

Smart Modeling for Digital Literary History



Christof Schöch, Maria Hinzmann,
Katharina Dietz, Julia Röttgermann, Anne Klee

Smart Data X Digital Humanities (DADH2020), Dec. 1-4, 2020

<https://mimotext.github.io/literary-history/>





Overview

1. Introduction: The MiMoText Project
2. Bibliographies
3. Scholarly Publications
4. Primary Sources
5. Conclusion: Bringing it all together

The MiMoText Project

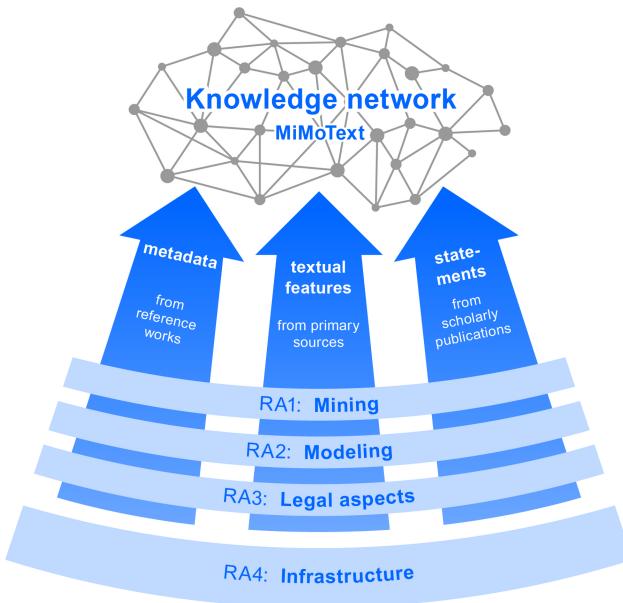
Context: Trier Center for Digital Humanities

The screenshot shows the homepage of the Trier Center for Digital Humanities. At the top, there is a dark navigation bar with links for Deutsch, English, Projects, What we Do, About Us, Publications, a search bar, and links for Partners, News, and Contact us. Below the navigation bar is a large banner featuring a woman in profile wearing a red patterned scarf, with the text "Trier Center for Digital Humanities" and "Kompetenzzentrum für elektronische Erschließungs- und Publikationsverfahren in den Geisteswissenschaften". To the right of the banner is a circular logo for "20 JAHRE TRIER CENTER FOR DIGITAL HUMANITIES". On the far right, there is a vertical bar with the "Universität Trier" logo. Below the banner, the page has a section titled "Projects" with four items:

- Digitising and Electronic Editing of Abraham Gotlob Werner's Correspondence**
Partners: Steering Committee: Professor Helmuth Albrecht (head), Hildegard Wiegel, MA MSt DPhil FSA (transcriptions and scholarly comments), both Institute for Industrial...
- ZHistLex – eHumanities-Centre for Historical Lexikography**
Management: Institut für Germanistik, Universität Gießen/ZMI • Partners: Akademie der Wissenschaften zu Göttingen, Akademie der Wissenschaften und der Literatur Mainz,...
- Lavater: Historical-Critical Edition**
Management: Dr. Ursula Caflisch-Schnetzer (University of Zurich), Prof. Dr. Davide Giurato (University of Zurich) • Partners: University of Zurich: S3iT (NIE-INE / DHLab,...
- Medulla Gestorum Trevrensum**
Partners: Prof. Dr. Michael Embach (Stadtbibliothek Trier), Prof. Dr. Wolfgang Schmid (Universität Trier, Fach Geschichte), Bayerische Staatsbibliothek München

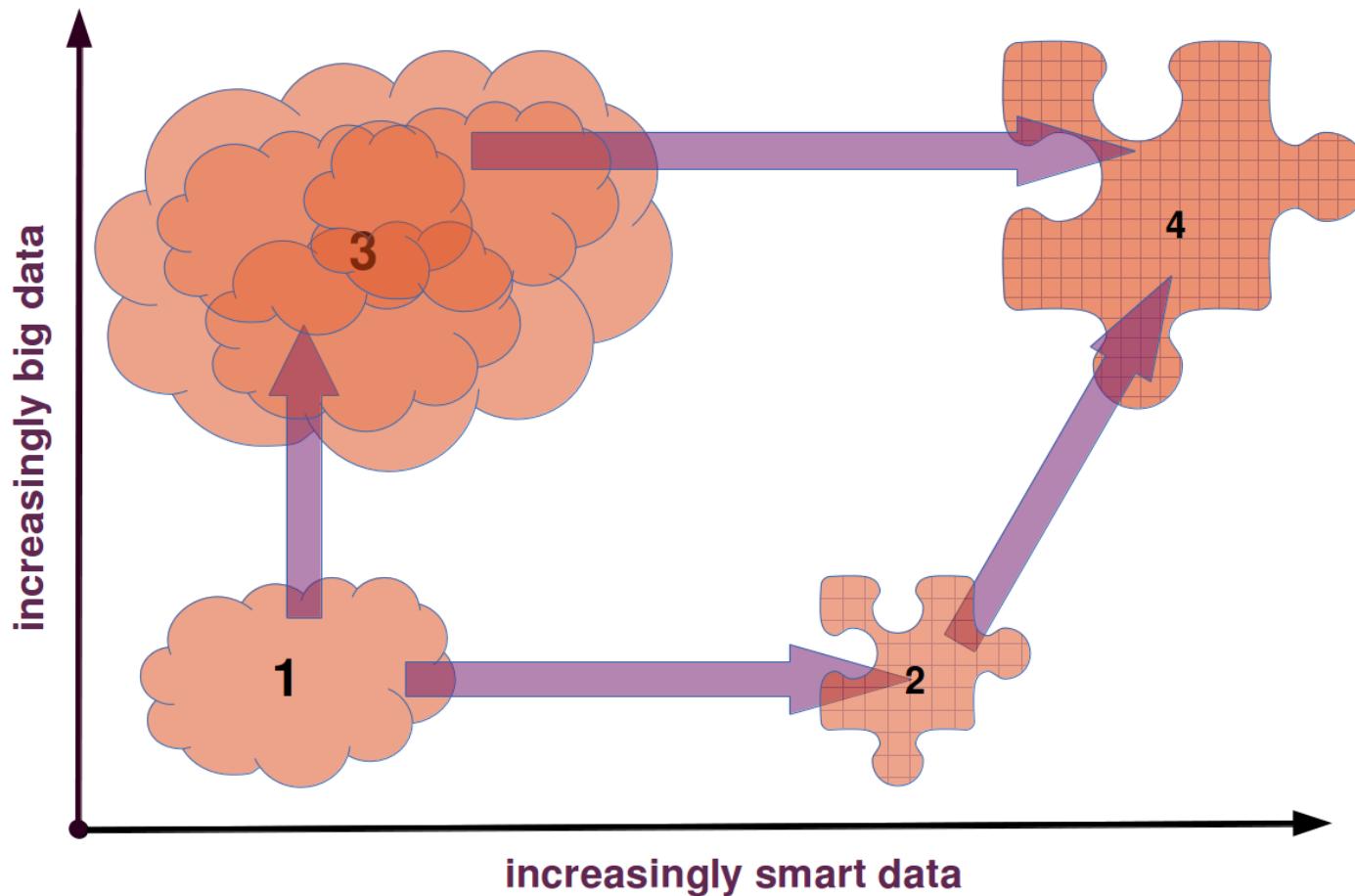
tcdh.uni-trier.de

MiMoText: overview



1. Bibliographies: metadata modeling
2. Primary Texts: analysis results
3. Scholarly Publications: statement extraction
4. Methods of Analysis (Machine Learning)

Smart Data, Smart Modeling



Schöch, "Big? Smart? Clean? Messy? Data in the Humanities" (2013)

(1) Metadata (Bibliographies)

Bibliographie du genre romanesque : Candide

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

Bibliographie modeled as RDF

```
1 <j.2:ListItem rdf:about="http://www.kompetenzzentrum.uni-trier.de/bgrf/9721">
2   <j.5:hasSequenceIdentifier>97.21</j.5:hasSequenceIdentifier>
3   <j.2:itemContent>
4     <j.7:BibliographicRecord rdf:about="http://www.kompetenzzentrum.uni-trier.de/bgrf/9721Record">
5       <j.7:references>
6         <j.4:Manifestation rdf:about="http://www.kompetenzzentrum.uni-trier.de/bgrf/9721Manifestation">
7           <j.4:embodimentOf>
8             <j.4:Expression rdf:about="http://www.kompetenzzentrum.uni-trier.de/bgrf/9721Expression">
9               <j.0:creator rdf:resource="http://www.viaf.org/viaf/54146831"/>
10              <j.0:language rdf:resource="http://id.loc.gov/vocabulary/iso639-2/fre"/>
11              <j.0:creator>DIDEROT, Denis</j.0:creator>
12              <j.0:title>Jacques le fataliste et son maître, par Diderot</j.0:title>
13            </j.4:Expression>
14          </j.4:embodimentOf>
15          <j.5:hasPageCount>xxii + 286, 320p.</j.5:hasPageCount>
16          <j.3:keyword>3e personne, avec dialogues, récits intercalés Ire personne, intervention du
narrateur; Jacques, son maître, personnages rencontrés sur le chemin; voyage, aventures
bouffonnes, galantes, romanesques; thèmes philosophiques, mise en cause des techniques du
romancier.</j.3:keyword>
17          <j.1:P30083>Buisson, an cinquième de la République,</j.1:P30083>
18          <j.1:P30137>Saint-Fargeau Tchemerzine Selon Tchemerzine, certains exemplaires portent: Jacques
la fataliste... Cet ouvrage a paru à la fin de 1796, date que donnent les bibliographies. Nous le
classons ici en raison de la date révolutionnaire que porte la page de titre.</j.1:P30137>
19          <j.1:P30197>2t. in-8,</j.1:P30197>
20          <j.1:P30270>BM BNt JP 8 vend. V Gay Q</j.1:P30270>
21          <j.1:P30088>Paris,</j.1:P30088>
22          </j.4:Manifestation>
23        </j.7:references>
24        <rdf:type rdf:resource="http://purl.org/spar/fabio/BibliographicMetadata"/>
25      </j.7:BibliographicRecord>
26    </j.2:itemContent>
27  </j.2:ListItem>
```

(2) Primary Sources

The "roman18" corpus

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

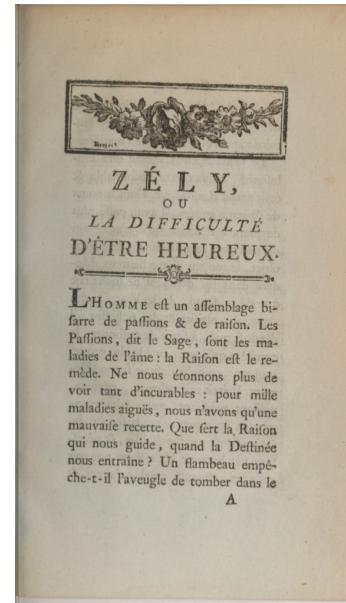
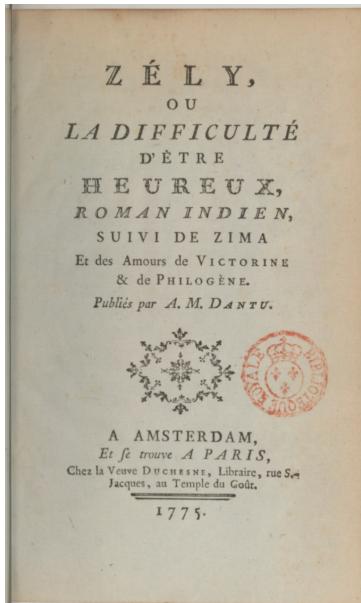
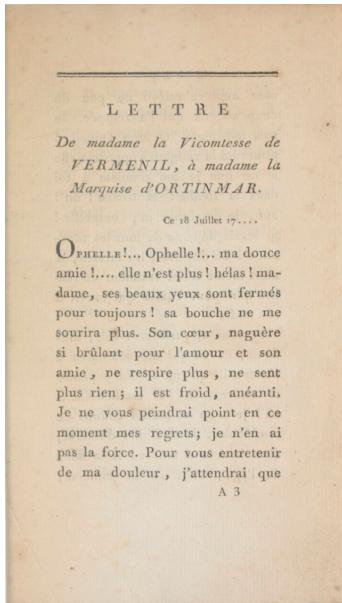
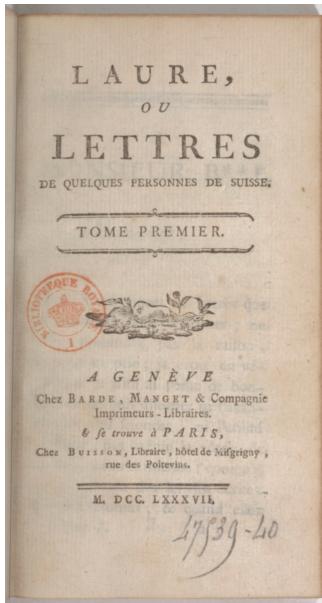


Language	Percentage
HTML	72.3%
Jupyter Notebook	26.6%
Python	1.1%

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

github.com/MiMoText/roman18

The "roman18" corpus: digitization



- Most important source: French National Library (PDF scans)
- Full-text digitization (Double Keying, OCR model)
- Simple encoding in XML-TEI (ELTeC schema)

Topic Modeling: Two major topics

"monarchy"

A word cloud centered around the theme of "monarchy". The most prominent words are "peuple" (people), "roi" (king), "grand" (great), "souverain" (sovereign), and "prince" (prince). Other visible words include "citoyen" (citizen), "gouvernement" (government), "ordre" (order), "patrie" (homeland), "porter" (bear), "rendre" (render), "ennemi" (enemy), "gloire" (glory), "monarque" (monarch), "sujet" (subject), "nation", "tendre" (tender), "troupe", and "empire". Smaller words like "point", "trône", and "gouvernement" are also present.

"love"

A word cloud centered around the theme of "love". The most prominent words are "coeur" (heart), "amour" (love), and "bonheur" (happiness). Other visible words include "âme" (soul), "voir" (see), "doux" (soft), "heureux" (happy), "sentir" (feel), "tendre" (tender), "temp", "jou", "aimer", "seul", "objet", "vie", "point", "jamais", "ami", and "sentiment". Smaller words like "jour" and "temp" are also present.

- Derived example statements
 - Candide ABOUT "monarchy"
 - Clarice ABOUT "love"

(3) Scholarly Publications

Statement Extraction

1. Named Entity Recognition (automatically)
2. Statement Annotation (manually, to create training data; shown here)
3. Statement Extraction (using Machine Learning, based on training data)

Named Entity Recognition

Candide [WORK] is Voltaire's [PERS] most widely read work, and was probably already during the author's [PERS] lifetime.

When it [WORK] first appeared in print in Geneva [LOC] in 1759 [TIME], it was immediately banned, but only with the result that it was reprinted thirteen times in the same year.

(Köhler)

- Recognition results (spaCy, F1-score)
 - PERS: 0.864-0.917
 - LOC: 0.829-0.837
 - WORK: 0.451-0.497

Statement Annotation

The screenshot shows a digital annotation interface. At the top, there are navigation controls (back, forward, search) and a page number (424). Below this is a text snippet in German. To the right, a search results panel is displayed, containing three entries for 'Candide' from Wikidata:

- [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

Below the search results, it says "50 items found". A search bar at the bottom contains the word "candide".

Der Auflösung aller sittlichen Bande in den höheren Schichten der Gesellschaft wird die Innigkeit des Familien-Alternative entgegengestellt.

Jean-François Marmontel

Wenn Marmontels Helden am Ende Lache und Glück finden, schreiben sie es gewöhnlich weniger einem v Schicksal als ihrem bürgerlichen Fleiß, ihrer Tüchtigkeit und Tugend zu danken.

Named entity → predicate → Named entity
"Bélisaire" → "Roman"

In seinem antikisierenden Roman

Bélisaire (1767) schließlich, der in den Augen der Zeitgenossen den

← (predicate)
hasAuthor → François Fénelon

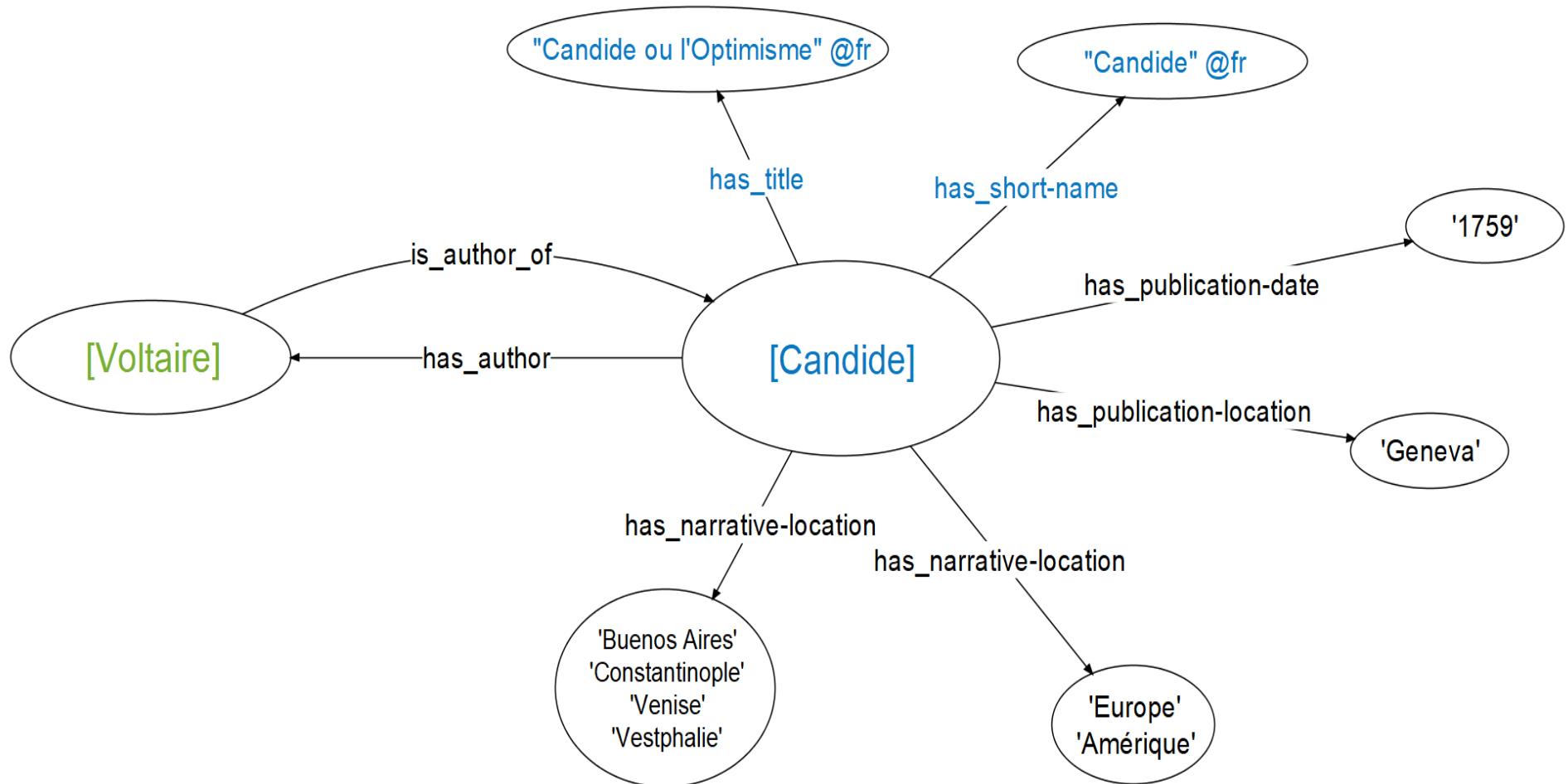
Fénelons übertrifft, wird der Leser aus der Privatsphäre herausgeführt und mit öffentlichen Bel Fragen konfrontiert.

Bringing it all together

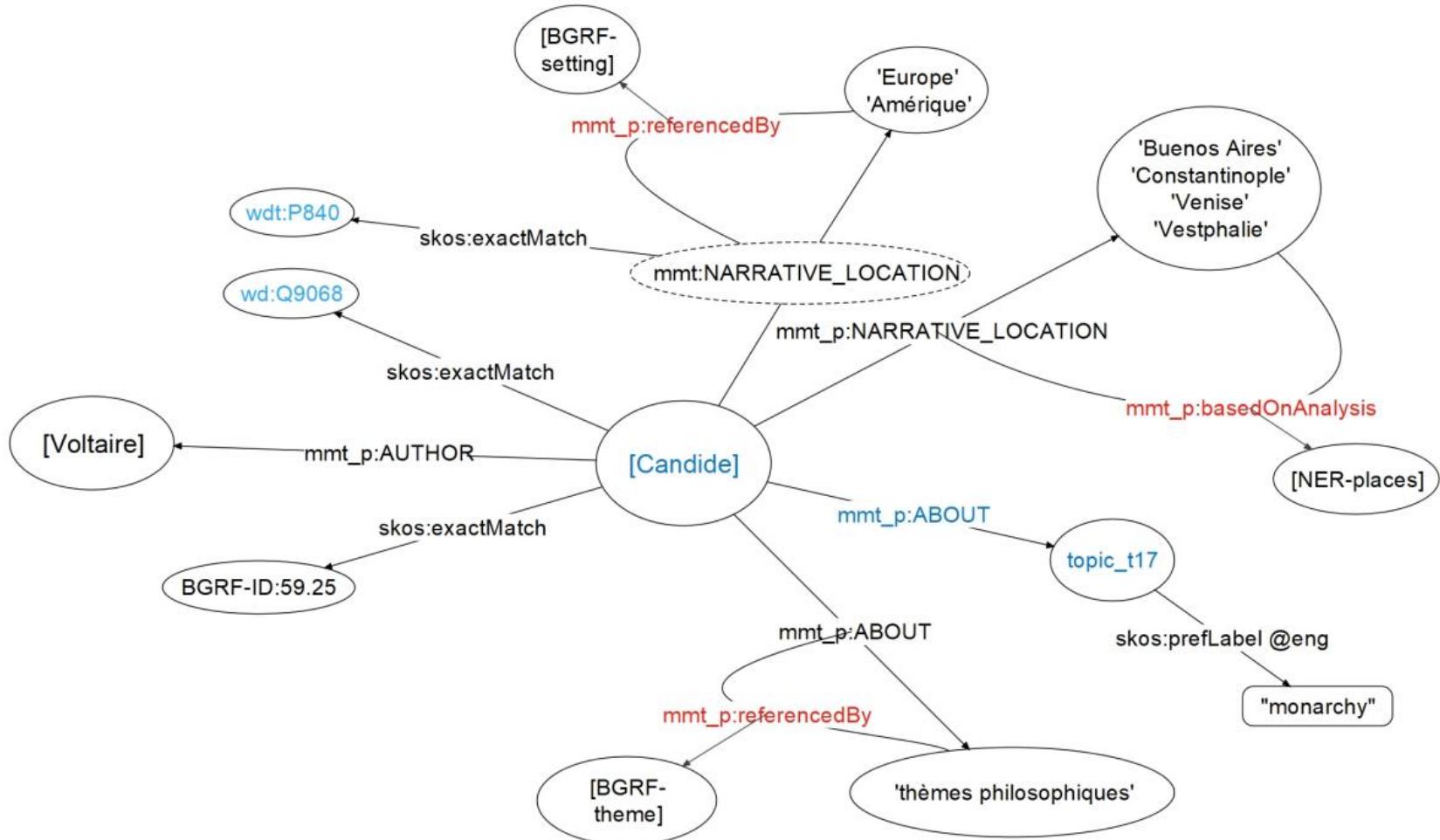
Modeling Challenges

- Of all the statements we could extract, which are the most relevant for literary history?
- How can we create consensus around the underlying domain ontology?
- How do we specifically model the information we extract?
- How can we best use the possibilities of LOD in our context?
- For example, how do we make the sources of information explicit?

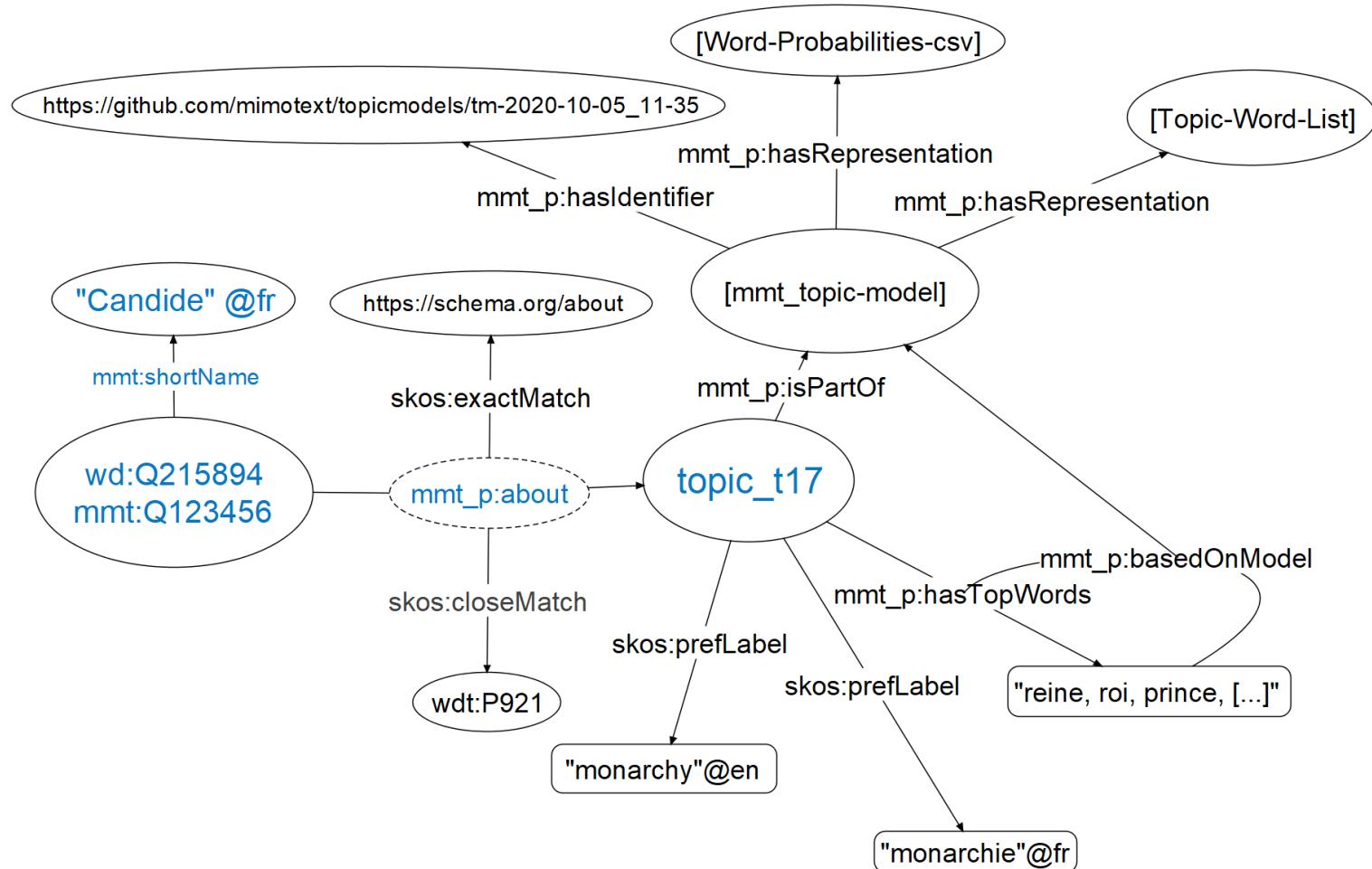
LOD for *Candide*: metadata



LOD for *Candide*: NARRATIVE_LOCATION



LOD for *Candide*: ABOUT



Thank you!

Questions or comments?

slides: mimotext.github.io/literary-history/

project: mimotext.uni-trier.de/