced and it will be shown to what extent benefits can be derived from these links.

## Workshop concept

The workshop covers basic knowledge and possibilities offered by the Linked Open Data paradigm. The multilingual knowledge graph MiMoTextBase on the domain of 18th century French literature created in the project as well as Wikidata will serve as illustrative examples.

## Learning Goals

The workshop aims to provide practical knowledge: How to write SPARQL queries? How to combine more than one knowledge base? What benefit can a knowledge graph offer regarding questions of literary history in particular and the humanities in general?

In the half-day workshop, the multilingual knowledge graph MiMoTextBase (Hinzmann et al. 2022a) of the project "Mining and Modeling Text" is introduced and usage scenarios are presented and visualized. In addition, participants will learn the basics of the query language SPARQL and formulate their own queries.

Specific learning objectives are the acquisition of basic knowledge about Semantic Web and RDF, Linked Open Data, Wikidata

Graph; deeper knowledge about SPARQL and the ability to formulate own SPARQL queries; familiarization with the software Wikibase and exploration of the visualization possibilities.

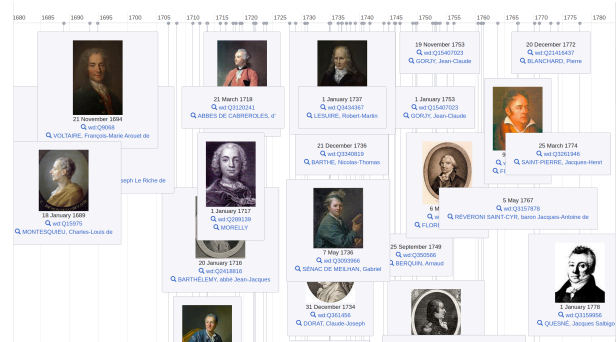


Fig. 3: Federated queries in SPARQL allow to query several knowledge graphs simultaneously. Example: <https://tinyurl.com/28x5ajjy>.

## Target audience and requirements

The workshop is directed at digital humanities scholars with an interest in Linked Open Data and SPARQL. No special prior knowledge is necessary. Participants need a laptop and an internet connection.

## Structure

The workshop consists of consecutive sessions, each of which combines input phases and practicing phases. A detailed tutorial page will be made available on GitHub Pages in advance. The participants (and all other interested parties) will be given access to it and it can also be useful in the follow-up for in-depth study (Hinzmann et al. 2022b).

The workshop itself will focus on three blocks in which the formulation of SPARQL queries will be practiced in an alternation of short impulses and more detailed hands-on sessions (for details see the schedule in the appendix). The three sections each have a different focus and are based on each other. In this way, even participants with no previous knowledge are gradually introduced to increasingly complex queries. The level of difficulty also progresses within the individual blocks, so that the focus will be on the individual writing and adaptation of sample queries and the clarification of all questions that arise in the process.

1. The first part focuses on queries about literary works. With regard to SPARQL, the central principles (such as writing simple triple patterns and possibilities of combining several triple patterns into increasingly complex queries) are dealt with here first. The benefit of such combination possibilities becomes particularly clear with the focus on the domain of the French Enlightenment novel given by the MiMoTextBase.

2. In the second part, we focus on Wikidata as the largest public knowledge graph, which can also be understood as a hub (Neubert 2017), and on authors as entities. Authors are relevant in all disciplines of the humanities and are an important hinge between different knowledge graphs. Related to the SPARQL syntax, we go a step further and integrate functions like OPTIONAL and FILTER

to extend the range of query possibilities. A start is made here with queries to authors of the MiMoTextBase domain. In the next step, the participants can explore the data of other person entities that are of interest to them in Wikidata.

3. The third part links the previous two parts on several levels. The focus is on 'federated queries', where we will concentrate on queries that span MiMoTextBase and Wikidata. The more detailed look at authors and literary works will be continued and deepened here. In this concluding part, the importance of standards and shared data models (ontologies or 'entity schemata') as well as the linking of resources will become particularly clear.<sup>3</sup> All authors within the MiMoTextBase, for which there are also Wikidata items, are linked to them via the property "exact match", which provides additional information and opens up various query options.<sup>4</sup> The Wikibase infrastructure also offers various exploration and visualization options, which are introduced as examples (marker clusters for geo-data, timelines for birth dates, etc.).

In the concluding discussion, we want to open a space for a critical perspective on developments in the area of the Semantic Web, for example, the question of which monopolization forces and market forces have an influence (van Hooland / Verborgh 2014: 247-48; Singhal 2012). Finally, the most important application possibilities and questions will be summarized and further resources (DuCharme 2013; Lincoln 2015; van Hooland / Verborgh 2014; Blaney 2017) as well as possibilities for cooperation will be addressed.

## Appendix

### Workshop organization

Maximum number of participants: 25. We need a room with WiFi and a projector. We would like to offer a hybrid scenario if that is possible.

### Workshop procedure (4h)

10 min. 20 min.	Welcome (introduction possibly via mentimeter) Introduction: input on Semantic Web, RDF, Linked Open Data for literary history using the Mining and Modeling Text project as an example. Wikidata & Wikibase Ecosystem, Multilingual Graph.
20 min.	SPARQL Part 1 (MiMoTextBase): (a) Input on SPARQL basics (interactive phase), SPARQL syntax, data visualization possibilities in Wikibase.
35 min.	(b) Practical part: Adapting existing and formulating simple, own SPARQL queries on the MiMoTextBase (breakout session or group work).
(15 min.)	Break
20 min.	SPARQL Part 2 (Wikidata): (a) Input: More advanced elements of SPARQL syntax such as GROUP BY and FILTER; data model for authors on Wikidata.
35 min.	(b) Practice: formulating slightly more complex queries on Wikidata.
(15 min.)	Break
20 min.	SPARQL Part 3 (Federated Queries): (a) Input: Advanced SPARQL queries: federated queries, defining prefixes, marker clusters and co.
35 min.	(b) Practice: Using federated queries and Co.
15 min.	Concluding discussion and recommending further resources

### Acknowledgements

"Mining and Modeling Text" (Trier University, Trier Center for Digital Humanities) is funded by the Rhineland-Palatinate Research Initiative 2019-2023.

### Contributors

The workshop will be conducted by members of the Linked Open Data project "Mining and Modeling Text". The interdisciplinary project has its own SPARQL endpoint within the infrastructure of a Wikibase instance.

- Tinghui Duan (duan@uni-trier.de; Trier Center for Digital Humanities), research interests: Computational Literary Studies and Computational Linguistics.
- Maria Hinzmann (hinzmannm@uni-trier.de; Trier Center for Digital Humanities | History Department / Digital Humanities; University of Wuppertal), research interests: Digital Literary and Cultural Studies, Linked Open Data.
- Anne Klee (klee@uni-trier.de; Trier Center for Digital Humanities), research interests: Digital Text Processing; Digital Lexicography.
- Johanna Konstanciak (konstanciak@uni-trier.de; Trier Center for Digital Humanities), research interests: Digital Text Processing; XML/Web-Technologies.
- Julia Röttgermann (roettger@uni-trier.de; Trier Center for Digital Humanities), research interests: French literature, Linked Open Data, Text mining.
- Christof Schöch (schoech@uni-trier.de; Trier Center for Digital Humanities), research interests: Computational Literary Studies.

### Notes

1. Some current examples of research projects that use Wikibase: Enslaved (Zhou et al. 2020), ArtBase (Rhizome 2021), Fact Grid (Simons 2022; cf. Brunner 2022).
2. For an introduction to knowledge graphs see Hogan et al. 2021.
3. This linkage corresponds to the 5th star in Berners-Lee's (2006) Linked Open Data model.
4. Cf. the corresponding query on the MiMoTextBase: <https://tinyurl.com/2cg2nhuq>.

## Bibliography

- Alves, Daniel (ed.) (2022): "IJHAC: A Journal of Digital Humanities. Special Issue: Linked Open Data in the Arts and the Humanities", in: IJHAC 16 (1): <https://www.eupublishing.com/doi/epdf/10.3366/ijhac.2022.0271> [20.04.2023].
- Arenas, Marcelo / Gutierrez, Claudio / Pérez, Jorge (2010): "On the Semantics of SPARQL", in: de Virgilio, Roberto / Giunchiglia, Fausto / Tanca, Letizia (eds.): *Semantic Web Information Management: A Model-Based Perspective*. Berlin, Heidelberg: Springer 281–307. DOI: 10.1007/978-3-642-04329-1\_13.
- Berners-Lee, Tim (2006): *Linked Data – Design Issues*. 27. Juli 2006. <https://www.w3.org/DesignIssues/LinkedData.html> [20.04.2023].
- Blaney, Jonathan (2017): "Introduction to the Principles of Linked Open Data.", in: *Programming Historian*. DOI: 10.46430/phen0068.

- Brunner, Katharina** (2022): FactGrid wants to become part of the Wikidata federation ecosystem – FactGrid. <https://blog.factgrid.de/archives/2922> [20.04.2023].
- Dörpinghaus, Jens** (2022): "Wissensgraphen: Interdisziplinäre Perspektiven für Linked Data in den Geistes- und Sozialwissenschaften", in: *Zeitschrift für digitale Geisteswissenschaften* 07. DOI: 10.17175/2022\_011.
- DuCharme, Bob** (2013): *Learning SPARQL*. Sebastopol, United States: O'Reilly Media.
- Hinzmann, Maria / Klee, Anne / Konstanciak, Johanna / Röttgermann, Julia / Schöch, Christof / Steffes, Moritz** (2022a): *MiMoTextBase*. <https://data.mimotext.uni-trier.de> [20.04.2023].
- Hinzmann, Maria / Klee, Anne / Konstanciak, Johanna / Röttgermann, Julia / Schöch, Christof / Steffes, Moritz** (2022b): *MiMoTextBase Tutorial*. [https://mimotext.github.io/MiMoTextBase\\_Tutorial/](https://mimotext.github.io/MiMoTextBase_Tutorial/) [20.04.2023].
- Hogan, Aidan et al.** (2021): "Knowledge Graphs", in: *Synthesis Lectures on Data, Semantics, and Knowledge* 12 (2): 1–257. DOI: 10.2200/S01125ED1V01Y202109DSK022.
- Hooland, Seth van / Verborgh, Ruben** (2014): *Linked Data for Libraries, Archives and Museums: How to clean, link and publish your metadata*. London: Facet Publishing.
- Ikonić Nešić, Milica / Stanković, Ranka / Rujević, Biljana** (2021): "Serbian ELTeC Sub-Collection in Wikidata", in: *Infothea* 21 (2): 60–86. DOI: 10.18485/infothea.2021.21.2.4.
- Lincoln, Matthew** (2015): "Using SPARQL to access Linked Open Data", in: *The Programming Historian*. DOI: 10.46430/phen0047.
- Neubert, Joachim** (2017): "Wikidata as a Linking Hub for Knowledge Organization Systems? Integrating an Authority Mapping into Wikidata and Learning Lessons for KOS Mappings." in: *Proceedings of the 17th European NKOS workshop*. <http://ceur-ws.org/Vol-1937/paper2.pdf> [20.04.2023].
- Ohmukai, Ikki / Yamada, Taizo** (eds.) (2022): *Digital Humanities 2022. Conference Abstracts. Responding to Asian Diversity*. Tokyo: ADHO. <https://dh2022.dhii.asia/dh2022bookofabsts.pdf> [20.04.2023].
- Prud'hommeaux, Eric / Buil-Aranda, Carlos** (2013): SPARQL 1.1 Federated Query, in: *W3C Recommendation*. <https://www.w3.org/TR/sparql11-federated-query/> [20.04.2023].
- Rhizome** (2021): "The ArtBase Relaunches: Welcome to Linked Open Data". *Rhizome*. <http://rhizome.org/editorial/2021/apr/26/the-artbase-relaunches-welcome-to-linked-open-data/> [20.04.2023].
- Sack, Harald / Alam, Mehwish** (2020): *Knowledge Graphs*. Potsdam. <https://open.hpi.de/courses/knowledge-graphs2020> [20.04.2023].
- Schöch, Christof / Hinzmann, Maria / Röttgermann, Julia / Klee, Anne / Dietz, Katharina** (2022): "Smart Modelling for Literary History", in: *IJHAC: International Journal of Humanities and Arts Computing [Special issue on Linked Open Data]* 16 (1): 78–93. DOI: 10.3366/ijhac.2022.0278.
- Simons, Olaf** (2022): *FactGrid*. Forschungszentrum Gotha der Universität Erfurt. [database.factgrid.de](https://database.factgrid.de) [20.04.2023].
- Singhal, Amit** (2012): "Introducing the Knowledge Graph: things, not strings." in: *Google (blog)*. <https://blog.google/products/search/introducing-knowledge-graph-things-not/> [20.04.2023].
- Thornton, Katherine / Seals-Nutt, Kenneth / Van Renmoortel, Marianne / Birkholz, Julie M. / De Potter, Pieterjan** (2021): "Linking Women Editors of Periodicals to the Wikidata Knowledge Graph", in: *Semantic Web Journal Special Issue Cultural Heritage* 2021. <http://www.semantic-web-journal.net/content/linking-women-editors-periodicals-wikidata-knowledge-graph>. [20.04.2023].
- Zhao, Fudie** (2022): "How to Critically Utilise Wikidata - A Systematic Review of Wikidata in DH Projects". in: Ohmukai, Ikki / Yamada, Taizo (eds.): *Digital Humanities 2022 - Conference Abstracts*. The University of Tokyo, Japan: DH2022 Local Organizing Committee. 608–610. <https://dh2022.dhii.asia/dh2022bookofabsts.pdf> [20.04.2023].
- Zhou, Lu / Shimizu, Cogan / Hitzler, Pascal / Sheill, Alicia M. / Estrecha, Seila G. / Foley, Catherine / Tarr, Duncan / Rehberger, Dean** (2020): "The Enslaved Dataset: A Real-world Complex Ontology Alignment Benchmark using Wikibase." in: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. New York, NY, USA: Association for Computing Machinery. 3197–3204. DOI: 10.1145/3340531.3412768.