Mining and Modeling Literary History and Historiography

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

https://mimotext.github.io/modeling/

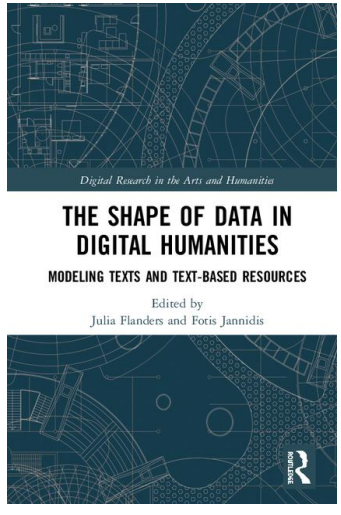
Vilnius, September 2020

Overview

1. Modeling in Digital Humanities
2. Introduction to 'Mining and Modeling Text'
3. Modeling Literary History
4. Modeling Literary Historiography
5. A network of information

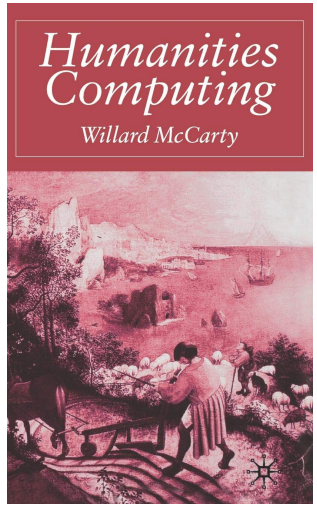
(1) Modeling in Digital Humanities

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 modeling of some segment of the world in such a way to make some aspects computable."

Willard McCarty, *Humanities Computing*, 2005



> *"The residue of uniqueness"*

- modeling as an iterative, knowledge-producing process)

Computational Modeling in DH

Computational Modeling in DH

- Data Modeling

Computational Modeling in DH

- Data Modeling
  - Conceptual Modeling: entities and relations (taxonomies, ontologies, vocabularies)

Computational Modeling in DH

- Data Modeling
  - Conceptual Modeling: entities and relations (taxonomies, ontologies, vocabularies)
  - Logical Modeling: schemas / data structures (e.g. in XML-TEI, RDF)

Computational Modeling in DH

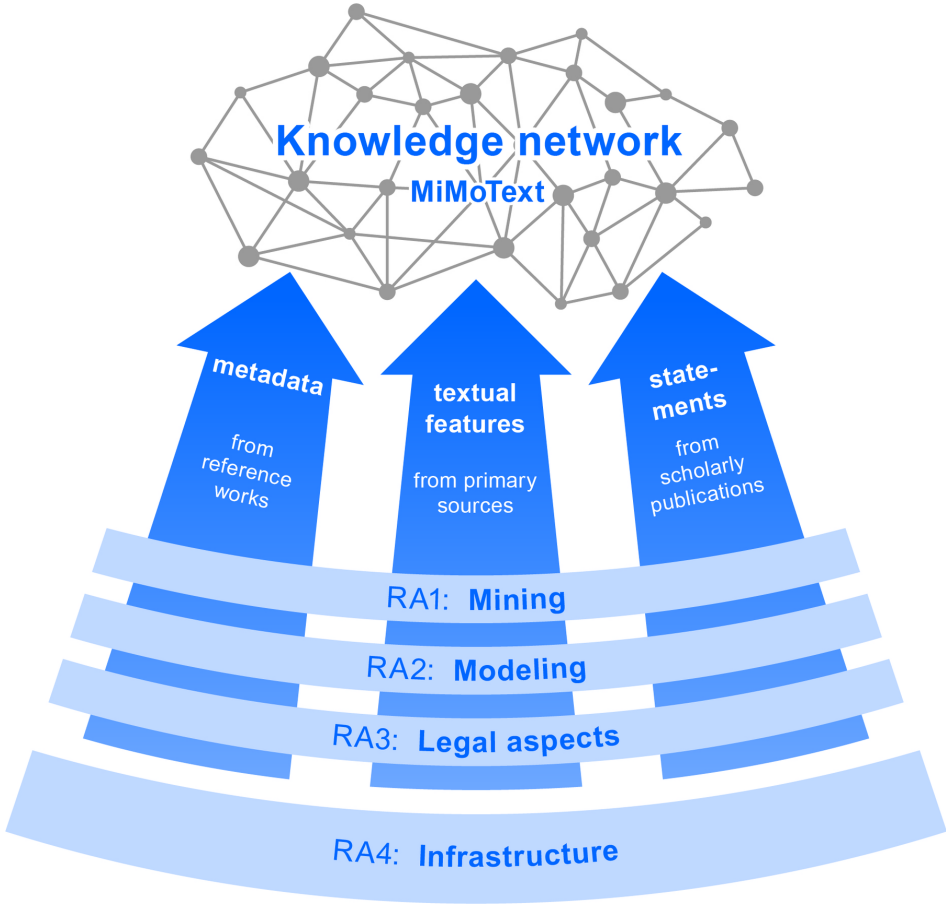
- Data Modeling
  - Conceptual Modeling: entities and relations (taxonomies, ontologies, vocabularies)
  - Logical Modeling: schemas / data structures (e.g. in XML-TEI, RDF)
- Statistical Modeling
  - Detecting trends and patterns (e.g. Topic Modeling, Linear Regression, Keyness etc.)

Computational Modeling in DH

- Data Modeling
  - Conceptual Modeling: entities and relations (taxonomies, ontologies, vocabularies)
  - Logical Modeling: schemas / data structures (e.g. in XML-TEI, RDF)
- Statistical Modeling
  - Detecting trends and patterns (e.g. Topic Modeling, Linear Regression, Keyness etc.)
- Language Models
  - Properties of Languages encoded in language models (e.g. Word Embedding Models)

(2) Introduction to 'Mining and Modeling Text'

MiMoText: overview



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

What information is relevant to literary history?

What information is relevant to literary history?

- Catalogues
  - Metadata: authors, works, publishers, etc.
  - Keywords in the *Bibliographie…*: setting, plot, protagonists, themes, style

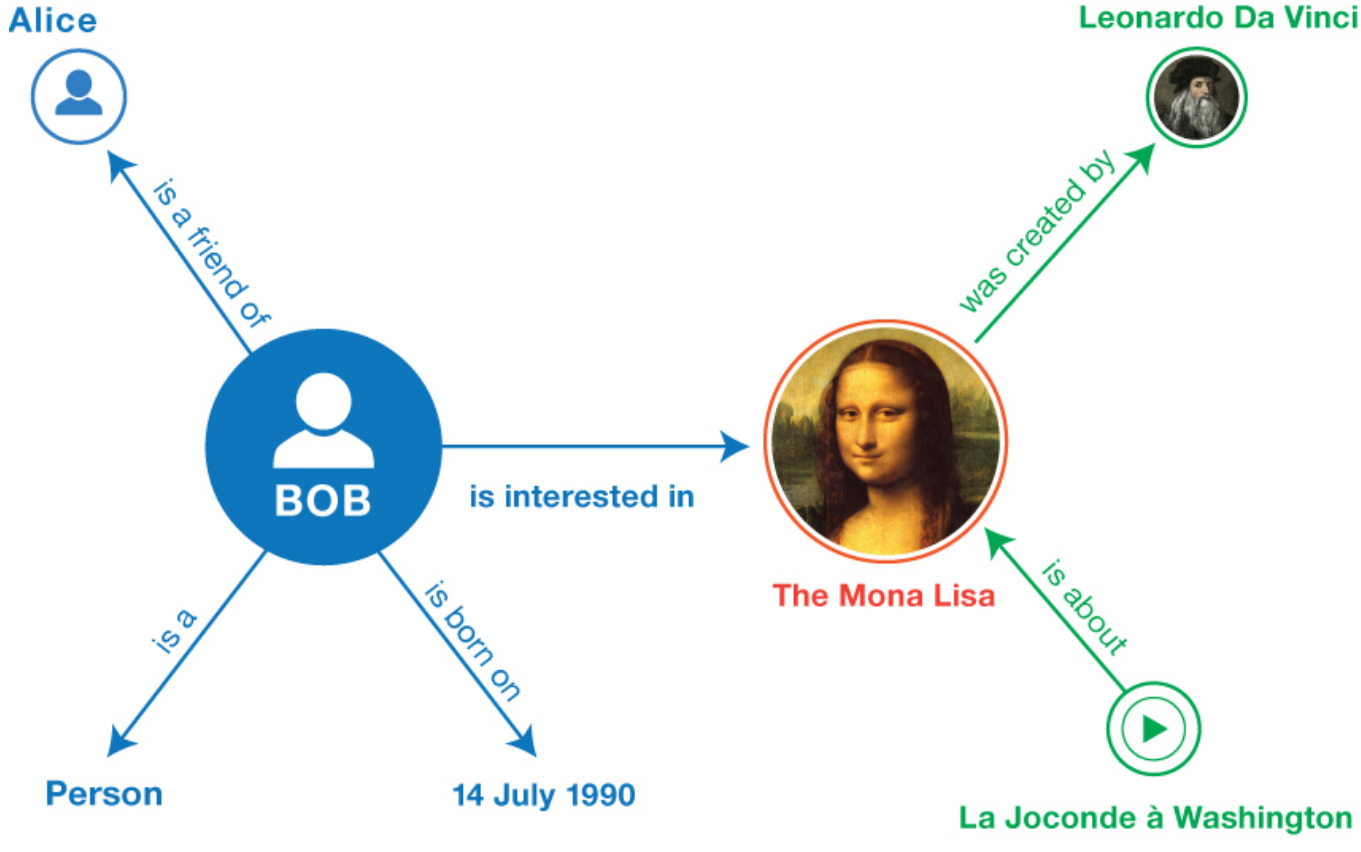
What information is relevant to literary history?

- Catalogues
  - Metadata: authors, works, publishers, etc.
  - Keywords in the *Bibliographie…*: setting, plot, protagonists, themes, style
- Literary histories
  - Plot, content, themes
  - Value judgements of authors and works
  - Relationships between authors and works

What information is relevant to literary history?

- Catalogues
  - Metadata: authors, works, publishers, etc.
  - Keywords in the *Bibliographie…*: setting, plot, protagonists, themes, style
- Literary histories
  - Plot, content, themes
  - Value judgements of authors and works
  - Relationships between authors and works
- Primary texts (novels)
  - Place names
  - Character names
  - Topics / themes
  - Foreign-language segments
  - Proportion of direct speech
  - Happy / sad ending?
  - etc.

How do we represent this information?
Linked Open Data / Triples



(3) Mining and Modeling Data

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)**

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>
```

Example: Mining primary texts (novels)

Example: Mining primary texts (novels)

- Aim: 200 volumes of novels (1750-1800)

Example: Mining primary texts (novels)

- Aim: 200 volumes of novels (1750-1800)
- Sources: double keying, OCR, portals

Example: Mining primary texts (novels)

- Aim: 200 volumes of novels (1750-1800)
- Sources: double keying, OCR, portals
- Corpus composition: decades and narrative forms

Example: Mining primary texts (novels)

- Aim: 200 volumes of novels (1750-1800)
- Sources: double keying, OCR, portals
- Corpus composition: decades and narrative forms
- Current pilot corpus: ca. 80 volumes (XML-TEI, ELTeC schema)

Example: Mining primary texts (novels)

- Aim: 200 volumes of novels (1750-1800)
- Sources: double keying, OCR, portals
- Corpus composition: decades and narrative forms
- Current pilot corpus: ca. 80 volumes (XML-TEI, ELTeC schema)
- Method of analysis: Topic Modeling

Topic Modeling

Topics

gene	0.04
dna	0.02
genetic	0.01
...	

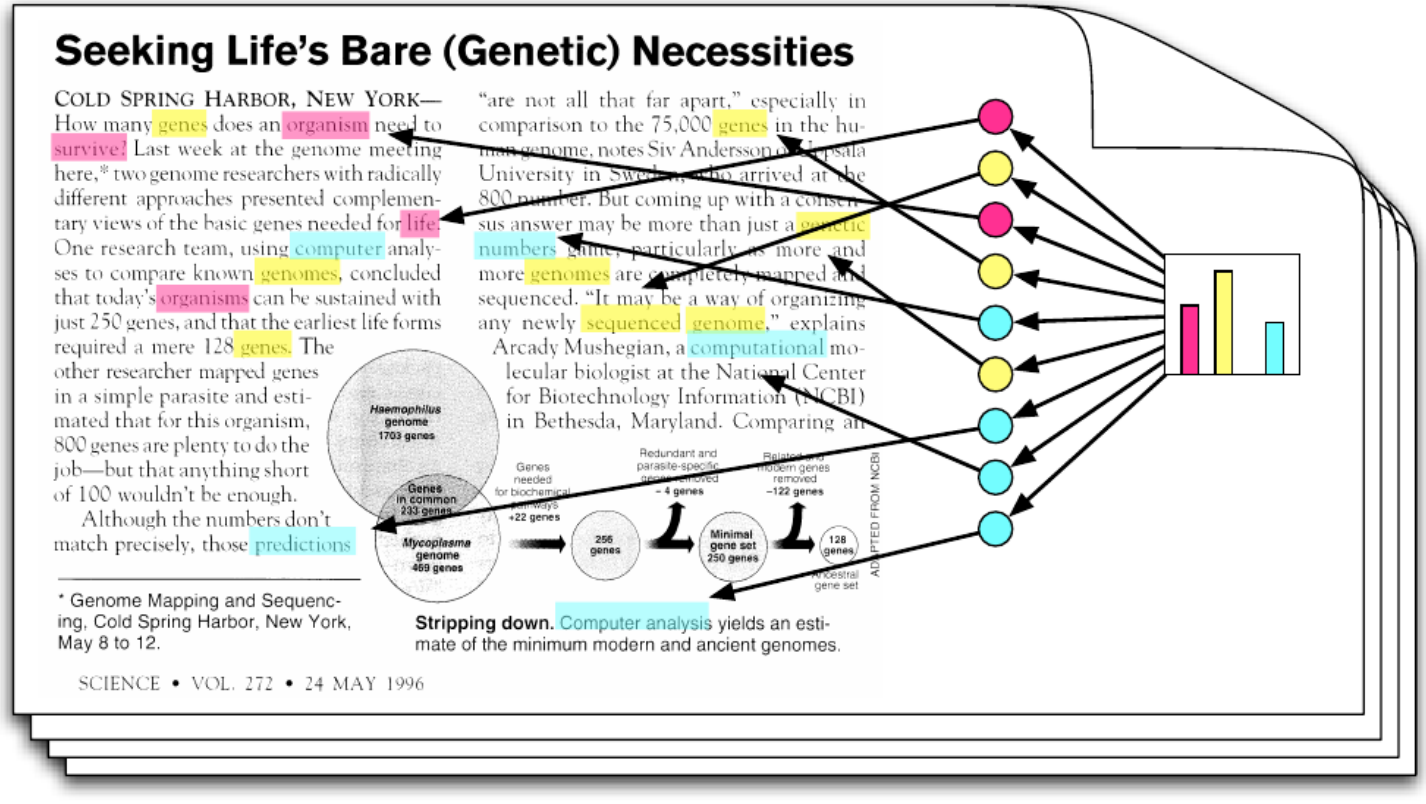
life	0.02
evolve	0.01
organism	0.01
...	

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

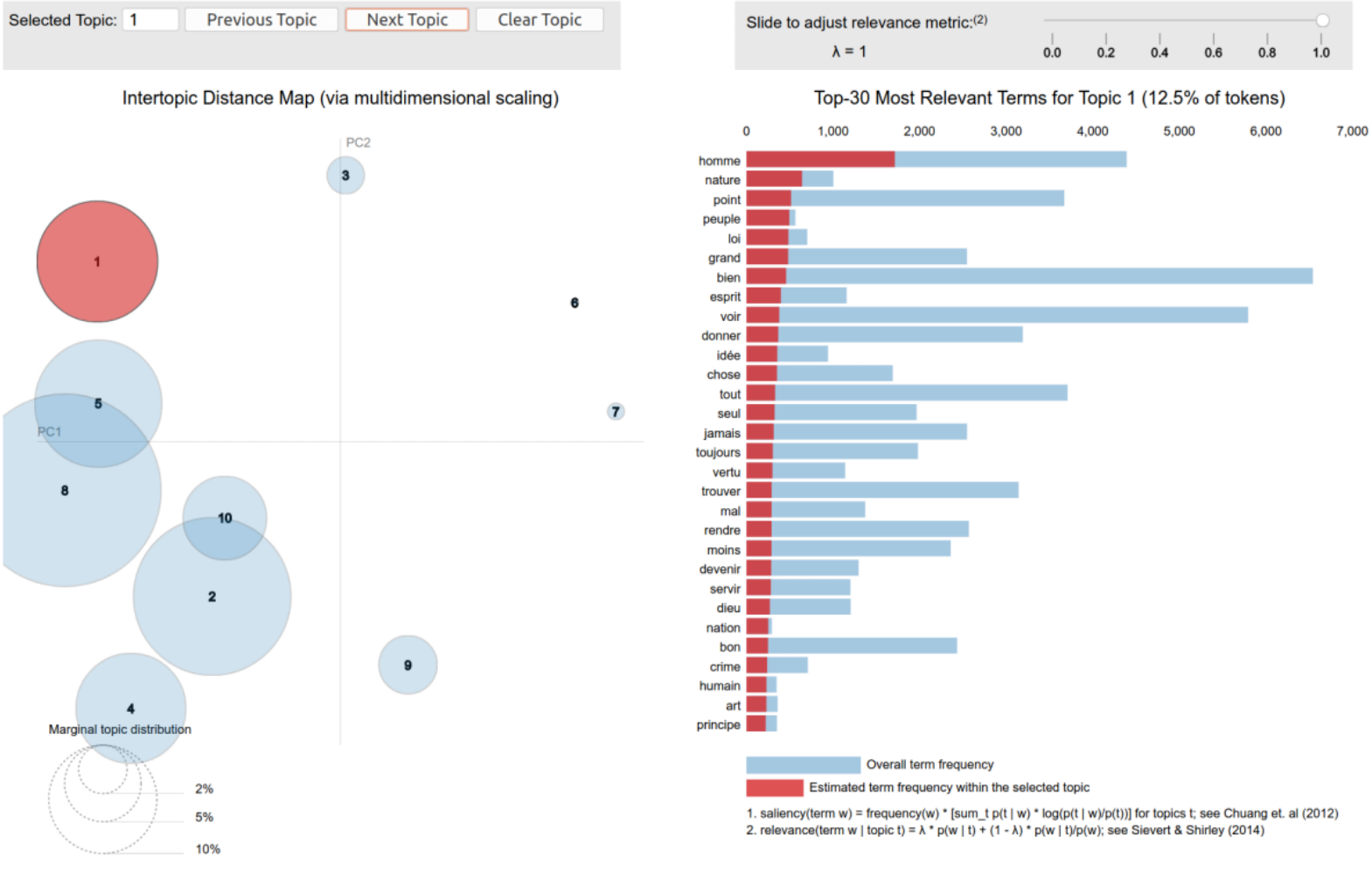
COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known 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.

First results



Modeling as Linked Open Data (LOD)

Modeling as Linked Open Data (LOD)

- predicate MAIN_SUBJECT

Modeling as Linked Open Data (LOD)

- predicate MAIN_SUBJECT
  - = Wikidata: P921 (cf. Dublin Core "Subject")

Modeling as Linked Open Data (LOD)

- predicate MAIN_SUBJECT
  - = Wikidata: P921 (cf. Dublin Core "Subject")
  - with threshold for inclusion and probability (and/or rank)

Modeling as Linked Open Data (LOD)

- predicate MAIN_SUBJECT
  - = Wikidata: P921 (cf. Dublin Core "Subject")
  - with threshold for inclusion and probability (and/or rank)
  - Candide MAIN_SUBJECT 'philosophy'

Modeling as Linked Open Data (LOD)

- predicate MAIN_SUBJECT
  - = Wikidata: P921 (cf. Dublin Core "Subject")
  - with threshold for inclusion and probability (and/or rank)
  - Candide MAIN_SUBJECT 'philosophy'
  - {Candide MAIN_SUBJECT 'philosophy'} PROBABILITY 0.34 / RANK 2

Modeling as Linked Open Data (LOD)

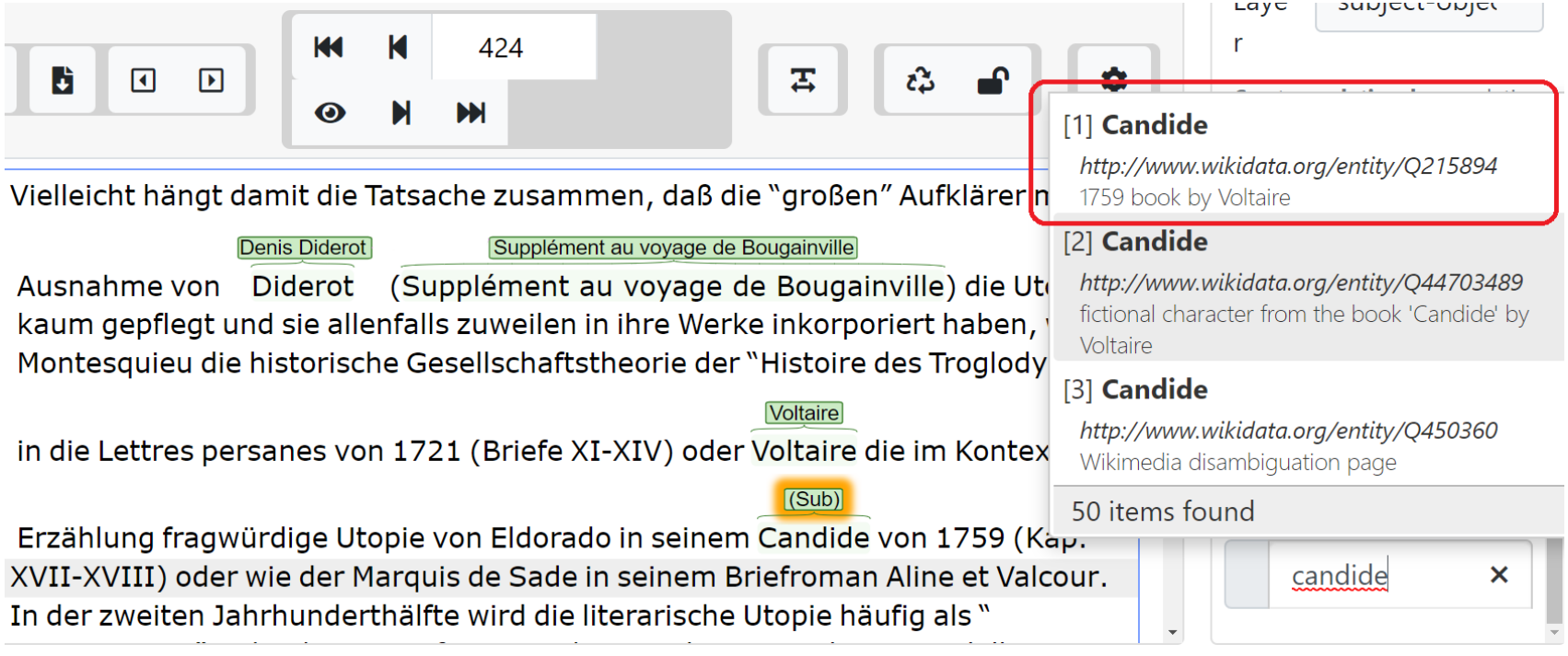
- predicate MAIN_SUBJECT
  - = Wikidata: P921 (cf. Dublin Core "Subject")
  - with threshold for inclusion and probability (and/or rank)
  - Candide MAIN_SUBJECT 'philosophy'
  - {Candide MAIN_SUBJECT 'philosophy'} PROBABILITY 0.34 / RANK 2
- Integration into the database of statements

(4) Modeling Literary Historiography

(A) Extracting Statements

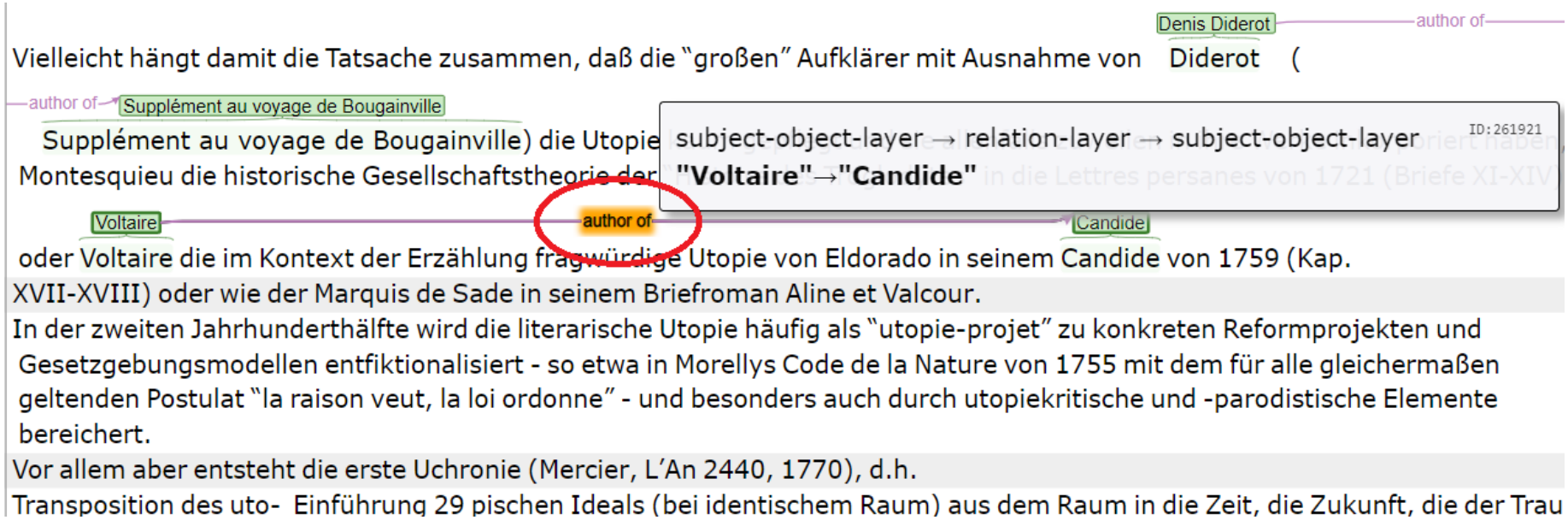
- Create training data by manual annotation (in INCePTION)
- Train and apply Machine Learning (in Python)

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

Machine Learning

- Material: sentences automatically annotated for named entities

Machine Learning

- Material: sentences automatically annotated for named entities
- Provide manual annotations of sentences (training and evaluation)

Machine Learning

- Material: sentences automatically annotated for named entities
- Provide manual annotations of sentences (training and evaluation)
- Learn patterns / probabilities for features indicative of a relation

Machine Learning

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

(B) Modeling Historiography

(B) Modeling Historiography

- Which types of statements are necessary?

(B) Modeling Historiography

- Which types of statements are necessary?
- How do we create scholarly consensus?

(B) Modeling Historiography

- Which types of statements are necessary?
- How do we create scholarly consensus?
- Meta-perspective on disciplinary discourse

Extract from literary history

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)

Statements (1)

Statements (1)

- Voltaire (viaf:36925746) AUTHOR_OF Candide (viaf:176620251)

Statements (1)

- Voltaire (viaf:36925746) AUTHOR_OF Candide (viaf:176620251)
- Candide PUBLICATION_DATE 1759

Statements (1)

- Voltaire (viaf:36925746) AUTHOR_OF Candide (viaf:176620251)
- Candide PUBLICATION_DATE 1759
- Candide PUBLICATION_LOCATION Geneva (tgn:7007279)

Statements (2)

Statements (2)

- Candide LEGAL_STATUS censored

Statements (2)

- Candide LEGAL_STATUS censored
- Candide RECEPTION_INTENSITY high

Statements (2)

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

Statements (2)

- Candide LEGAL_STATUS censored
- Candide RECEPTION_INTENSITY high
- Candide RECEPTION_TIME immediate;long-term
- Candide GENRE novel; satire; utopia

Statements (2)

- Candide LEGAL_STATUS censored
- Candide RECEPTION_INTENSITY high
- Candide RECEPTION_TIME immediate;long-term
- Candide GENRE novel; satire; utopia
- Candide NARRATIVE_LOCATION Lisbon; Eldorado; Constantinople

Statements (2)

- Candide LEGAL_STATUS censored
- Candide RECEPTION_INTENSITY high
- Candide RECEPTION_TIME immediate;long-term
- Candide GENRE novel; satire; utopia
- Candide NARRATIVE_LOCATION Lisbon; Eldorado; Constantinople
- Voltaire INFLUENCED_BY Leibniz

Special kinds of statements

Special kinds of statements

- Source:
  {Voltaire AUTHOR_OF Candide} SOURCE Köhler_1984

Special kinds of statements

- Source:
  {Voltaire AUTHOR_OF Candide} SOURCE Köhler_1984

- Qualifier:
  {Candide LEGAL_STATUS censored} TEMPORAL_SCOPE 1759-1765

Special kinds of statements

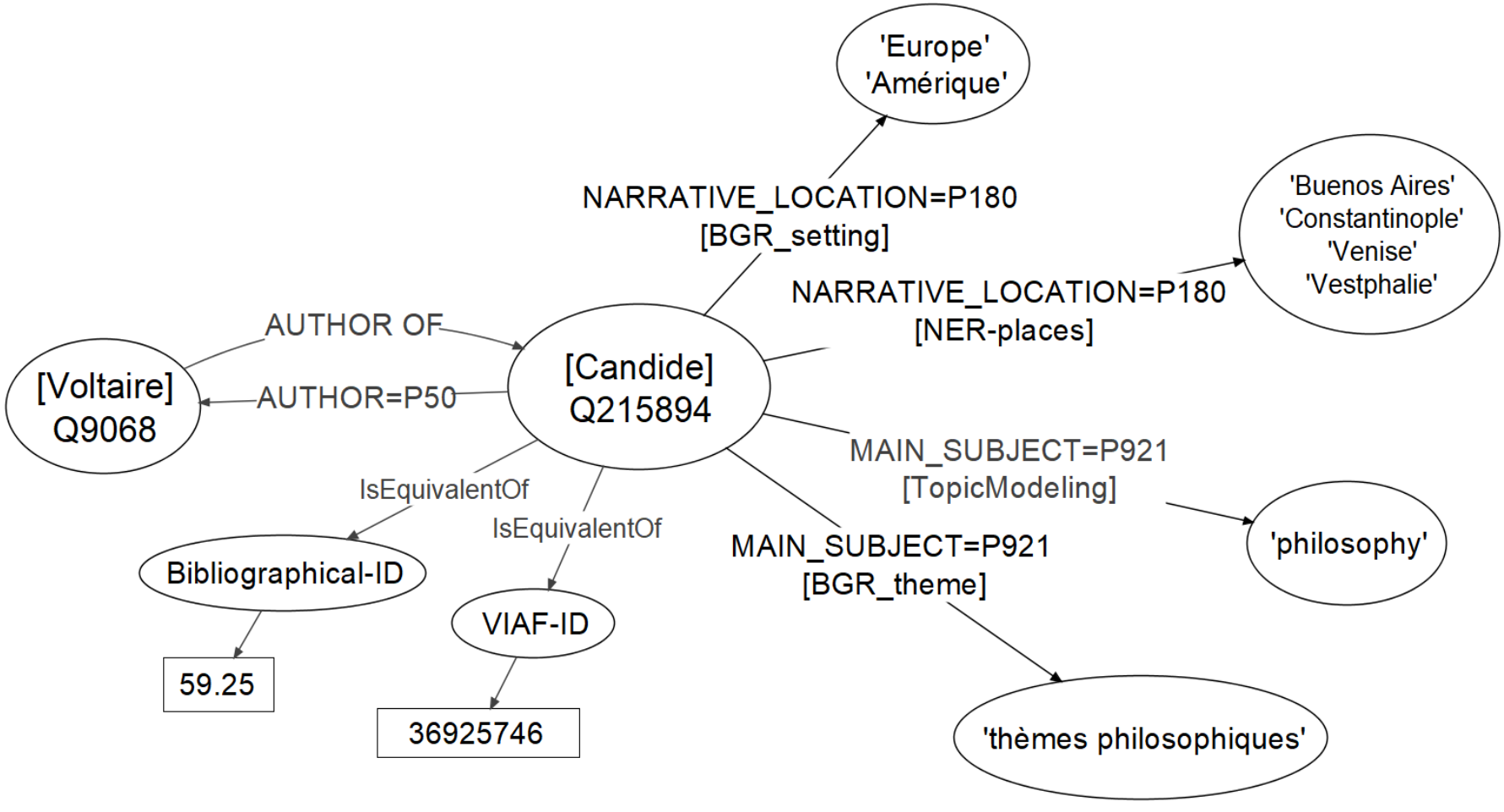
- Source:
  {Voltaire AUTHOR_OF Candide} SOURCE Köhler_1984

- Qualifier:
  {Candide LEGAL_STATUS censored} TEMPORAL_SCOPE 1759-1765

- Implicit:
  Köhler_1984 MAIN_SUBJECT Voltaire; Candide; Geneva

(5) A network of information

Bringing it all together


'Europe'
'Amérique'
NARRATIVE_LOCATION=P180
[BGR_setting]
'Buenos Aires'
'Constantinople'
'Venise'
'Vestphalie'
NARRATIVE_LOCATION=P180
[NER-places]
AUTHOR OF
[Voltaire]
Q9068
AUTHOR=P50
[Candide]
Q215894
MAIN_SUBJECT=P921
[TopicModeling]
'philosophy'
IsEquivalentOf
IsEquivalentOf
Bibliographical-ID
VIAF-ID
MAIN_SUBJECT=P921
[BGR_theme]
59.25
36925746
'thèmes philosophiques'

Turn publications into data (retrospectively)

Turn publications into data (retrospectively)

- Our goal: "Wikidata for Literary History"
  - Information system for literary history
  - Linked Open Data; SPARQL-Endpoint for queries; search ∓ browse

Turn publications into data (retrospectively)

- Our goal: "Wikidata for Literary History"
  - Information system for literary history
  - Linked Open Data; SPARQL-Endpoint for queries; search ∓ browse
- But:
  - much more specific focus (novels 1750-1800)
  - better coverage (authors, works)
  - greater density of statements
  - systematic set of types of statements
  - many usage scenarios for literary historiography

Consider publications as data (prospectively)

Consider publications as data (prospectively)

- Digital and Open Access

Consider publications as data (prospectively)

- Digital and Open Access
- Publications as maschine-readable data

Consider publications as data (prospectively)

- Digital and Open Access
- Publications as maschine-readable data
  - Rich metadata

Consider publications as data (prospectively)

- Digital and Open Access
- Publications as maschine-readable data
  - Rich metadata
  - Explicit, semantic text encoding

Consider publications as data (prospectively)

- Digital and Open Access
- Publications as maschine-readable data
  - Rich metadata
  - Explicit, semantic text encoding
  - Encoding and identification of named entities (authority files)

Consider publications as data (prospectively)

- Digital and Open Access
- Publications as maschine-readable data
  - Rich metadata
  - Explicit, semantic text encoding
  - Encoding and identification of named entities (authority files)
  - Key statements formulated as LOD

Consider publications as data (prospectively)

- Digital and Open Access
- Publications as maschine-readable data
  - Rich metadata
  - Explicit, semantic text encoding
  - Encoding and identification of named entities (authority files)
  - Key statements formulated as LOD
  - Everything in open file formats / standards

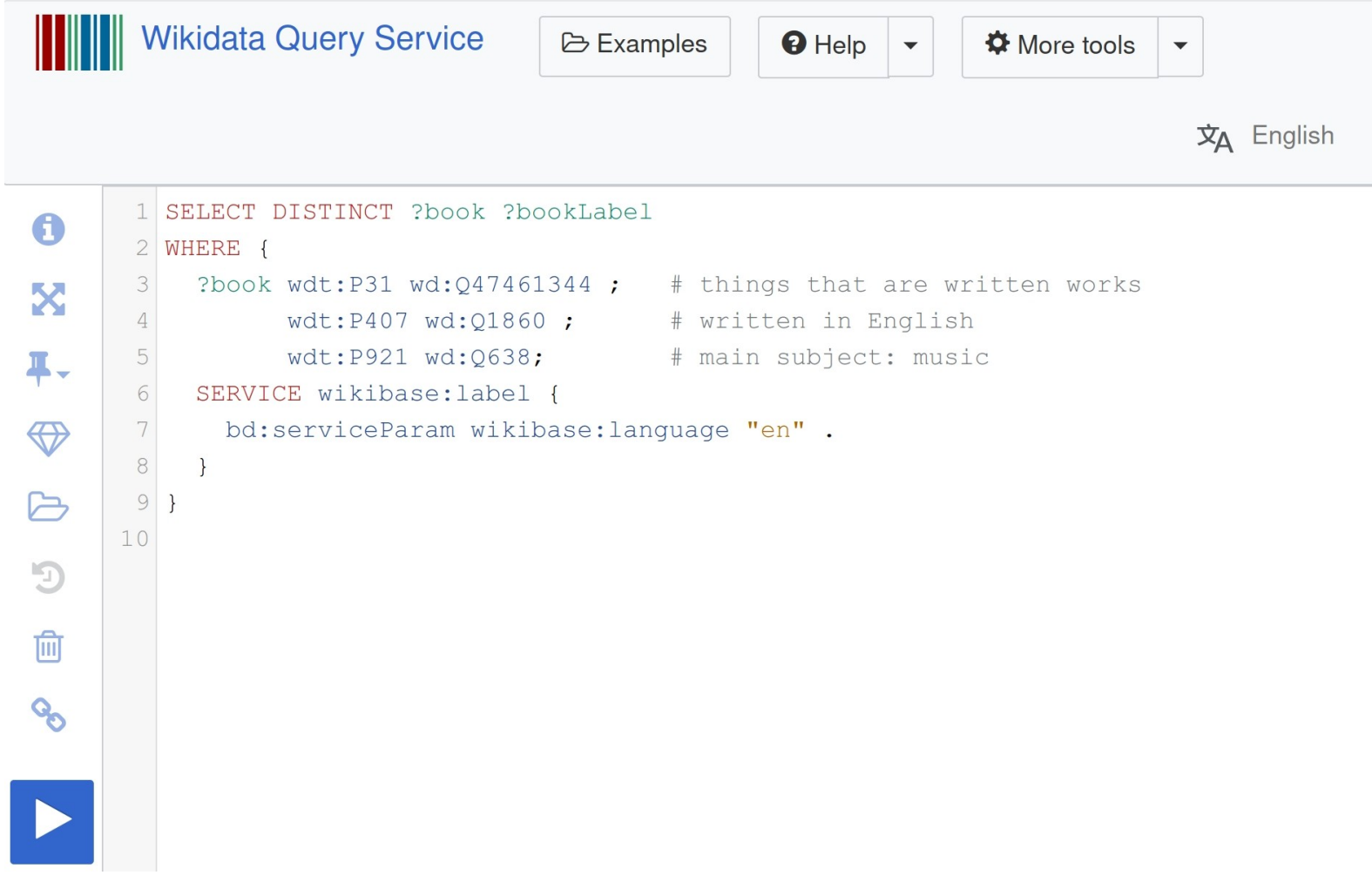
Thank you!

Questions or comments?

slides: https://mimotext.github.io/modeling/
project: https://mimotext.uni-trier.de/

Bonus slides

Wikidata query: literary works about music



Wikidata query: written works in French about philosophy