Towards an ontology for literary history: issues of complexity and scale when constructing the MiMoTextBase

MiMoText

Christof Schöch and Maria Hinzmann
with Julia Röttgermann, Tinghui Duan, Anne Klee, Johanna Konstancziak, Matthias Bremm

https://mimotext.uni-trier.de/en

Ontologies for Narrative and Fiction, Groningen | July 3-4, 2023

Kompetenzzentrum
Trier Center For Digital Humanities (TCDH)

Universität Trier

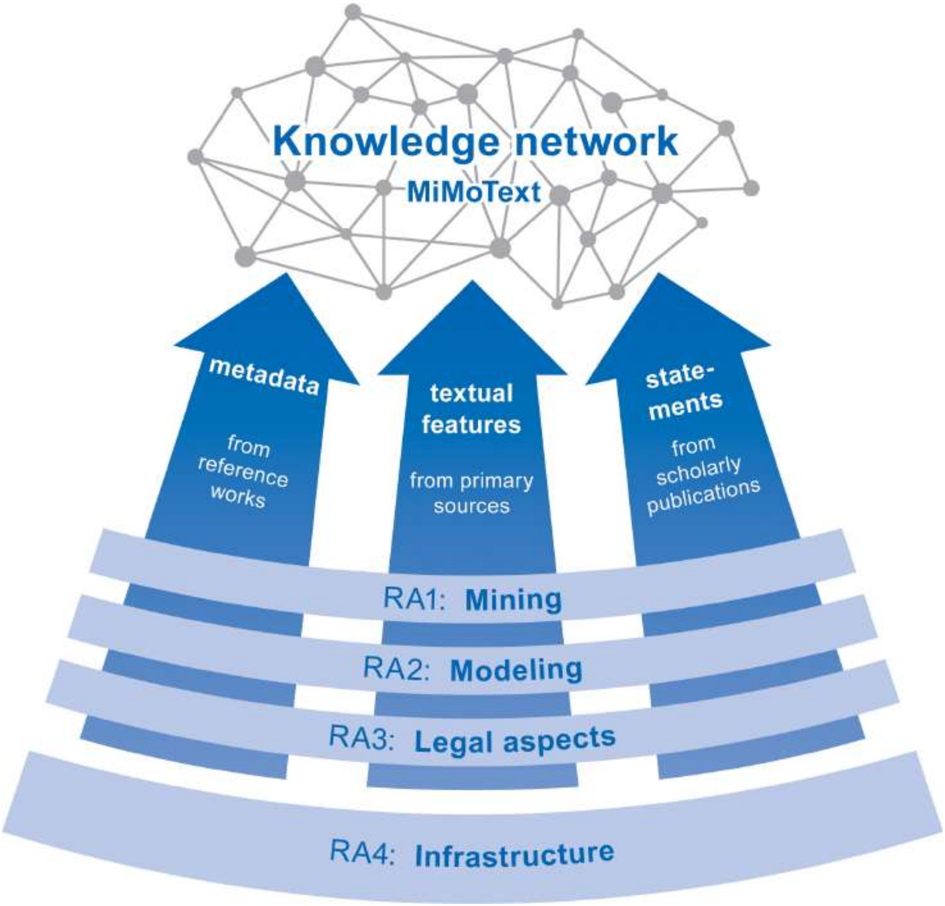
RheinlandPfalz
MINISTERIUM FÜR
WISSENSCHAFT, WEITERBILDUNG
UND KULTUR

Structure

1. Mining and Modeling Text:
   Linked Open Literary History
2. Ontology Design:
   Modules - Dimensions - Ecosystem
3. Conclusion

(1) Mining and Modeling Text: Linked Open Literary History

MiMoText in a nutshell

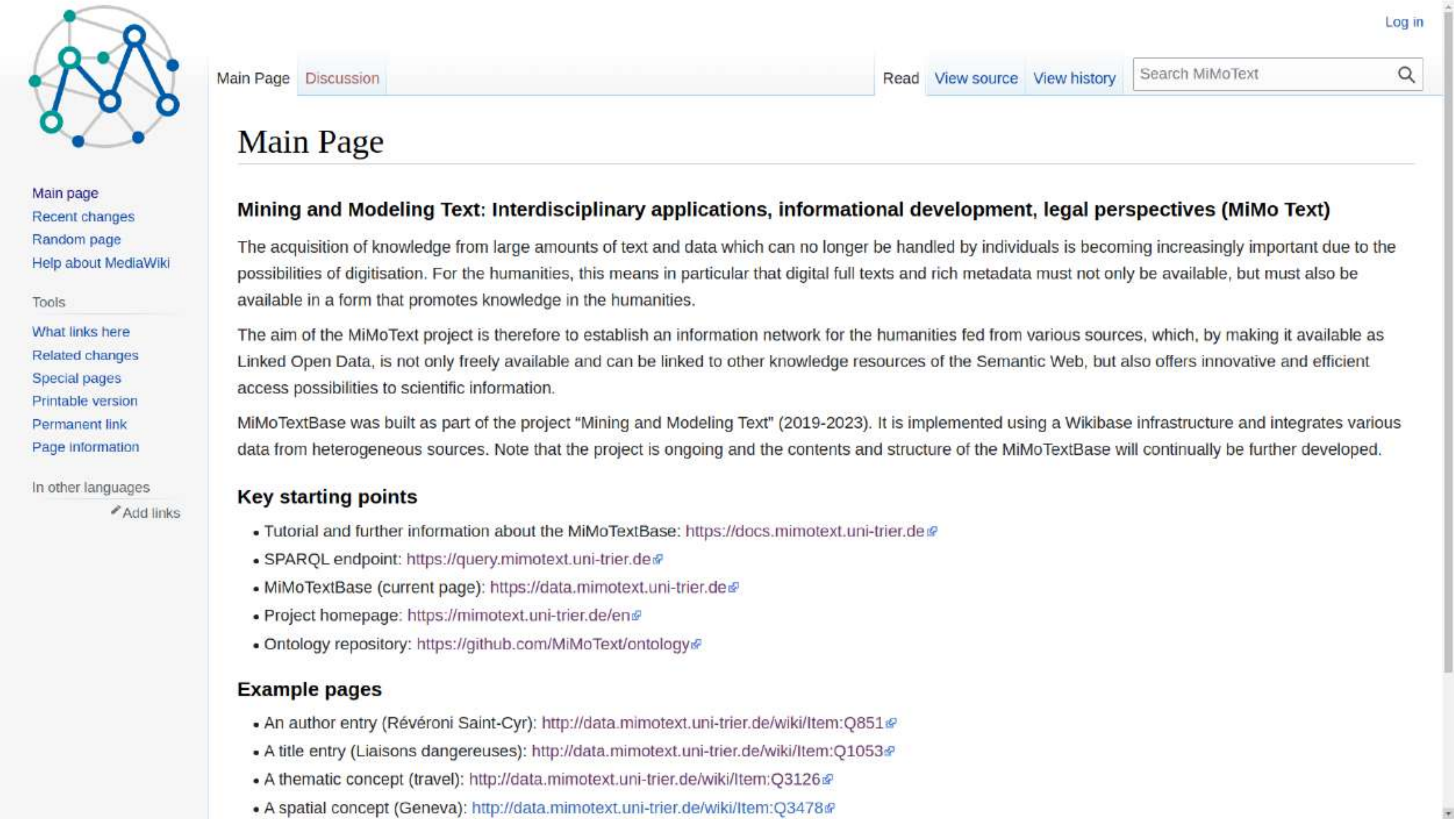


https://mimotext.uni-trier.de/en

Aims of the project

- Our goal: "Wikidata for literary history"
  - An information system for literary history
  - LOD-based, with exploratory interface and SPARQL-endpoint
  - Sort of an "atomization" of literary history into many small statements
  - Held together by taxonomies, ontologies, authority files
- Unlike Wikidata:
  - Much more focused on one domain (French novel 1750-1800)
  - Better coverage for this domain
  - Higher density of assertions for this domain
  - Based on explicit data modeling
  - Facilitates advanced analysis scenarios

Result: the MiMoTextBase



Main Page

Mining and Modeling Text: Interdisciplinary applications, informational development, legal perspectives (MiMo Text)

The acquisition of knowledge from large amounts of text and data which can no longer be handled by individuals is becoming increasingly important due to the possibilities of digitisation. For the humanities, this means in particular that digital full texts and rich metadata must not only be available, but must also be available in a form that promotes knowledge in the humanities.

The aim of the MiMoText project is therefore to establish an information network for the humanities fed from various sources, which, by making it available as Linked Open Data, is not only freely available and can be linked to other knowledge resources of the Semantic Web, but also offers innovative and efficient access possibilities to scientific information.

MiMoTextBase was built as part of the project "Mining and Modeling Text" (2019-2023). It is implemented using a Wikibase infrastructure and integrates various data from heterogeneous sources. Note that the project is ongoing and the contents and structure of the MiMoTextBase will continually be further developed.

Key starting points

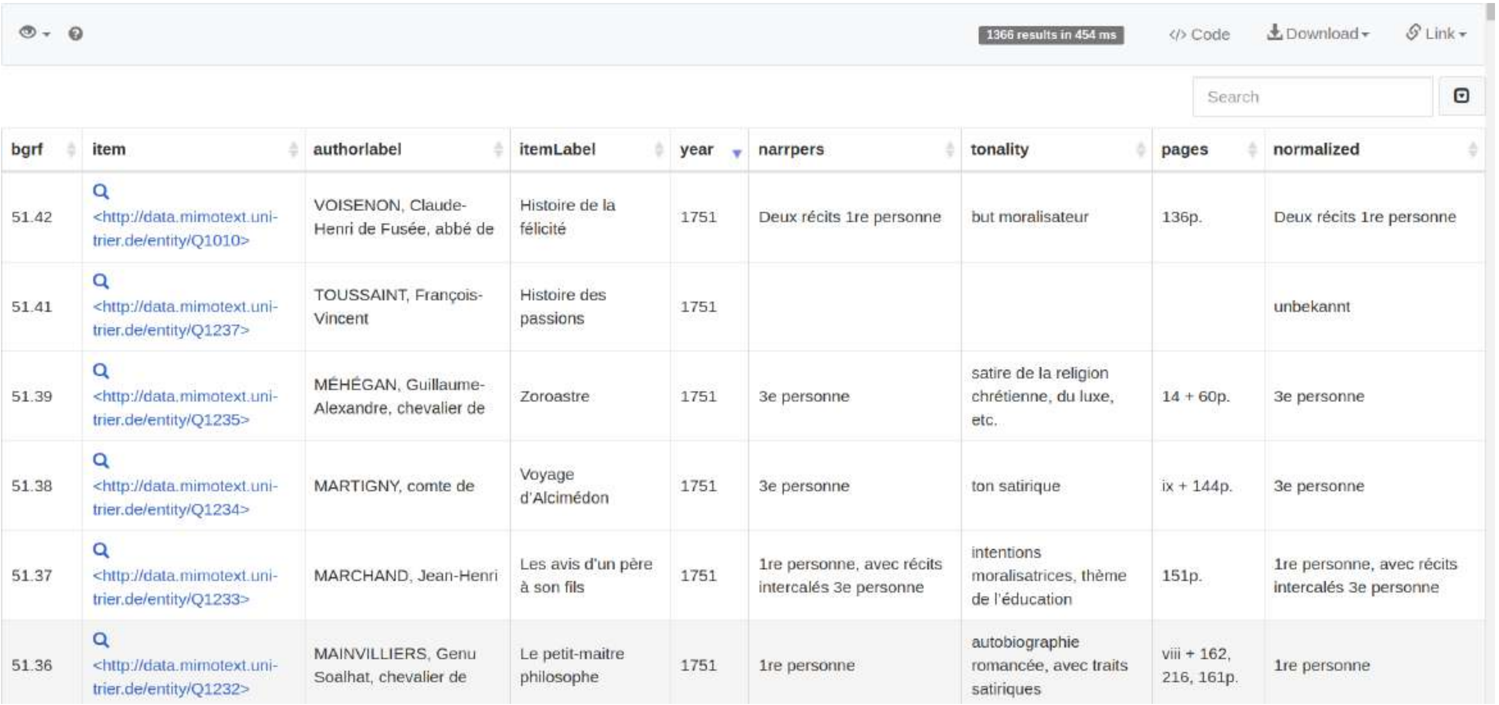
- Tutorial and further information about the MiMoTextBase: https://docs.mimotext.uni-trier.de
- SPARQL endpoint: https://query.mimotext.uni-trier.de
- MiMoTextBase (current page): https://data.mimotext.uni-trier.de
- Project homepage: https://mimotext.uni-trier.de/en
- Ontology repository: https://github.com/MiMoText/ontology

Example pages

- An author entry (Réveroni Saint-Cyr): http://data.mimotext.uni-trier.de/wiki/Item:Q851
- A title entry (Liaisons dangereuses): http://data.mimotext.uni-trier.de/wiki/Item:Q1053
- A thematic concept (travel): http://data.mimotext.uni-trier.de/wiki/Item:Q3126
- A spatial concept (Geneva): http://data.mimotext.uni-trier.de/wiki/Item:Q3478

- http://data.mimotext.uni-trier.de

The SPARQL endpoint



bgrf	item	authorlabel	itemLabel	year	narrpers	tonality	pages	normalized
51.42	<http://data.mimotext.uni-trier.de/entity/Q1010>	VOISENON, Claude-Henri de Fusée, abbé de	Histoire de la félicité	1751	Deux récits 1re personne	but moralisateur	136p.	Deux récits 1re personne
51.41	<http://data.mimotext.uni-trier.de/entity/Q1237>	TOUSSAINT, François-Vincent	Histoire des passions	1751				unbekannt
51.39	<http://data.mimotext.uni-trier.de/entity/Q1235>	MÉHÉGAN, Guillaume-Alexandre, chevalier de	Zoroastre	1751	3e personne	satire de la religion chrétienne, du luxe, etc.	14 + 60p.	3e personne
51.38	<http://data.mimotext.uni-trier.de/entity/Q1234>	MARTIGNY, comte de	Voyage d'Alcimédon	1751	3e personne	ton satirique	ix + 144p.	3e personne
51.37	<http://data.mimotext.uni-trier.de/entity/Q1233>	MARCHAND, Jean-Henri	Les avis d'un père à son fils	1751	1re personne, avec récits intercalés 3e personne	intentions moralisatrices, thème de l'éducation	151p.	1re personne, avec récits intercalés 3e personne
51.36	<http://data.mimotext.uni-trier.de/entity/Q1232>	MAINVILLIERS, Genu Soalhat, chevalier de	Le petit-maitre philosophe	1751	1re personne	autobiographie romancée, avec traits satiriques	viii + 162, 216, 161p.	1re personne

- SPARQL = SPARQL Protocol and RDF Query Language
- Used to formulate complex queries on LOD
- https://query.mimotext.uni-trier.de

Some example queries

- Simple queries
  - List of novels with information from BGRF
  - The number of works written by each author (first 25)
  - The themes of the novels, in French and in English

- Queries with visualization
  - Number of novels published per year
  - The authors (by date of birth, with portrait)
  - The narrative form of the novels (and their prevalence)
  - Book history: formats per year

- Federated queries
  - The narrative locations in all novels (map)

- Compare information from two sources
  - Themes derived from topic modeling compared to themes according to BGRF
  - Combined: themes by BGRF (string, label, Q1) vs. from topic modeling (label, Q21)

(2) Ontology Design:
Modules - Dimensions - Ecosystem

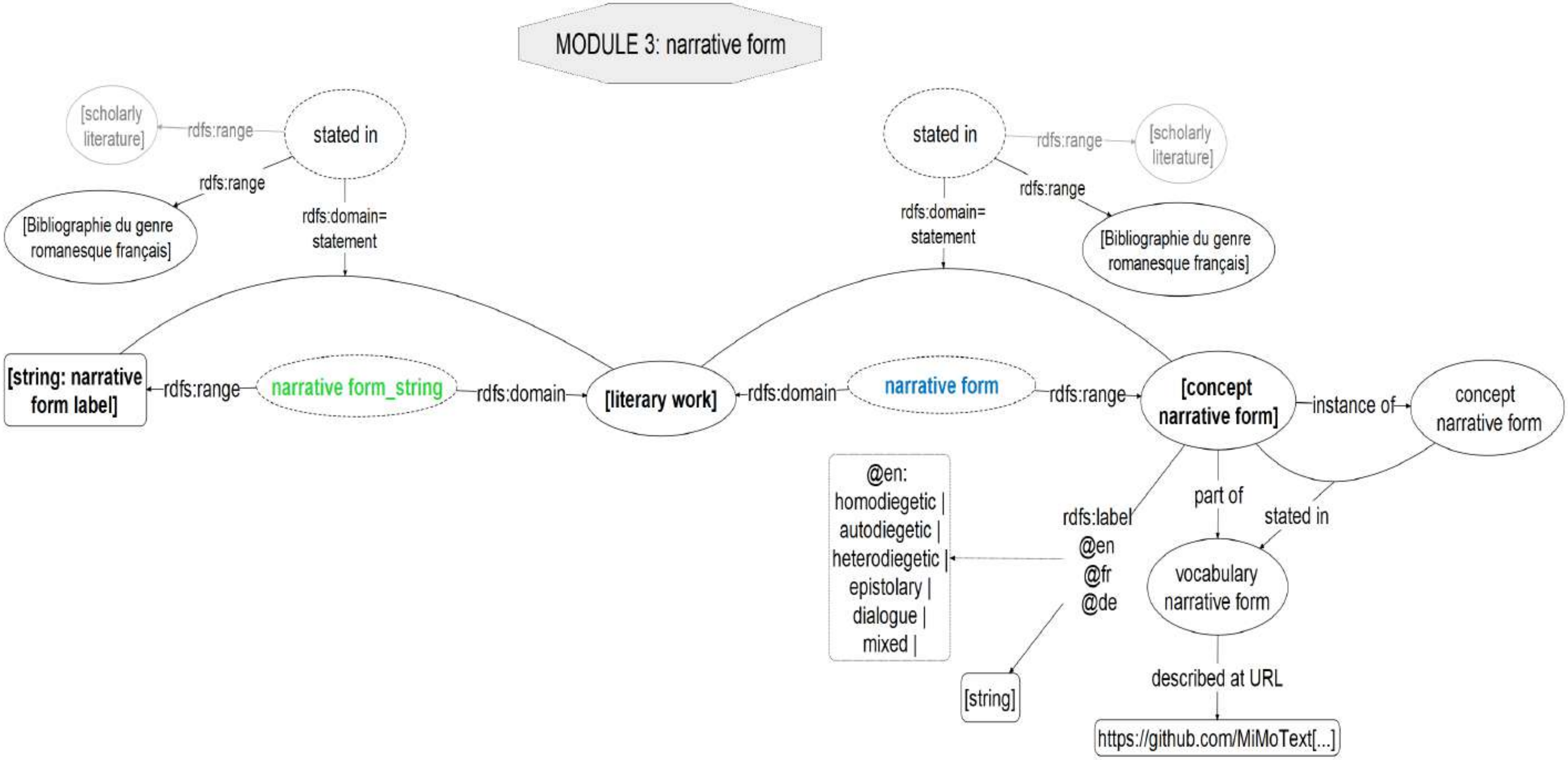
Modules

Overview

- Module 1: theme
- Module 2: space
- Module 3: narrative form
- Module 4: literary work
- Module 5: author
- Module 6: mapping
- Module 7: referencing
- Module 8: versioning & publication
- Module 9: terminology
- Module 10: bibliography
- Module 11: scholarly work

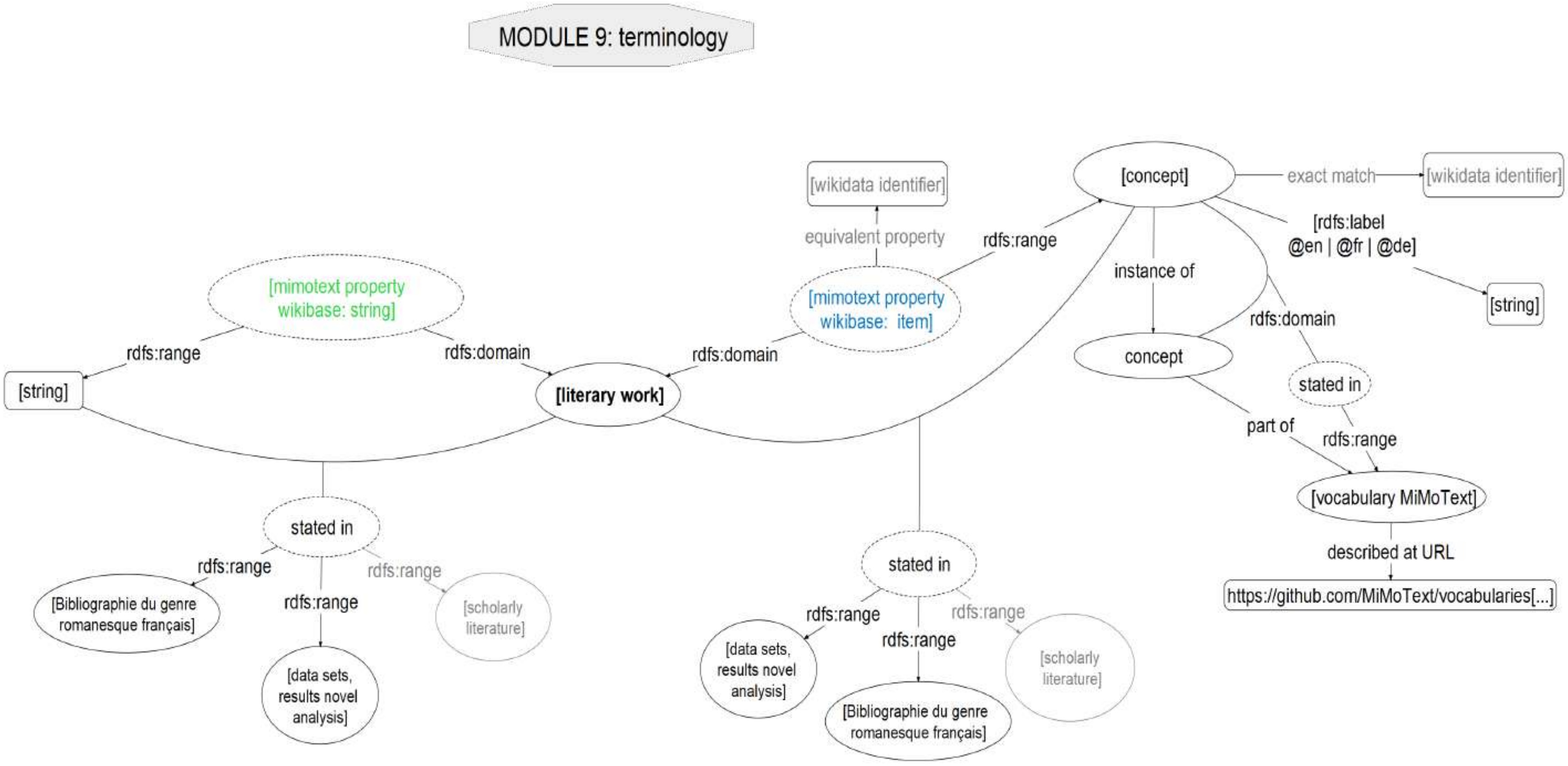
Overview Visualizations Modules 1-11

Module 3: Narrative form



- Cf. Calvo Tello (2021) adapting Genette (1979)
- See Balancing: https://github.com/MiMoText/balance_novels

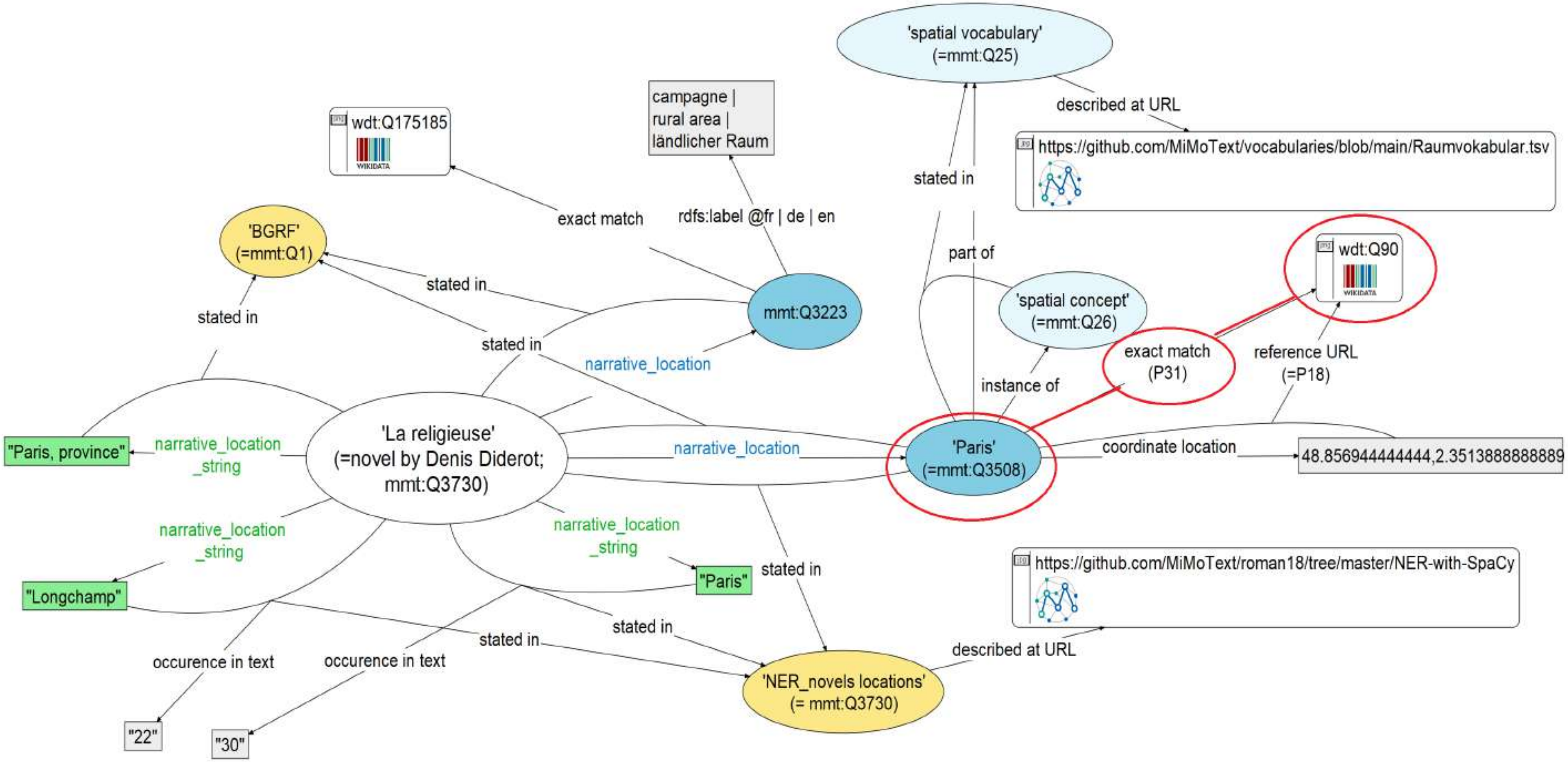
Module 9: terminology



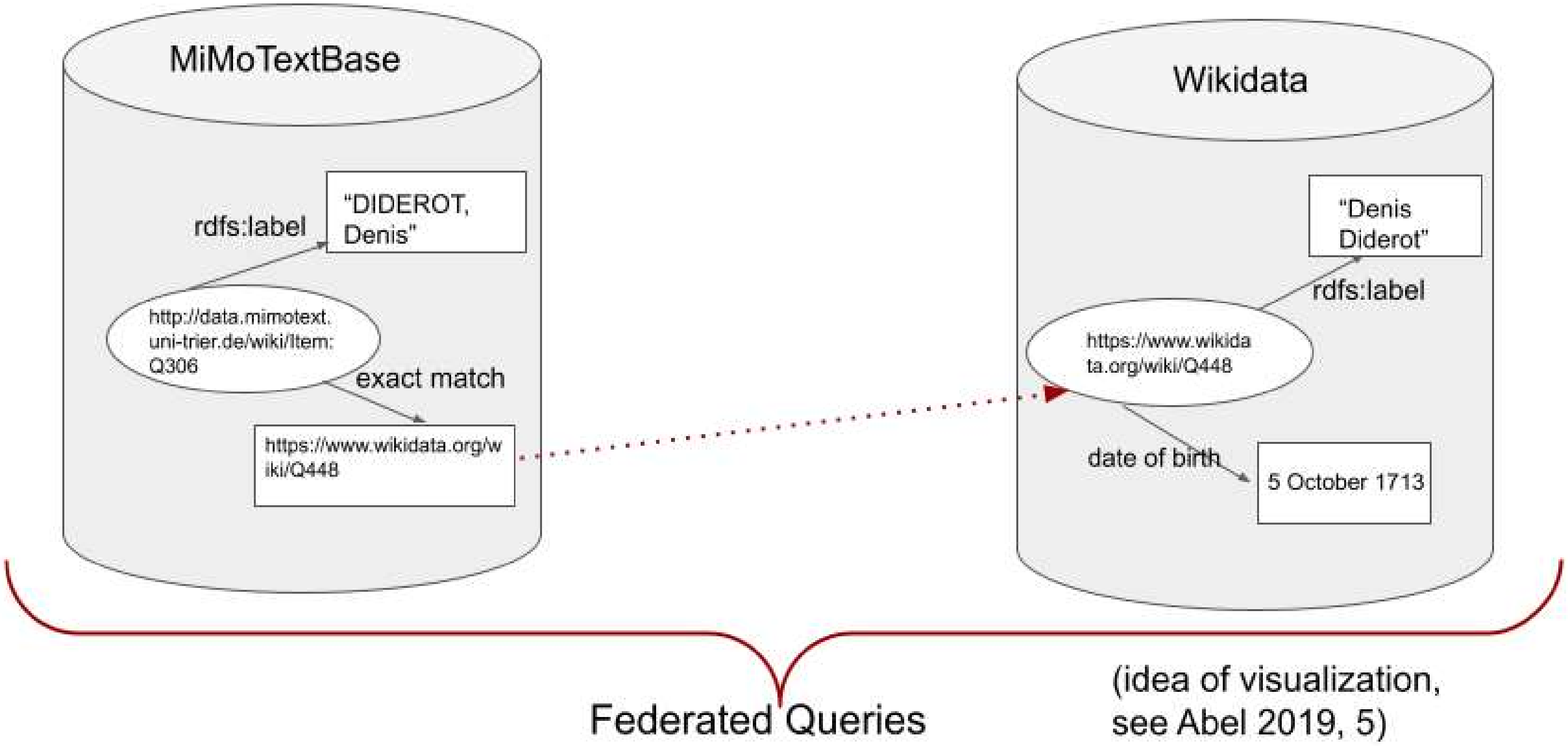
Controlled Vocabularies: https://github.com/MiMoText/vocabularies

Dimensions

Representing ‘fictionality’ - modeling ‘narrative locations’

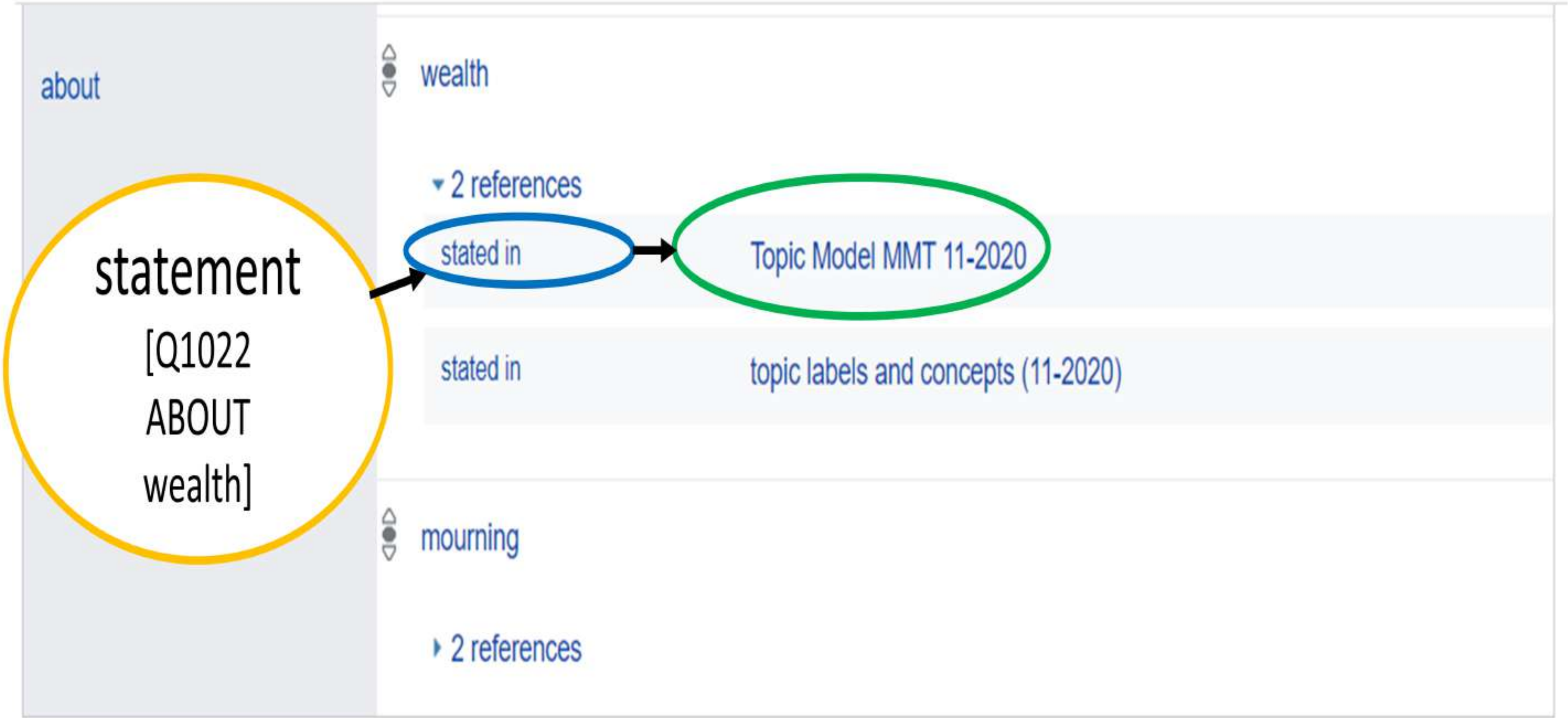


Alignment with Wikidata: enabling ‘federated queries’



(idea of visualization, see Abel 2019, 5)

Reification (1)



about

wealth

2 references

stated in → Topic Model MMT 11-2020

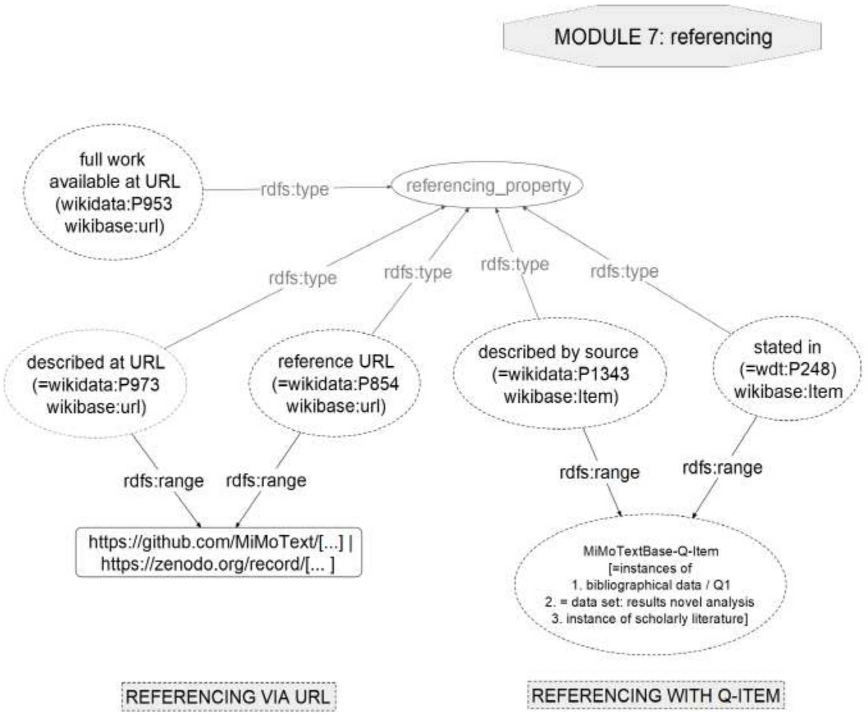
stated in topic labels and concepts (11-2020)

mourning

2 references

statement
[Q1022
ABOUT
wealth]

Reification (2)



MODULE 7: referencing

full work available at URL (wikidata:P953 wikibase:url)

rdfs:type

referencing_property

rdfs:type

rdfs:type

rdfs:type

rdfs:type

described at URL (=wikidata:P973 wikibase:url)

reference URL (=wikidata:P854 wikibase:url)

described by source (=wikidata:P1343 wikibase:Item)

stated in (=wdt:P248) wikibase:Item

rdfs:range

rdfs:range

rdfs:range

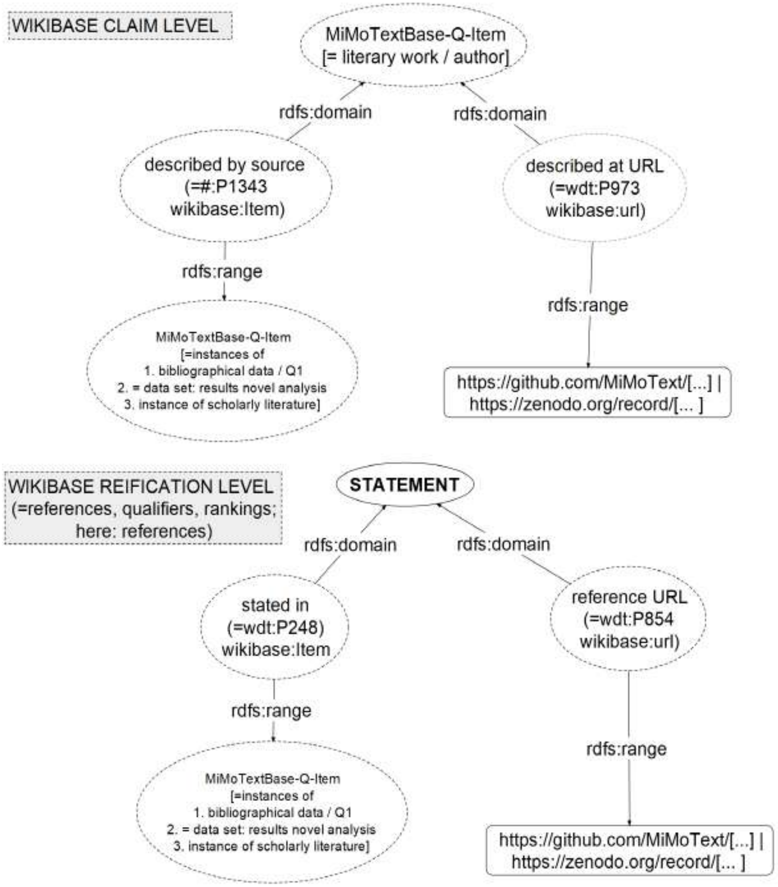
rdfs:range

https://github.com/MiMoText/[...] | https://zenodo.org/record/[...]

MiMoTextBase-Q-Item [=instances of 1. bibliographical data / Q1 2. = data set: results novel analysis 3. instance of scholarly literature]

REFERENCING VIA URL

REFERENCING WITH Q-ITEM



WIKIBASE CLAIM LEVEL

MiMoTextBase-Q-Item [= literary work / author]

rdfs:domain

rdfs:domain

described by source (=#:P1343 wikibase:Item)

described at URL (=wdt:P973 wikibase:url)

rdfs:range

rdfs:range

MiMoTextBase-Q-Item [=instances of 1. bibliographical data / Q1 2. = data set: results novel analysis 3. instance of scholarly literature]

https://github.com/MiMoText/[...] | https://zenodo.org/record/[...]

WIKIBASE REIFICATION LEVEL (=references, qualifiers, rankings; here: references)

STATEMENT

rdfs:domain

rdfs:domain

stated in (=wdt:P248) wikibase:Item

reference URL (=wdt:P854 wikibase:url)

rdfs:range

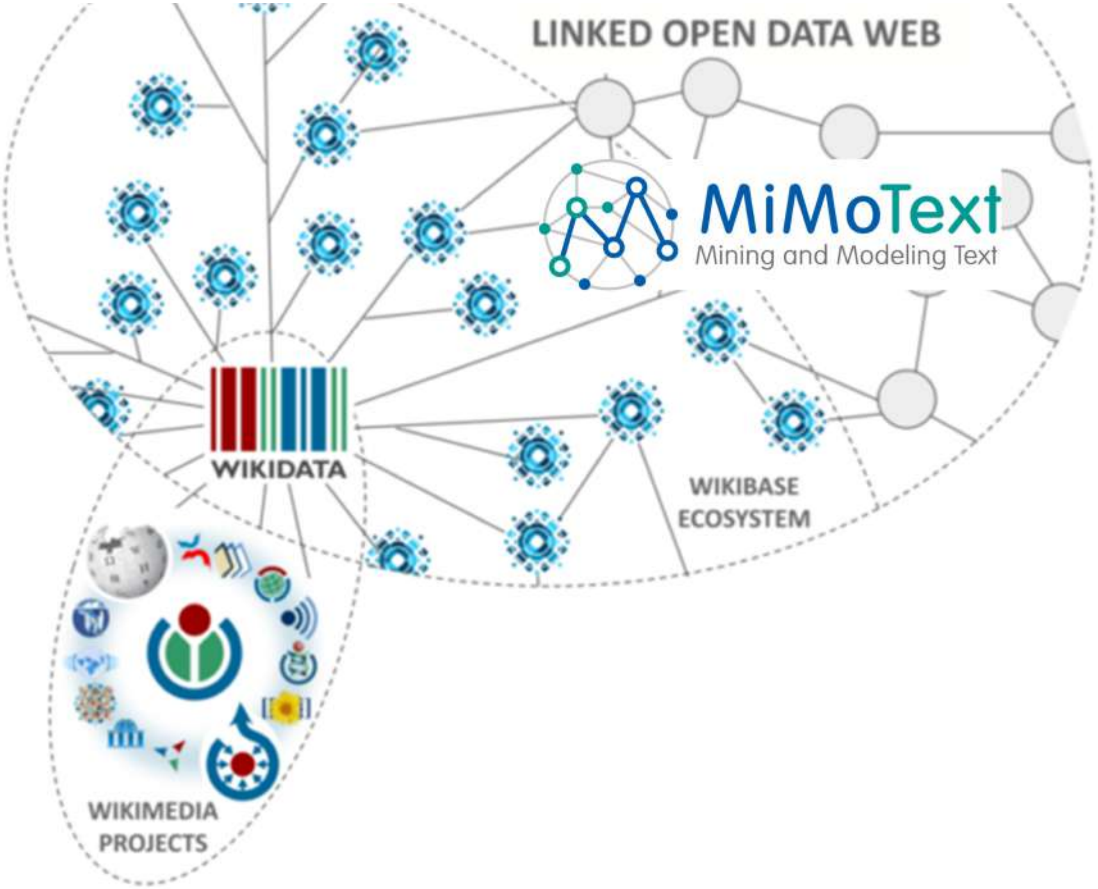
rdfs:range

MiMoTextBase-Q-Item [=instances of 1. bibliographical data / Q1 2. = data set: results novel analysis 3. instance of scholarly literature]

https://github.com/MiMoText/[...] | https://zenodo.org/record/[...]

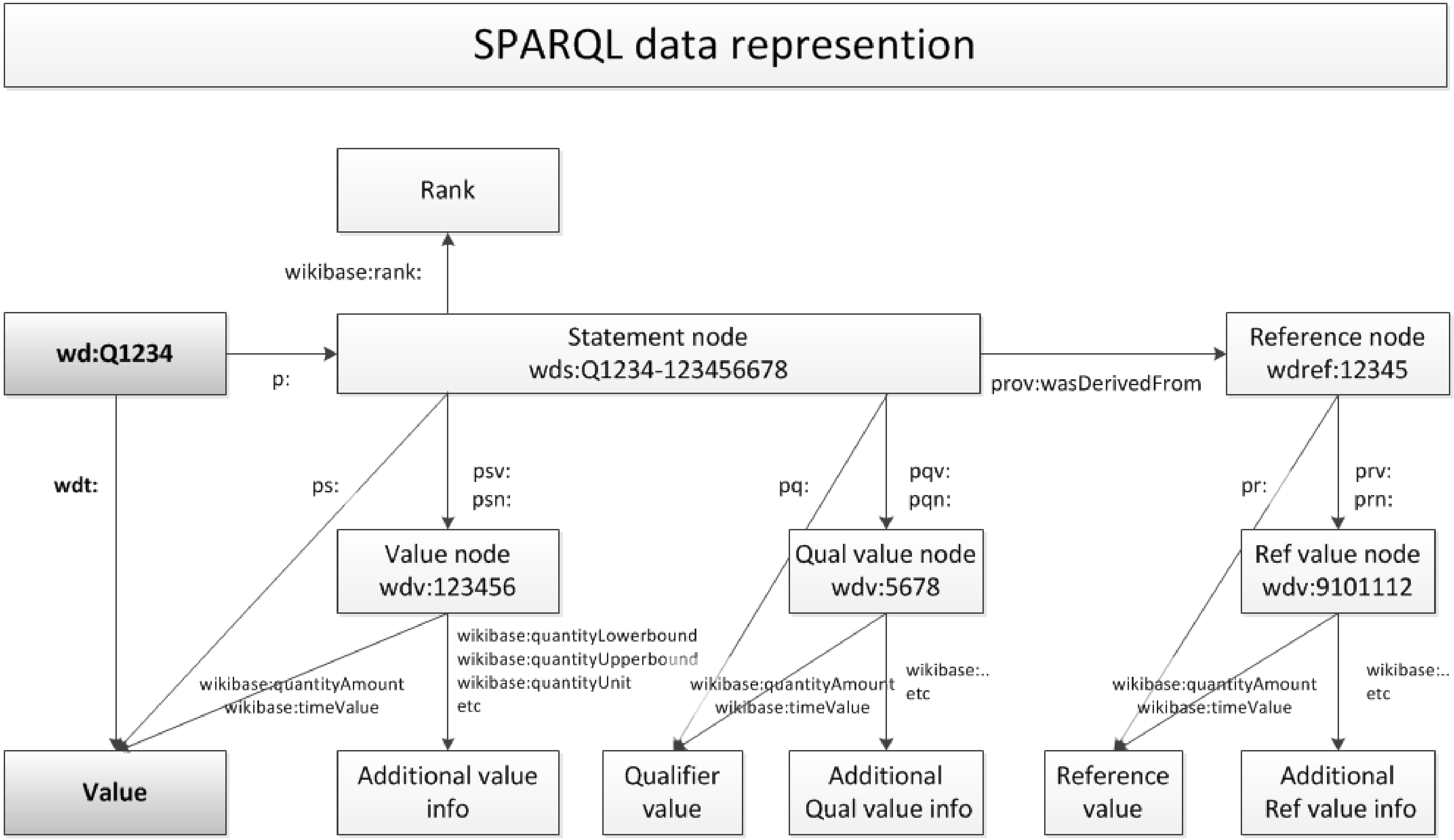
Wikibase ecosystem

MiMoTextBase as part of the Wikibase ecosystem



A view of the MiMoTextBase within the Wikimedia Linked Open Data web. Credit original visualization: Dan Shick (WMDE) / CC-BY-SA 4.0

Wikibase Data model (1)



Source: UserHenkvD: SPARQL data representation, as used by Wikidata Query Service. 2017. CC BY-SA 4.0

Wikibase Data model (2)

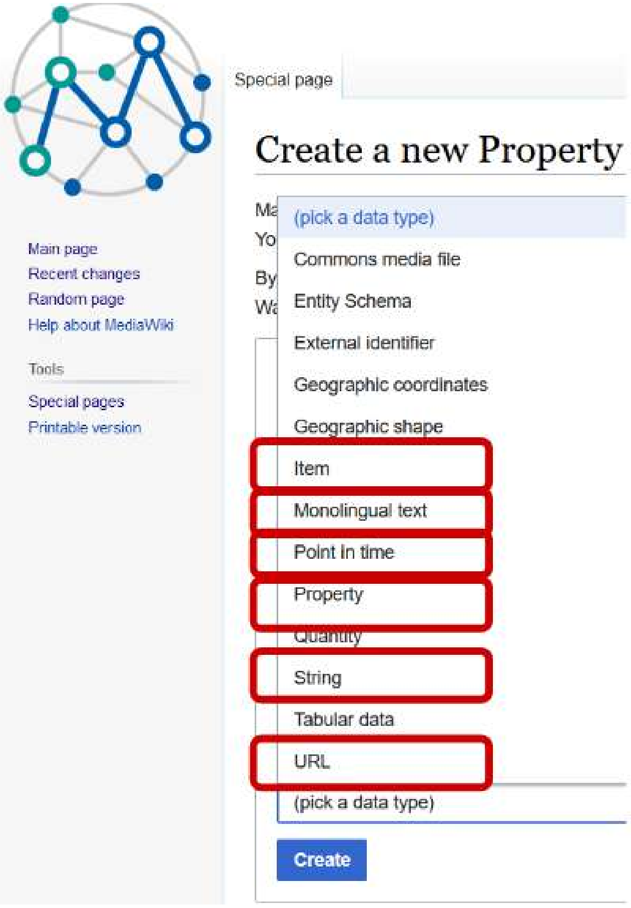
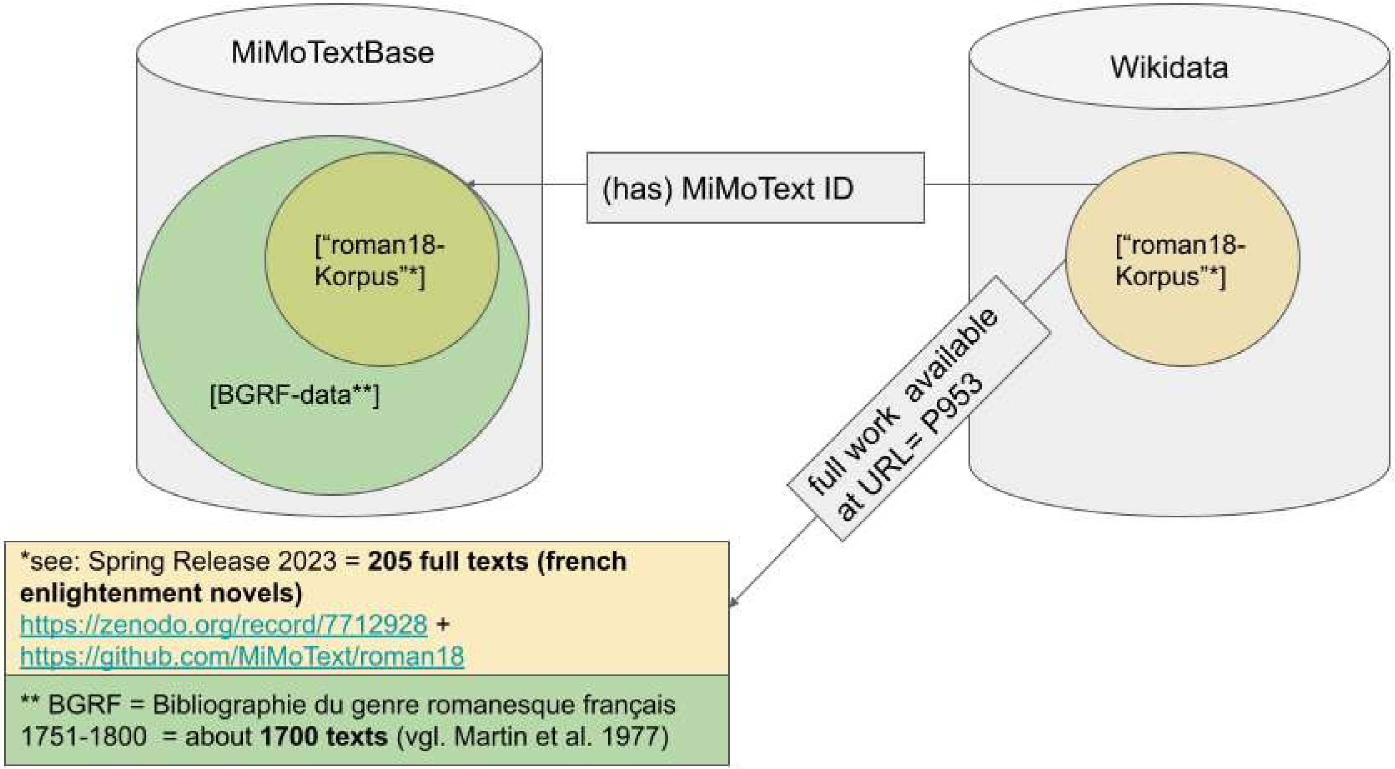


Fig.: Property data types in the MiMoTextBase (red)

Potentials (1)

- Wikidata as a “linking hub” (Neubert 2017)
- Large amount of data across domains & disciplines
- Open Access, Open Science, Open Knowledge (Schöch 2021)
- Multilingualism
- Visualization in the DockerWikibaseQueryService
- Linking entities & enabling federated queries
- Advantages of alignment within the same infrastructure and contributing data directly to Wikidata

Potentials (2)



*see: Spring Release 2023 = **205 full texts (french enlightenment novels)** https://zenodo.org/record/7712928 + https://github.com/MiMoText/roman18

** BGRF = Bibliographie du genre romanesque français 1751-1800 = about **1700 texts** (vgl. Martin et al. 1977)

Fig.: Connecting MiMoTextBase - the roman18 corpus - and Wikidata (future work)

Limitations

- no systematic ontology
- specific data model which is not directly interoperabel with OWL standard
- problem of semantic expressivity (Sack 2022)
- loss of reasoning potential / possibilities
- biases and dominances (e.g. English language) in reality (despite awareness and initiatives)

(3) Conclusion

Some of the challenges we have encountered

- Modeling meta-assertions
  => more or less solved
- Lack of consensus on fundamental assertions
  => need to coordinate broadly
- Modeling and need for formal ontologies
  => Documentation, but not in OWL
- ...

Some advantages of linked open literary history data

- Ability to connect heterogeneous data sources
- Allows to model, gather and compare contradicting information
- Makes the process of constructing knowledge transparent (sources)
- Allows to re-use information already present elsewhere (federated queries)
- Has been an immense learning opportunity for the whole team
- ...

Many thanks!

MiMoText

To learn more

- Tutorial: https://docs.mimotext.uni-trier.de
- Visualizations: mimotext.github.io/MiMoTextBase_Tutorial/visualizations.html
- SPARQL endpoint: https://query.mimotext.uni-trier.de
- MiMoTextBase: https://data.mimotext.uni-trier.de
- MiMoText Ontology: https://github.com/MiMoText/ontology
- Reference publication: 'Smart Modeling for Digital Literary History'
- References & readings: Zotero

Link to this page https://mimotext.github.io/lod-lithist/eng.html#/4/3

Back Matter

Slides: https://mimotext.github.io/lod-listhist/eng.html
Project: https://mimotext.uni-trier.de/en