Linked Open Data for Literary History: Constructing, Querying and Using the MiMoTextBase

Maria Hinzmann and Tinghui Duan

with Matthias Bremm, Anne Klee, Johanna Konstanciak,
Julia Röttgermann and Christof Schöch

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

Slides: https://mimotext.github.io/lod-lithist/berlin23.html

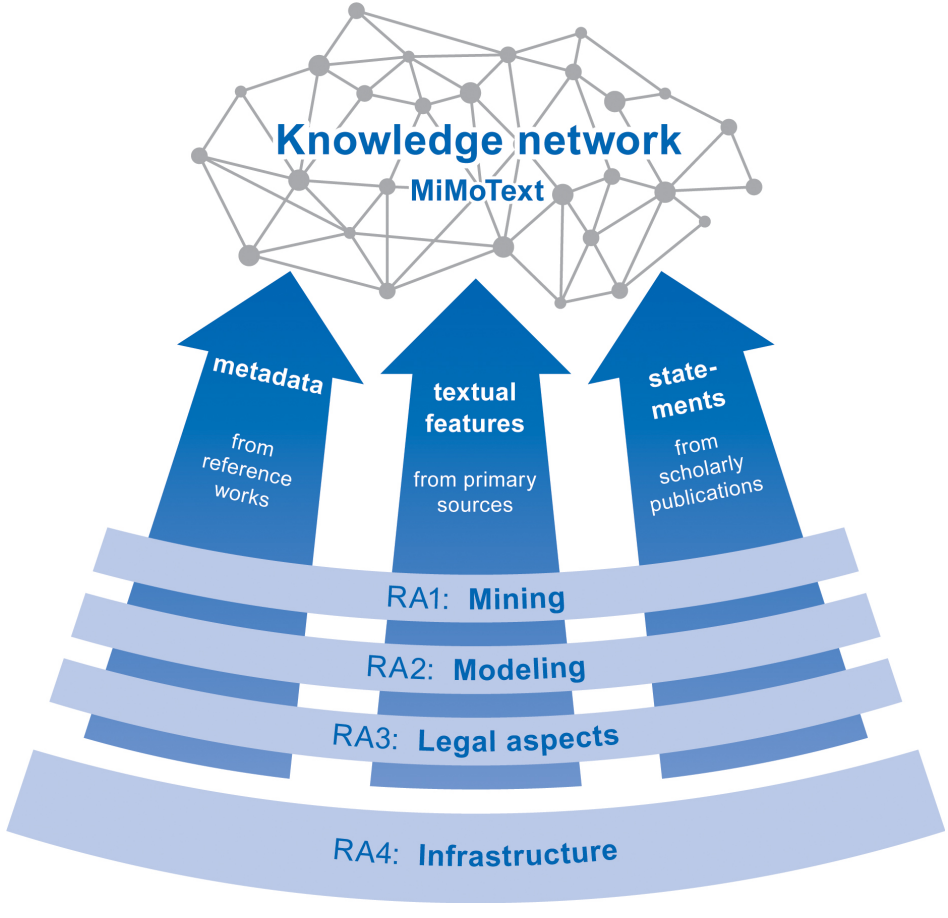
Workshop FU Berlin, 10th October 2023
Wikipedia, Wikidata and Wikibase: Usage Scenarios for Literary Studies

Structure

1. Introduction
2. Mining
3. Modeling
4. MiMoTextBase & Wikidata
5. Examples

(1) Introduction

MiMoText in a nutshell



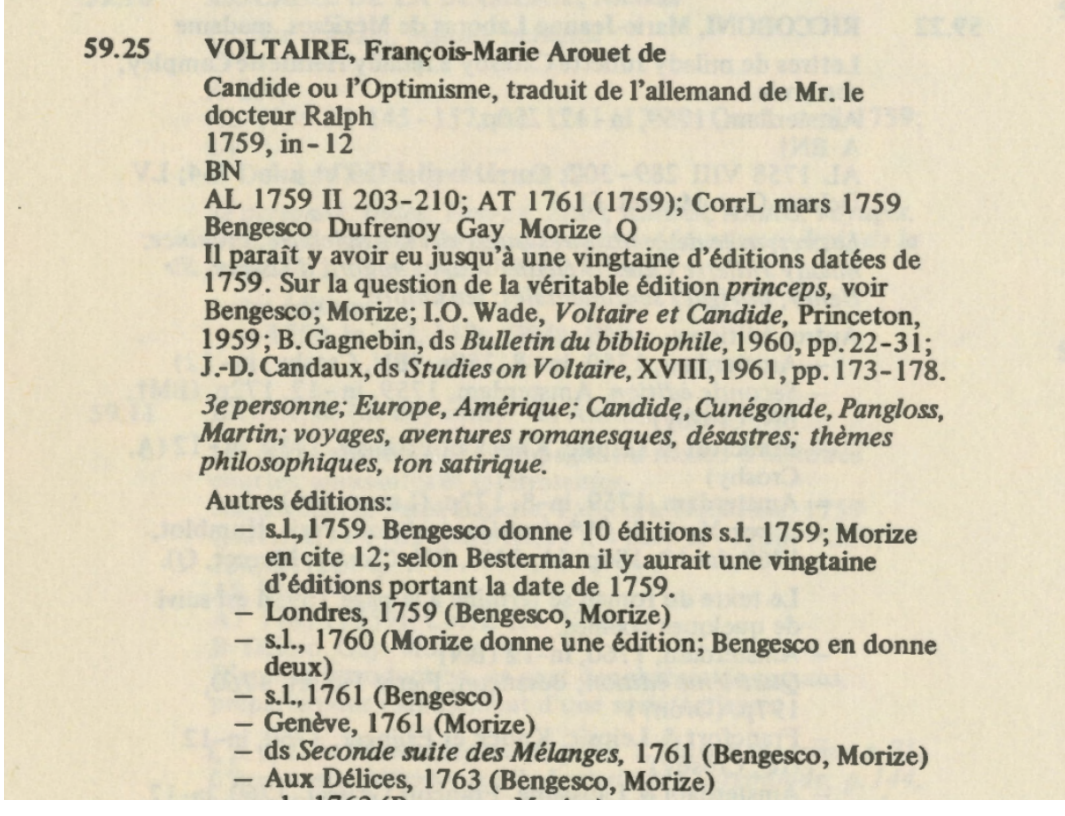
https://mimotext.uni-trier.de/en

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
- Compared to 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 more explicit data modeling

(2) Mining

Source type 1: Bibliographic data



59.25 **VOLTAIRE, François-Marie Arouet de**

Candide ou l'Optimisme, traduit de l'allemand de Mr. le docteur Ralph
1759, in-12
BN
AL 1759 II 203-210; AT 1761 (1759); CorrL mars 1759
Bengesco Dufrenoy Gay Morize Q
Il paraît y avoir eu jusqu'à une vingtaine d'éditions datées de 1759. Sur la question de la véritable édition *princeps*, voir Bengesco; Morize; I.O. Wade, *Voltaire et Candide*, Princeton, 1959; B. Gagnebin, ds *Bulletin du bibliophile*, 1960, pp. 22-31; J.-D. Candaux, ds *Studies on Voltaire*, XVIII, 1961, pp. 173-178.

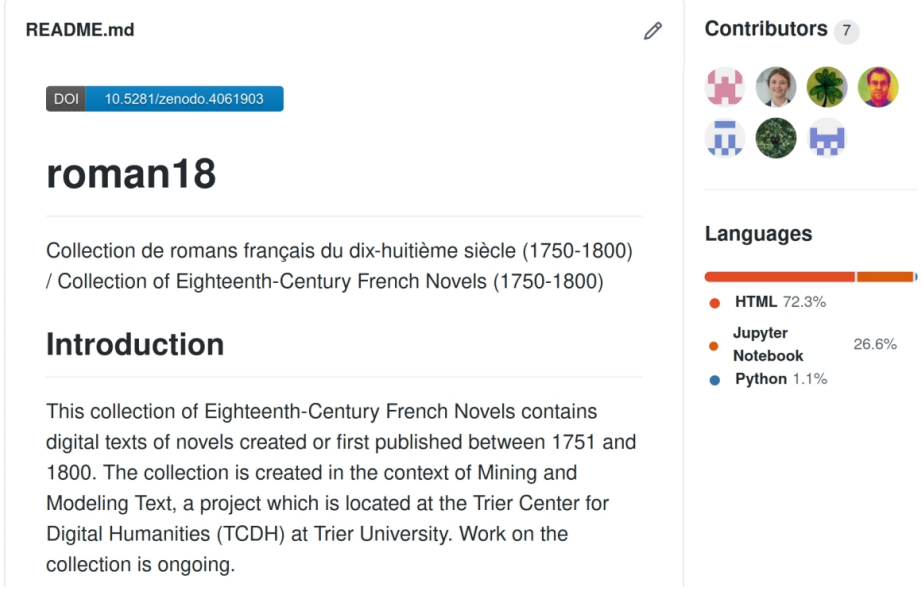
3e personne; Europe, Amérique; Candide, Cunégonde, Pangloss, Martin; voyages, aventures romanesques, désastres; thèmes philosophiques, ton satirique.

Autres éditions:
- s.l., 1759. Bengesco donne 10 éditions s.l. 1759; Morize en cite 12; selon Besterman il y aurait une vingtaine d'éditions portant la date de 1759.
- Londres, 1759 (Bengesco, Morize)
- s.l., 1760 (Morize donne une édition; Bengesco en donne deux)
- s.l., 1761 (Bengesco)
- Genève, 1761 (Morize)
- ds *Seconde suite des Mélanges*, 1761 (Bengesco, Morize)
- Aux Délices, 1763 (Bengesco, Morize)

Martin / Mylne / Frautschi: *Bibliographie du genre romanesque français, 1751-1800*, 1977

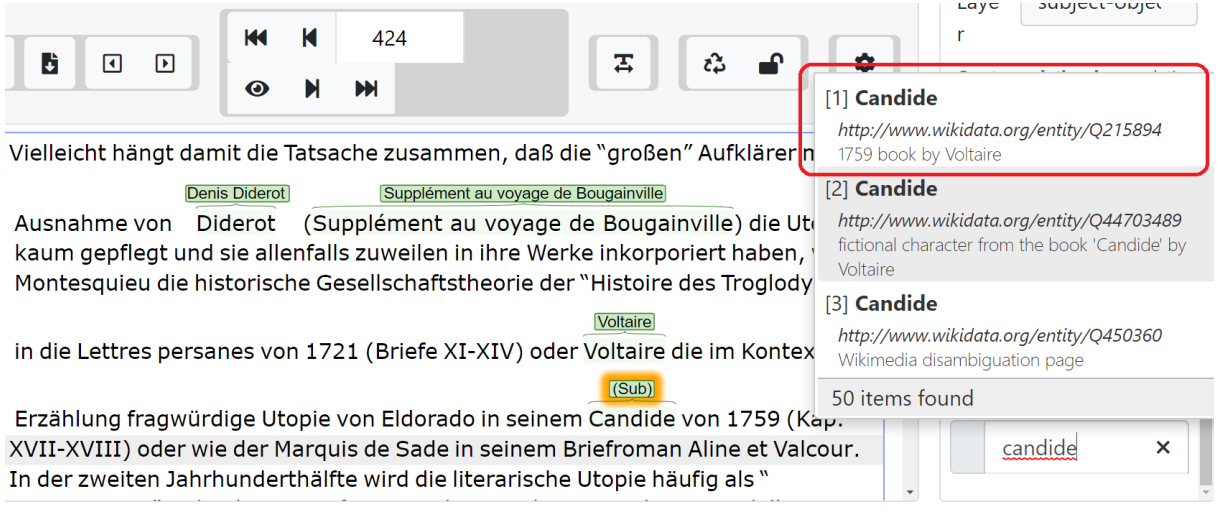
Source type 2: Primary literature (novels)

- corpus of 205 french novels (1750-1800)
- in XML-TEI, with metadata, following the ELTeC schema
- methods: e.g. Topic Modeling, NER, stylometry



Collection of Eighteenth-Century French Novels (1750-1800)

Source type 3: Scholarly publications



- manual annotation (e.g. statements about themes)
- annotation tool INCEpTION linked with Wikidata and MiMoTextBase
- pipeline (export INCEpTION annotations => import Wikibase statements)

(3) Modeling

Ontology

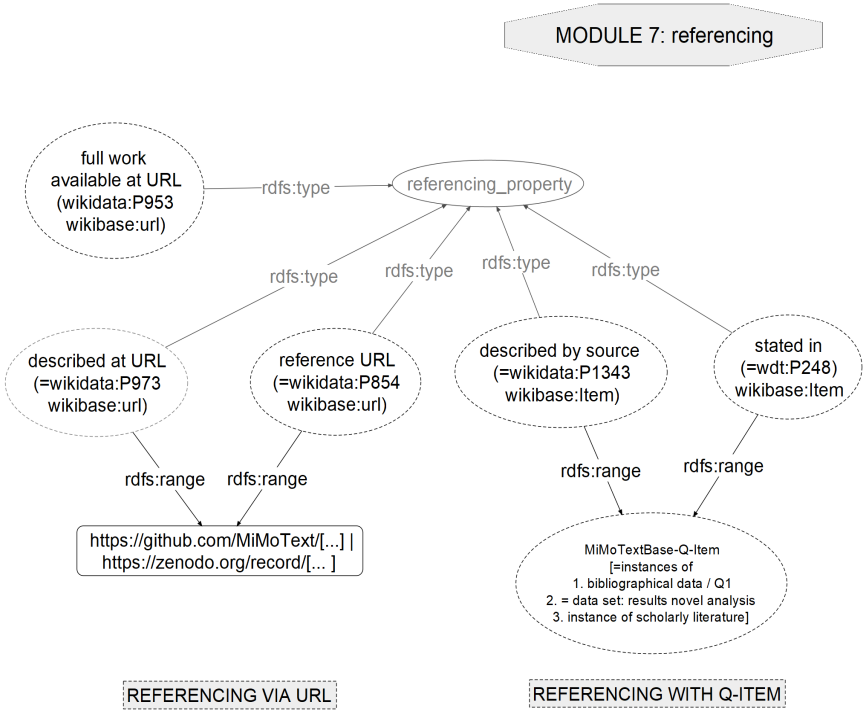
- Überblick
  - Module 1: theme
  - Module 2: space
  - Module 3: narrative form
  - Module 4: literary work
  - Module 5: author
  - Module 6: mapping
  - Module 7: referencing
  - Module 8: versioning & publication
  - Module 9: terminology
  - Module 10: bibliography
  - Module 11: scholarly work

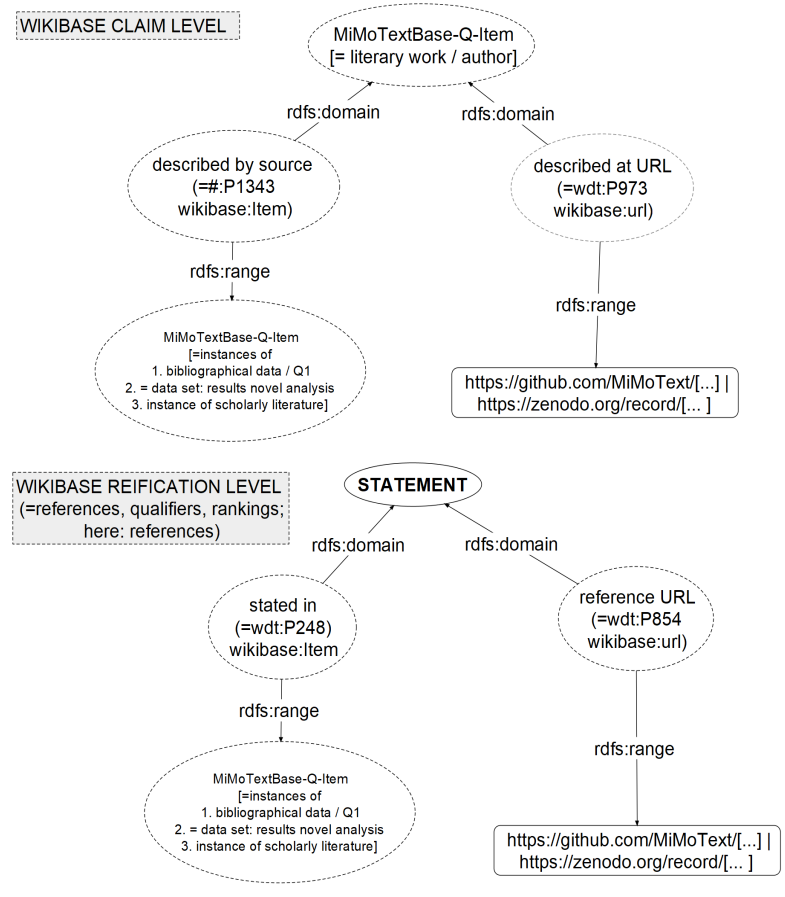
(https://github.com/MiMoText/ontology)

Reification (1)

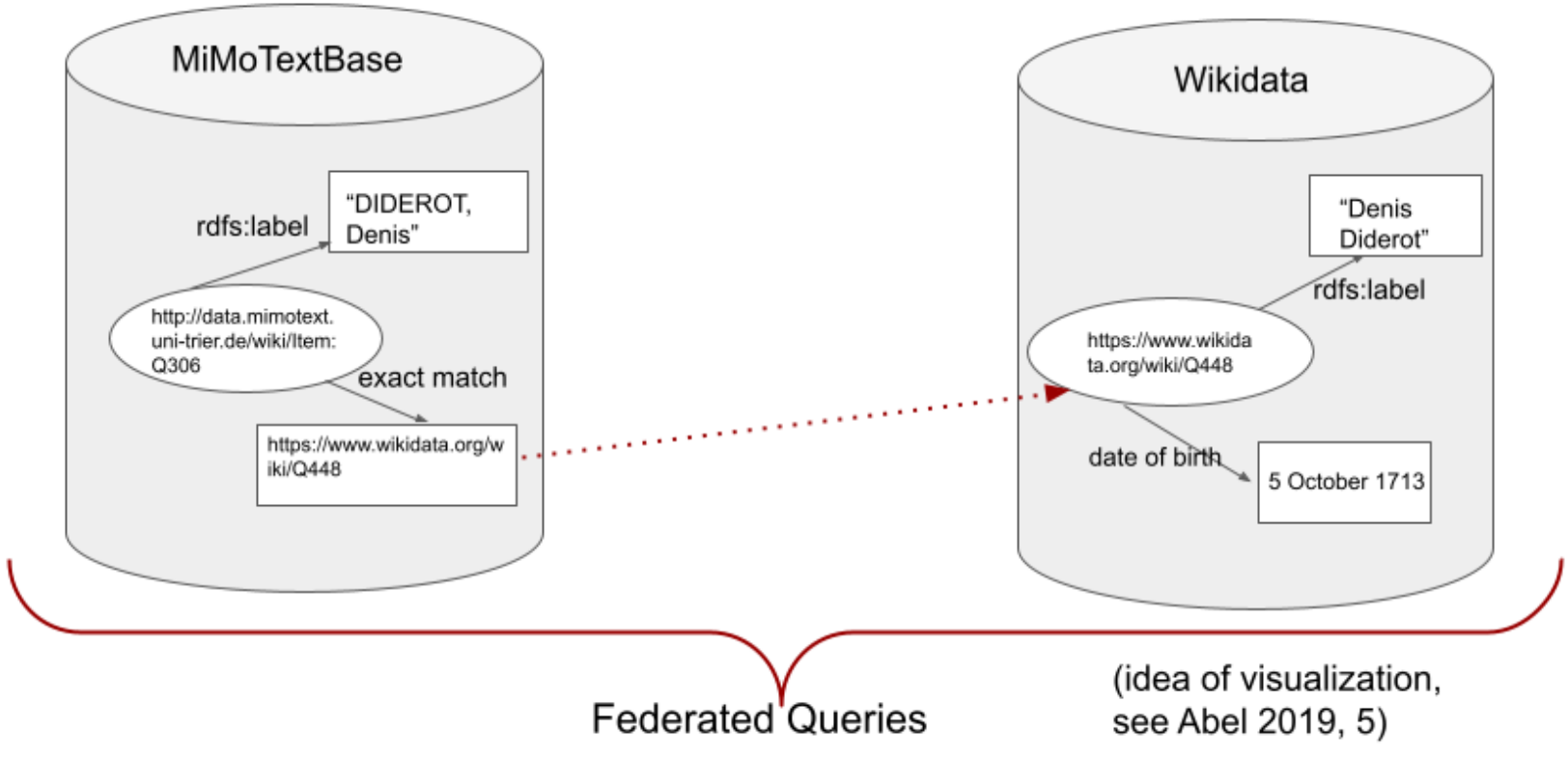
about	
	libertinism
	▾ 2 references
	stated in — Bibliographie du genre romanesque français
	stated in — BGRF_matching-table (03-2022)
	correspondence
	▾ 2 references
	stated in — Topic Model MMT 11-2020
	stated in — topic labels and concepts (11-2020)

Reification (2)





Alignment with Wikidata: enabling 'federated queries'



(4) MiMoTextBase & Wikidata

What is Wikibase?

WIKIPEDIA
The Free Encyclopedia

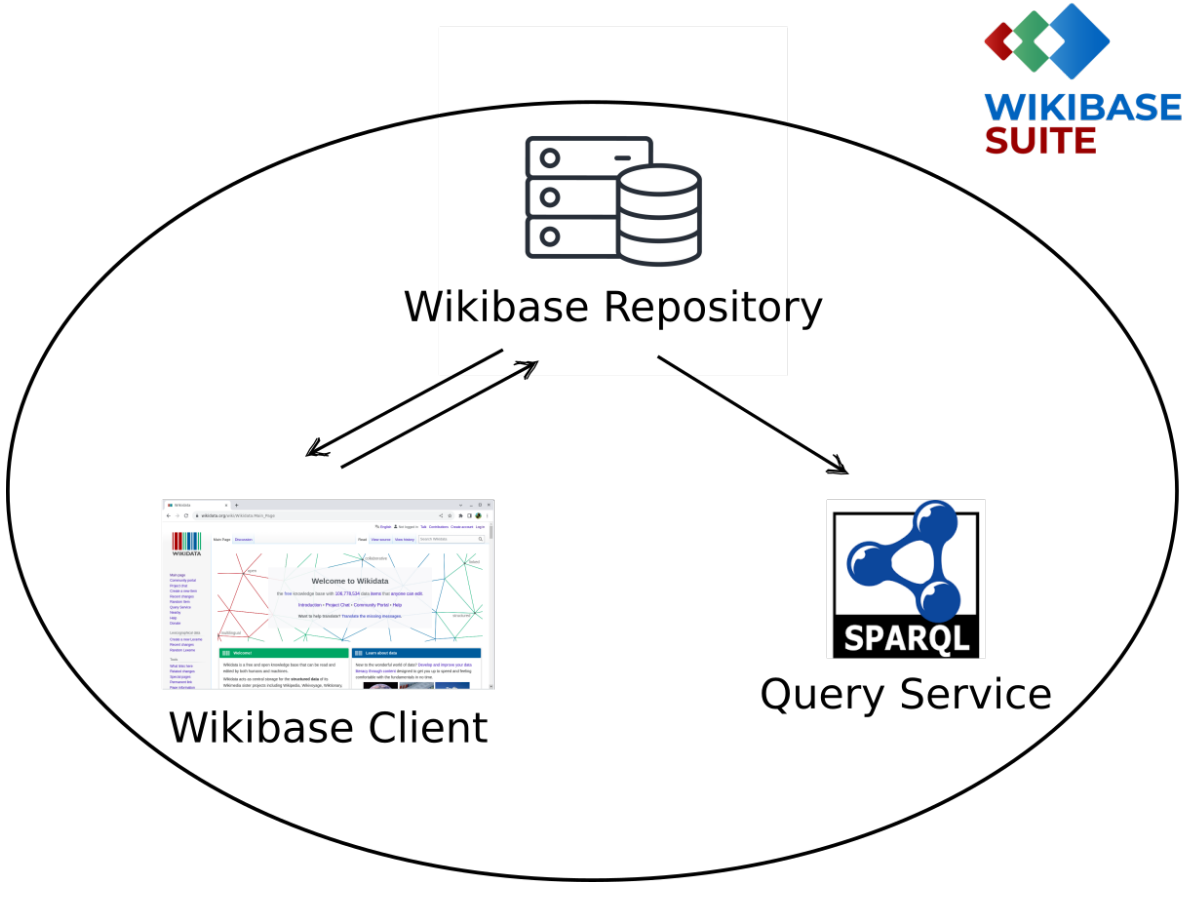
since 2001

since 2012

WIKIDATA

WIKIBASE

Wikibase as infrastructure



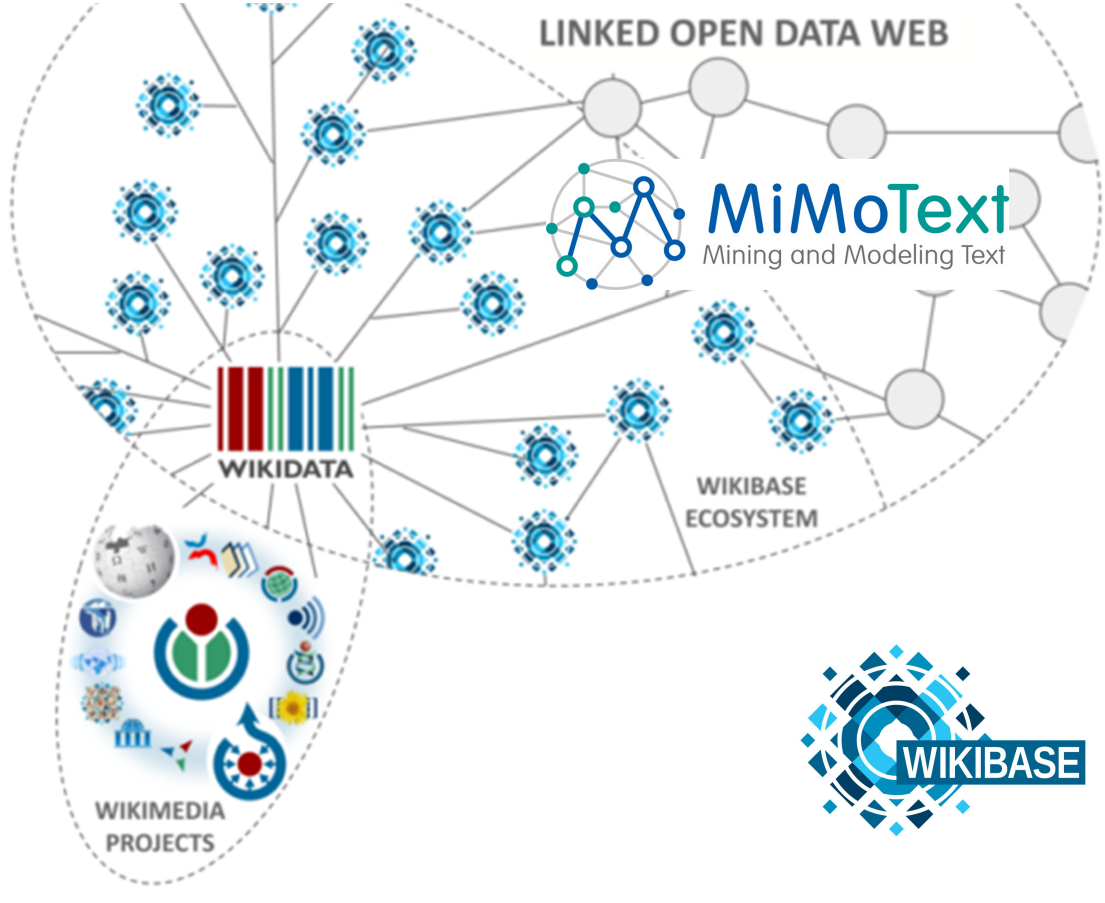
Why using Wikibase?

- It’s a FOSS! (free open source software)
- Flexible data model
- User-friendly interface and different import possibilities
- Easily linkable to other Wikibase-instances
- Multilingualism

Some example instances of Wikibase

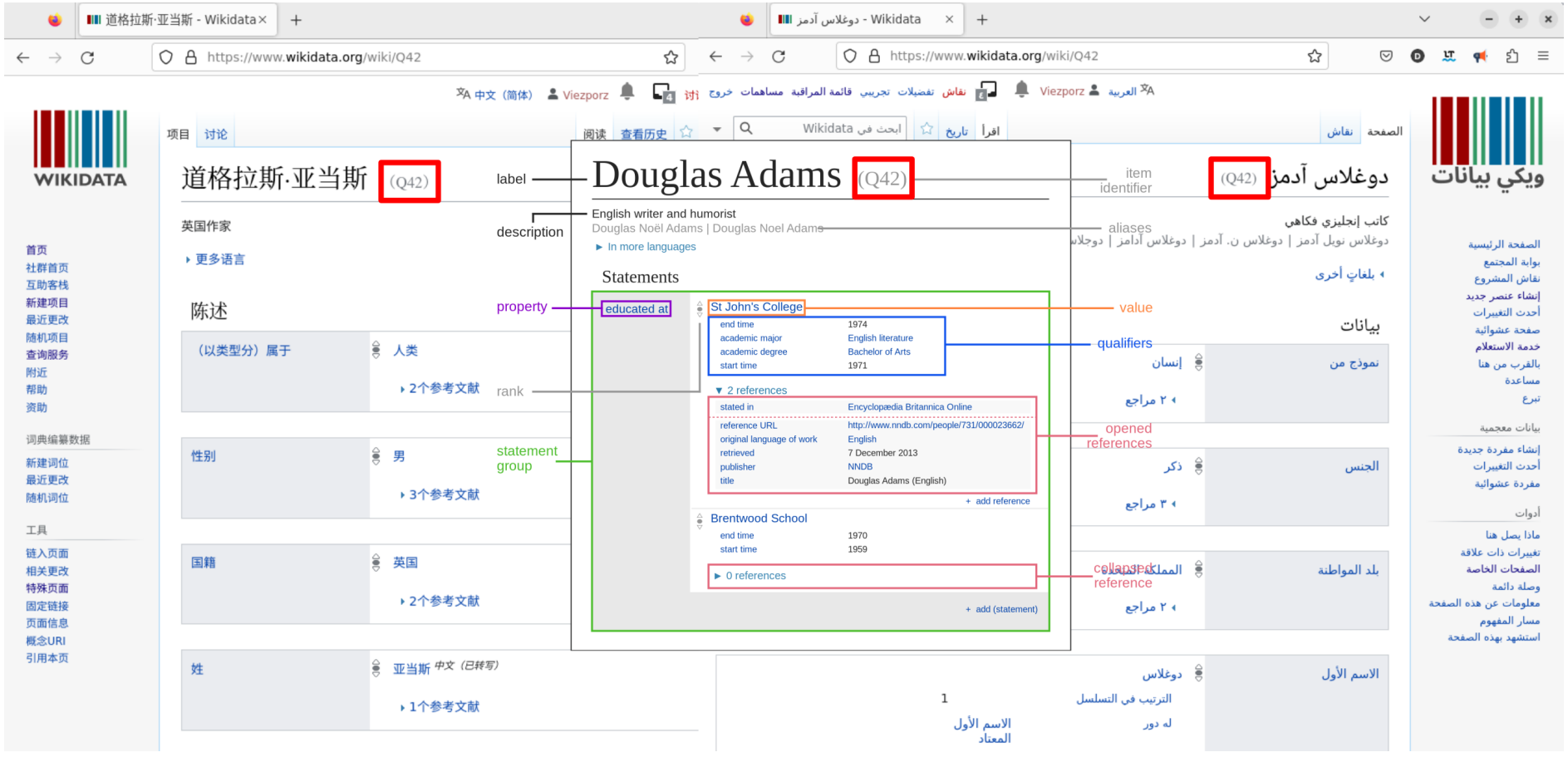
RHIZOME

MiMoTextBase as part of the Wikibase ecosystem

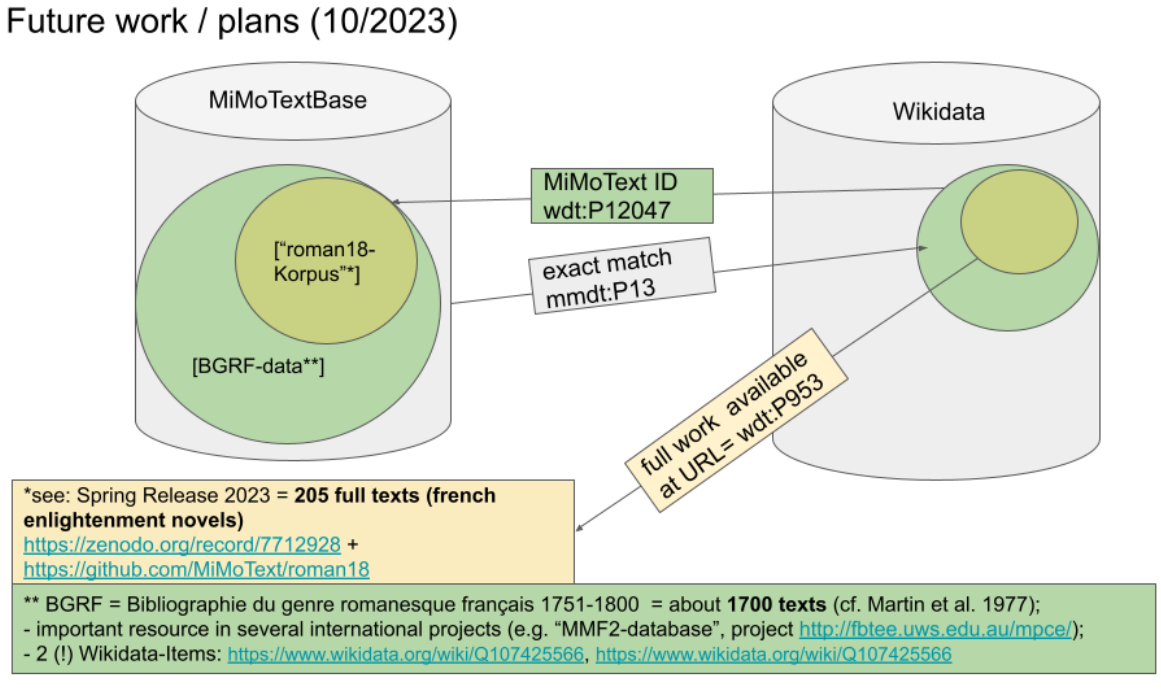


A view of the MiMoTextBase within the Wikimedia Linked Open Data web.
Credit original visualization: Dan Shick (WMDE) / CC-BY-SA 4.0

Multilingualism facilitates international collaboration



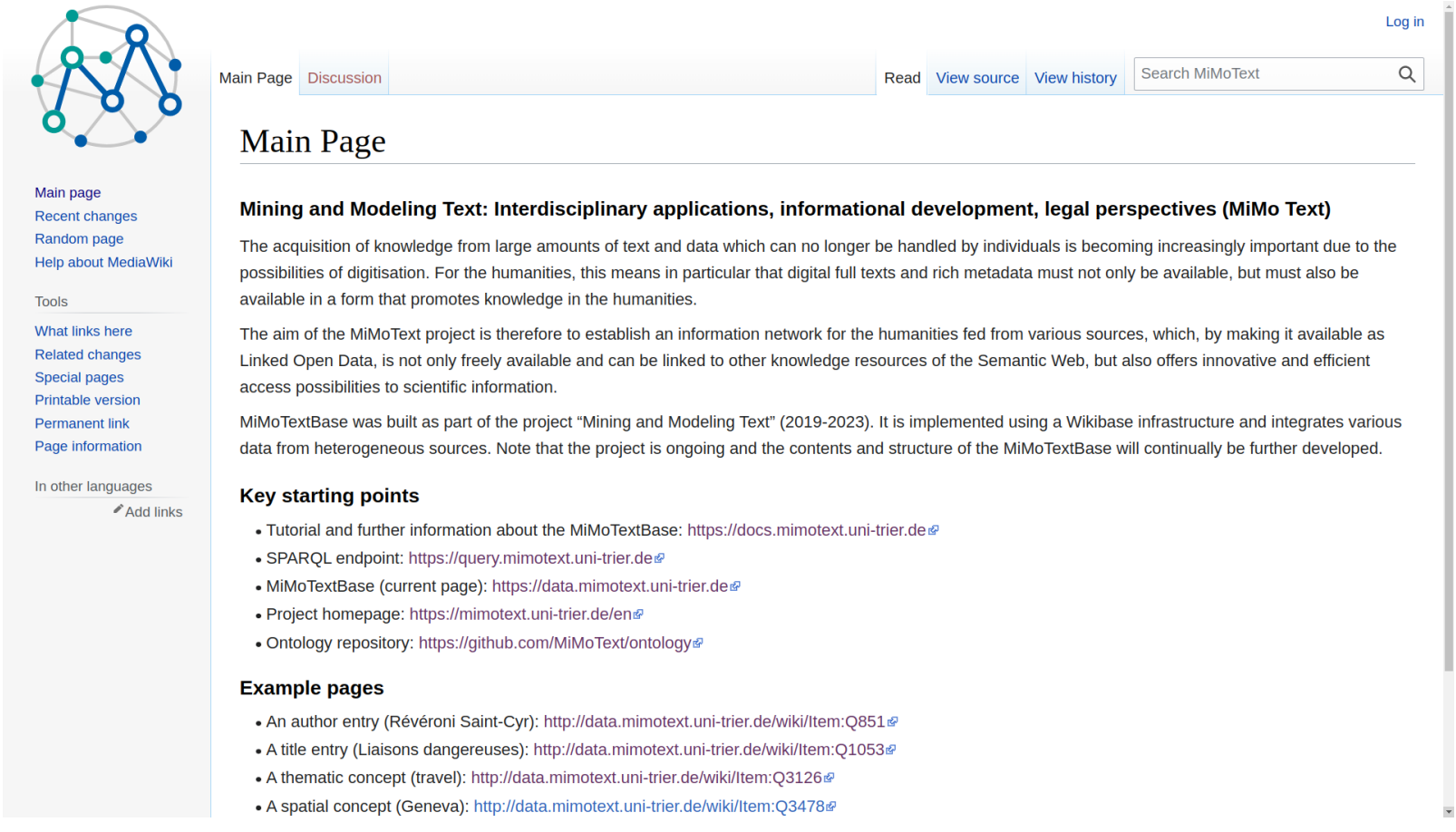
Connecting MiMoTextBase & Wikidata



- Contribution of MiMoText to Wikidata
  - Adding new Items to Wikidata
  - MiMoText ID as a new property (external identifier) to Wikidata
  - Linking novel items on Wikidata to fulltexts on Github

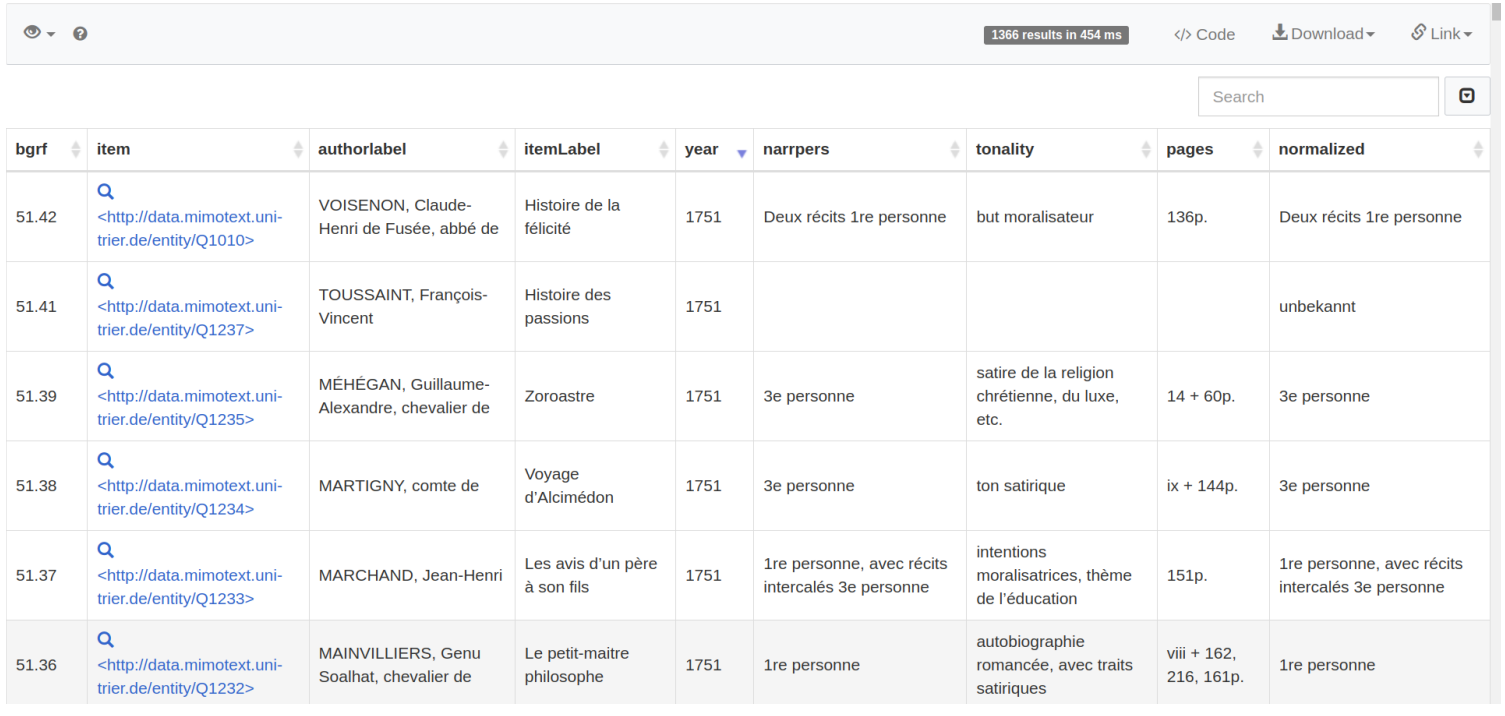
(5) Examples

Result: the MiMoTextBase



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

The SPARQL endpoint



1366 results in 454 ms Code Download Link

Search

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)
  - Alternative authorlabels via skos:altLabel
  - Linking with catalogue data via 'BNF identifier'
  - Linking relations between authors via 'influenced by'
- Compare information from two sources
  - Themes derived from topic modeling compared to themes according to BGRF
  - Combined: themes by BGRF vs. from topic modeling

Thank you!

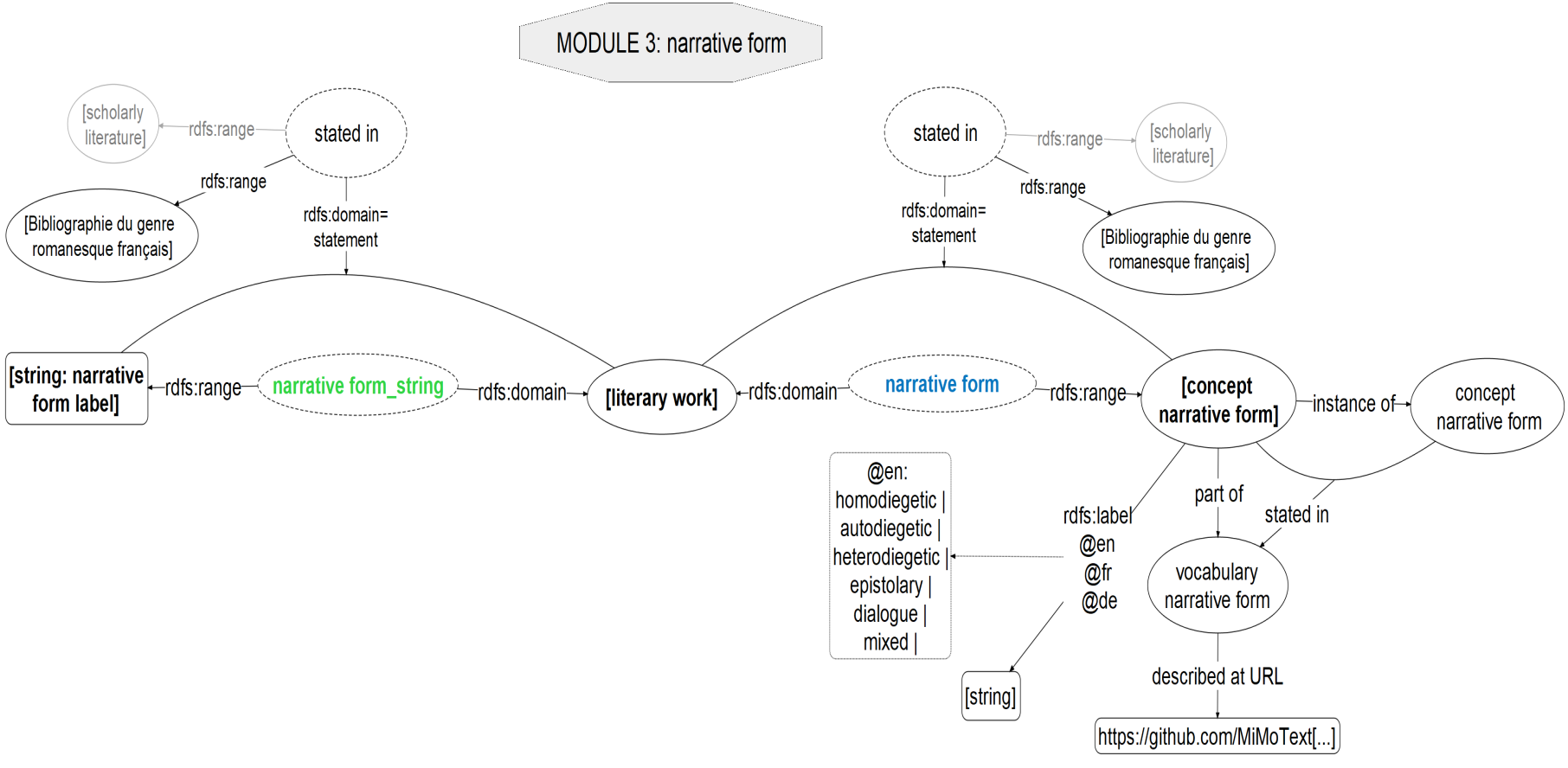
To learn more

- Tutorial: https://docs.mimotext.uni-trier.de
- 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'
- Overview visualizations WDQS

Link to this page https://mimotext.github.io/lod-lithist/berlin23.html#/6/4

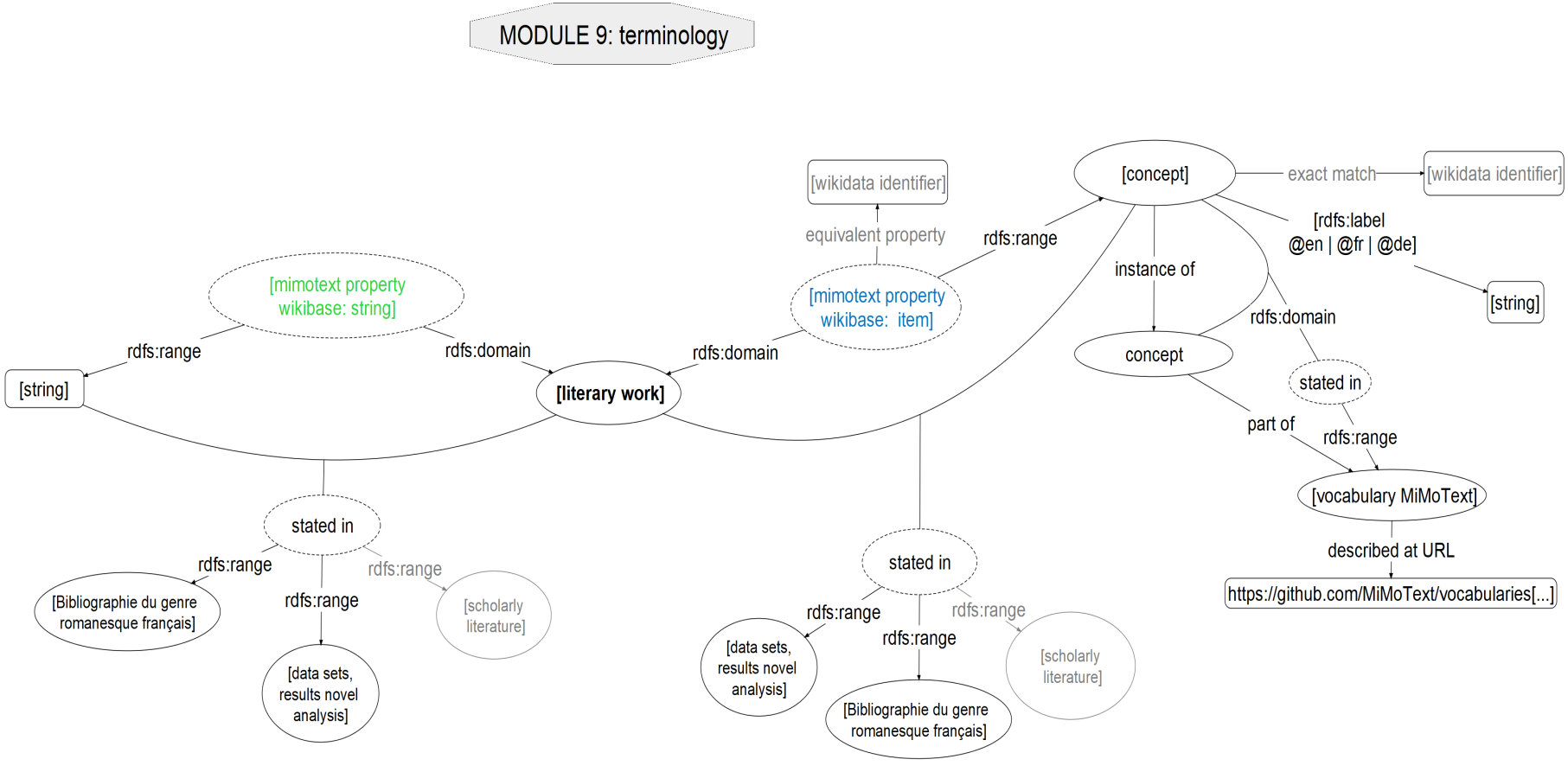
Bonus slides

Module 3: Narrative form



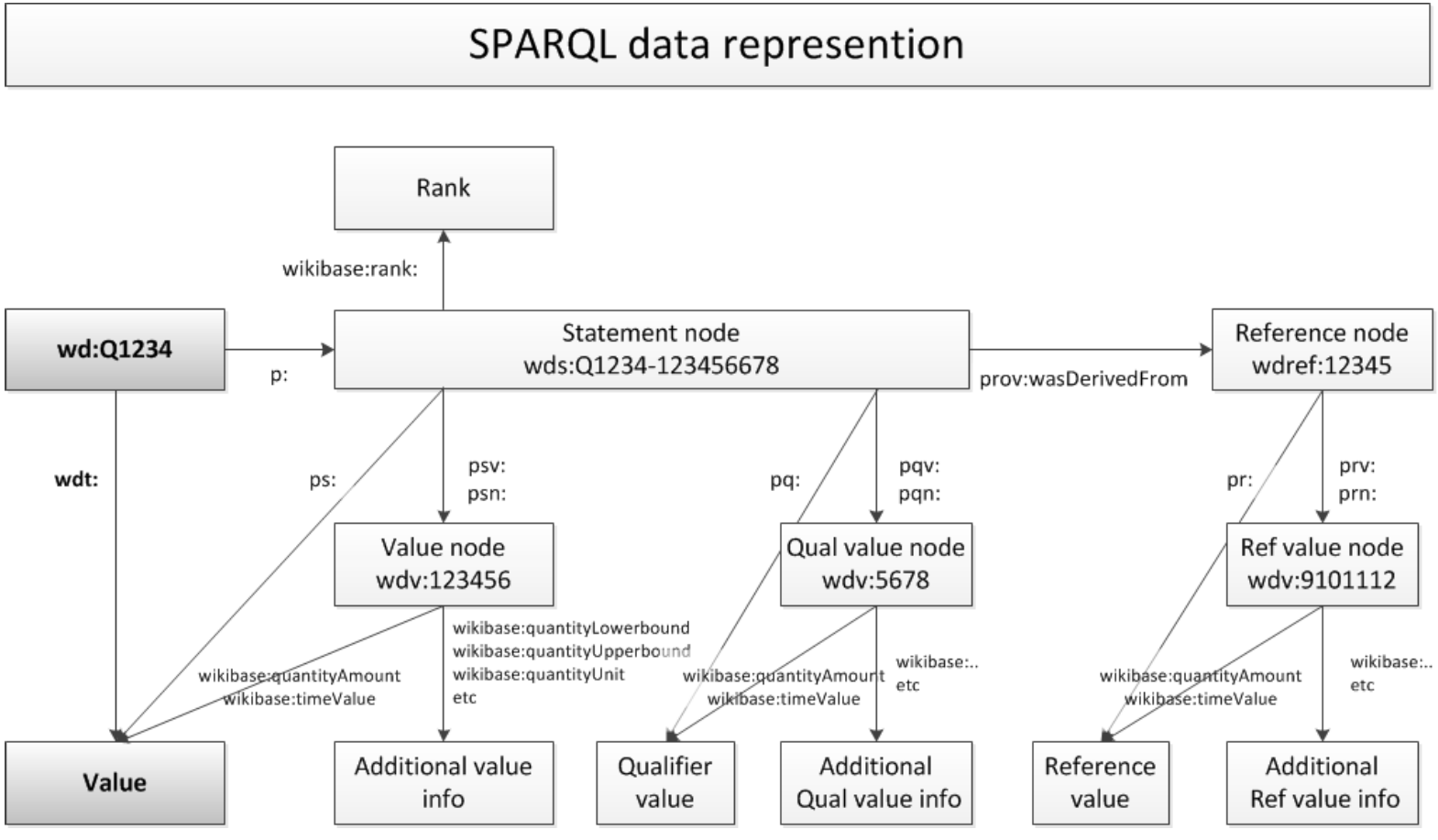
- Cf. Calvo Tello (2021) adapting Genette (1979)
- See Balancing: https://github.com/MiMoText/balance_novels

Module 9: terminology



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

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)

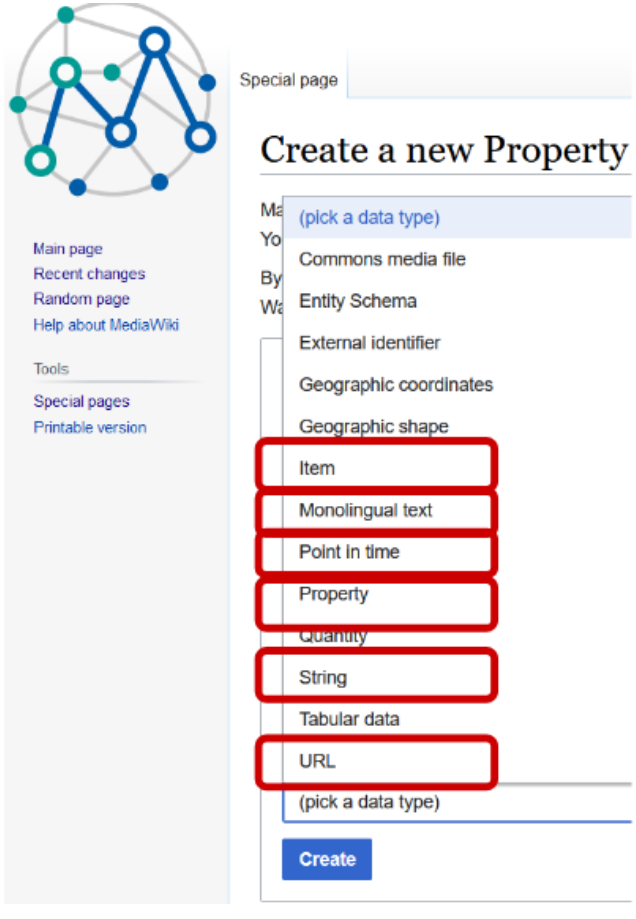


Fig.: Property data types in the MiMoTextBase (red)

Potentials

- 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

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
- ...

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)

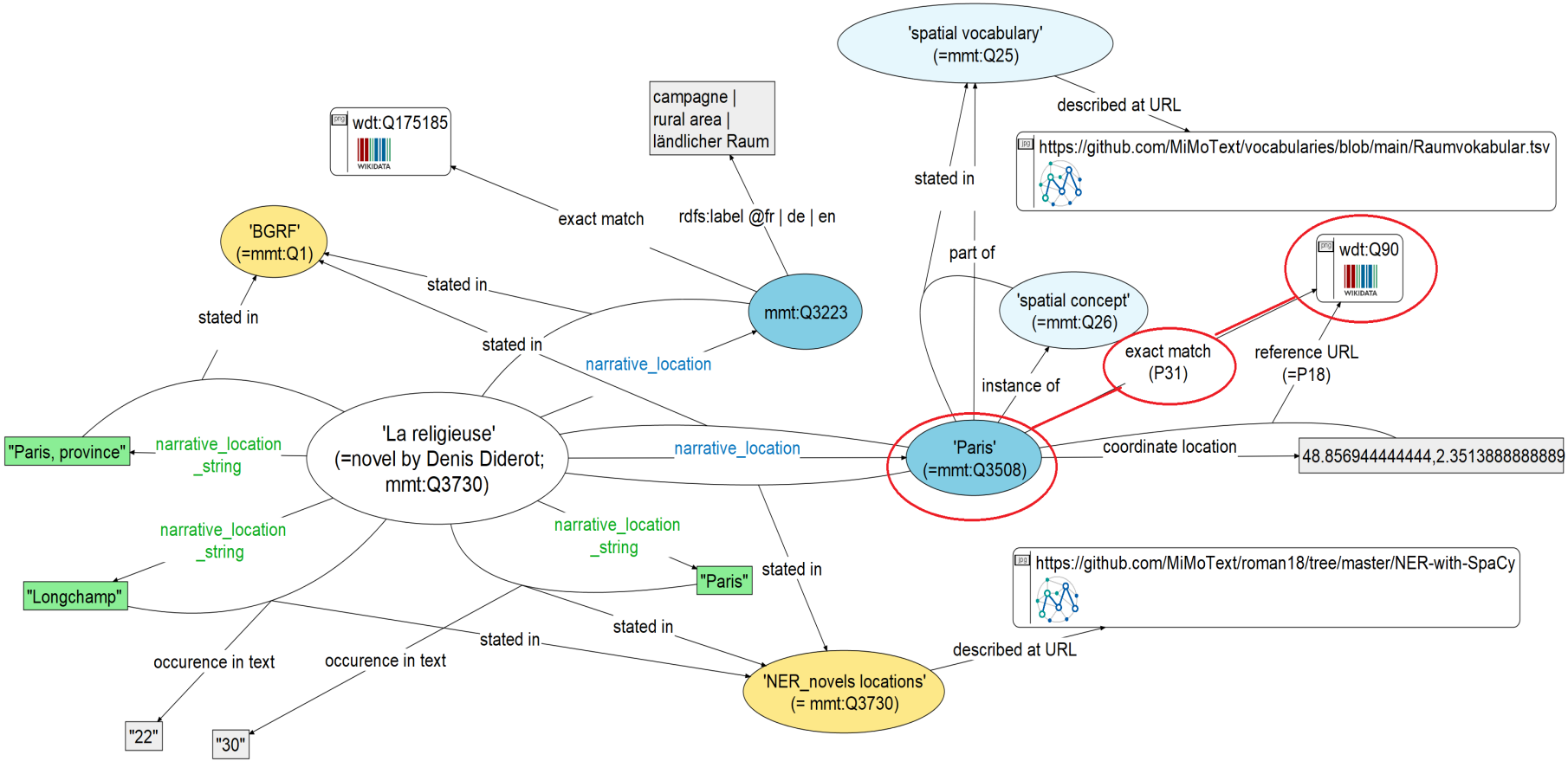
Some of the challenges we have encountered

- 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
- ...

MiMoTextBase - overview

- 331671 Triple
- 1750 literary works written by 622 authors
- 1178 elements in different controlled vocabularies => 919 ‘exact’ or ‘close’ matches to Wikidata-Items (query: https://tinyurl.com/26zkbn9j)

Combining data from different sources (narrative locations)



Back Matter

Thanks!

Slides: https://mimotext.github.io/lod-listhist/berlin23.html
Project: https://mimotext.uni-trier.de/en
