

# 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

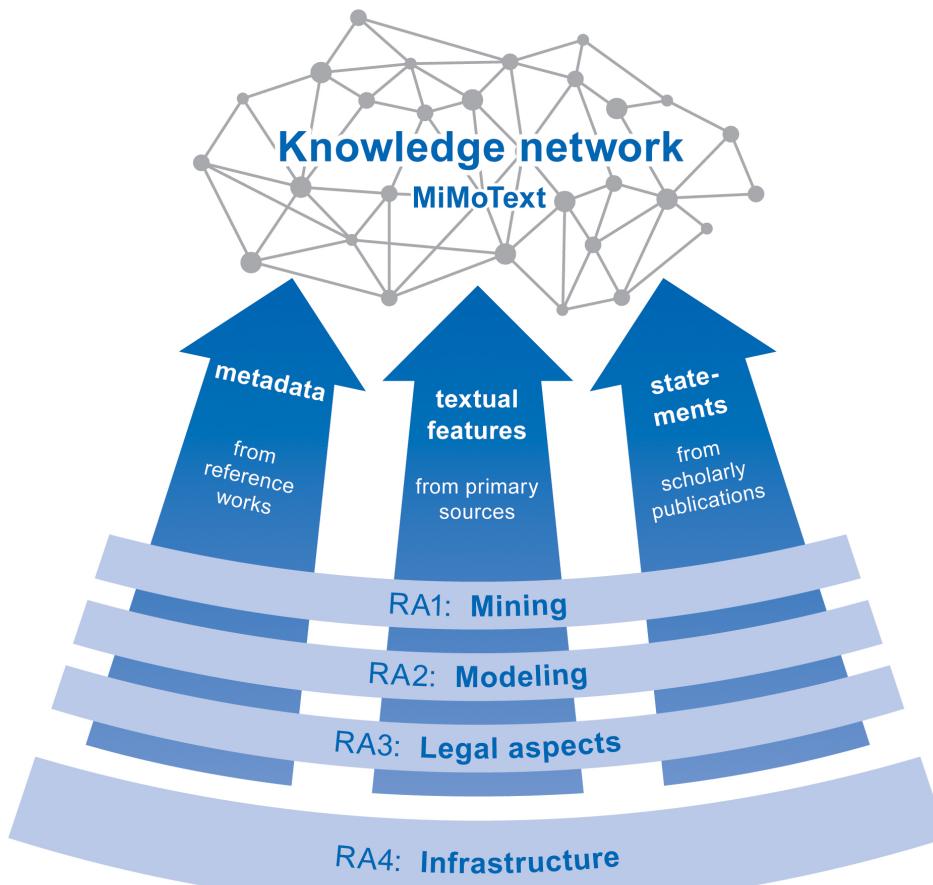


# Structure

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

# (1) Introduction

# MiMoText in a nutshell



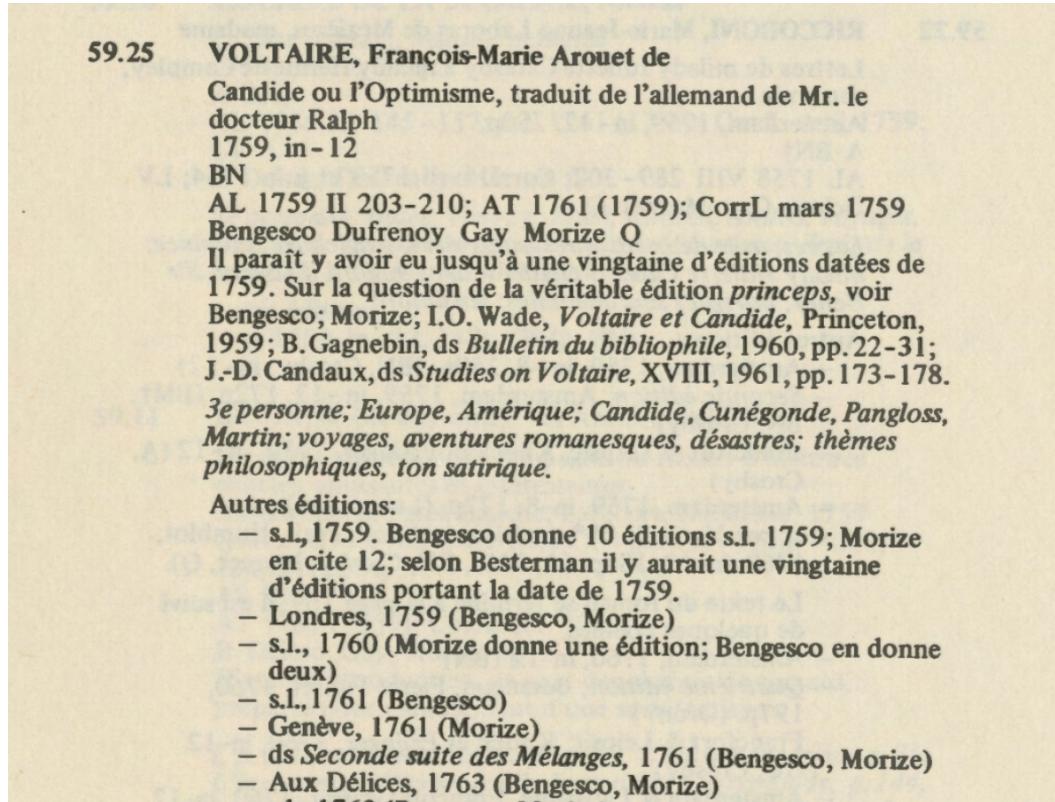
<https://mimotext.uni-trier.de/en>

# 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



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

# Source type 2: Primary literature (novels)

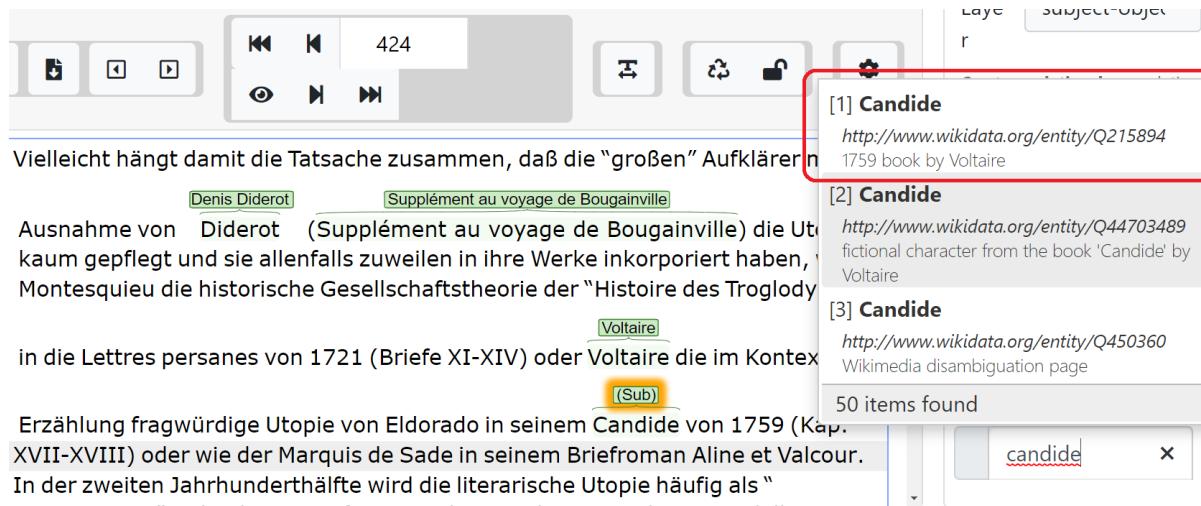
- corpus of 205 french novels (1750-1800)
- in XML-TEI, with metadata, following the ELTeC schema
- methods: e.g. Topic Modeling, NER, stylometry

The screenshot shows a GitHub repository page for 'roman18'. The left sidebar contains a 'README.md' file, a 'DOI' link (10.5281/zenodo.4061903), and a 'Contributors' section with 7 contributors. The main content area includes an 'Introduction' section and a 'Languages' section. The 'Languages' section features a horizontal bar chart and a table:

Language	Percentage
HTML	72.3%
Jupyter Notebook	26.6%
Python	1.1%

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)

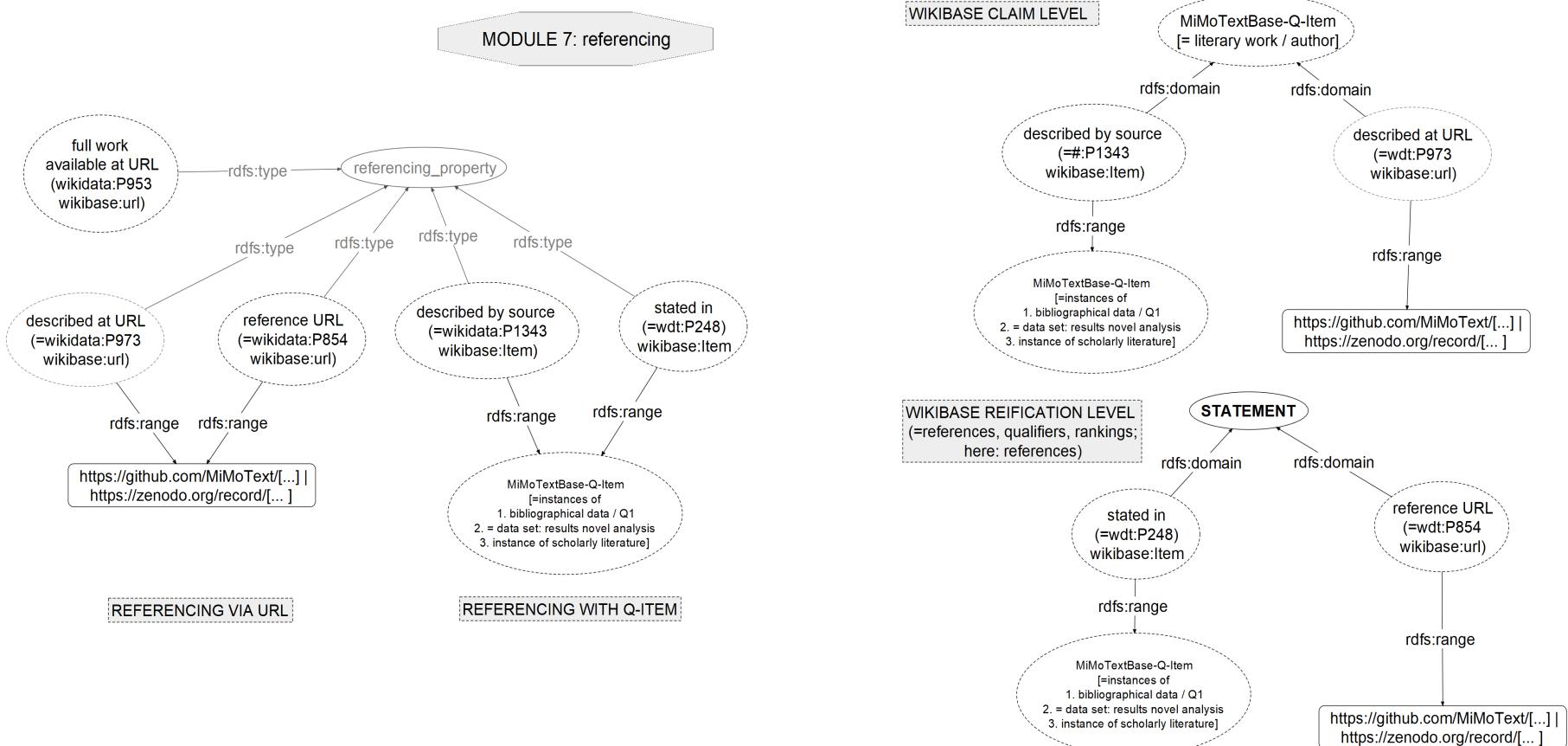
correspondence

▼ 2 references

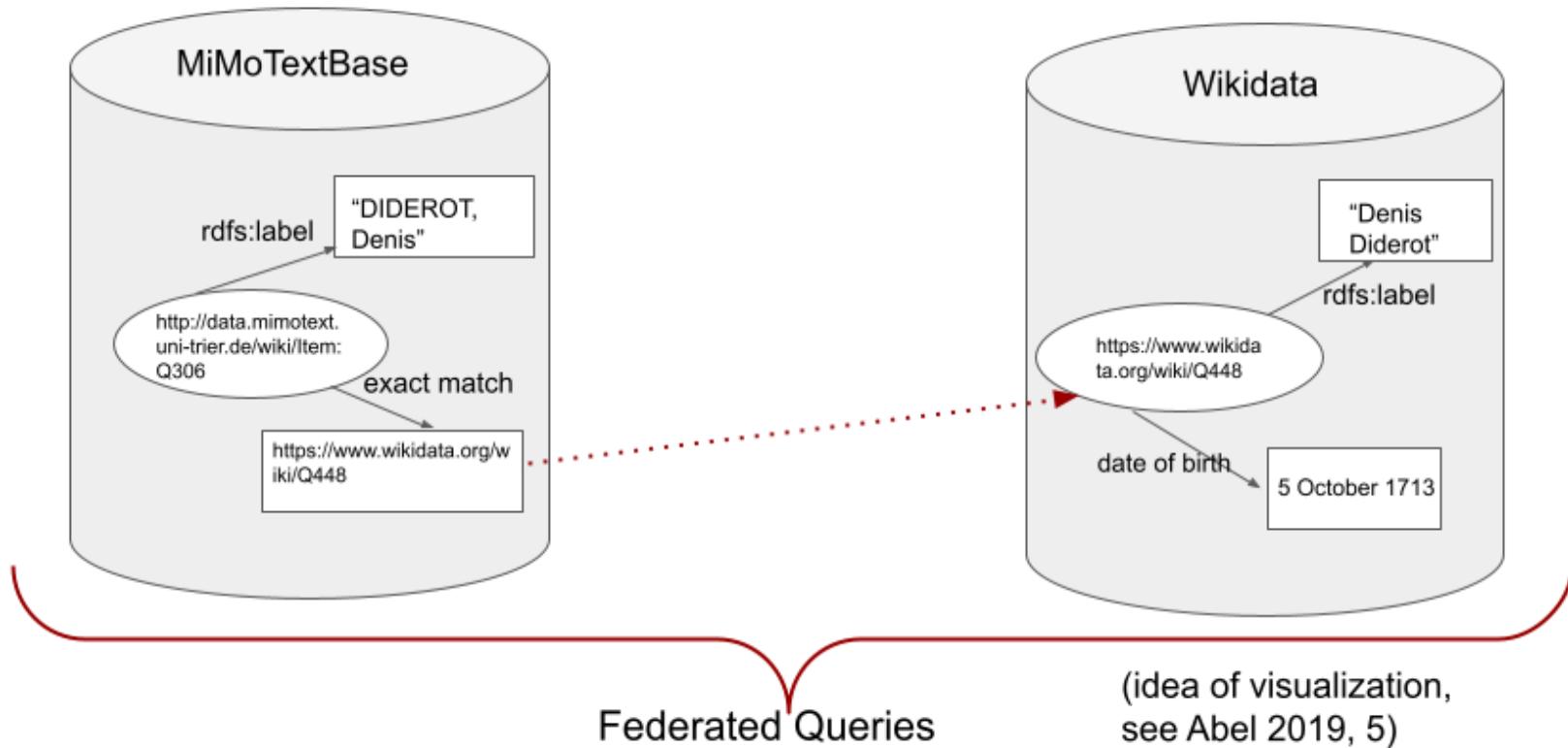
stated in      Topic Model MMT 11-2020

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

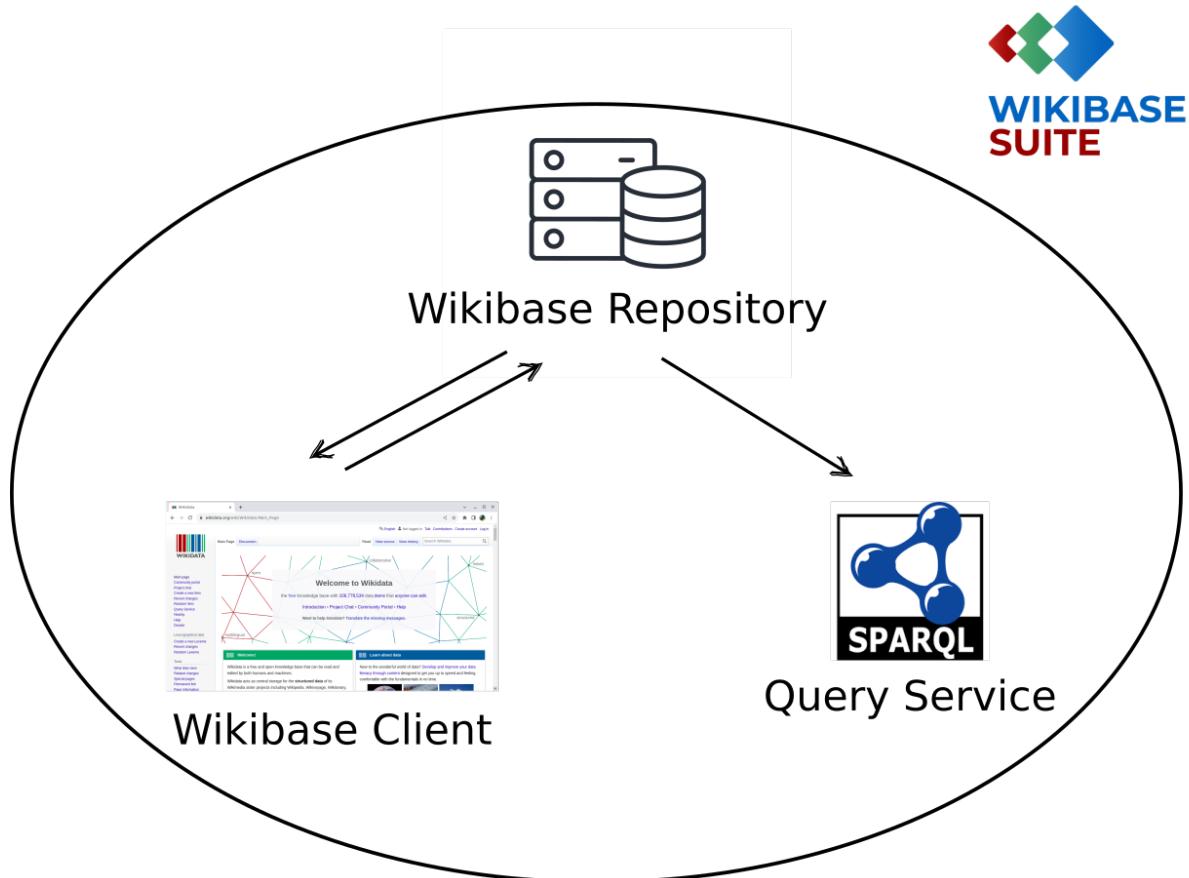
The screenshot shows the Douglas Adams page on Wikipedia. A red box highlights the "Wikidata" sidebar on the right side of the page. The sidebar contains the following information:

<b>Born</b>	Douglas Noel Adams 11 March 1952 Cambridge, England
<b>Died</b>	11 May 2001 (aged 49) Montecito, California, US
<b>Resting place</b>	Highbury Cemetery, London, England
<b>Occupation</b>	Author • screenwriter • essayist • humorist • satirist • dramatist
<b>Alma mater</b>	St John's College, Cambridge
<b>Genre</b>	Science fiction, comedy, satire
<b>Notable work</b>	<i>The Hitchhiker's Guide to the Galaxy</i> <i>Infogard</i> (1983)
<b>Notable awards</b>	
<b>Spouse</b>	Jane Henson (m. 1991)
<b>Children</b>	1
<b>Signature</b>	
<b>Website</b>	<a href="http://dougladasdams.com/">dougladasdams.com/</a>

since 2012



# 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

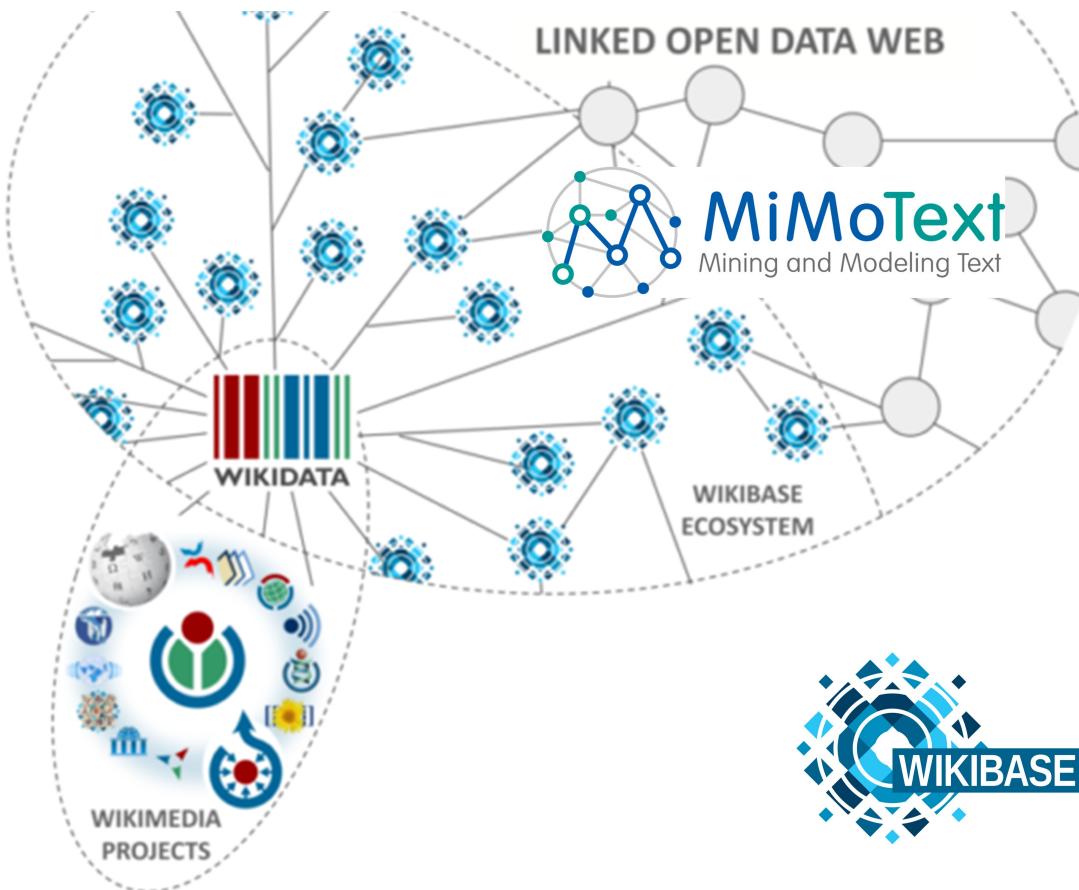
# Some example instances of Wikibase



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

The screenshot shows three Wikidata pages for the same entity, Douglas Adams (Q42), displayed in three different languages: Chinese (简体), English, and Arabic. The pages are part of a multilingual interface where the same data is presented in multiple languages simultaneously.

**Left Panel (Wikidata Home):** Contains links to various Wikidata features like '首页' (Home), '社群首页' (Community Home), and '词典编纂数据' (Data Curation).

**Top Navigation:** Shows tabs for '中文 (简体)' (Chinese Simplified), 'Viezporz' (User), and 'Douglas Adams - Wikidata' (Arabic).

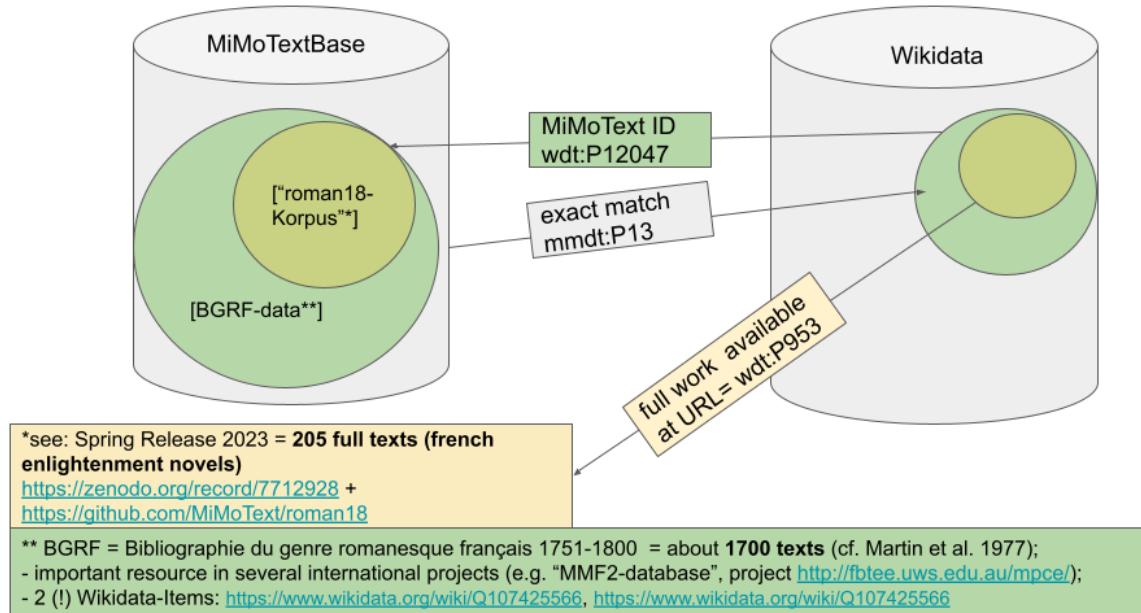
**Entity Pages:**

- Chinese Page:** Shows '道格拉斯·亚当斯 (Q42)' as the label. Below it, there's a summary box for '英国作家' (British Writer) and detailed sections for '陈述' (Statements), '性别' (Gender), '国籍' (Nationality), and '姓' (Name).
- English Page:** Shows 'Douglas Adams (Q42)' as the label. It provides a detailed description as an 'English writer and humorist' and lists 'Statements' related to his education at 'St John's College' and 'Brentwood School'.
- Arabic Page:** Shows 'دوغلاس آدمز (Q42)' as the label. It includes a summary box for 'كاتب إنجليزي فكاهي' (English writer and humorist) and lists 'بيانات' (Statements) for his education.

**Right Panel (Wikidata Features):** Includes sections for 'صفحة نقاش' (Discussion Page), 'الصفحة نقاش' (Discussion Page), 'بيانات' (Statements), 'مراجع' (References), 'الجنس' (Gender), 'بلد المواطن' (Place of Birth), 'الاسم الأول' (First Name), and 'الترتيب في التسلسل' (Order in Sequence). A sidebar on the right lists various Wikidata features in both Chinese and English.

# Connecting MiMoTextBase & Wikidata

## Future work / plans (10/2023)



- 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



Main page  
Recent changes  
Random page  
Help about MediaWiki

Tools  
What links here  
Related changes  
Special pages  
Printable version  
Permanent link  
Page information

In other languages  
Add links

Main Page Discussion Read View source View history Search MiMoText

## 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

A screenshot of a web-based SPARQL endpoint interface. At the top, there are navigation icons (refresh, help) and a status bar indicating "1366 results in 454 ms". To the right are links for "Code", "Download", and "Link". Below the header is a search bar with a magnifying glass icon. The main area is a table with the following columns: bgrf, item, authorlabel, itemLabel, year, narrpers, tonality, pages, and normalized. The table contains six rows of data, each corresponding to a book entry with details like title, author, year, and genre.

bgrf	item	authorlabel	itemLabel	year	narrpers	tonality	pages	normalized
51.42	<a href="#">Q &lt;http://data.mimotext.uni-trier.de/entity/Q1010&gt;</a>	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	<a href="#">Q &lt;http://data.mimotext.uni-trier.de/entity/Q1237&gt;</a>	TOUSSAINT, François-Vincent	Histoire des passions	1751				unbekannt
51.39	<a href="#">Q &lt;http://data.mimotext.uni-trier.de/entity/Q1235&gt;</a>	MÉHÉGAN, Guillaume-Alexandre, chevalier de	Zoroastre	1751	3e personne	satire de la religion chrétienne, du luxe, etc.	14 + 60p.	3e personne
51.38	<a href="#">Q &lt;http://data.mimotext.uni-trier.de/entity/Q1234&gt;</a>	MARTIGNY, comte de	Voyage d'Alcimédon	1751	3e personne	ton satirique	ix + 144p.	3e personne
51.37	<a href="#">Q &lt;http://data.mimotext.uni-trier.de/entity/Q1233&gt;</a>	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	<a href="#">Q &lt;http://data.mimotext.uni-trier.de/entity/Q1232&gt;</a>	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!



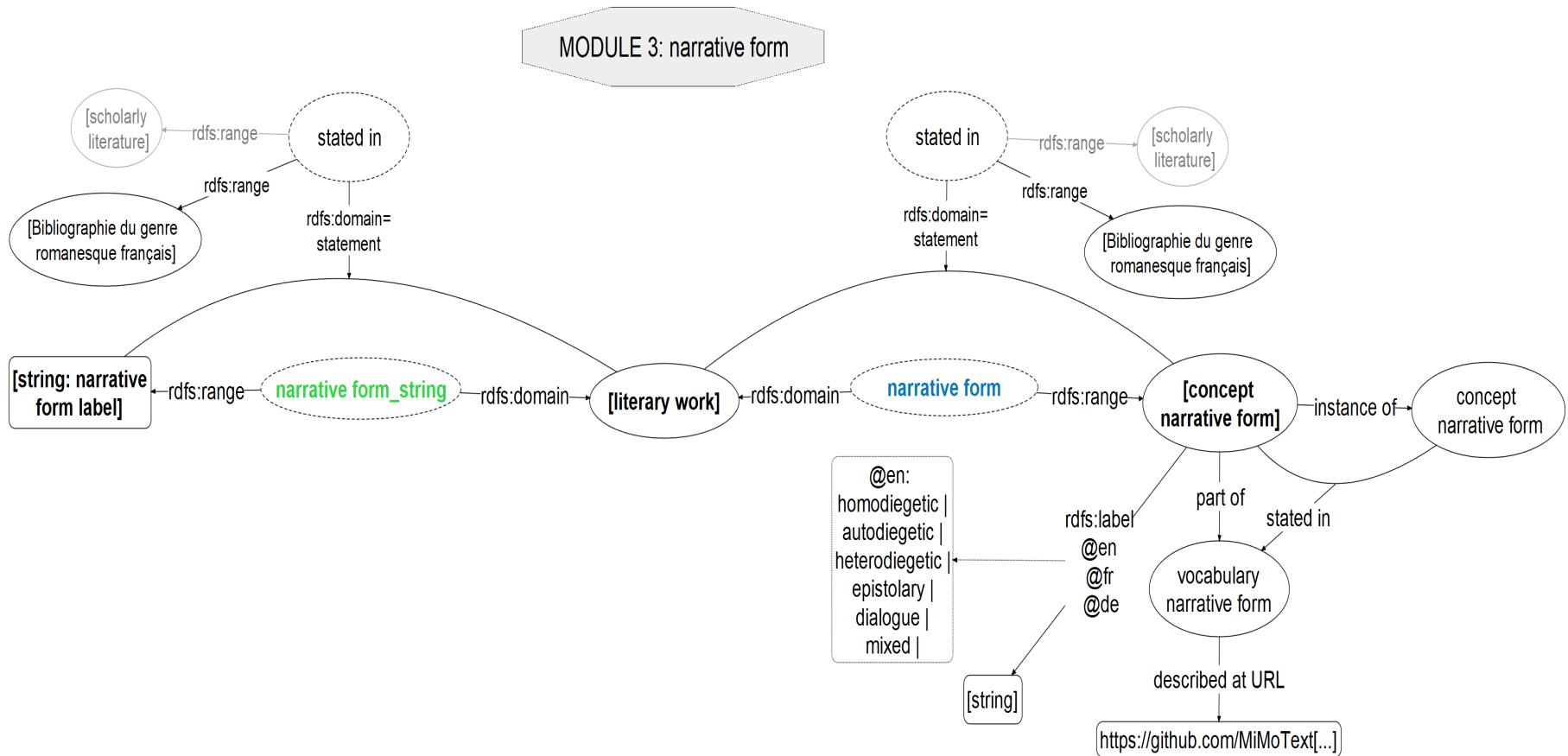
## To learn more

- 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>

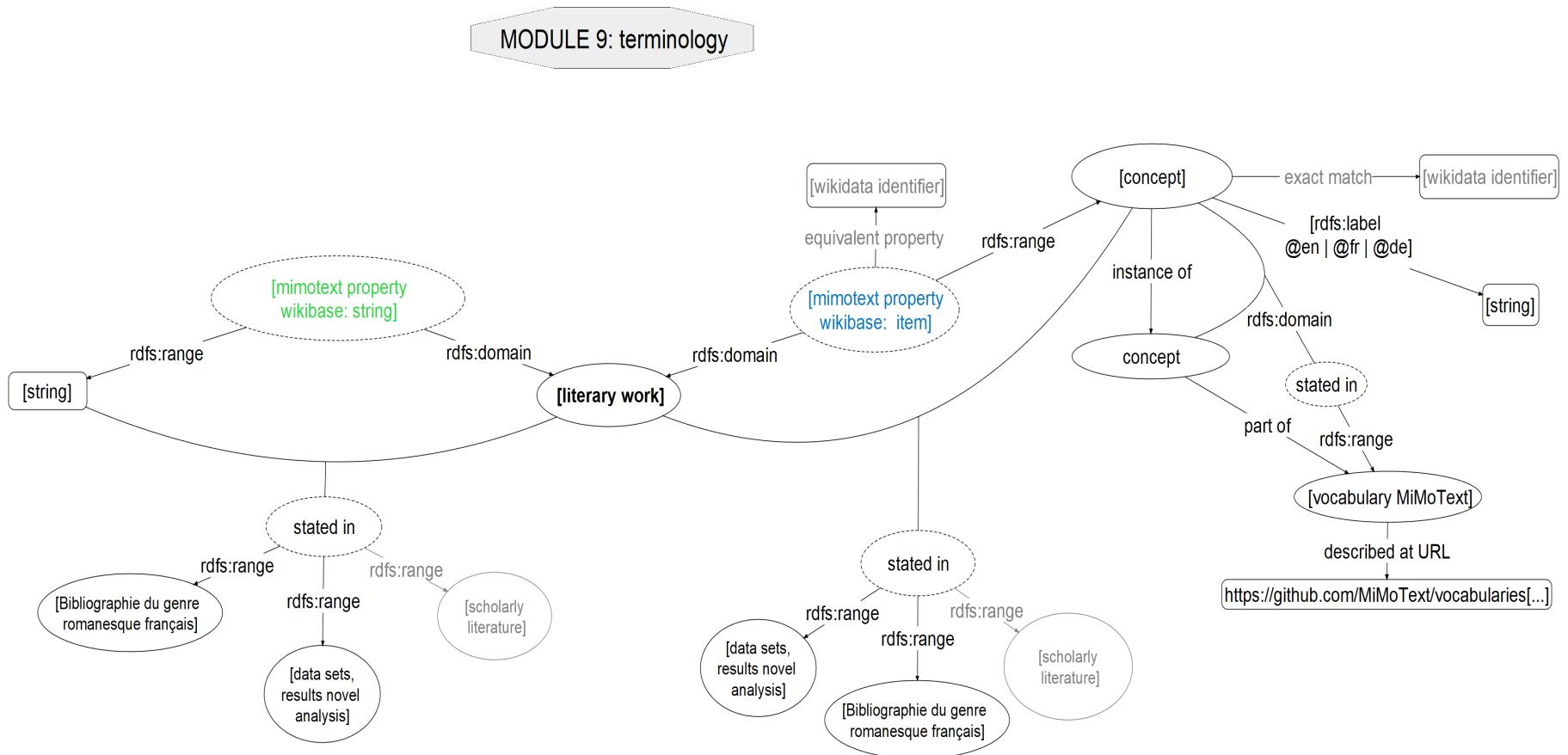
# Bonus slides

# Module 3: Narrative form



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

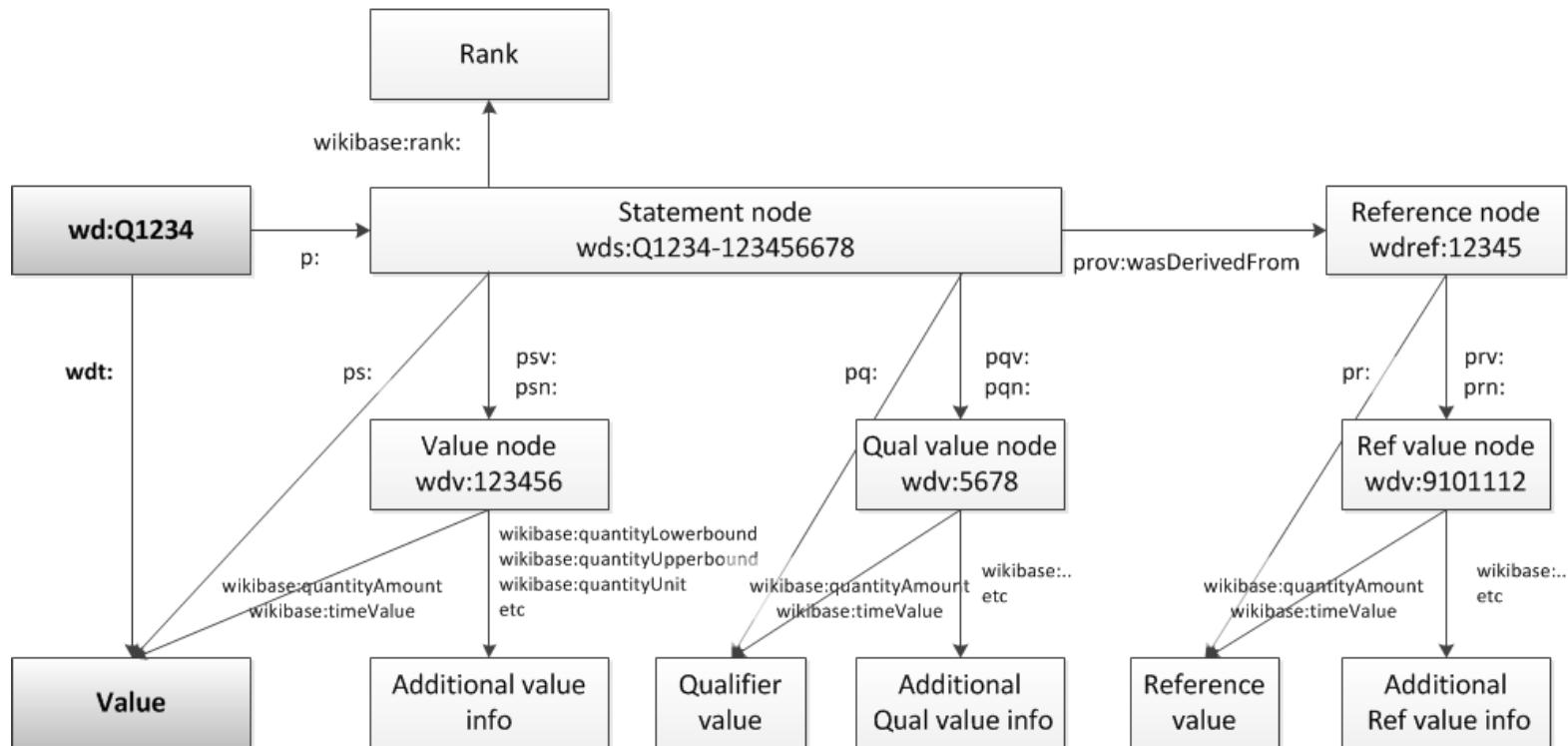
# Module 9: terminology



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

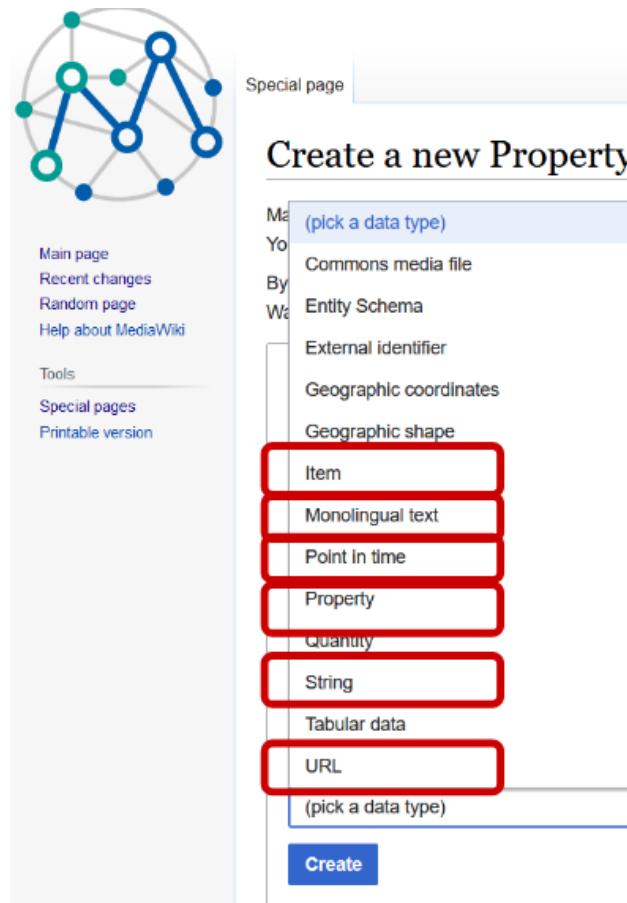
# Wikibase Data model (1)

## SPARQL data representation



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

# Wikibase Data model (2)



The screenshot shows the 'Create a new Property' form on a MediaWiki page. The top navigation bar includes links for 'Special page', 'Main page', 'Recent changes', 'Random page', 'Help about MediaWiki', 'Tools', 'Special pages', and 'Printable version'. The main form has a title 'Create a new Property' and a dropdown menu labeled '(pick a data type)' containing several options: 'Commons media file', 'Entity Schema', 'External identifier', 'Geographic coordinates', 'Geographic shape', 'Item' (highlighted with a red border), 'Monolingual text', 'Point in time', 'Property' (highlighted with a red border), 'Quantity', 'String', 'Tabular data', and 'URL' (highlighted with a red border). Below the dropdown is another '(pick a data type)' field and a blue 'Create' button.

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 interoperable 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)

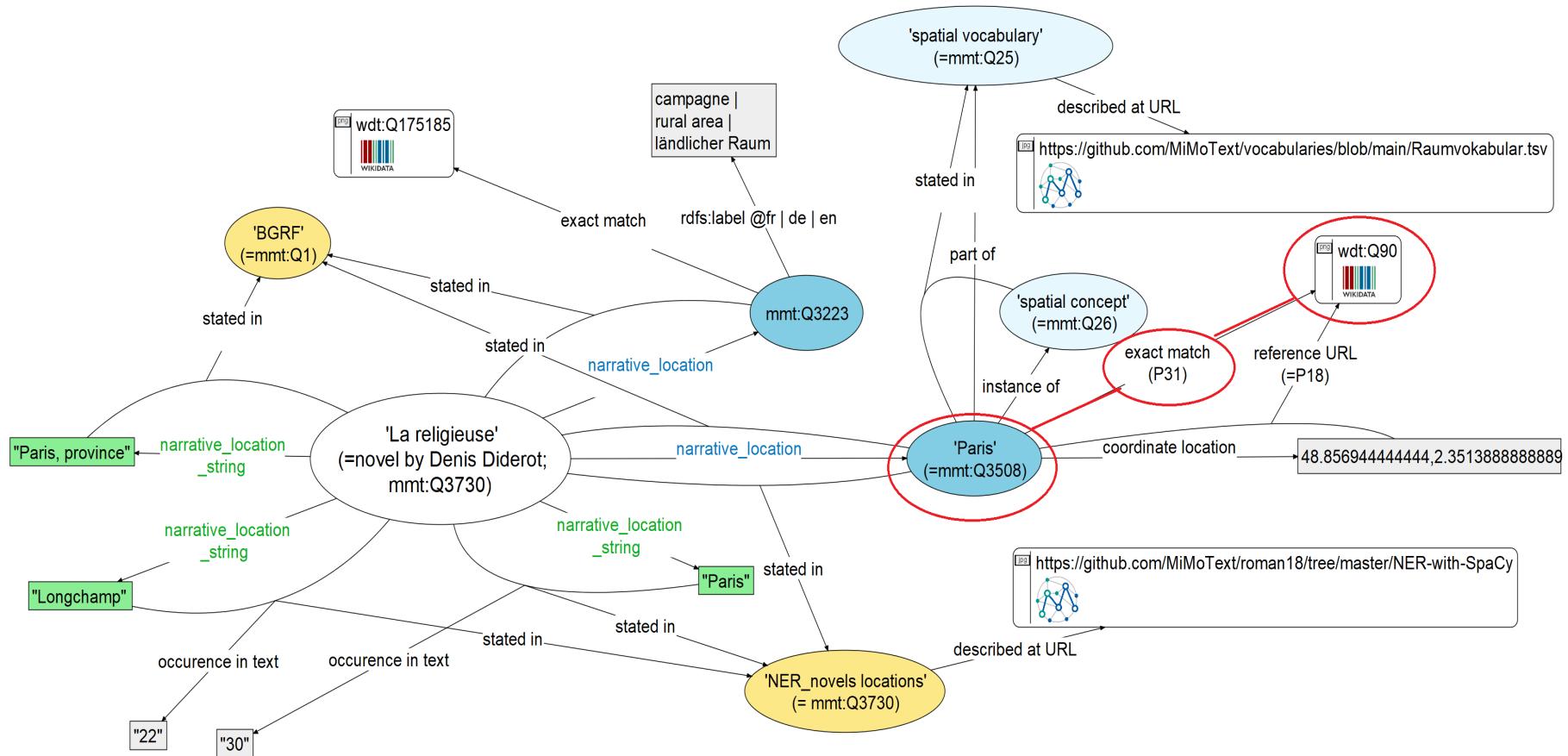
# Some of the challenges we have encountered

- 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>

Licence: Creative Commons Attribution (CC BY), 2023

---