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

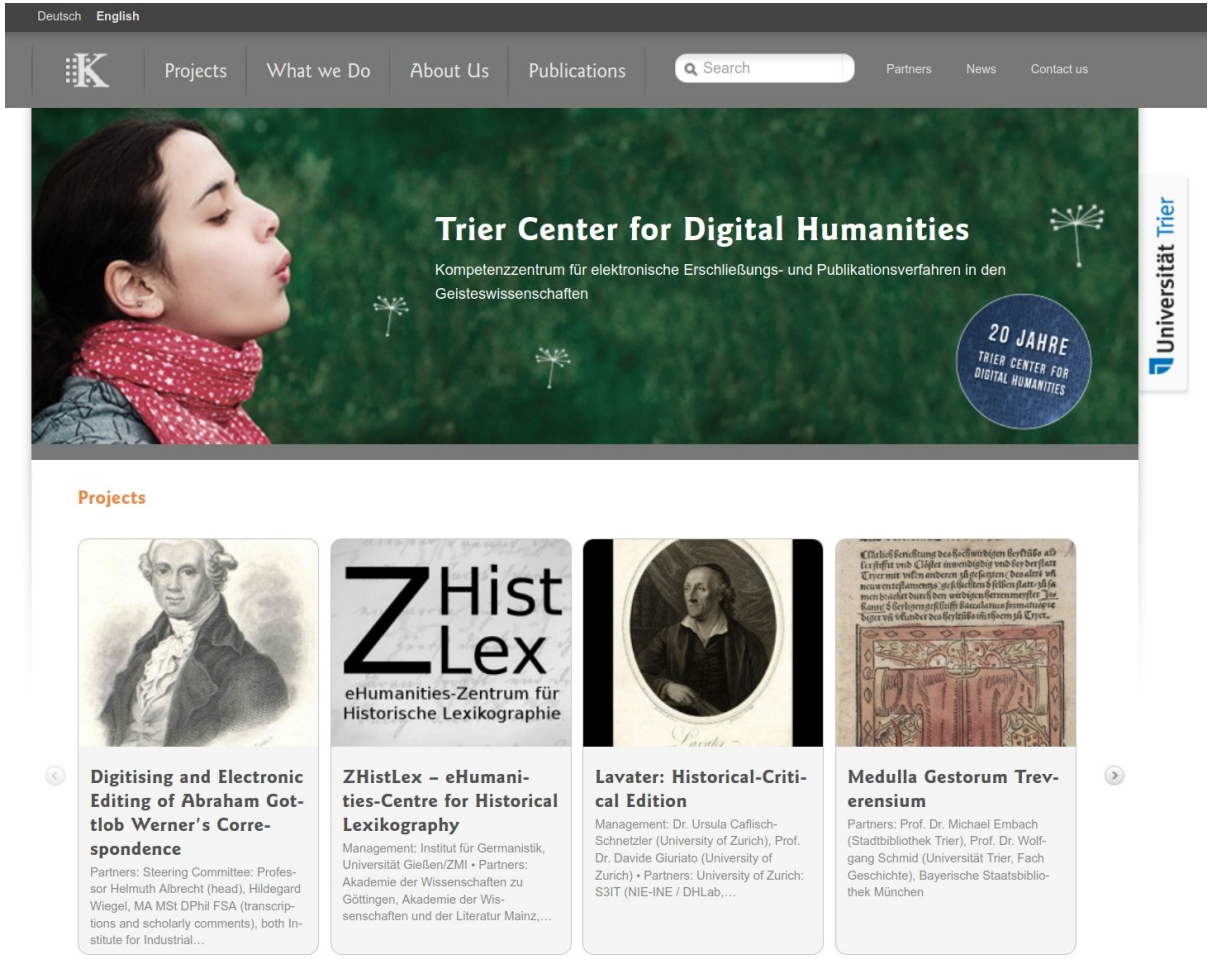
https://mimotext.github.io/literary-history/

Overview

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

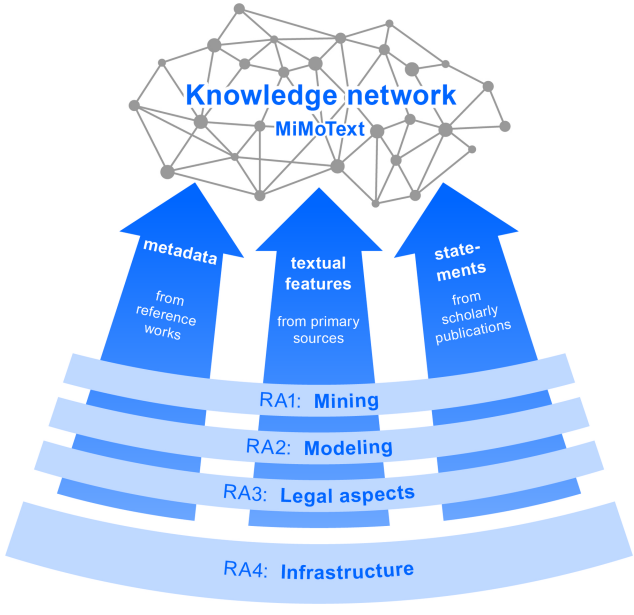
The MiMoText Project

Context: Trier Center for Digital Humanities



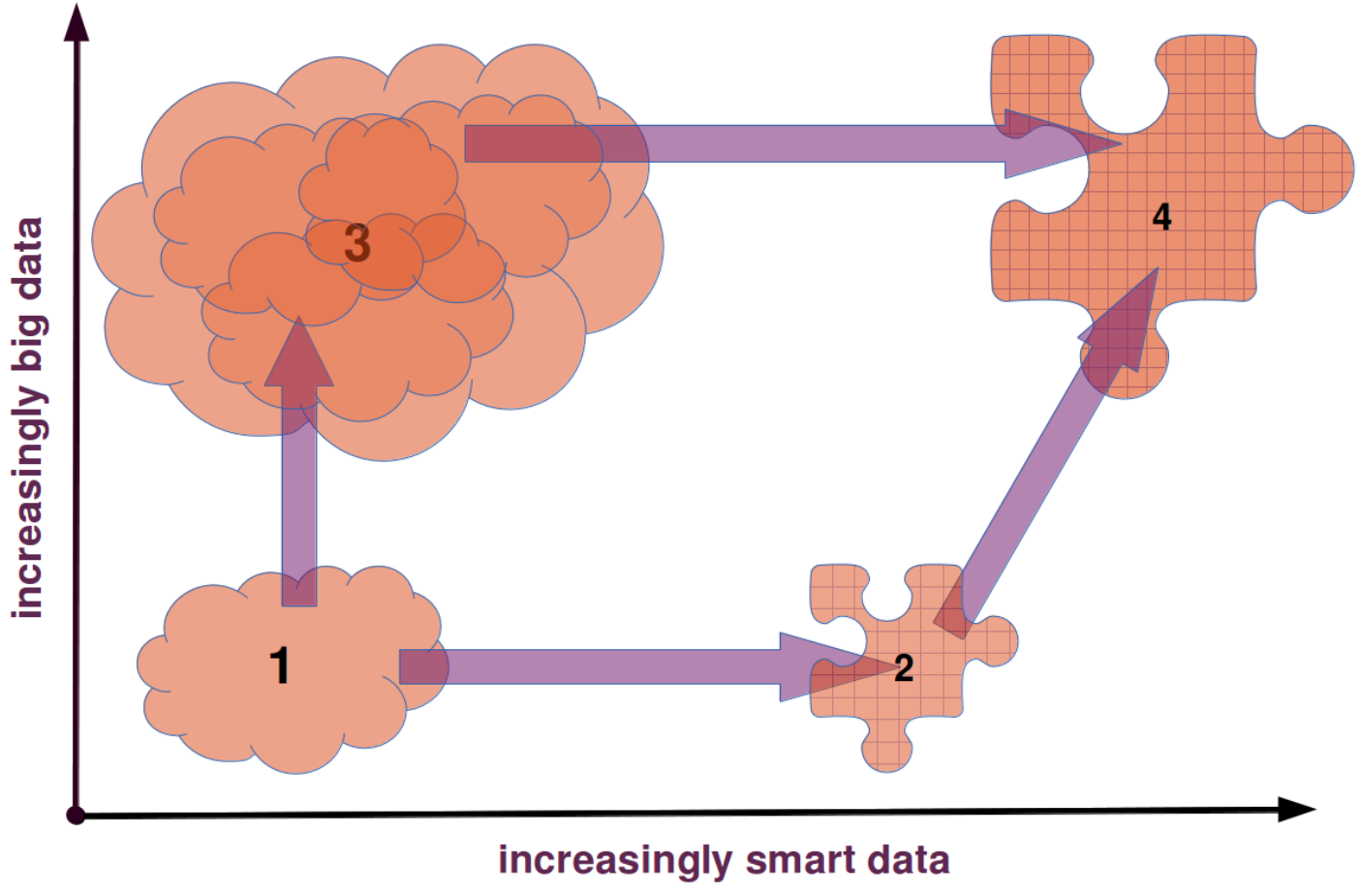
tcdh.uni-trier.de

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)

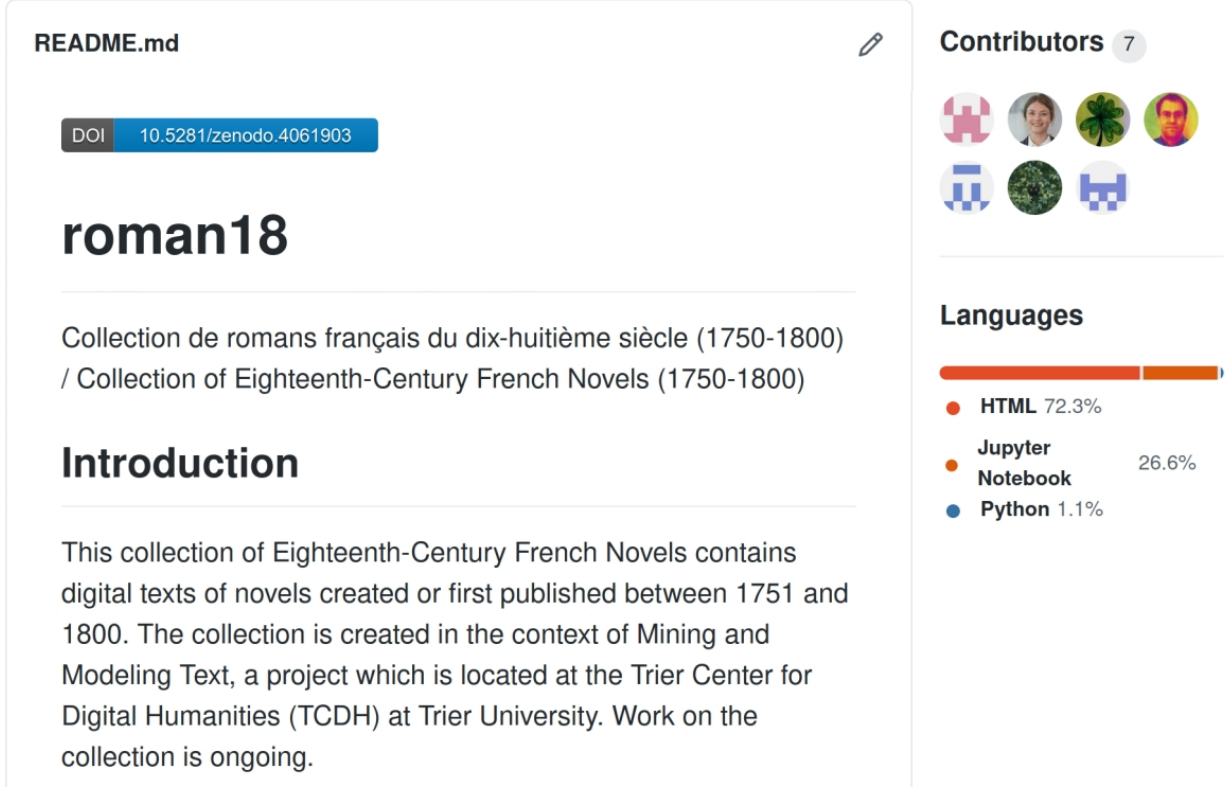
Martin / Mylne / Frautschi: *Bibliographie du genre romanesque français, 1751-1800*, 1977

Bibliographie modeled as RDF

```
<j.2:ListItem rdf:about="http://www.kompetenzzentrum.uni-trier.de/bgrf/9721">
  <j.5:hasSequenceIdentifier>97.21</j.5:hasSequenceIdentifier>
  <j.2:itemContent>
    <j.7:BibliographicRecord rdf:about="http://www.kompetenzzentrum.uni-trier.de/bgrf/9721Record">
      <j.7:references>
        <j.4:Manifestation rdf:about="http://www.kompetenzzentrum.uni-trier.de/bgrf/9721Manifestation">
          <j.4:embodimentOf>
            <j.4:Expression rdf:about="http://www.kompetenzzentrum.uni-trier.de/bgrf/9721Expression">
              <j.0:creator rdf:resource="http://www.viaf.org/viaf/54146831"/>
              <j.0:language rdf:resource="http://id.loc.gov/vocabulary/iso639-2/fre"/>
              <j.0:creator>DIDEROT, Denis</j.0:creator>
              <j.0:title>Jacques le fataliste et son maître, par Diderot</j.0:title>
            </j.4:Expression>
          </j.4:embodimentOf>
          <j.5:hasPageCount>xxii + 286, 320p.</j.5:hasPageCount>
          <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>
          <j.1:P30083>Buisson, an cinquième de la République,</j.1:P30083>
          <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 bibliographes. Nous le
          classons ici en raison de la date révolutionnaire que porte la page de titre.</j.1:P30137>
          <j.1:P30197>2t. in-8,</j.1:P30197>
          <j.1:P30270>BM BNt JP 8 vend. V Gay Q</j.1:P30270>
          <j.1:P30088>Paris,</j.1:P30088>
        </j.4:Manifestation>
      </j.7:references>
      <rdf:type rdf:resource="http://purl.org/spar/fabio/BibliographicMetadata"/>
    </j.7:BibliographicRecord>
  </j.2:itemContent>
</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

HTML 72.3%

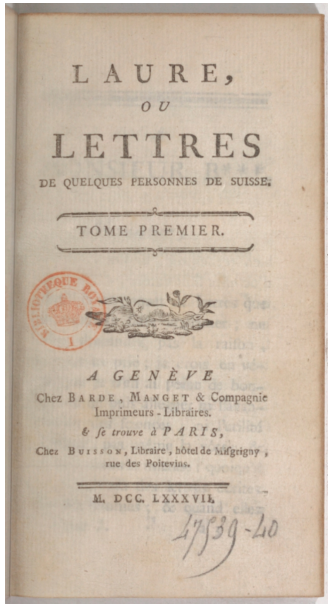
Jupyter Notebook 26.6%

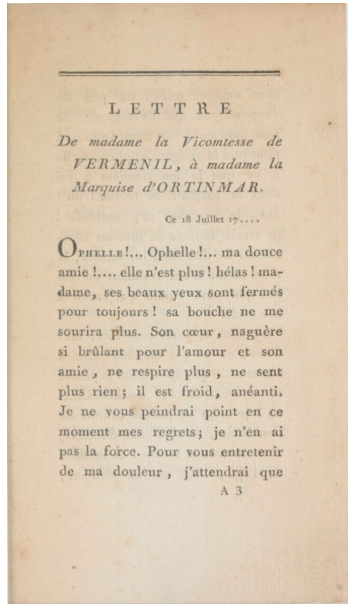
Python 1.1%

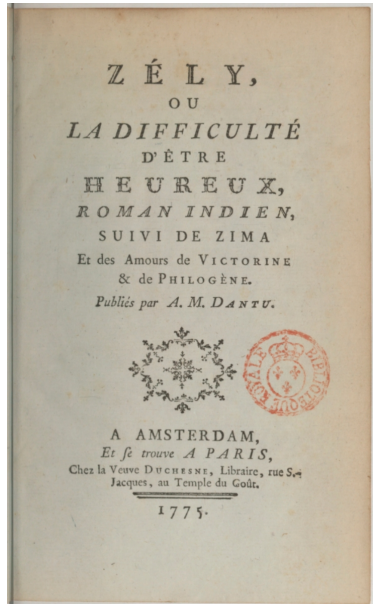
Collection of Eighteenth-Century French Novels (1750-1800)

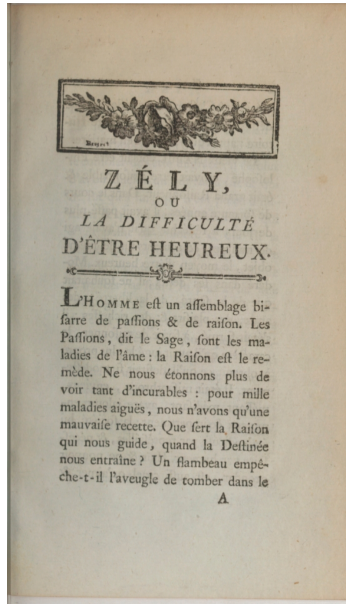
github.com/MiMoText/roman18

The "roman18" corpus: digitization


LAURE,
OU
LETTRES
DE QUELQUES PERSONNES DE SUISSE.
TOME PREMIER.
A GENÈVE
Chez BARDE, MANGET & Compagnie
Imprimeurs-Libraires.
& se trouve à PARIS,
Chez BUISSON, Libraire, hôtel de Mésgrigny, rue des Poitevins.
M. DCC. LXXXVII.


LETTRE
De madame la Vicomtesse de VERMENIL, à madame la Marquise d'ORTINMAR.
Ce 18 Juillet 17....
OPHELLE!... Ophelle!... ma douce amie!.... elle n'est plus! hélas! madame, ses beaux yeux sont fermés pour toujours! sa bouche ne me sourira plus. Son cœur, naguère si brûlant pour l'amour et son amie, ne respire plus, ne sent plus rien; il est froid, anéanti. Je ne vous peindrai point en ce moment mes regrets; je n'en ai pas la force. Pour vous entretenir de ma douleur, j'attendrai que
A 3


ZÉLY,
OU
LA DIFFICULTÉ
D'ÊTRE
HEUREUX,
ROMAN INDIEN,
SUIVI DE ZIMA
Et des Amours de VICTORINE
& de PHILOGÈNE.
Publiés par A. M. DANTU.
A AMSTERDAM,
Et se trouve A PARIS,
Chez la Veuve DUCHESNE, Libraire, rue S. Jacques, au Temple du Goût.
1775.


ZÉLY,
OU
LA DIFFICULTÉ
D'ÊTRE HEUREUX.
L'HOMME est un assemblage bizarre de passions & de raison. Les Passions, dit le Sage, sont les maladies de l'âme: la Raison est le remède. Ne nous étonnons plus de voir tant d'incurables: pour mille maladies aiguës, nous n'avons qu'une mauvaise recette. Que sert la Raison qui nous guide, quand la Destinée nous entraîne? Un flambeau empêche-t-il l'aveugle de tomber dans le
A

- 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"

"love"

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

(3) Scholarly Publications

Statement Extraction

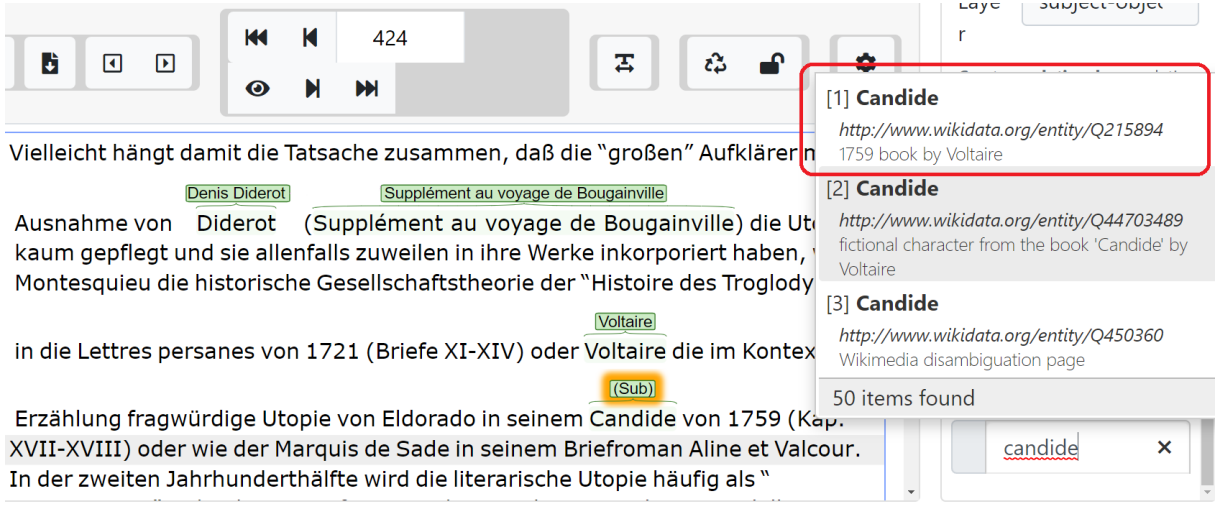
1. Named Entity Recognition (automatically)
2. Statement Annotation (manually, to create training data; shown here)
3. Statement Extraction (using Machine Learning, based on training data)

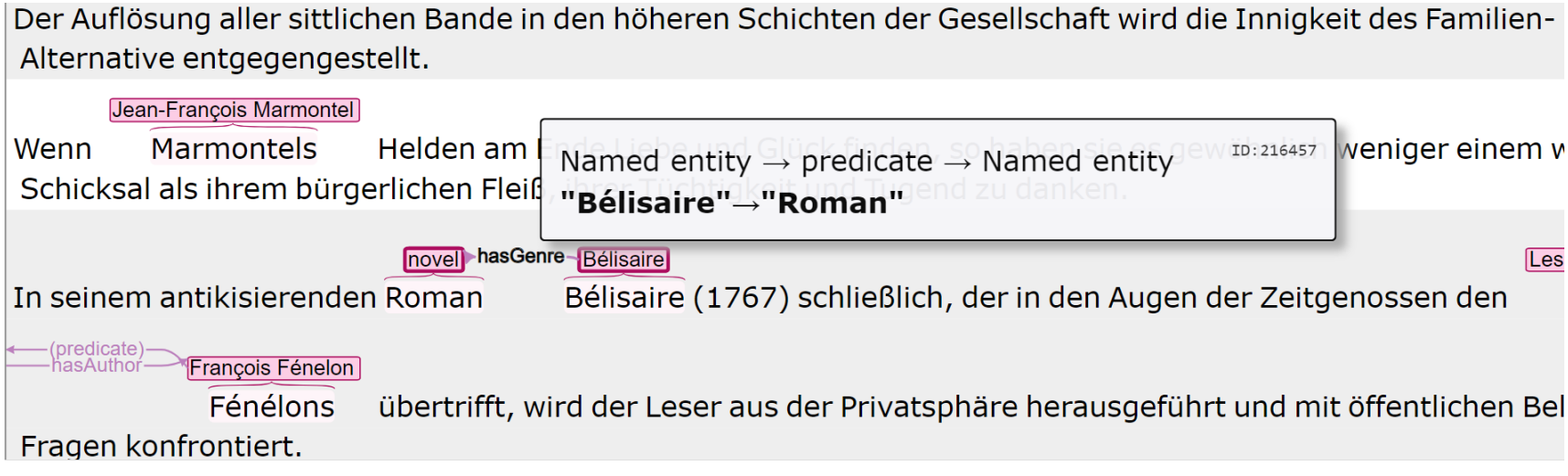
Named Entity Recognition

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



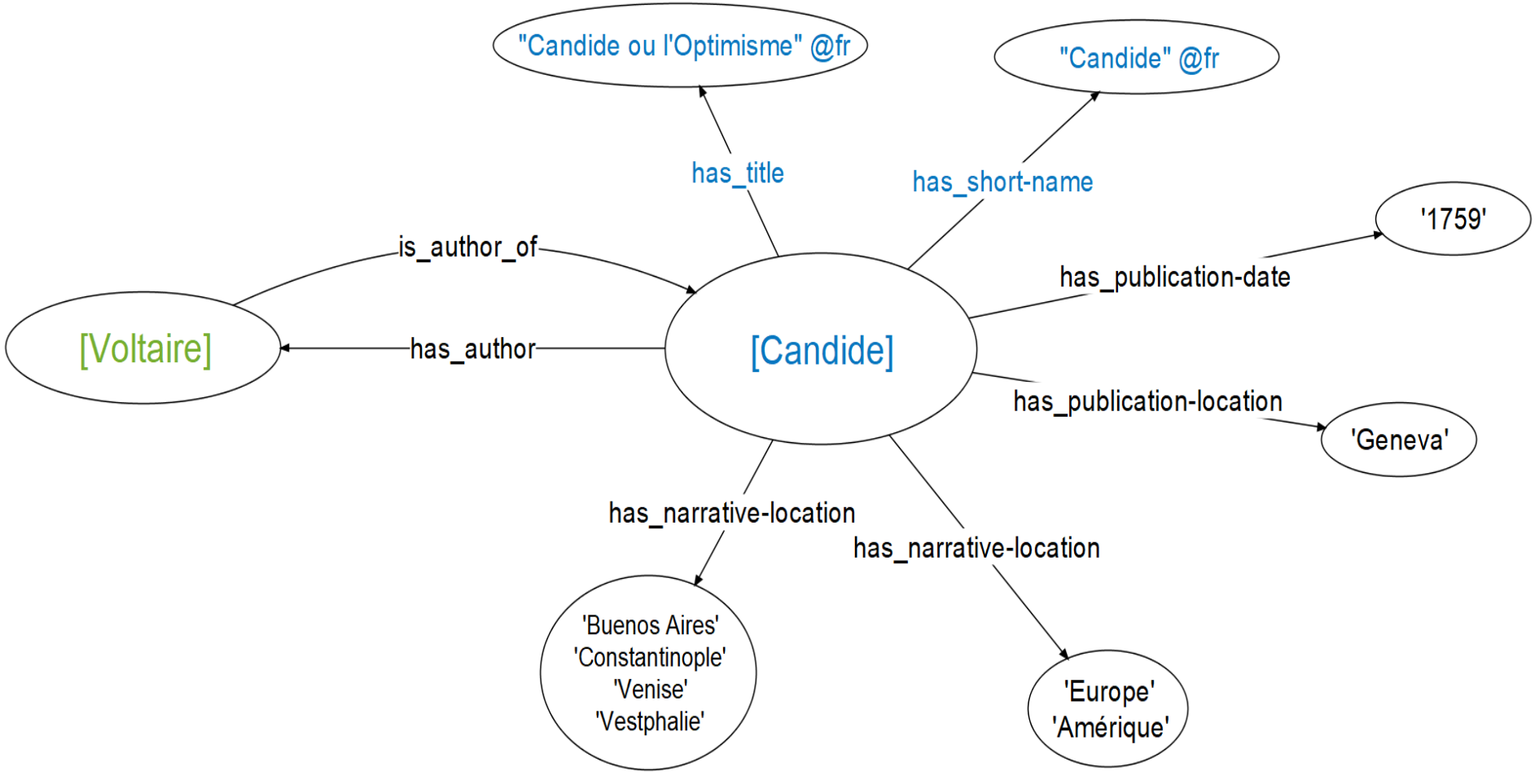


Bringing it all together

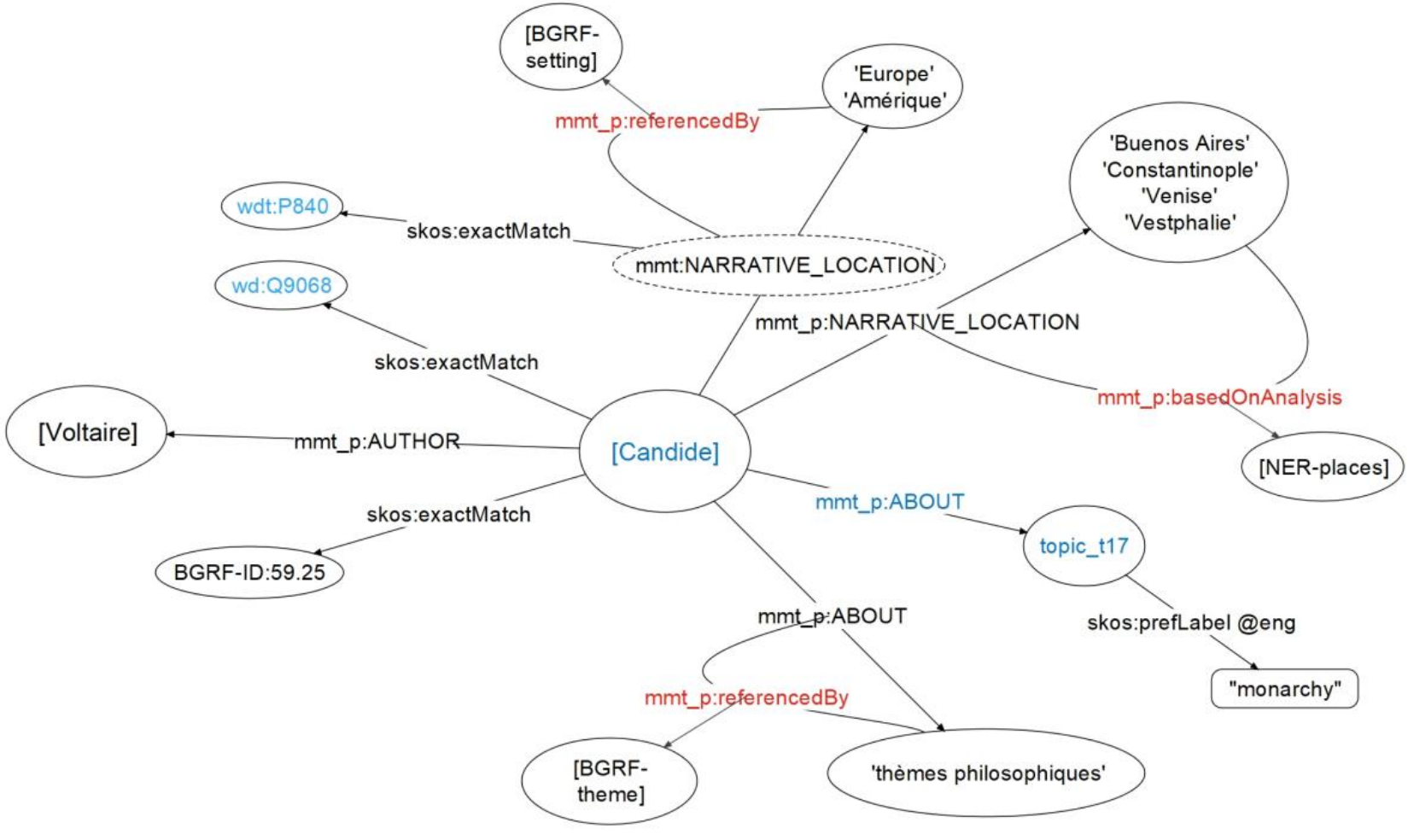
Modeling Challenges

- 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?

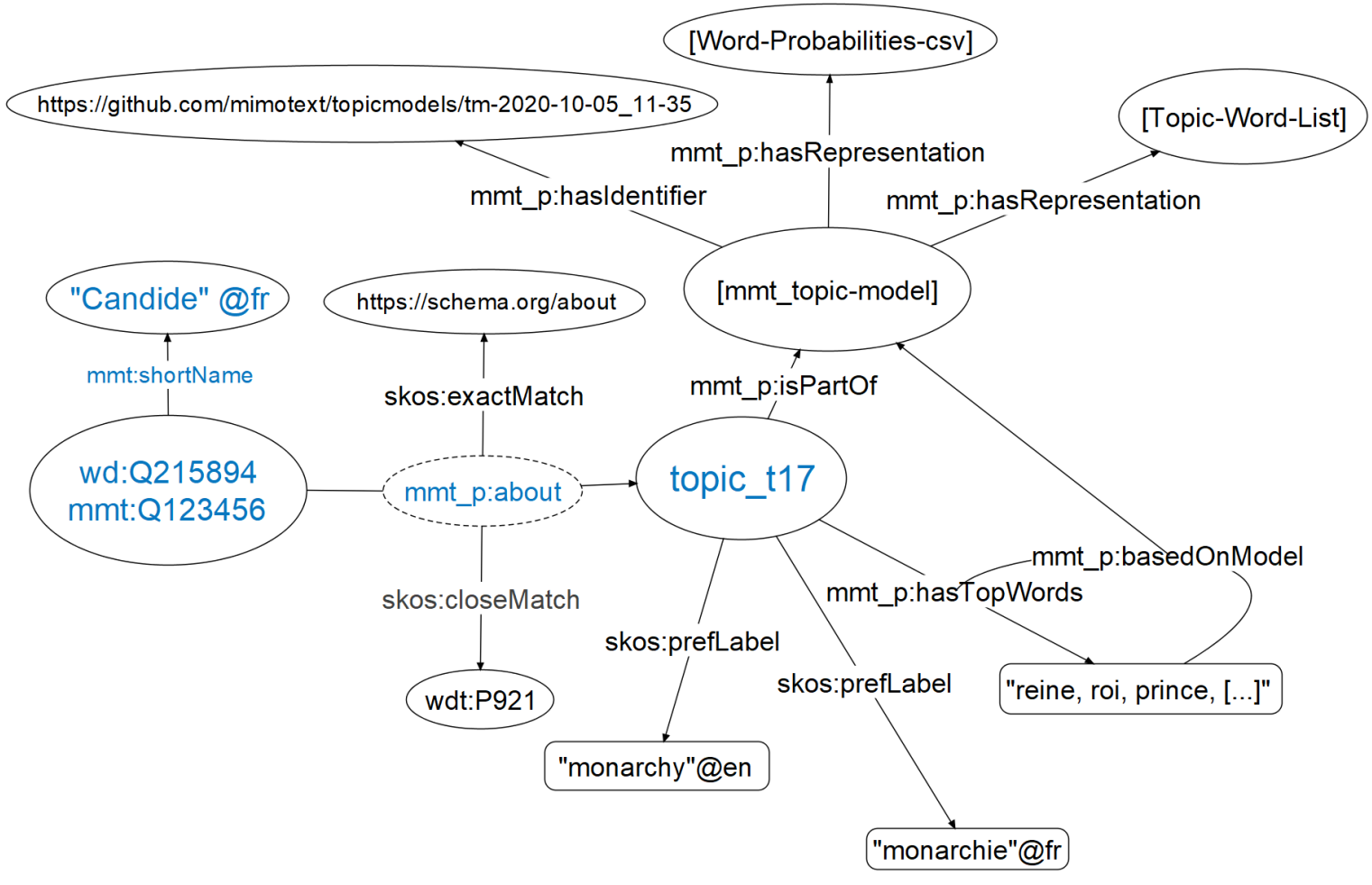
LOD for *Candide*: metadata



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/