UNIVERSITÄT TRIER

Kompetenzzentrum
Trier Center for Digital Humanities

Bigger Smarter Data

Extracting, Modeling and Linking Data for Literary History

Christof Schöch
(Trier University, Germany)

Korea University
Seoul, South Korea

23 May 2024

https://dhtrier.quarto.pub/ku – CC BY

Introduction

Thanks

Korea University, as well as KADH (Korean Association for Digital Humanities).

The Ministry for Research and Education in Rhineland-Palatinate, Germany, for funding this research (Mining and Modeling Text, 2019-2023)

Thanks to all the project contributors: Maria Hinzmann, Matthias Bremm, Tinghui Duan, Anne Klee, Johanna Konstanciak, Julia Röttgermann and many others.

Overview

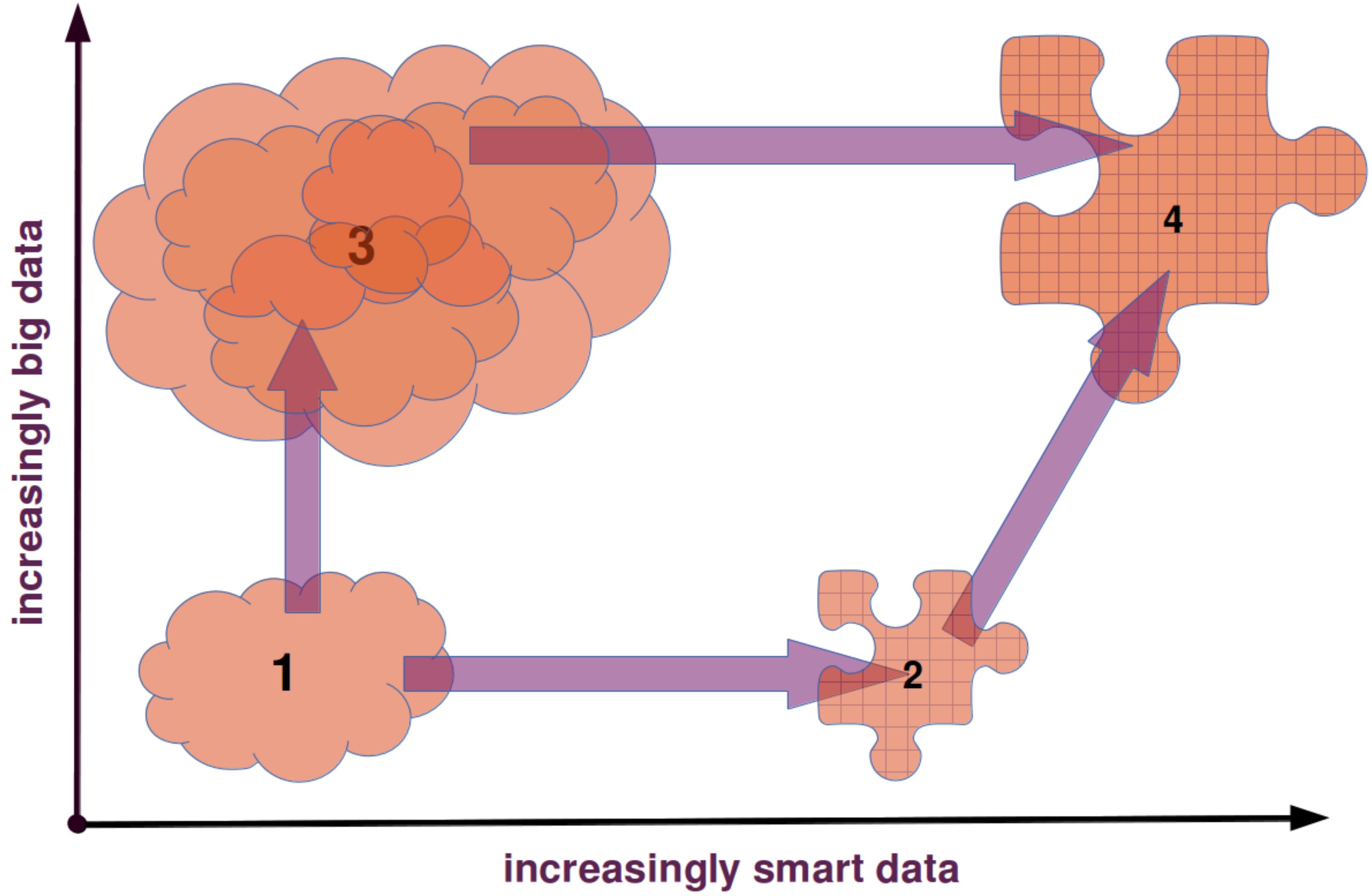
- 1 – Introduction
- 2 – Bigger Smarter Data and Linked Open Data
- 3 – Mining: Information Retrieval
- 4 – Modeling: Linked Open Data
- 5 – Results: Queryable Database
- 6 – Conclusion

Bigger Smarter Data:
Linked Open Data

Three Modes of Data (and Digital Humanities)

- Qualitative DH:
    - Datasets are typically small, curated, heavily annotated, flawless, specialized (‘smart data’)
    - Prototype: digital scholarly editions, e.g. *Faust Edition*
- Quantitative DH:
    - Datasets are typically large, scraped, unannotated, with errors and biases, generic (‘big data’)
    - Prototype: Large Language Models trained on lots of text, e.g. ChatGPT

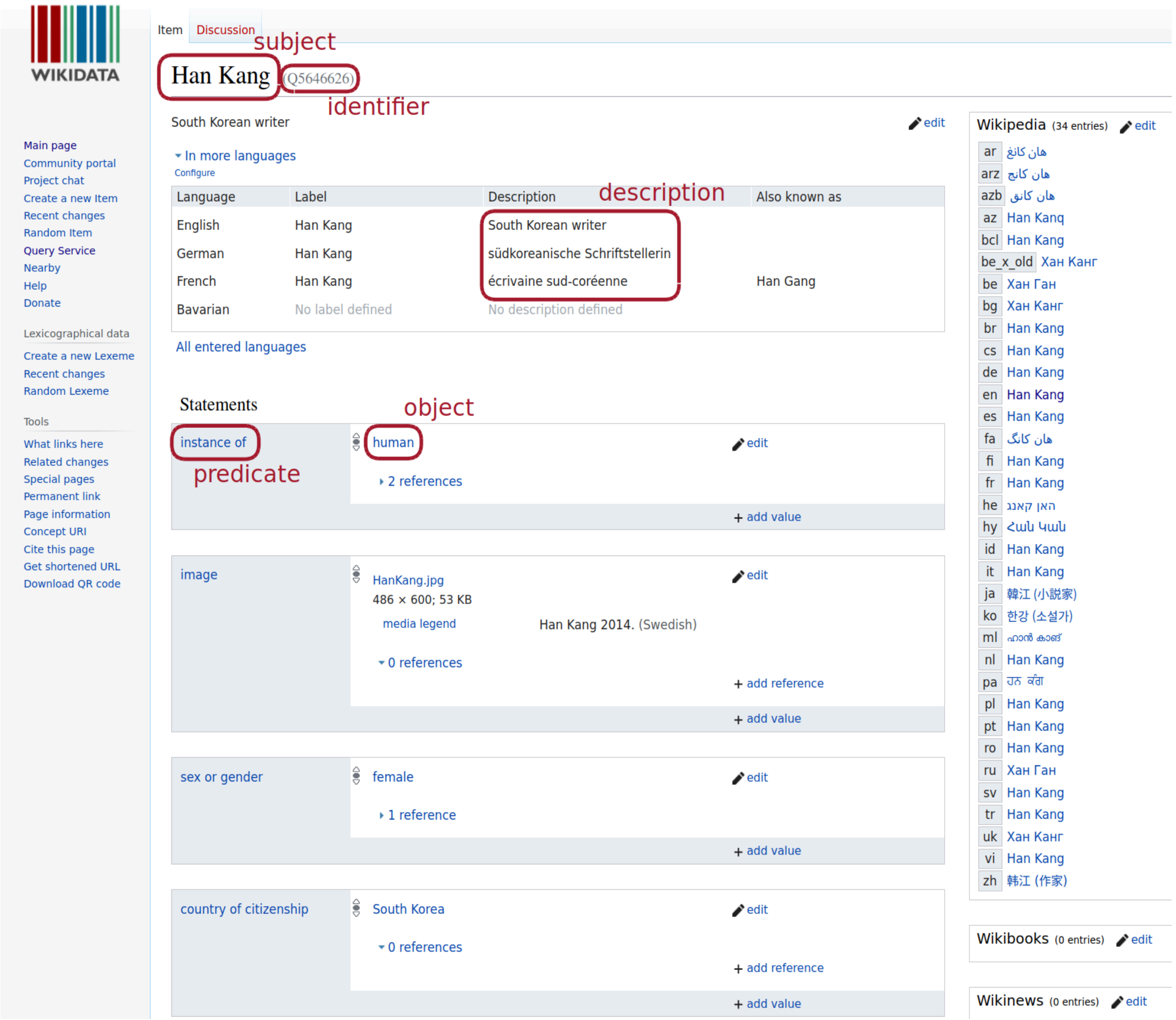
Third Way: Bigger Smarter Data



Background: What is Machine Learning?

- Fundamentally, ML involves detecting relations between features and labels
  - Features that we can observe in data
  - Labels, classes, or values that are relevant to our research
- We use this approach primarily for information retrieval
  - We start from a text collection
  - We may annotate part of the data, or use an existing model
  - And then derive labels, classes, values from the text

Background: What is Linked Open Data?



WIKIDATA

- Main page
- Community portal
- Project chat
- Create a new Item
- Recent changes
- Random Item
- Query Service
- Nearby
- Help
- Donate

Lexicographical data

- Create a new Lexeme
- Recent changes
- Random Lexeme

Tools

- What links here
- Related changes
- Special pages
- Permanent link
- Page information
- Concept URI
- Cite this page
- Get shortened URL
- Download QR code

Item Discussion

subject

Han Kang (Q5646626)

identifier

South Korean writer

edit

In more languages

Configure

description

Language	Label	Description	Also known as
English	Han Kang	South Korean writer	
German	Han Kang	südkoreanische Schriftstellerin	
French	Han Kang	écrivaine sud-coréenne	Han Gang
Bavarian	No label defined	No description defined	

All entered languages

Statements

object

instance of — human — edit

predicate

2 references

+ add value

image — HanKang.jpg — edit

486 × 600; 53 KB

media legend — Han Kang 2014. (Swedish)

0 references

+ add reference

+ add value

sex or gender — female — edit

1 reference

+ add value

country of citizenship — South Korea — edit

0 references

+ add reference

+ add value

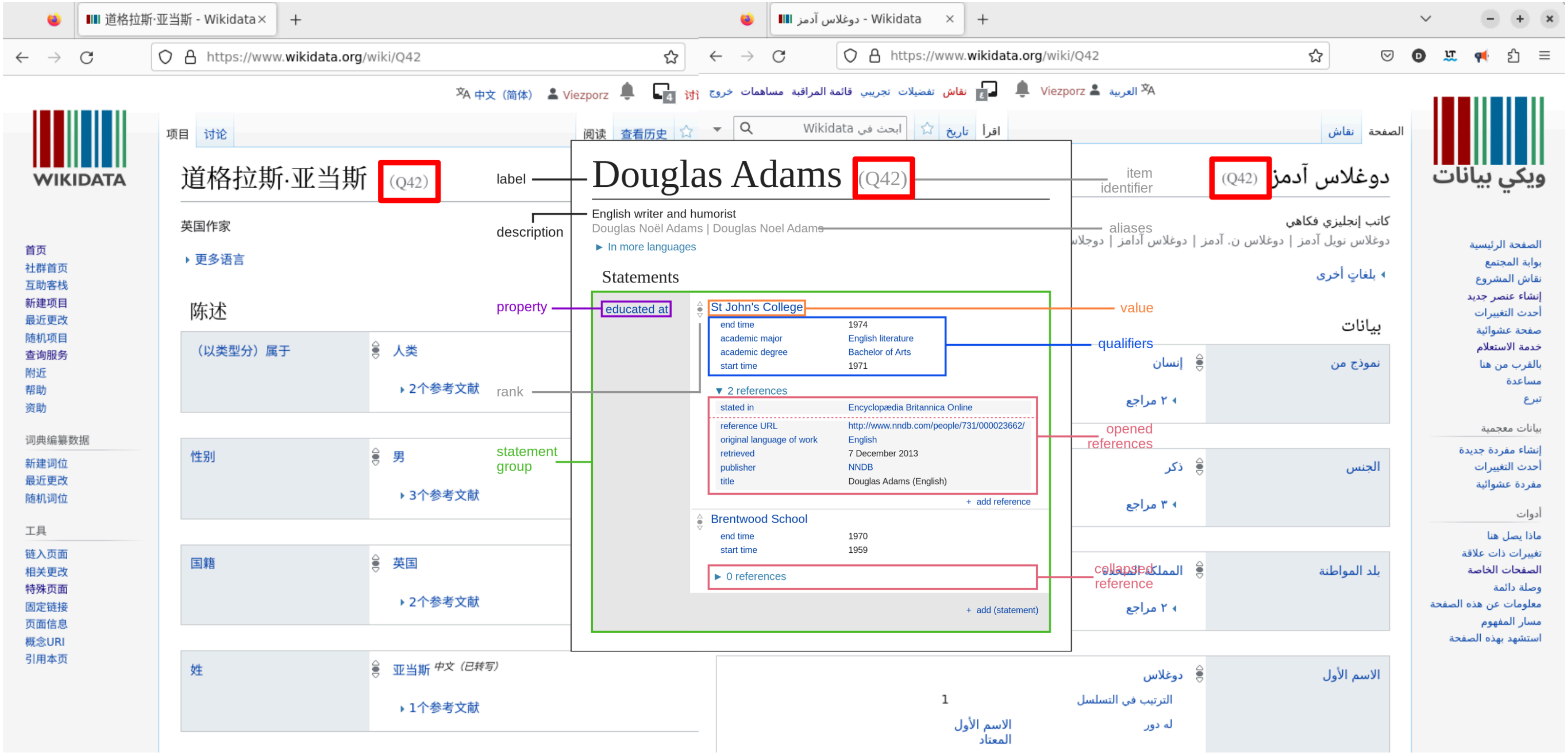
Wikipedia (34 entries) edit

ar	هان كانغ
arz	هان كانج
azb	هان كانق
az	Han Kanq
bcl	Han Kang
be_x_old	Хан Канг
be	Хан Ган
bg	Хан Канг
br	Han Kang
cs	Han Kang
de	Han Kang
en	Han Kang
es	Han Kang
fa	هان کانگ
fi	Han Kang
fr	Han Kang
he	האן קאנג
hy	Հան Կան
id	Han Kang
it	Han Kang
ja	韓江 (小説家)
ko	한강 (소설가)
ml	ഹാൻ കാങ്
nl	Han Kang
pa	ਹਨ ਕੰਗ
pl	Han Kang
pt	Han Kang
ro	Han Kang
ru	Хан Ган
sv	Han Kang
tr	Han Kang
uk	Хан Канг
vi	Han Kang
zh	韩江 (作家)

Wikibooks (0 entries) edit

Wikinews (0 entries) edit

Multilingualism



道格拉斯·亚当斯 - Wikidata

https://www.wikidata.org/wiki/Q42

文A 中文（简体） Viezporz 讨i

项目 讨论 阅读 查看历史

WIKIDATA

首页
社群首页
互助客栈
新建项目
最近更改
随机项目
查询服务
附近
帮助
资助

词典编纂数据

新建词位
最近更改
随机词位

工具

链入页面
相关更改
特殊页面
固定链接
页面信息
概念URI
引用本页

道格拉斯·亚当斯 (Q42)

英国作家

‣ 更多语言

陈述

（以类型分）属于	人类 ‣ 2个参考文献
性别	男 ‣ 3个参考文献
国籍	英国 ‣ 2个参考文献
姓	亚当斯 中文（已转写） ‣ 1个参考文献

Douglas Adams (Q42)

English writer and humorist

Douglas Noël Adams | Douglas Noel Adams

► In more languages

Statements

educated at

St John's College

end time	1974
academic major	English literature
academic degree	Bachelor of Arts
start time	1971

▼ 2 references

stated in	Encyclopædia Britannica Online
reference URL	http://www.nndb.com/people/731/000023662/
original language of work	English
retrieved	7 December 2013
publisher	NNDB
title	Douglas Adams (English)

+ add reference

Brentwood School

end time	1970
start time	1959

► 0 references

+ add (statement)

دوغلاس آدمز - Wikidata

https://www.wikidata.org/wiki/Q42

العربية Viezporz نقاش تفضيلات تجريبي قائمة المراقبة مساهمات خروج

الصفحة نقاش اقرأ تاريخ ابحث في Wikidata

ويكي بيانات

الصفحة الرئيسية
بوابة المجتمع
نقاش المشروع
إنشاء عنصر جديد
أحدث التغييرات
صفحة عشوائية
خدمة الاستعلام
بالقرب من هنا
مساعدة
تبرع

بيانات معجمية

إنشاء مفردة جديدة
أحدث التغييرات
مفردة عشوائية

أدوات

ماذا يصل هنا
تغييرات ذات علاقة
الصفحات الخاصة
وصلة دائمة
معلومات عن هذه الصفحة
مسار المفهوم
استشهد بهذه الصفحة

دوغلاس آدمز (Q42)

كاتب إنجليزي فكاهي

دوغلاس نويل آدمز | دوغلاس ن. آدمز | دوغلاس آدامز | دوجلاس

‹ بلغاتٍ أخرى

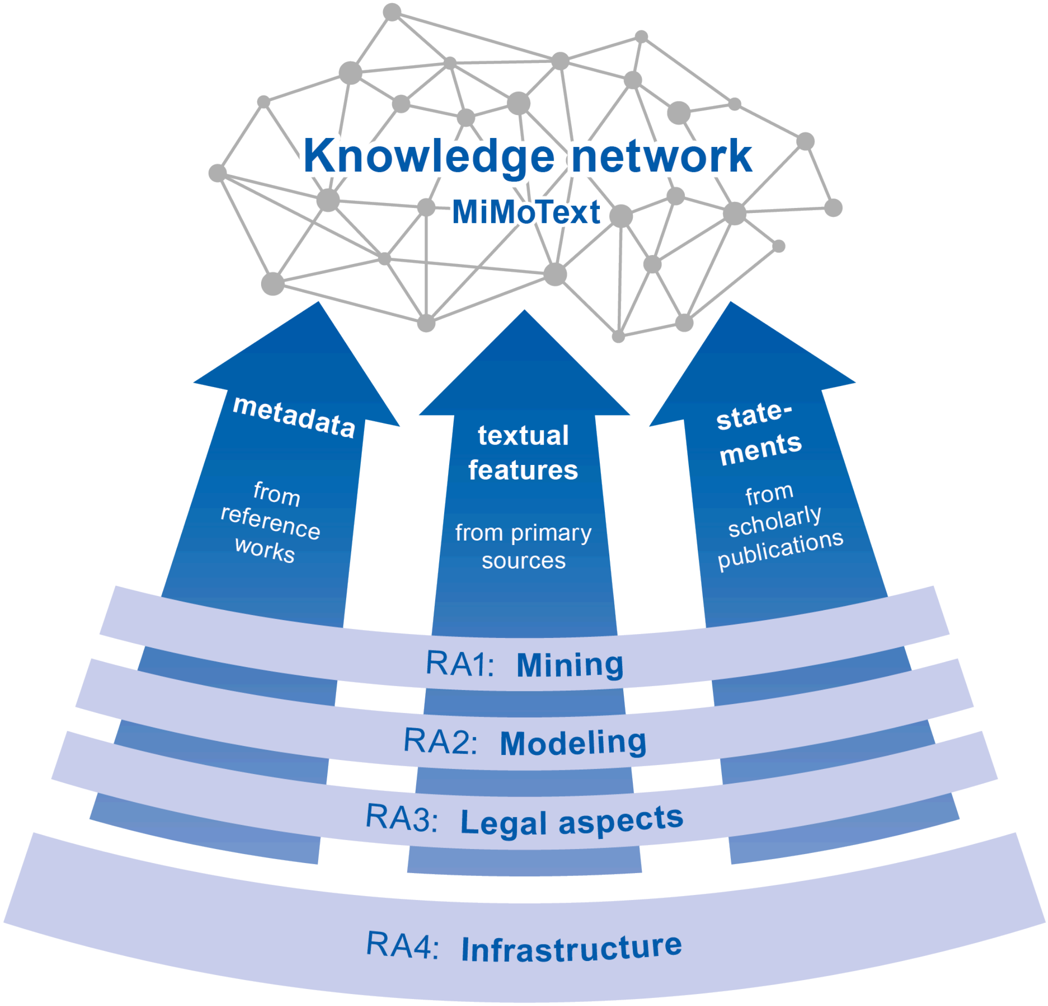
بيانات

نموذج من	إنسان ‹ ٢ مراجع
الجنس	ذكر ‹ ٣ مراجع
بلد المواطنة	المملكة المتحدة ‹ ٢ مراجع
الاسم الأول	دوغلاس الترتيب في التسلسل 1 له دور الاسم الأول المعتاد

Background: What is Literary History?

- Goals of literary history
    - Collecting and documenting knowledge of literary history
    - Providing explanations for the development of literature
- Organizational principles
    - Nations, periods, movements/currents, genres
    - Authors and works
    - Similarities and differences, continuities and change
- Functions
    - Explanations for literary developments
    - a. by cultural or socio-historical context
    - b. by internal dynamics of the literary system

The project 'Mining and Modeling Text'



Literary History in Linked Open Data

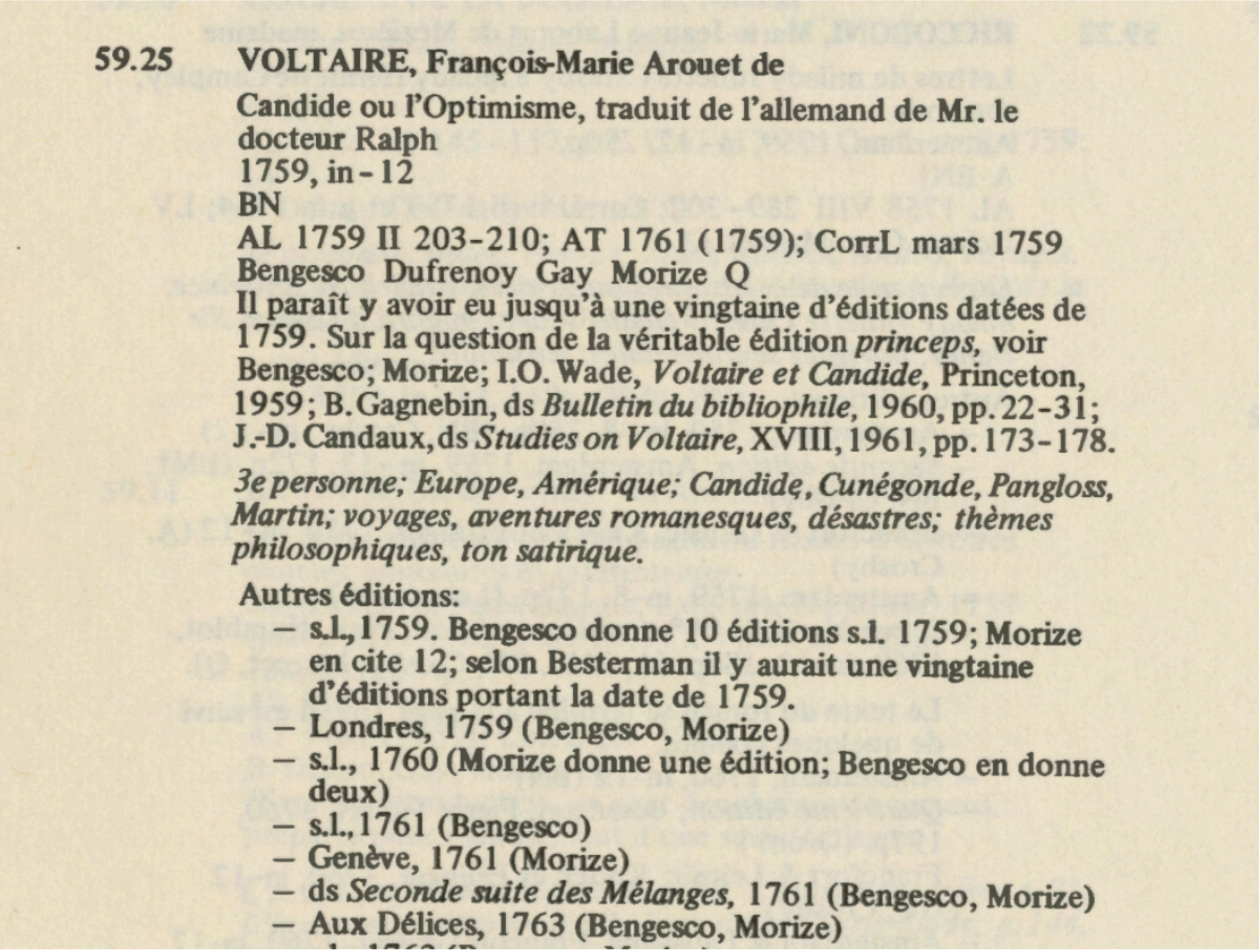
- Building blocks
  - Subjects, including persons (author, etc.) and works (primary text, scholarly literature, etc.)
  - Objects, including works, but also themes, locations, protagonists, literary genre, etc.
  - Predicates, as required, including: author_of, about, sameAs etc.
  - Qualifications, e.g: Source (with type, date, URL)
- Some exemplary statement types
  - Bibliographic: `[person] author_of [work]`
  - Contentual: `[work] about [theme]`
  - Formal: `[work] narrative_form [type]`
  - and many more.

Wikidata for Literary History

- Idea: Create a “Wikidata for the history of literature”
  - Literary history information system
  - LOD-based, with explorative interface and SPARQL endpoint
  - Approach of an “atomization” of the historical knowledge
  - Linking with other knowledge systems (taxonomies, standard data, knowledge bases)
  - Key values: human and machine readable, open, collaborative, multilingual
- Compared to Wikidata
  - Focused on one domain (French novel, 1750-1800)
  - Better coverage / higher density of information for this domain
  - Development of a systematic ontology
  - much smaller: 300k vs. 1.5 billion statements

Mining: Information Extraction

Pillar 1: *Bibliographie du genre romanesque français*



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

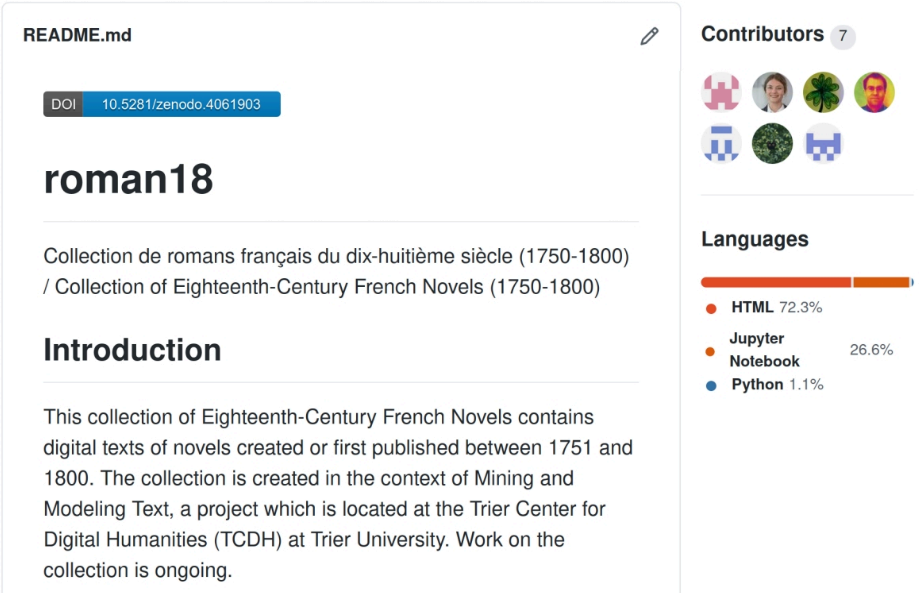
Il paraît y avoir eu jusqu'à une vingtaine d'éditions datées de 1759. Sur la question de la véritable édition *princeps*, voir Bengesco; Morize; I.O. Wade, *Voltaire et Candide*, Princeton, 1959; B. Gagnebin, ds *Bulletin du bibliophile*, 1960, pp. 22-31; J.-D. Candaux, ds *Studies on Voltaire*, XVIII, 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)

Pillar 2: primary literature (novels)



README.md

DOI 10.5281/zenodo.4061903

roman18

Collection de romans français du dix-huitième siècle (1750-1800) / Collection of Eighteenth-Century French Novels (1750-1800)

Introduction

This collection of Eighteenth-Century French Novels contains digital texts of novels created or first published between 1751 and 1800. The collection is created in the context of Mining and Modeling Text, a project which is located at the Trier Center for Digital Humanities (TCDH) at Trier University. Work on the collection is ongoing.

Contributors 7

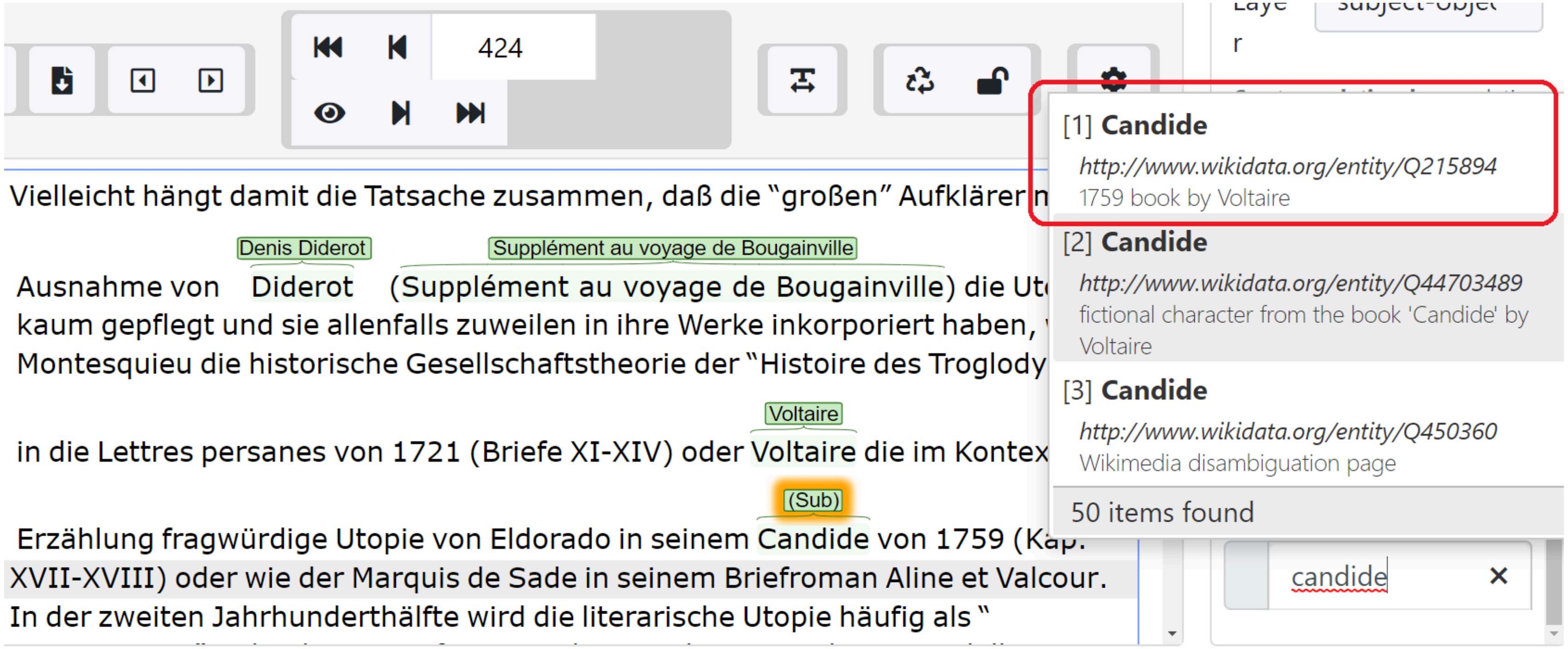
Languages

- ● HTML 72.3%
- ● Jupyter Notebook 26.6%
- ● Python 1.1%

- Corpus of 200 French novels (1750-1800)
- Coding: in XML-TEI, with metadata, according to ELTeC schema
- Analysis methods: Topic modeling, NER, stylometry, etc.

Collection of Eighteenth-Century French novels (1750-1800), ed. Julia Röttgermann. See

Pillar 3: Scholarly Literature



Vielleicht hängt damit die Tatsache zusammen, daß die "großen" Aufklärer n

Ausnahme von Diderot (Supplément au voyage de Bougainville) die Ut
kaum gepflegt und sie allenfalls zuweilen in ihre Werke inkorporiert haben,
Montesquieu die historische Gesellschaftstheorie der "Histoire des Troglody

in die Lettres persanes von 1721 (Briefe XI-XIV) oder Voltaire die im Kontex

Erzählung fragwürdige Utopie von Eldorado in seinem Candide von 1759 (Kap.
XVII-XVIII) oder wie der Marquis de Sade in seinem Briefroman Aline et Valcour.
In der zweiten Jahrhunderthälfte wird die literarische Utopie häufig als "

[1] **Candide**
http://www.wikidata.org/entity/Q215894
1759 book by Voltaire

[2] **Candide**
http://www.wikidata.org/entity/Q44703489
fictional character from the book 'Candide' by Voltaire

[3] **Candide**
http://www.wikidata.org/entity/Q450360
Wikimedia disambiguation page

50 items found

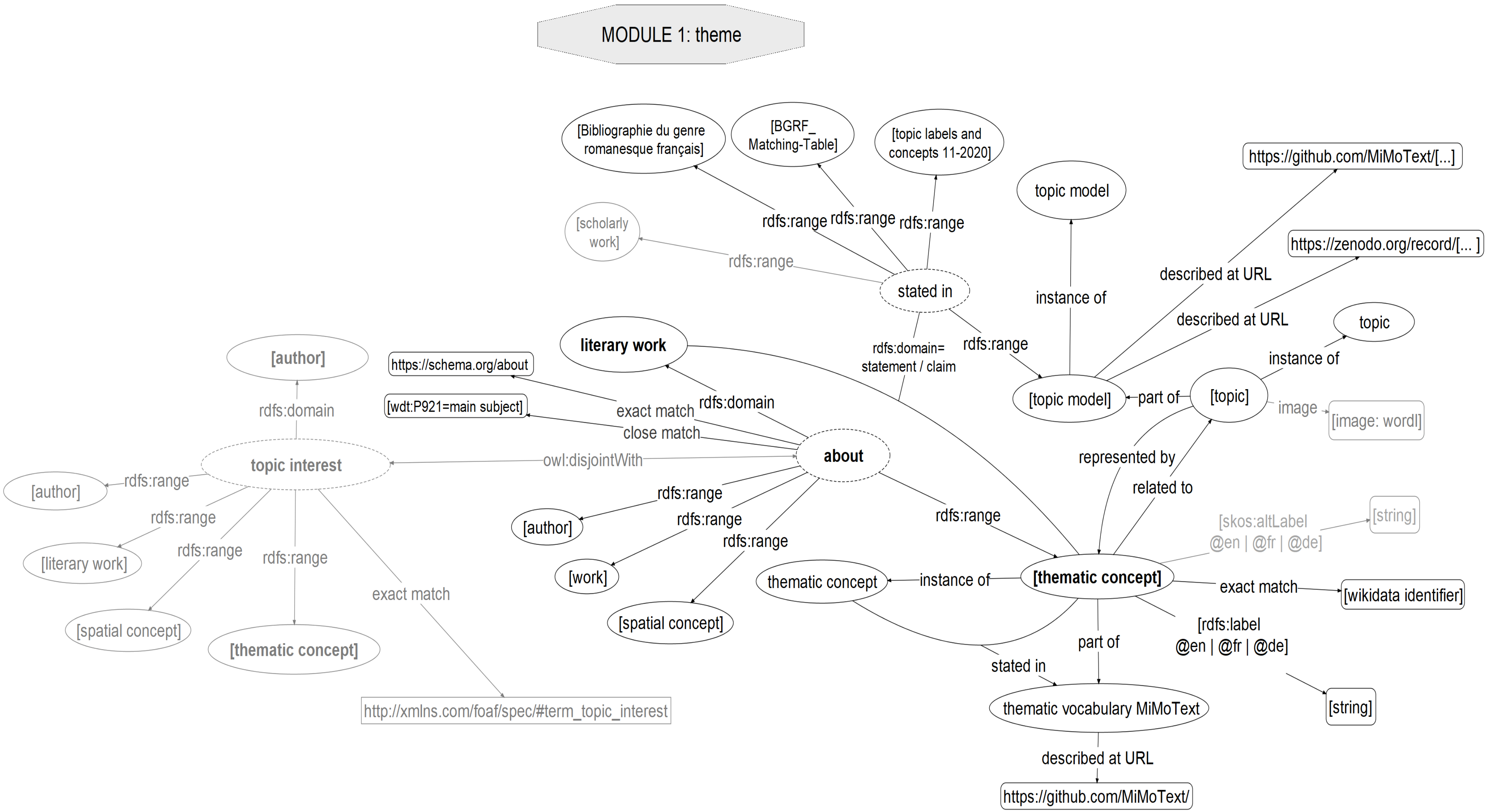
- Annotation Guidelines (based on the data model)
- Manual annotations (using INCEpTION)
- Linking of INCEpTION with MiMoTextBase and Wikidata => disambiguation
- Creation of statements about authors and works (genres, themes, etc.)

Modeling: Data Modeling

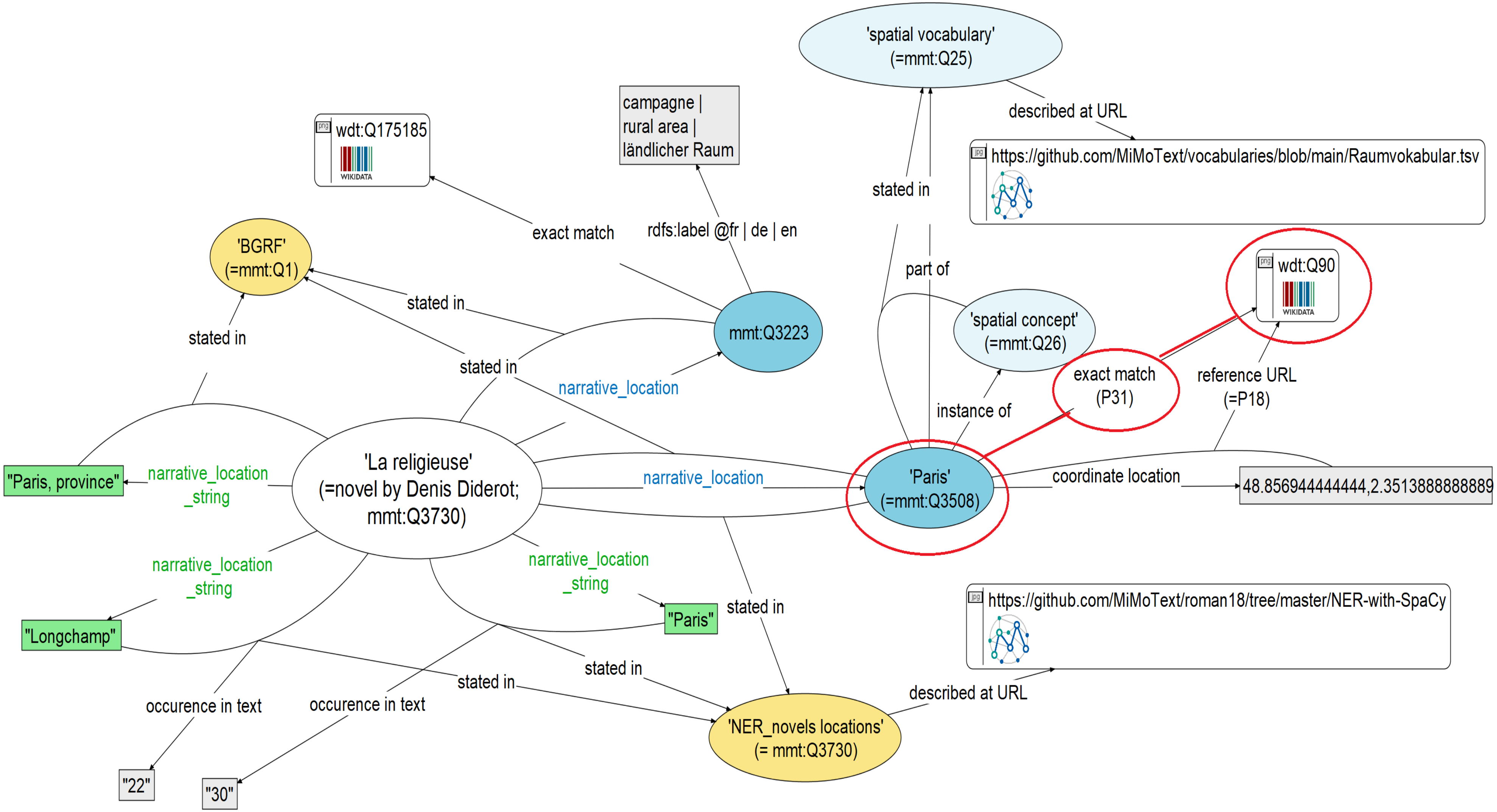
Modular Data Model

- 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 literature

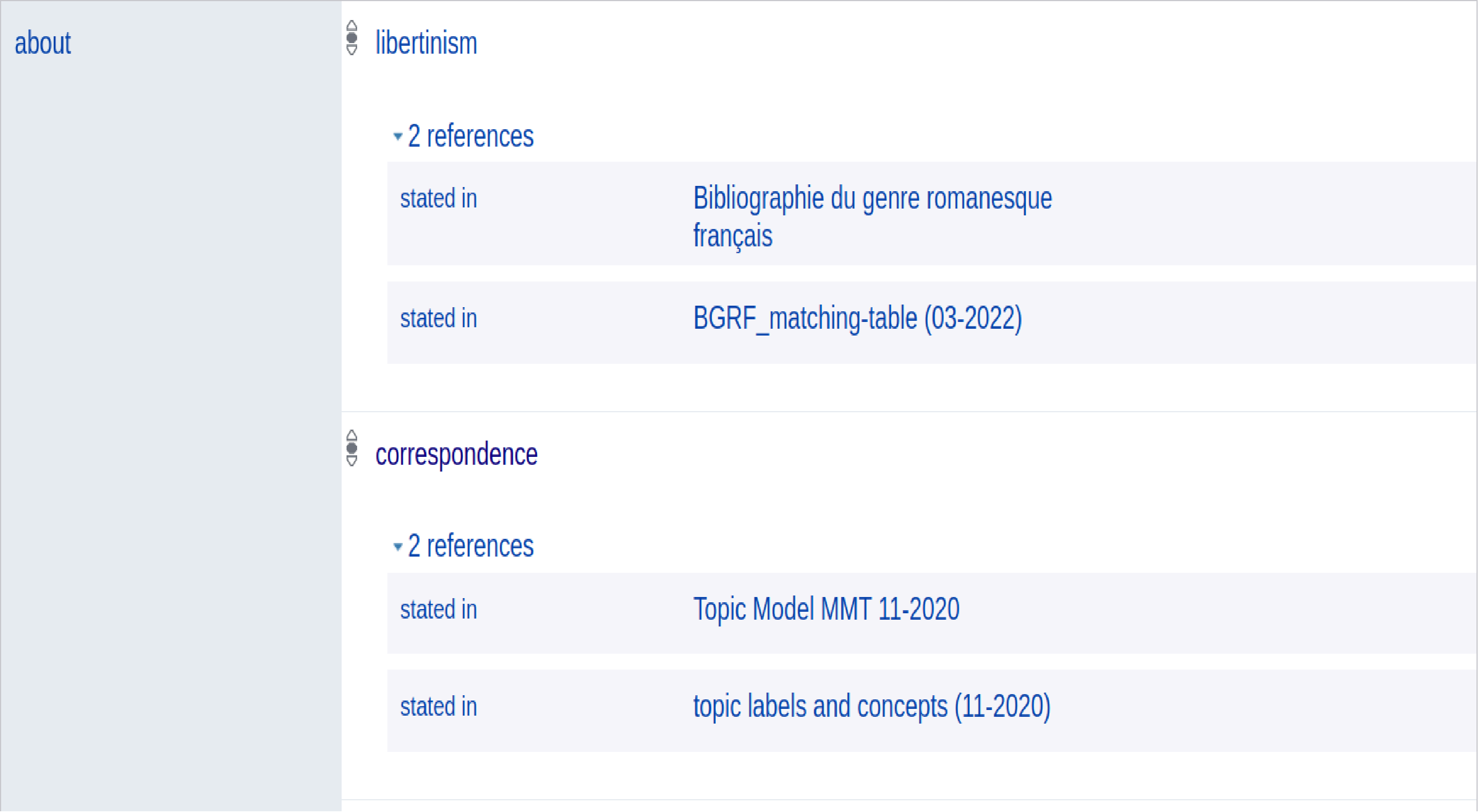
Example: The module on themes



Example: The module on narrative location



Meta-Statements



about

libertinism

▾ 2 references

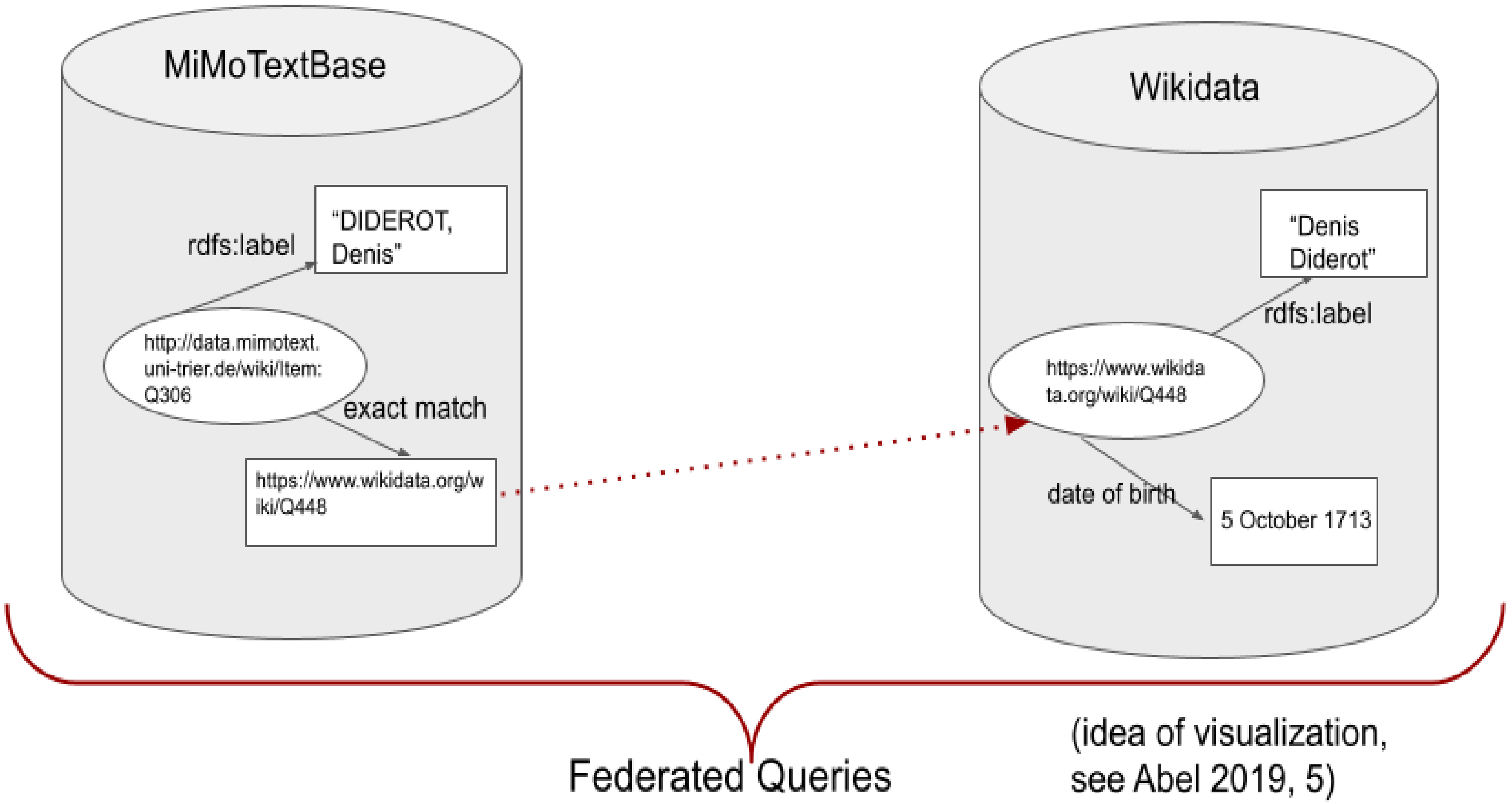
stated in	Bibliographie du genre romanesque français
stated in	BGRF_matching-table (03-2022)

correspondence

▾ 2 references

stated in	Topic Model MMT 11-2020
stated in	topic labels and concepts (11-2020)

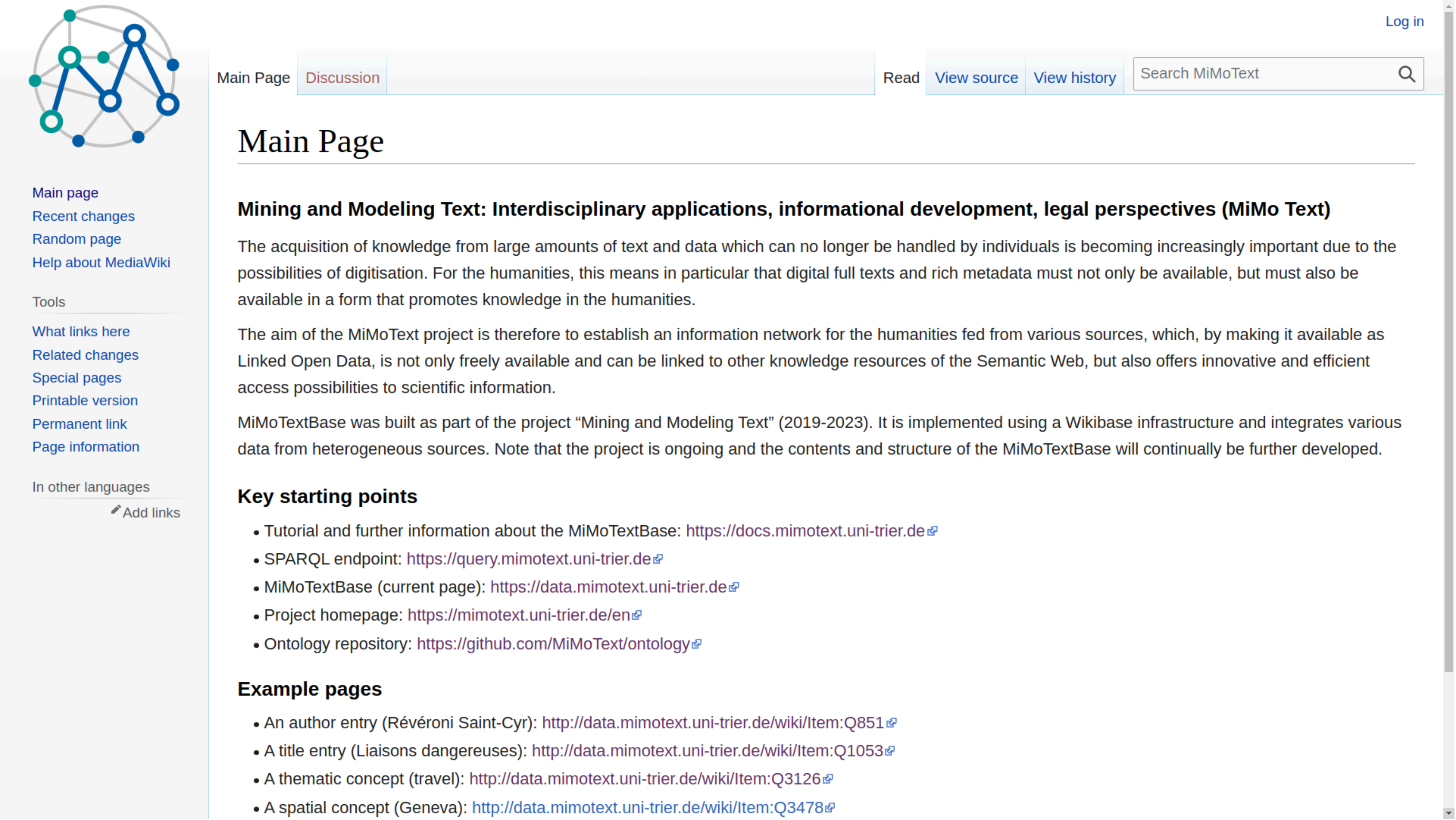
Linking with Wikidata for 'federated queries'



(idea of visualization, see Abel 2019, 5)

Result: Queryable Database

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.

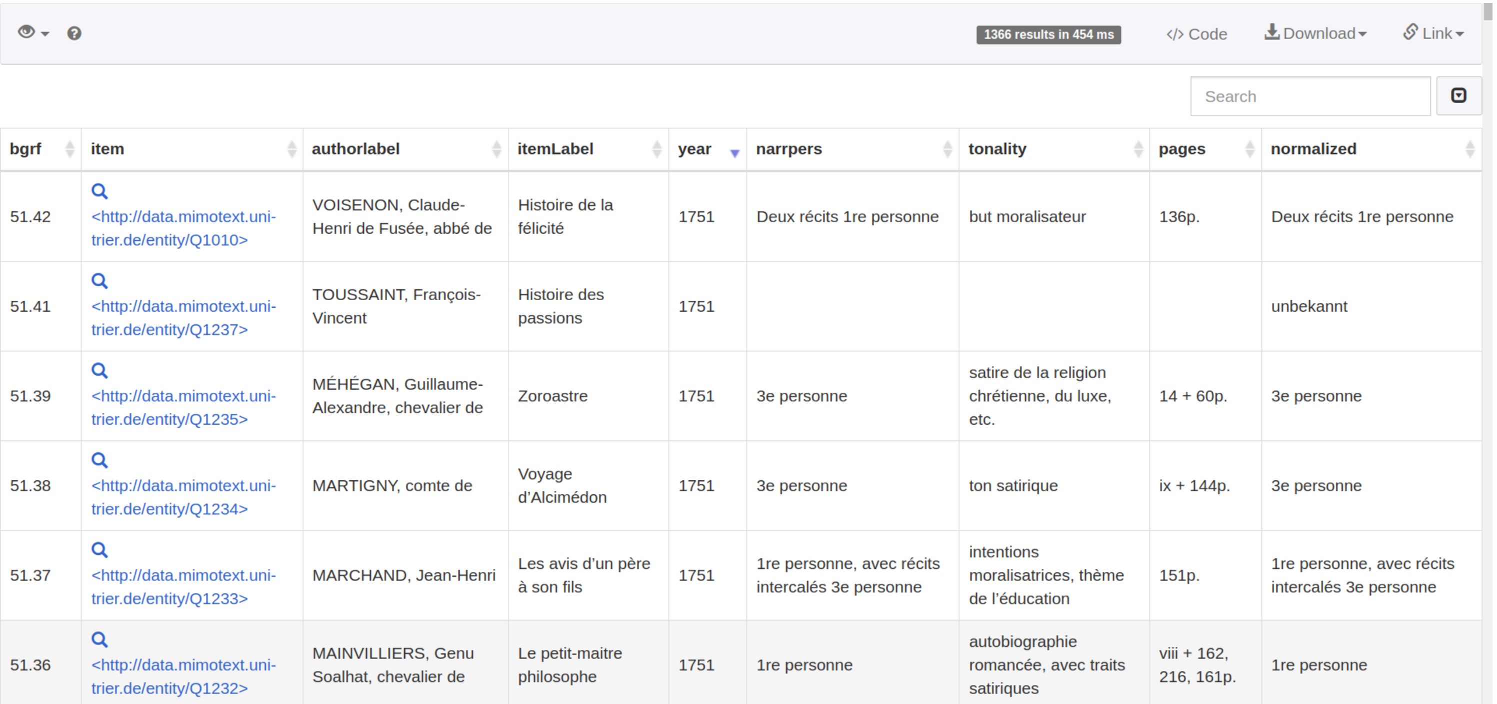
Key starting points

- 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 (Révéroni 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

SPARQL endpoint

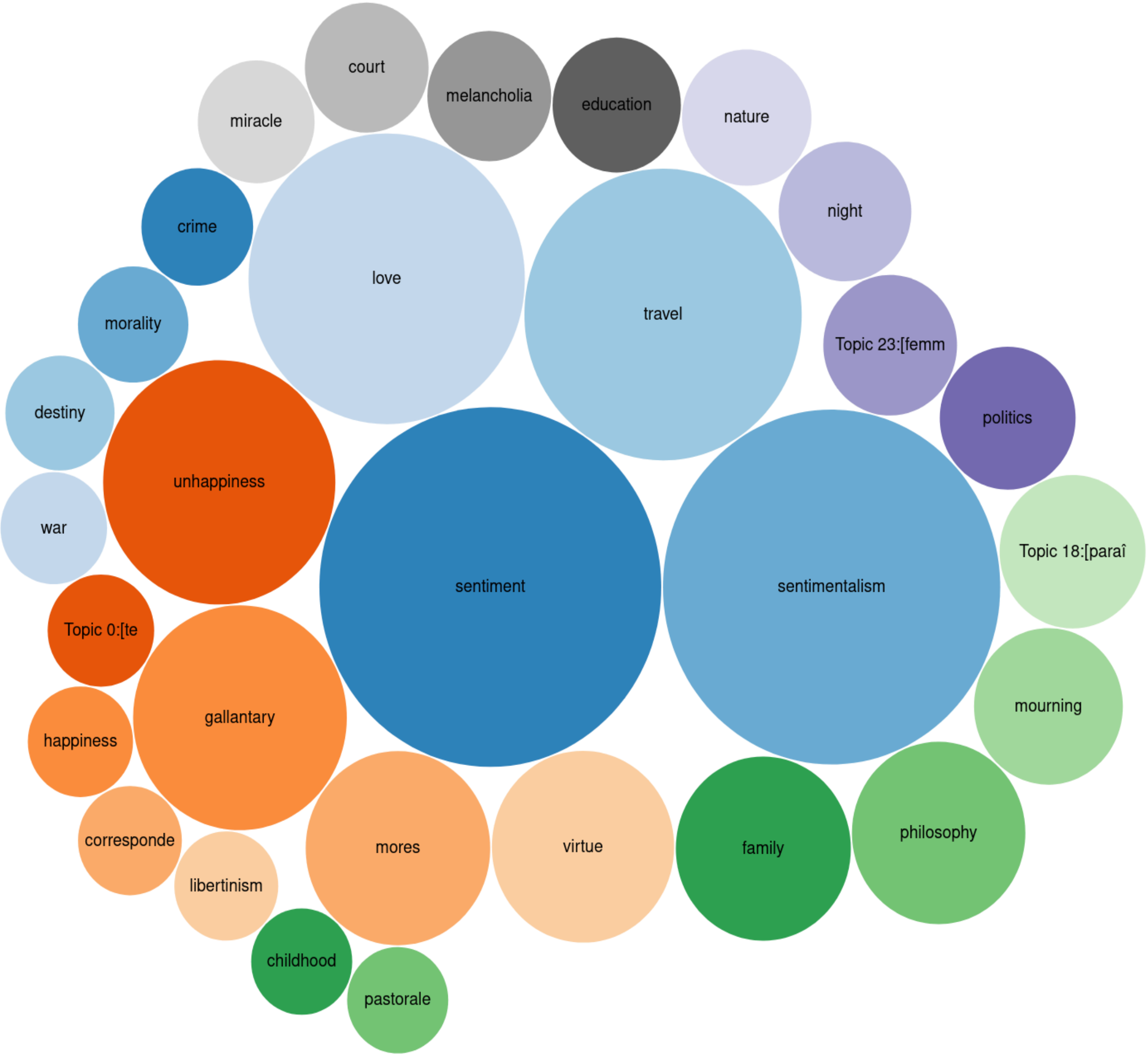


1366 results in 454 ms

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 queries

MiMoText Base: Query for themes in novels



Some sample queries: simple queries

- List of novels with information from BGRF
- Number of works per author (first 25))
- Themes of novels, in French and in English

Example queries: visualizations

- Number of novels per year
- Narrative forms over time (decades)
- Book history: print formats over time (5 years)

Sample queries: networked and federated

- Link with catalogue data from French National Library (using BNF id)
- Narrative locations of novels (map)
- Authors by birth year, with portrait)
- Alternative author names from Wikidata infobox
- Network of influences between authors (using ‘influenced by’)
- Querying MiMoText from Wikidata (it works both ways)
- Novels and basic information, from Wikidata

Sample queries: comparative queries

- Themes from topic modeling compared to themes in BGRF
- Themes from BGRF vs. Topic Modeling (in one query)

Conclusion

Opportunities & challenges

- Opportunities
  - Linking heterogeneous data from different types of sources
  - Modeling, collecting and comparing contradictory statements
  - Transparency in knowledge production (sources)
- Challenges
  - Lack of consensus on relevant statement types in the discipline
  - Complexity reduction (triple structure)
  - Interoperability (tension ‘Wikiverse’ vs. OWL standard)

Lessons Learned

- Federated queries
  - Central element of the LOD vision
  - => Making it happen is not trivial (data model, infrastructure)
- Modeling meta statements
- Very important: perspectives / statements, not facts
- => Very different approaches in different technical contexts
- Exchange across communities
  - Literary Studies vs. Digital Humanities vs. Wikiverse
  - => is essential but needs more development
- There is still so much to do!
  - => We are continuing this effort in a new project called ‘Linked Open Data in the Humanities’ (LODinG)

Many thanks for your kind attention

Further resources

- 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 of visuals: mimotext.github.io/MiMoTextBase_Tutorial/visualizations.html

References

Martin, Mylne, and Frautschi. 1977. *Bibliographie Du Genre Romanesque Français, 1751-1800*. Mandell.

Röttgermann, Julia. 2024. “The Collection of Eighteenth-Century French Novels 1751-1800.” *Journal of Open Humanities Data* 10 (1): 31. https://doi.org/10.5334/johd.201.

Schöch, Christof. 2013. “Big? Smart? Clean? Messy? Data in the Digital Humanities.” *Journal of Digital Humanities* 2 (3): 1–19. https://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the-humanities/.

Schöch, Christof, Maria Hinzmann, Julia Röttgermann, Katharina Dietz, and Anne Klee. 2022. “Smart Modelling for Literary History.” *International Journal of Humanities and Arts Computing* 16 (1): 78–93. https://doi.org/10.3366/ijhac.2022.0278.