How Could Digital Literary Historiography Work? Some Lessons Learned in the MiMoText Project

Christof Schöch, with contributions from Maria Hinzmann, Julia Röttgermann, Katharina Dietz, Andreas Lüschow and Anne Klee

DHLab@GS, Univ. of Texas at Austin | October 26, 2020
https://mimotext.github.io/literary-history/

Overview

1. Introduction: Literary Historiography
2. Metadata (Bibliographies)
3. Data (Corpora)
4. Model (Ontology)
5. Methods (Machine Learning)
    1. Unsupervised: Topic Modeling
    2. Supervised: Statement Extraction
6. Conclusion

Literary Historiography

How Does Literary Historiography Work?

How Does Literary Historiography Work?

- Goals
  - Recover and document the facts of literary history
  - Provide explanations for the evolution of literature

How Does Literary Historiography Work?

- Goals
  - Recover and document the facts of literary history
  - Provide explanations for the evolution of literature
- Sources
  - literary works (and other primary sources)
  - scholarly publications; above all: literary histories

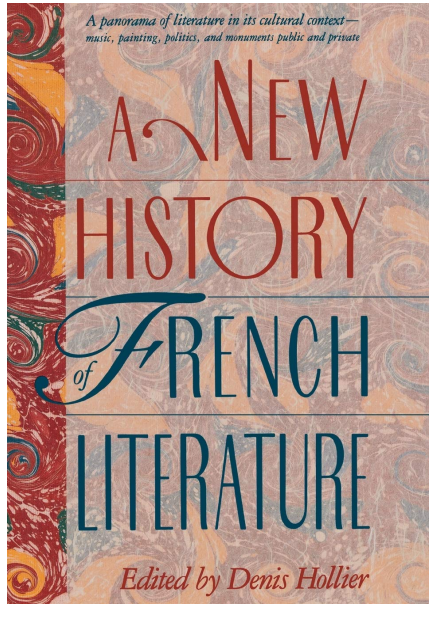
How Does Literary Historiography Work?

- Goals
  - Recover and document the facts of literary history
  - Provide explanations for the evolution of literature
- Sources
  - literary works (and other primary sources)
  - scholarly publications; above all: literary histories
- Organizing principles
  - countries, periods, movements, genres
  - similarities and dissimilarities
  - continuity and change

How Does Literary Historiography Work?

- Goals
  - Recover and document the facts of literary history
  - Provide explanations for the evolution of literature
- Sources
  - literary works (and other primary sources)
  - scholarly publications; above all: literary histories
- Organizing principles
  - countries, periods, movements, genres
  - similarities and dissimilarities
  - continuity and change
- Explanation of literary evolution
  - Changing cultural and socio-historical context
  - Inner dynamics of the literary system

(Theory of) Literary Historiography



- Jan-Dirk Müller, "Literaturgeschichte / Literaturgeschichtsschreibung" (1982)
- Claus Uhlig, "Current Models and Theories of Literary Historiography" (1987)
- David Hollier, *A New History of French Literature* (1989)
- David Perkins, *Is Literary History Possible?* (1992)
- Jan Borkowski and Philipp David Heine, "Ziele der Literaturgeschichtsschreibung" (2013)

How can a *Digital* Literary Historiography Work?

How can a *Digital* Literary Historiography Work?

- Data-driven literary history
  - Reconstruct factoids about literature
  - Reconstruct statements from literary historiography

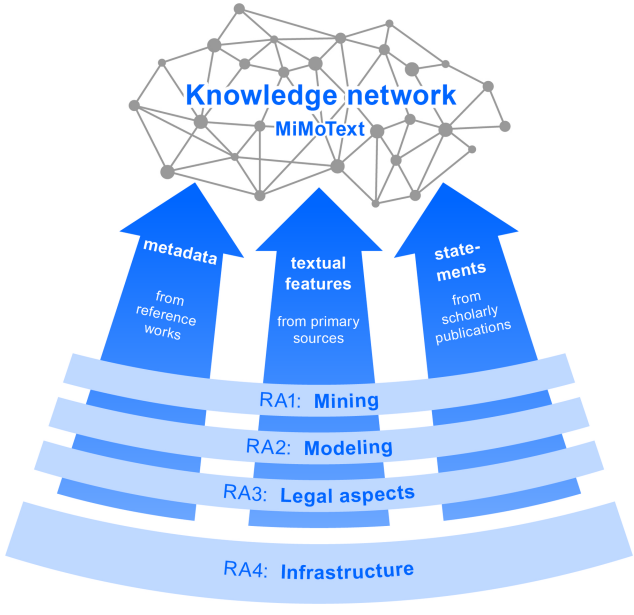
How can a *Digital* Literary Historiography Work?

- Data-driven literary history
  - Reconstruct factoids about literature
  - Reconstruct statements from literary historiography
- Coherence through modeling
  - Join all statements in a shared data model
  - Coherence as an emergent property

How can a *Digital* Literary Historiography Work?

- Data-driven literary history
  - Reconstruct factoids about literature
  - Reconstruct statements from literary historiography
- Coherence through modeling
  - Join all statements in a shared data model
  - Coherence as an emergent property
- Two inspirations
  - Hollier: fragmentation pushed to the limit
  - Moretti: Distant Reading (original sense)

MiMoText: overview



1. Metadata (bibliographies)
2. Data (corpora)
3. Data Model (ontology of literary history)
4. Methods of Analysis (Machine Learning)

(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>
```

What's in the bibliography?

What's in the bibliography?

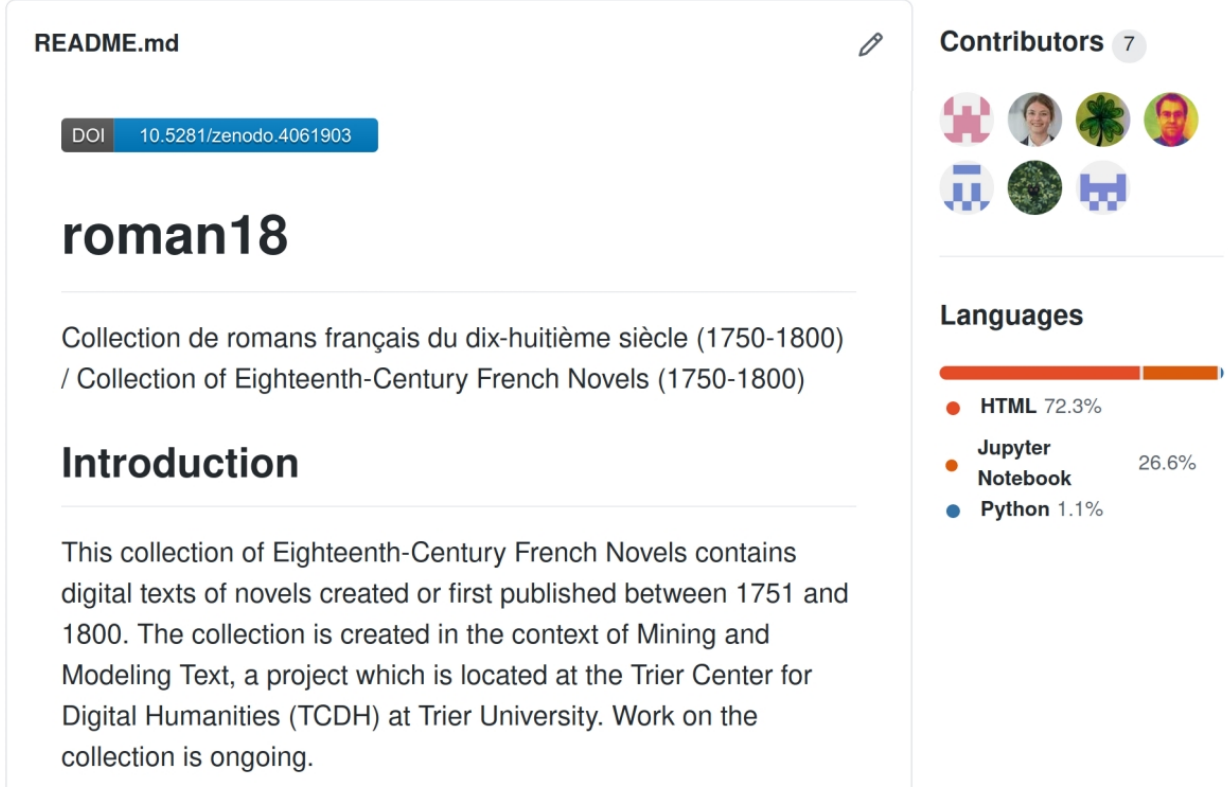
- Summary statistics
  - ~1100 different authors
  - ~2600 entries (novels)
  - ~58.000 triples (22/novel)

What's in the bibliography?

- Summary statistics
  - ~1100 different authors
  - ~2600 entries (novels)
  - ~58.000 triples (22/novel)
- Further information
  - ~720 novels in first person
  - ~920 in third person
  - 2210 entries with notes on content

(2) Data (Corpora)

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

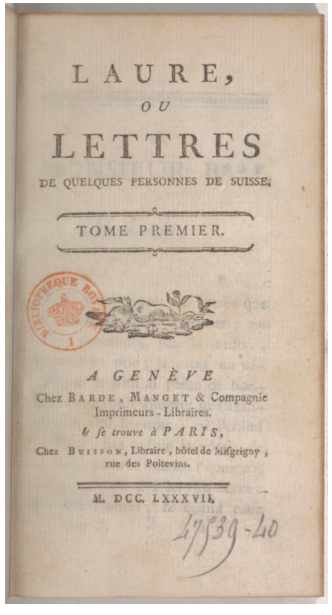
HTML 72.3%

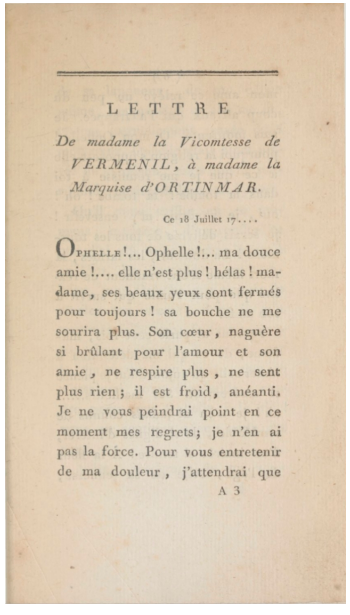
Jupyter Notebook 26.6%

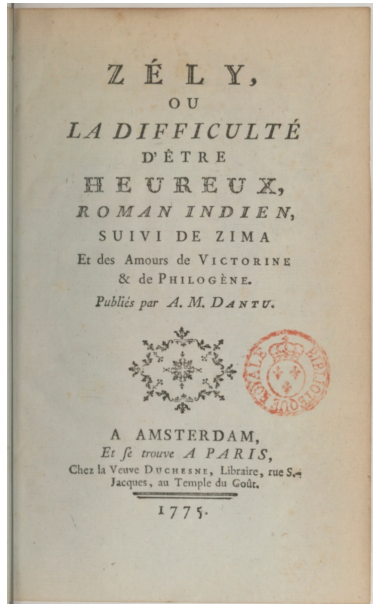
Python 1.1%

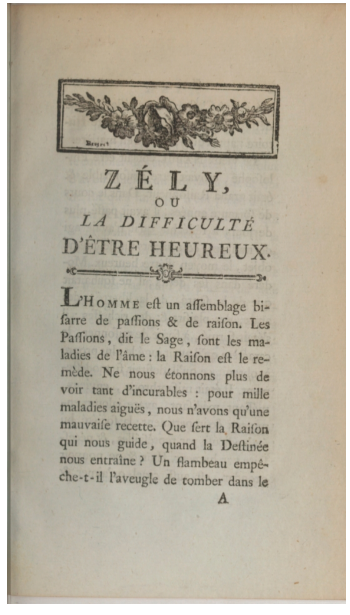
Collection de romans français du dix-huitième siècle (1750-1800) / Collection of Eighteenth-Century French Novels (1750-1800)

Sources


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 Mesgrigny, 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

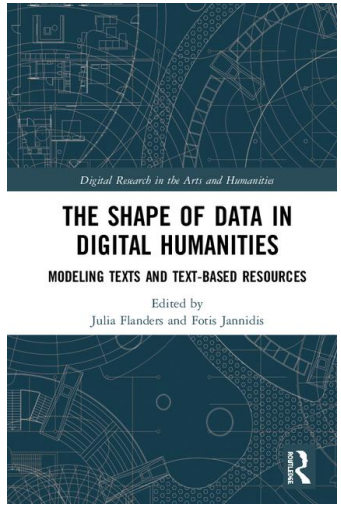
- Various platforms for full texts
- Double-keying based on facsimiles
- Specifically-trained OCR model (OCR4all)
- => Need to unify the text formats

Text Encoding (XML-TEI)

- Simple encoding in XML-TEI (ELTeC schema)
- Metadata and basic text structure
- Modernization: only automatically / "on-the-fly"

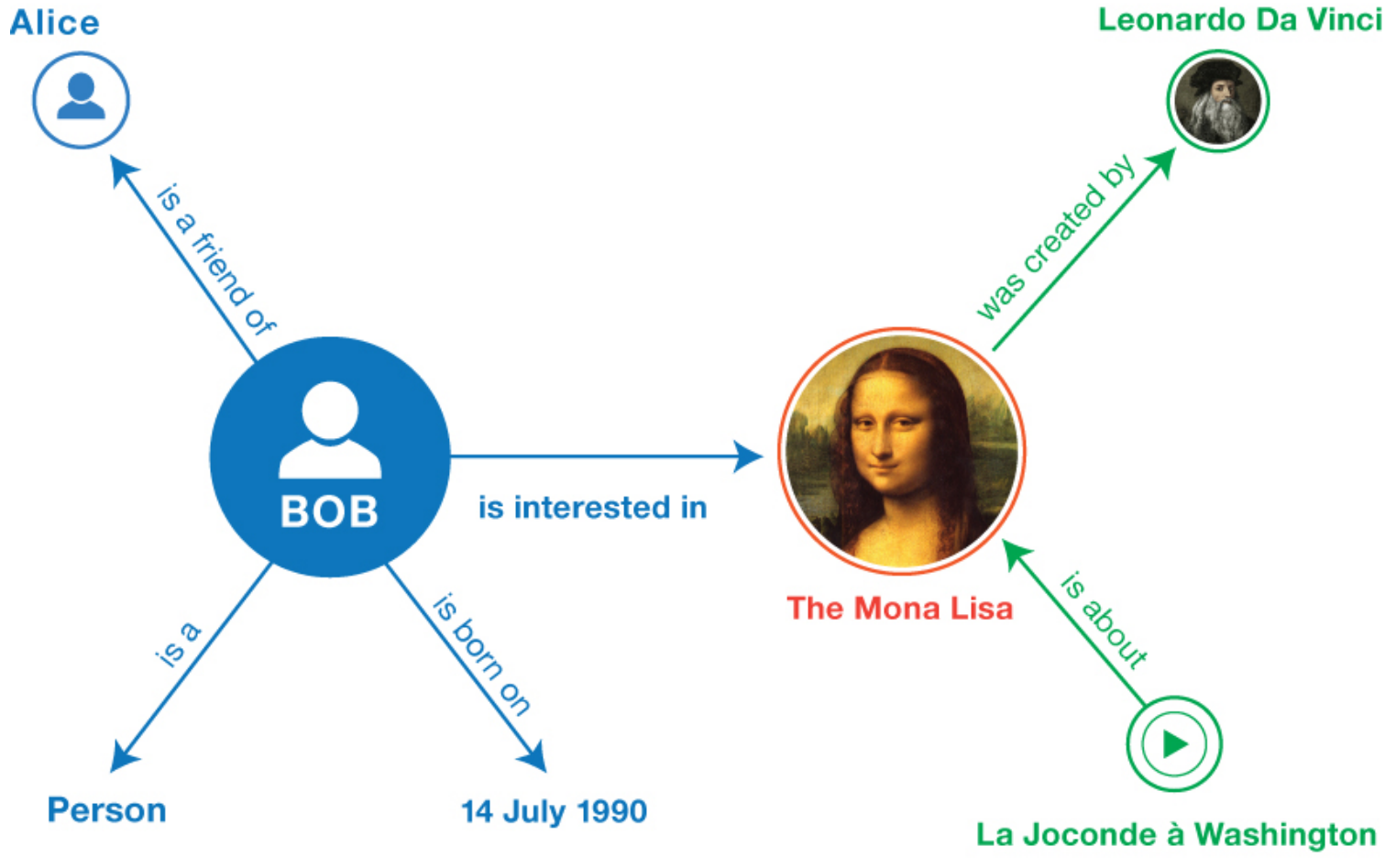
(3) Models (Ontology)

Jannidis & Flanders, *The Shape of Data in DH*, 2019



"The term 'data modeling' in computer science is most typically used in a fairly restrictive sense for the modeling of relational databases, while the digital humanities has a more general understanding of the term: data modeling is the [digital representation] of some segment of the world in such a way to make some aspects computable."

Design decision: Linked Open Data



What are relevant statements?

- More precisely: what are the relevant entities and predicates
- Beyond bibliographical metadata
- Which types of statements are necessary? Which might be useful?
- How do we determine / create scholarly consensus around this?
- Meta-perspective on disciplinary discourse

Fundamental subjects

- Person (role: author, publisher, historical figure)
- Publication (type: primary source; scholarly publication)

Fundamental statements (1)

- person AUTHOR_OF publication
- publication PLACE_OF_PUBLICATION place
- publication DATE_OF_PUBLICATION year
- publication ABOUT keyword

Fundamental statements (2)

- publication NARRATIVE_LOCATION place
- publication NARRATIVE_TIME time period
- publication EXTENT_WORDS number of words
- publication EXTENT_CHAPTERS number of chapters

Fundamental statements (3)

- person/publ. DESCRIBED_AS (adjective)
- person/publ. MEMBER_OF_GROUP movement
- publication INSTANCE_OF_GENRE literary genre
- person/publ. SIMILAR_TO person/publ.
- person/publ. DISSIMILAR_TO person/publ.
- person/publ. INFLUENCED_BY person/publ.

(4a) Unsupervised Methods: Topic Modeling (Primary Sources)

Topic Modeling

Topics

gene	0.04
dna	0.02
genetic	0.01
...	

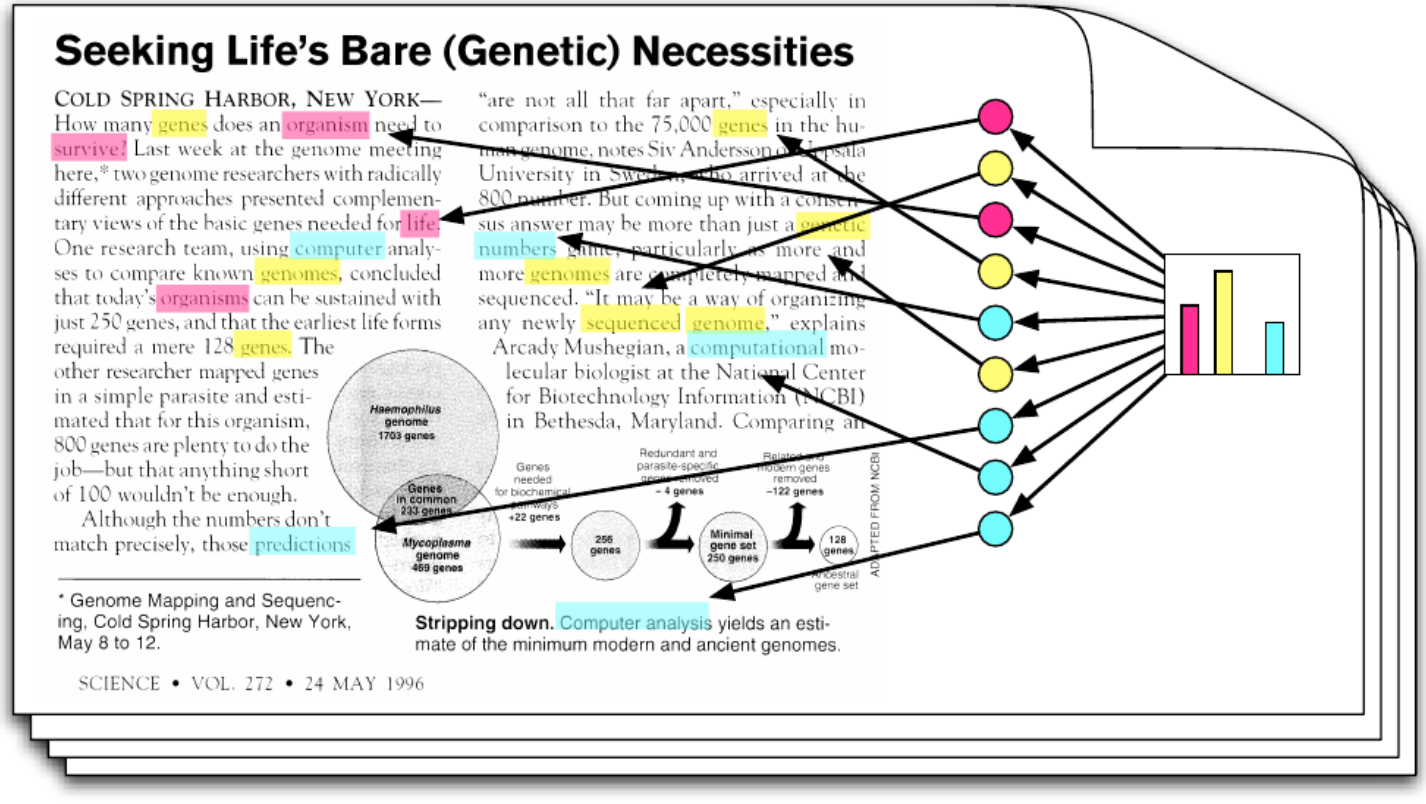
life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
...	

Documents

Topic proportions and assignments



Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Two major topics

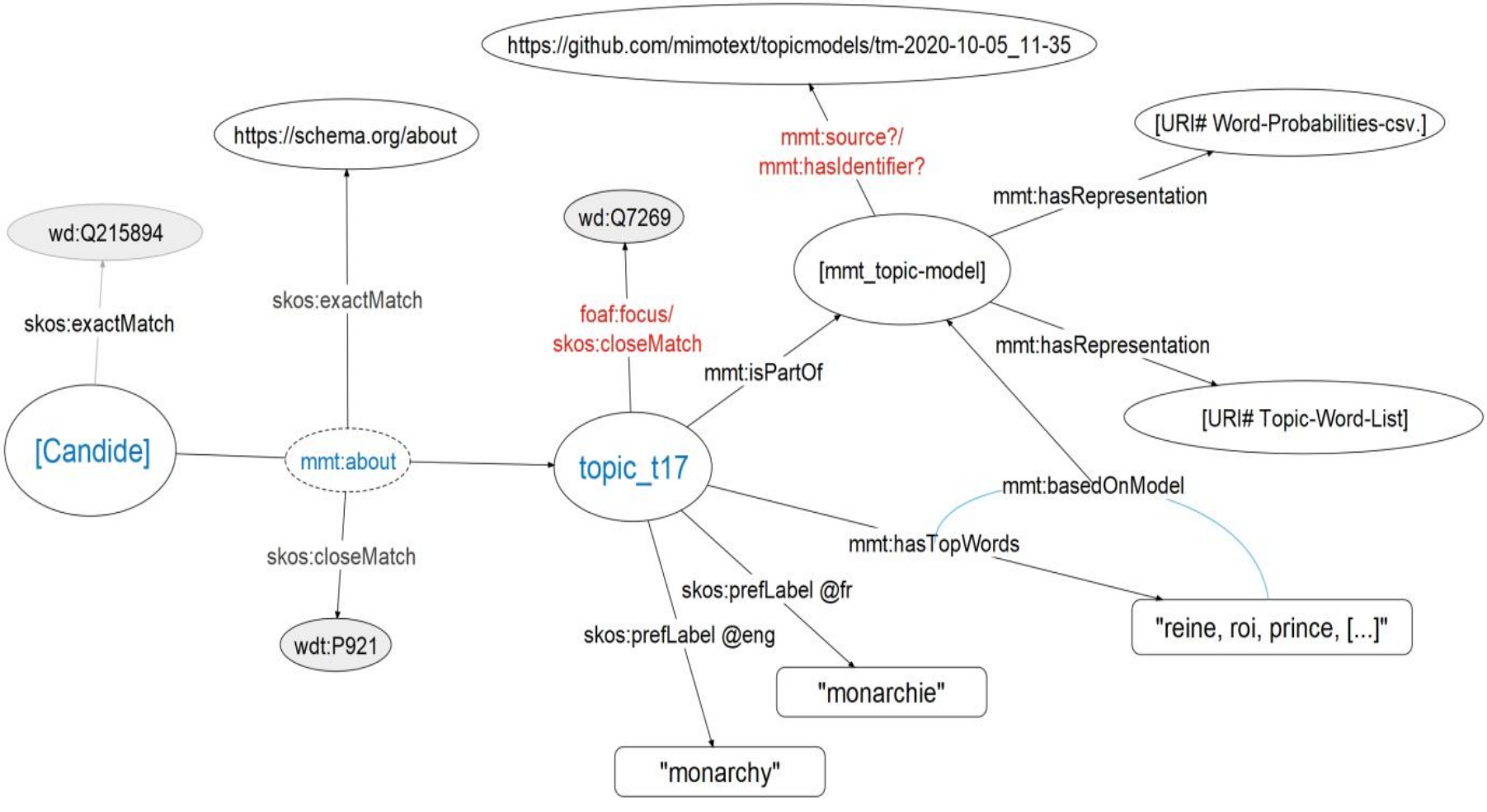
"monarchy"

"love"

Derived statements

- Candide MAIN_SUBJECT "monarchy"
- Clarice MAIN_SUBJECT "love"

Topic Modeling as LOD

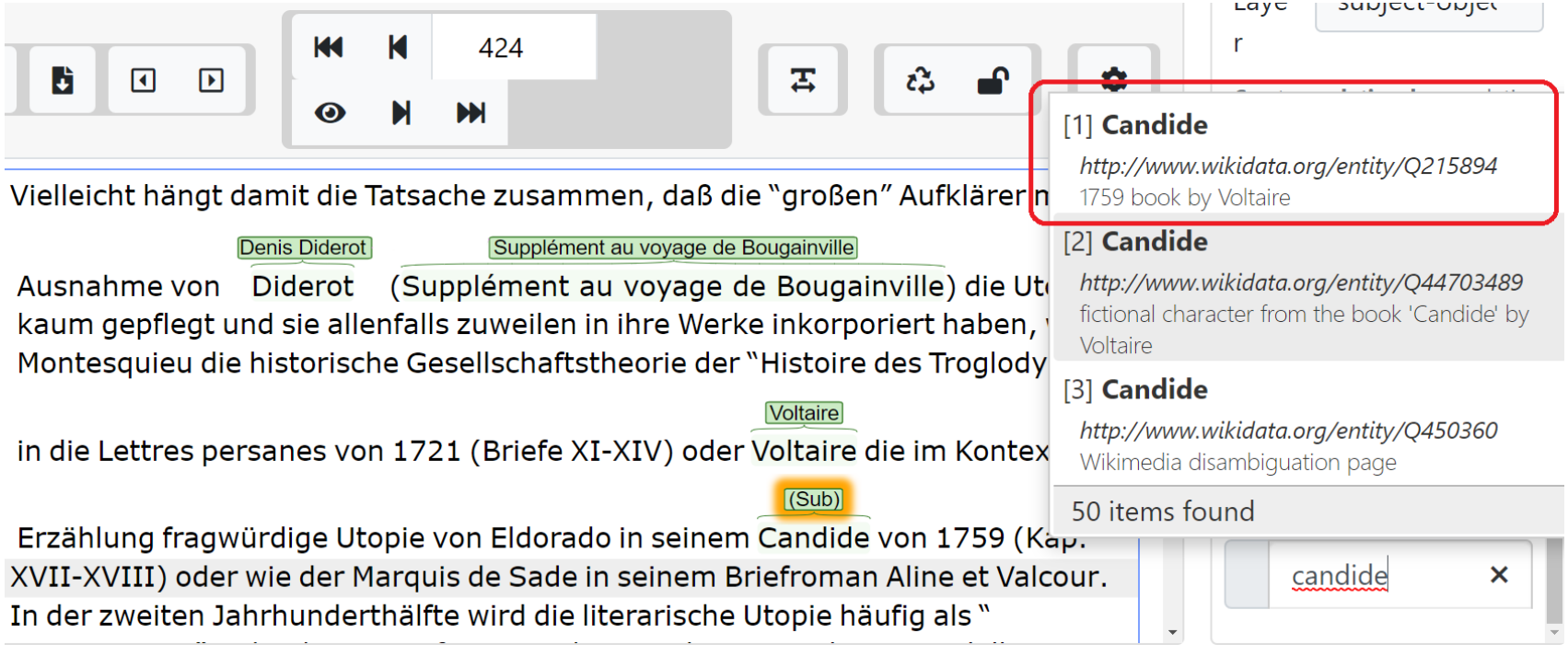


(4b) Supervised Machine Learning: Statement Extraction (Scholarly Publications)

Basic setup for statement extraction

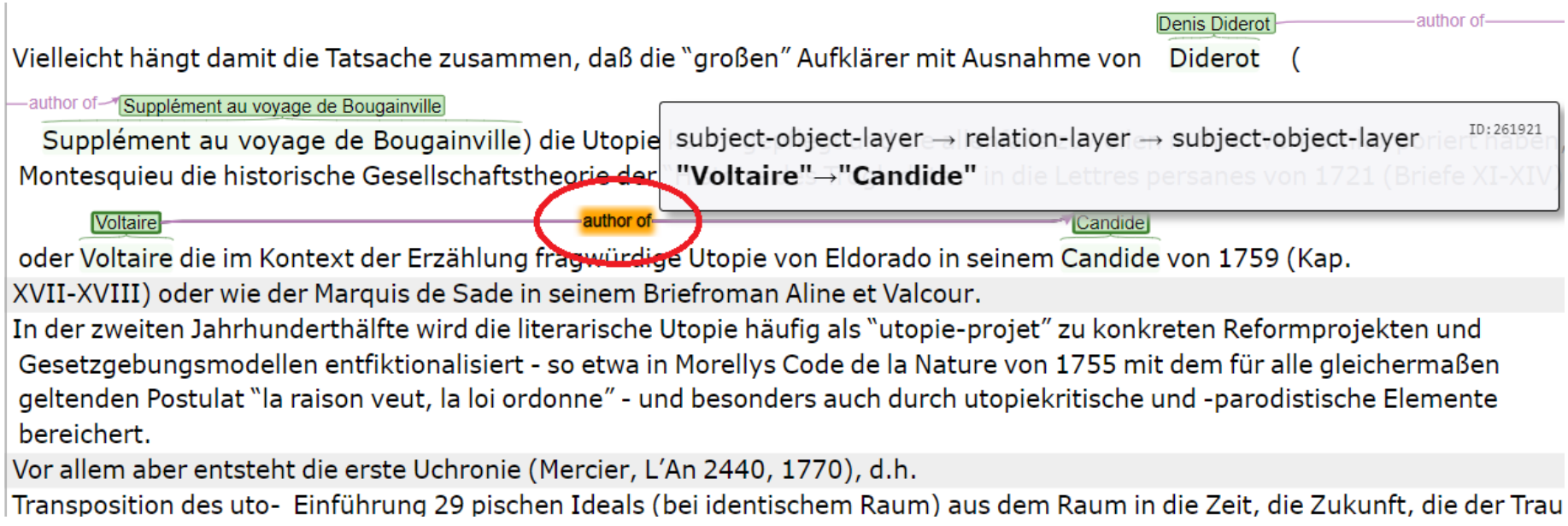
1. Create annotation guidelines (based on data model)
2. Perform manual annotation to create training data (in INCePTION)
3. Train and apply Machine Learning (in Python)

Annotate subject-object layer



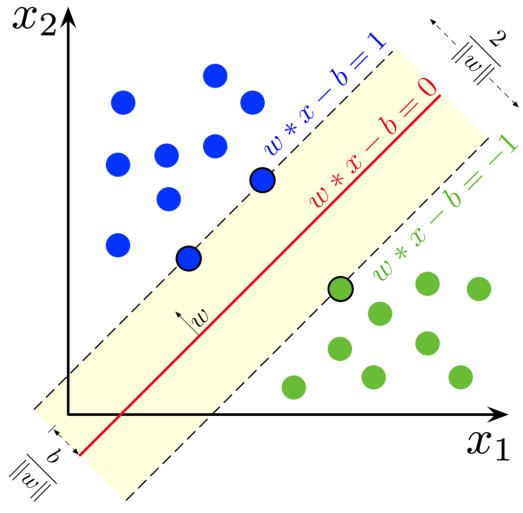
- here: named entities: authors, works
- disambiguation of entities via Wikidata IDs

Annotate relation layer



- here: "author_of" relation (Wikidata: inverse of P50)
- Statements/ LOD triples: 'author AUTHOR_OF work'
- Training data: sentences + statements

Machine Learning



- Material: sentences automatically annotated for named entities
- Further linguistic annotation (feature engineering)
- Provide manual annotations of sentences (training and evaluation)
- Learn patterns / probabilities for features indicative of a relation
- Generate relation annotations for all sentences

Image source: https://commons.wikimedia.org/wiki/File:SVM_margin.png (CC BY)

Example from literary historiography

Candide is Voltaire's most widely read work and was probably already during the author's lifetime. When it first appeared in print in Geneva in 1759, it was immediately banned, but only with the result that it was reprinted thirteen times in the same year. (Erich Köhler, Aufklärung II, 1984; translation: DeepL)

Bibliographic statements

Bibliographic statements

- Voltaire AUTHOR_OF Candide

Bibliographic statements

- Voltaire AUTHOR_OF Candide
- Candide PUBLICATION_DATE 1759

Bibliographic statements

- Voltaire AUTHOR_OF Candide
- Candide PUBLICATION_DATE 1759
- Candide PUBLICATION_LOCATION Geneva

More statements

More statements

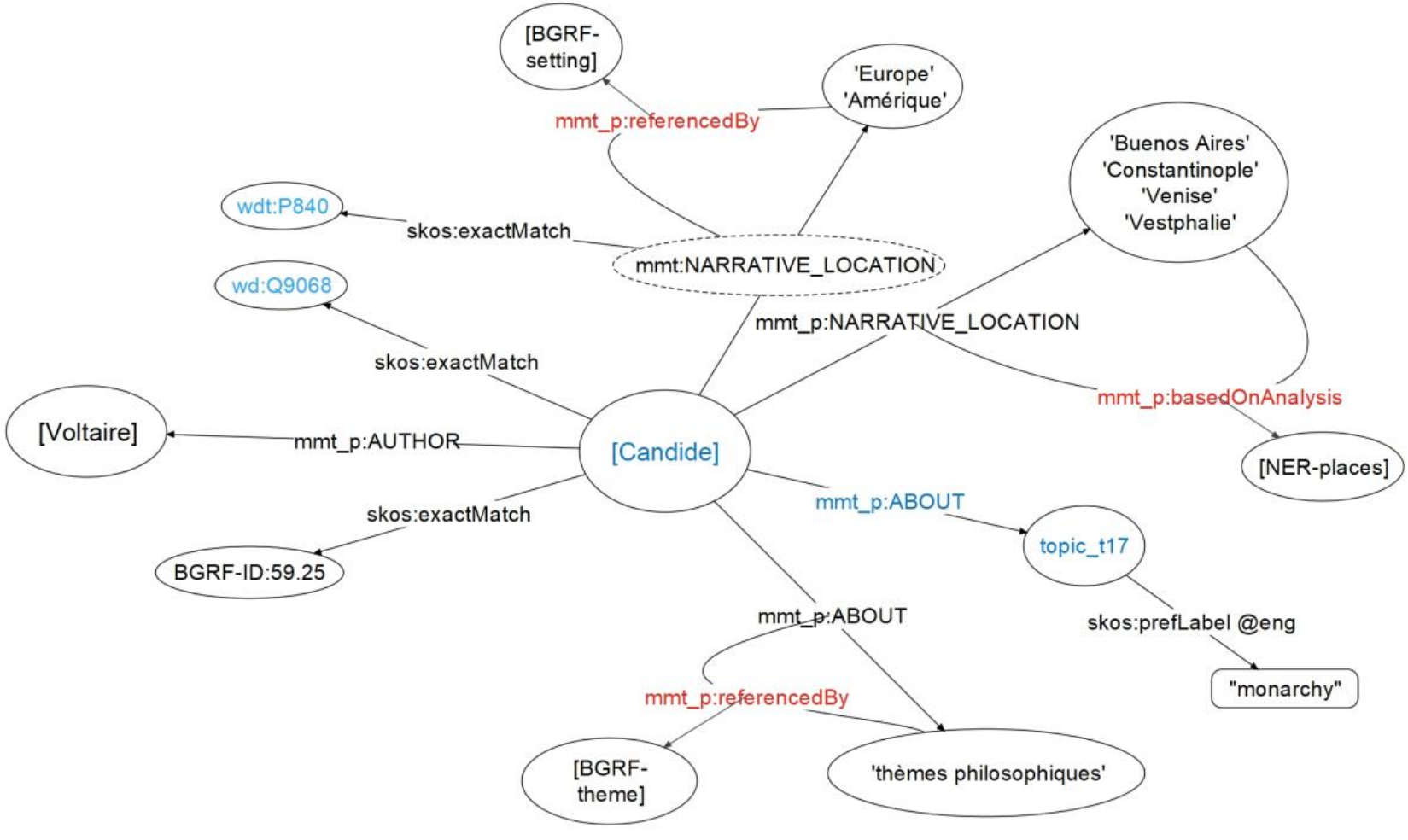
- Candide LEGAL_STATUS censored
- Candide RECEPTION_INTENSITY high
- Candide RECEPTION_TIME immediate;long-term

More statements

- Candide LEGAL_STATUS censored
- Candide RECEPTION_INTENSITY high
- Candide RECEPTION_TIME immediate;long-term
- Candide GENRE novel; satire; utopia
- Candide NARRATIVE_LOCATION Europe; Amérique
- Voltaire INFLUENCED_BY Leibniz

Conclusion

Bringing it all together: A Network of Information



Thank you!

Questions or comments?

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

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