

Mining and Modeling Text: Linked Open Data für die Literaturgeschichtsschreibung



Maria Hinzmann & Christof Schöch, mit Beiträgen von Andreas Lüschow, Julia Röttgermann, Katharina Dietz und Anne Klee

<https://mimotext.github.io/lod-lithist>

Forschungskolloquium Digital History | Berlin | 27.01.2021



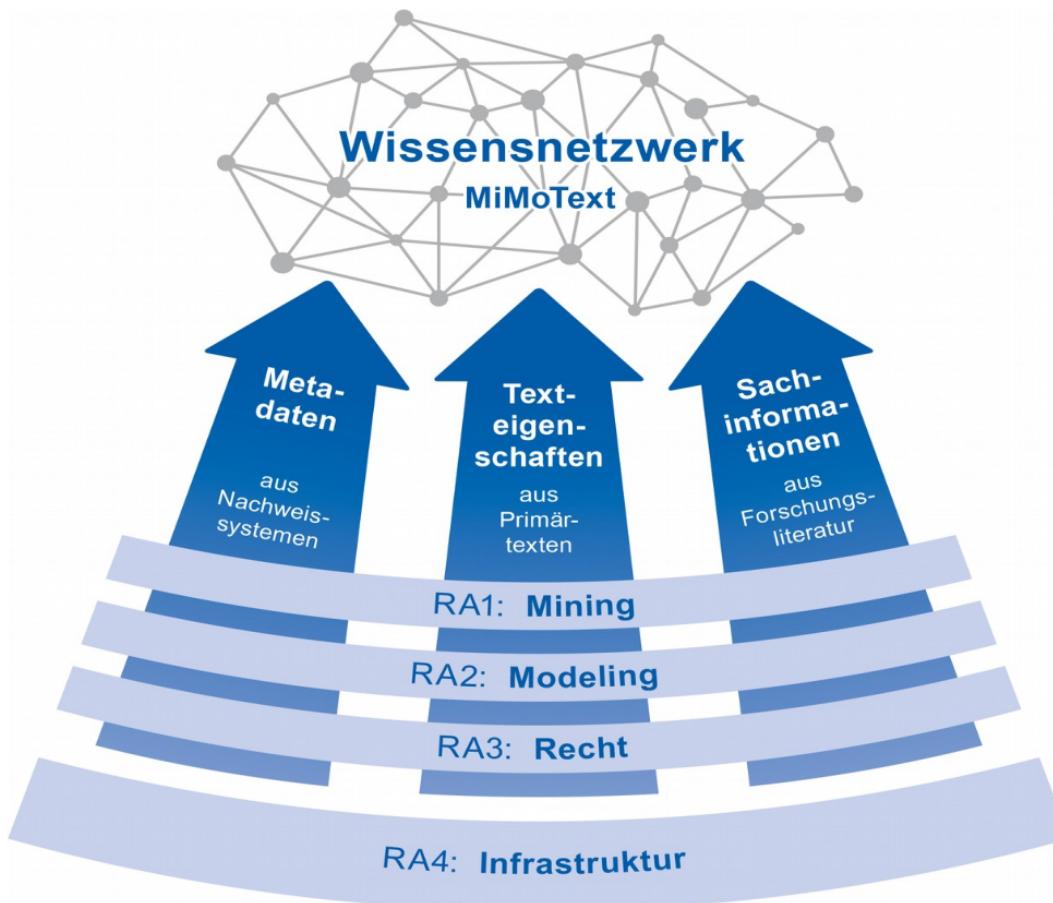


Gliederung

1. Einführung in MiMoText
2. Literaturgeschichte als LOD
3. Mining - Extrahieren von Informationen
4. Modeling - Repräsentation und Vernetzung von Wissen
5. Fazit - LOD für die Literaturgeschichtsschreibung

(1) Einführung in 'Mining and Modeling Text'

MiMoText: Überblick



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

Was sind (literaturhistorisch) relevante (und extrahierbare) Informationen?

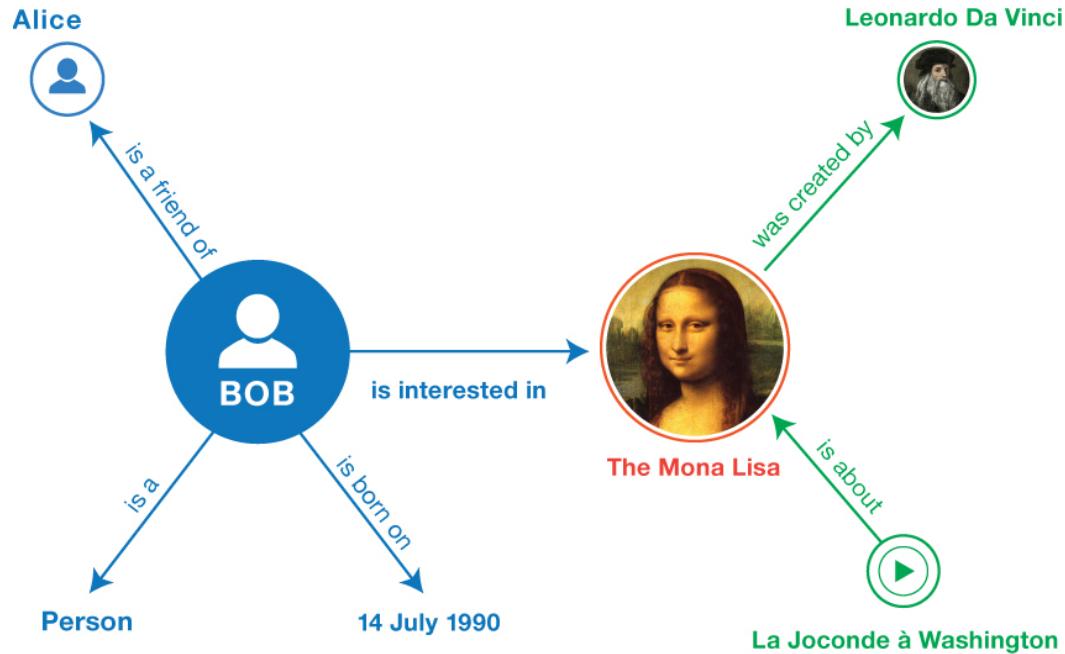
- Nachweissysteme
 - Metadaten: Autoren, Werke, Verlage, etc.
 - Keywords in der *Bibliographie*...: Setting, Themen, Protagonisten
- Primärtexte
 - Handlungsorte
 - Figurennamen
 - Topics
 - Anteil direkter Rede
 - Textlänge
 - uvm.
- Fachliteratur
 - Handlung, Inhalte, Themen
 - Wertungen von Autor:innen und Werken
 - Beziehungen zwischen Autor:innen und Werken

Was ist oder wie funktioniert Literaturgeschichte?

- Ziele
 - Sammeln und Dokumentieren literaturgeschichtlicher Fakten
 - Liefern von Erklärungen für die Entwicklung von Literatur
- Organisationsprinzipien
 - Nationen, Perioden, Bewegungen/Strömungen, Genres
 - Ähnlichkeiten und Unterschiede
 - Kontinuitäten und Wandel
 - Funktionen
- Erklärungen für literarische Entwicklungen
 - kultureller oder soziohistorischer Kontext
 - innere Dynamiken des literarischen Systems

(2) Literaturgeschichte als LOD

Modellierungsansatz: Linked Open Data (LOD)



Quelle:
<https://www.w3.org/TR/rdf11-primer/>

Grundprinzip

- Linked Open Data
 - große Menge einfacher Aussagen
 - Subjekt, Prädikat, Objekt (Tripel)
- zentrale literaturgeschichtliche Subjekte
 - Personen (Autor:in, Herausgeber:in, Verleger:in etc.)
 - Publikationen (Primärtext, Fachliteratur etc.)

Aussagetypen (1)

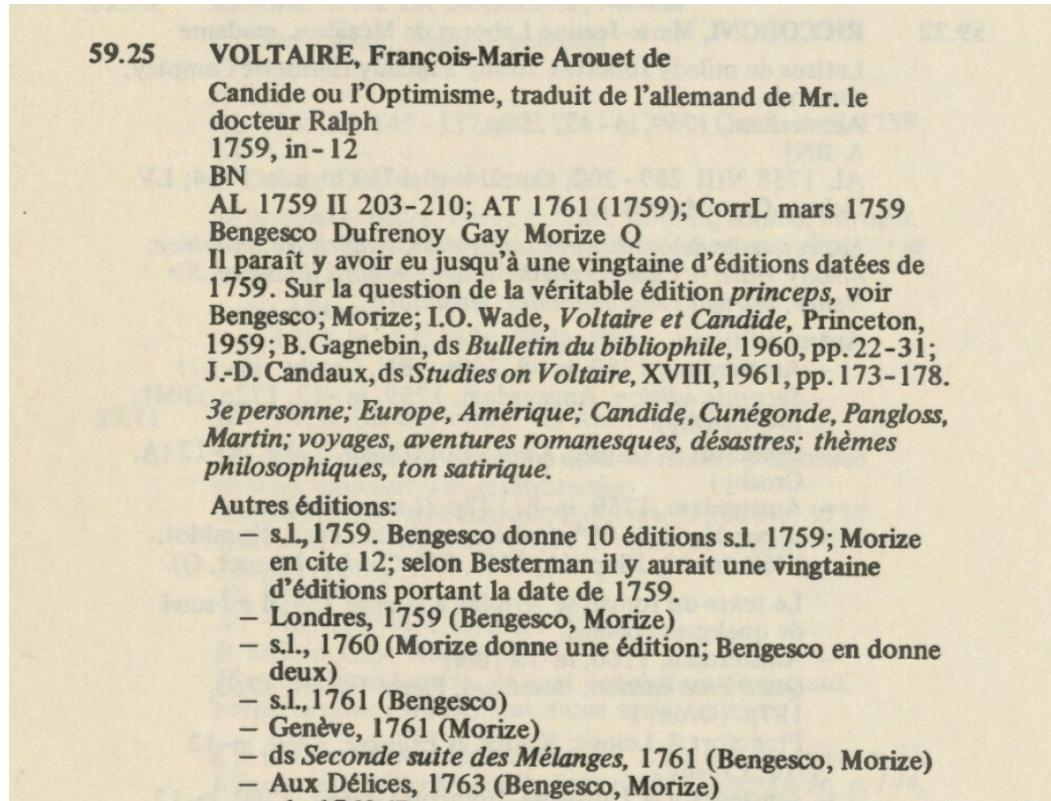
- person AUTHOR_OF publication
- publication PLACE_OF_PUBLICATION place
- publication DATE_OF_PUBLICATION year

Aussagetypen (2)

- publication NARRATIVE_LOCATION place
- publication EXTENT_WORDS number of words
- person/publ. DESCRIBED_AS (adjective | noun)
- person/publ. SIMILAR_TO person/publ.
- person/publ. DISSIMILAR_TO person/publ.
- person/publ. INFLUENCED_BY person/publ.
- publication ABOUT keyword

(3) Mining - Extrahieren von Informationen

Säule 1: Bibliographie du genre romanesque français



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

Welche Metadaten enthält diese Bibliographie?

- zusammenfassende Statistik
 - ~1100 verschiedene Autor:innen
 - ~2600 Einträge (Romane)
 - ~58.000 Triples (ca. 22 pro Roman)
- weitere Informationen
 - ~720 Romane in der 1. Person
 - ~920 in 3. Person
 - 2210 Einträge mit Angaben zum Inhalt

Säule 2: Primärliteratur (Romane)

- Pilotkorpus: ca. 100 französische Romane (1750-1800)
- Kodierung: in XML-TEI, mit Metadaten, nach ELTeC-Schema
- Analyse: Topic Modeling zur Identifikation von Themen

Das "roman18"-Korpus

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	Usage (%)
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)

Topic Modeling - Erste Ergebnisse

Selected Topic: 1 Previous Topic Next Topic Clear Topic

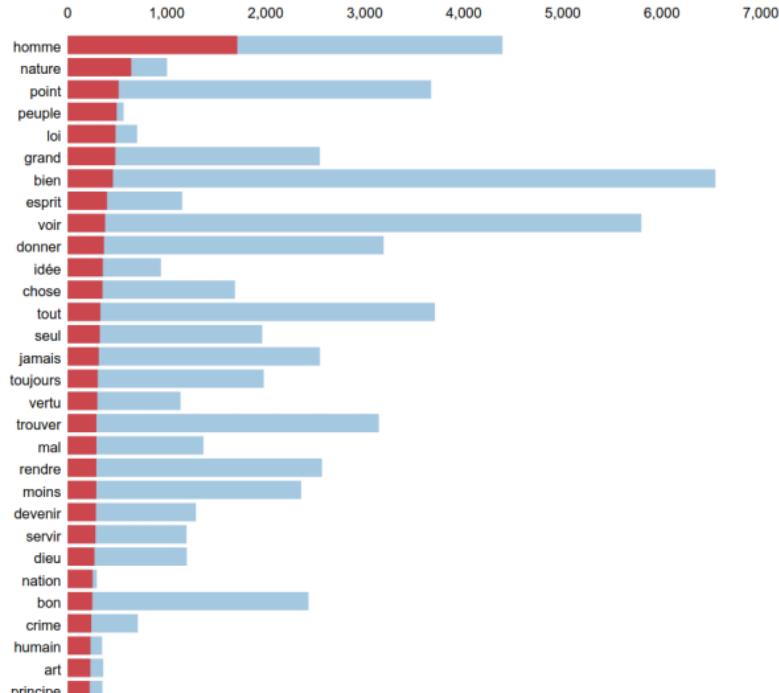


Slide to adjust relevance metric:⁽²⁾

$\lambda = 1$

0.0 0.2 0.4 0.6 0.8 1.0

Top-30 Most Relevant Terms for Topic 1 (12.5% of tokens)



Overall term frequency

Estimated term frequency within the selected topic

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)

2. relevance(term w | topic t) = $\lambda * p(w | t) + (1 - \lambda) * p(w | t)/p(w)$; see Sievert & Shirley (2014)

Zwei "major topics"

"monarchy"

A word cloud centered around the theme of monarchy. The most prominent words are 'peuple' (people), 'roi' (king), 'grand' (great), 'citoyen' (citizen), 'prince' (prince), and 'souverain' (sovereign). Other visible words include 'ordre' (order), 'gouvernement' (government), 'ennemi' (enemy), 'glorie' (glory), 'monarque' (monarch), 'sujet' (subject), 'nation', 'tendre' (tender), 'troupe' (troop), and 'empire'. Smaller words like 'point', 'trône', and 'rendre' 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), 'temp', 'sentir', 'doux', 'heureux', 'ami', 'sentiment', 'tendre', 'jou', 'aimer', 'seul', 'objet', 'vie', 'point', and 'jamais'. Smaller words like 'plaisir' and 'rendre' are also present.

Abgeleitete Aussagen ('statements')

- Candide IS_ABOUT "monarchie"
- Clarice IS_ABOUT "amour"

Säule 3: Sekundärliteratur (literaturwiss. Fachtexte)

3 Elemente für Extraktion von Aussagen aus Sekundärliteratur:

- Annotationsguidelines (basierend auf Datenmodell)
- manuelle Annotation: Generierung von Traingsdaten (in INCEpTION)
- Training: Machine Learning (in Python)

Annotationen auf 'Named Entity'-Layer

The screenshot shows a digital annotation interface. At the top, there's a toolbar with various icons for file operations like download, upload, and search, followed by a page number (424) and navigation controls (back, forward, search, etc.). To the right of the page number is a dropdown menu labeled "Subject-Object". Below the toolbar, a text snippet from a document is displayed:

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

Denis Diderot Supplément au voyage de Bougainville

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

Voltaire

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

(Sub)

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 "

On the right side of the interface, a sidebar displays search results for the term "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 term "candide".

- Annotationen von Autoren und Werken (Entitäten -> Subjekt-/ObjektPosition in Aussagen)
- Disambiguierung von Entitäten über Knowledgebase-Anbindung (hier: Wikidata)

Annotation auf 'Predicate'/Relation-Layer

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

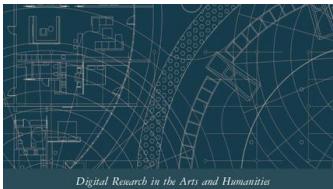
Jean-François Marmontel
Wenn Marmontels Helden am Endeliche und Glück finden, so haben sie es gewonnen. ID: 216457 weniger einem v Schicksal als ihrem bürgerlichen Fleiß, ihrer Tüchtigkeit und Tugend zu danken.
"Bélisaire" → "Roman"

In seinem antikisierenden Roman novel hasGenre Bélisaire
Bélisaire (1767) schließlich, der in den Augen der Zeitgenossen den Les
(predicate)
← hasAuthor François Fénelon
Fénelons übertrifft, wird der Leser aus der Privatsphäre herausgeführt und mit öffentlichen Bel
Fragen konfrontiert.

- "hasGenre" als Property bzw. Prädikat
- Subjekt verknüpft mit Wikidata-Identifier für "Bélisaire" (Q5005038)
- Objekt verknüpft mit Wikidata-Identifier für "novel" (Q8261)

(4) Modellierung - Repräsentation und Vernetzung von Wissen

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



THE SHAPE OF DATA IN DIGITAL HUMANITIES

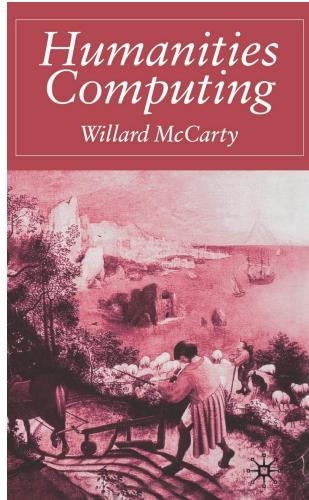
MODELING TEXTS AND TEXT-BASED RESOURCES

Edited by
Julia Flanders and Fotis Jannidis



"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



"recursive modeling":
Modellierung als ein iterativer Prozess

Säule 1: Bibliographie als RDF

```
<j:2:nextItem>
<j:2:ListItem rdf:about="http://www.kompetenzzentrum.uni-trier.de/bgrf/5925">
  <j:5:hasSequenceIdentifier>59.25</j:5:hasSequenceIdentifier>
  <j:2:itemContent>
    <j:7:BibliographicRecord
      rdf:about="http://www.kompetenzzentrum.uni-trier.de/bgrf/5925Record">
        <j:7:references>
          <j:4:Manifestation
            rdf:about="http://www.kompetenzzentrum.uni-trier.de/bgrf/5925Manifestation">
              <j:4:embodimentOf>
                <j:4:Expression
                  rdf:about="http://www.kompetenzzentrum.uni-trier.de/bgrf/5925Expression">
                    <j:0:language rdf:resource="http://id.loc.gov/vocabulary/iso639-2/fre"/>
                    <j:0:creator>VOLTAIRE, François-Marie Aronnet de</j:0:creator>
                    <j:0:title>Candide ou l'Optimisme, traduit de l'allemand de Mr. le docteur
                      Ralph</j:0:title>
                    </j:4:Expression>
                  </j:4:embodimentOf>
                  <j:3:keyword>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; L.O. Wade,
                    Voltaire et Candide, Princeton, 1959; B.Gagnebin, ds Bulletin du bibliophile, 1960,
                    pp. 22-31; J.-D.Candaux, ds 5ftrdreson Voltaire.'XNYtt., 1961, pp. 173-178.
                    3epersonne; Europe, Amérique; Candide, Cunégonde, Pangloss, Martin; voyages,
                    aventures romanesques, désastres; thèmes philosophiques, ton satirique.</j:3:keyword>
                  <j:1:P30197>in-12</j:1:P30197>
                  <j:1:P30011>1759</j:1:P30011>
                  <j:1:P30270>BN AL 1759 II 203-210; AT 1761 (1759); CorrL mars 1759 Bengesco
                    Dufrenoy Gay Morize Q</j:1:P30270>
                </j:4:Manifestation>
              </j:7:references>
              <rdf:type rdf:resource="http://purl.org/spar/fabio/BibliographicMetadata"/>
            </j:7:BibliographicRecord>
          </j:2:itemContent>
```

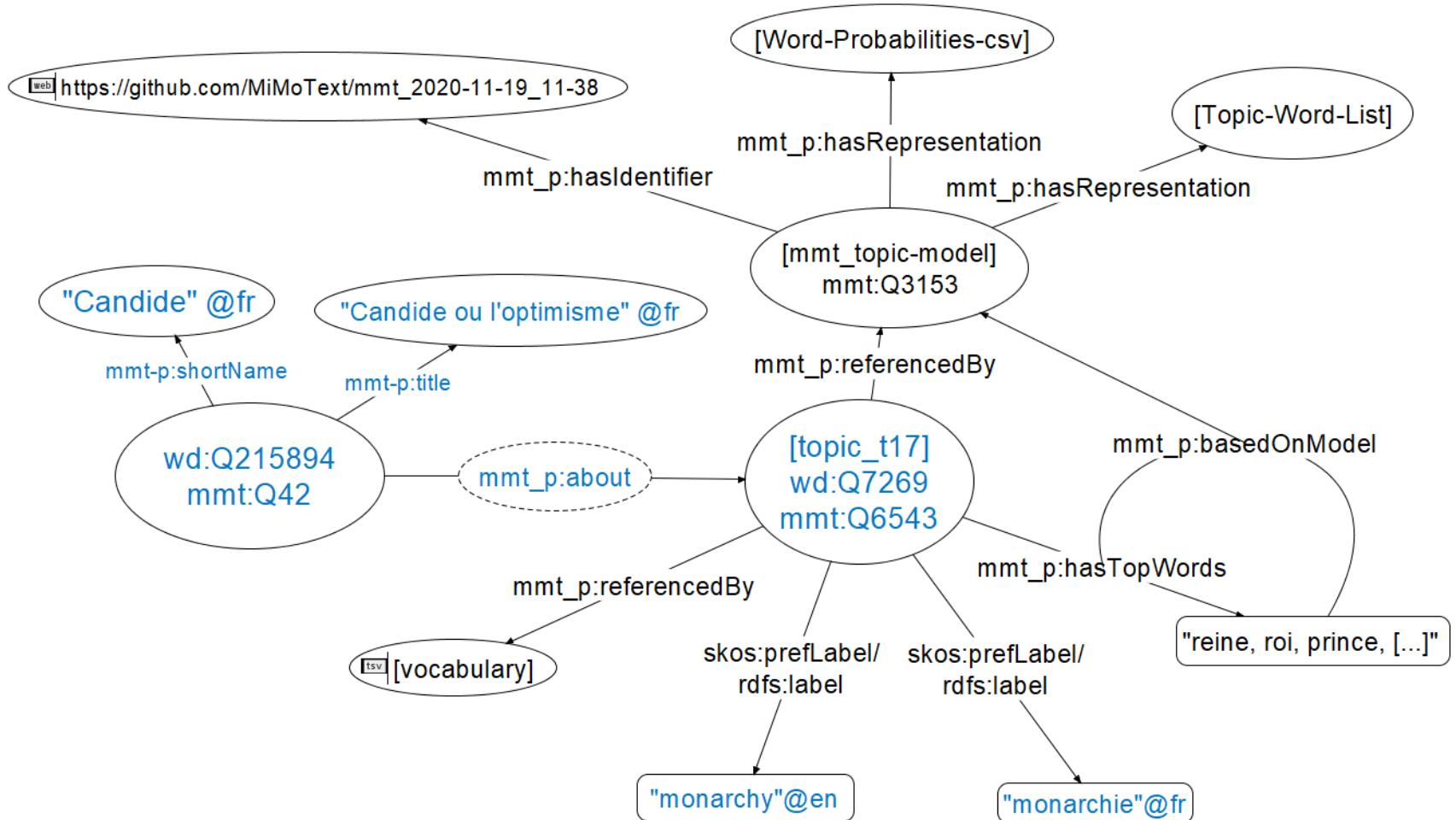
Säule 1: Statements aus Bibliographie

- Candide NARRATIVE_LOCATION "Europe"
- Candide NARRATIVE_LOCATION "Amérique"
- Candide IS_ABOUT "thèmes philosophiques"

Säule 2: Statements aus Topic Modeling

- Beispiel Topic Modeling
 - Subjekt: *Candide*
 - Prädikat: IS_ABOUT ([schema.org/about](#); Wikidata "main_subject")
 - Objekt: Topic "philosophie"
- LOD-Statements
 - Candide IS_ABOUT 'philosophie'
 - Candide IS_ABOUT 'monarchie'

Säule 2: Romananalysen (Topic Modeling) als LOD



Säule 3: Literaturgeschichtsschreibung

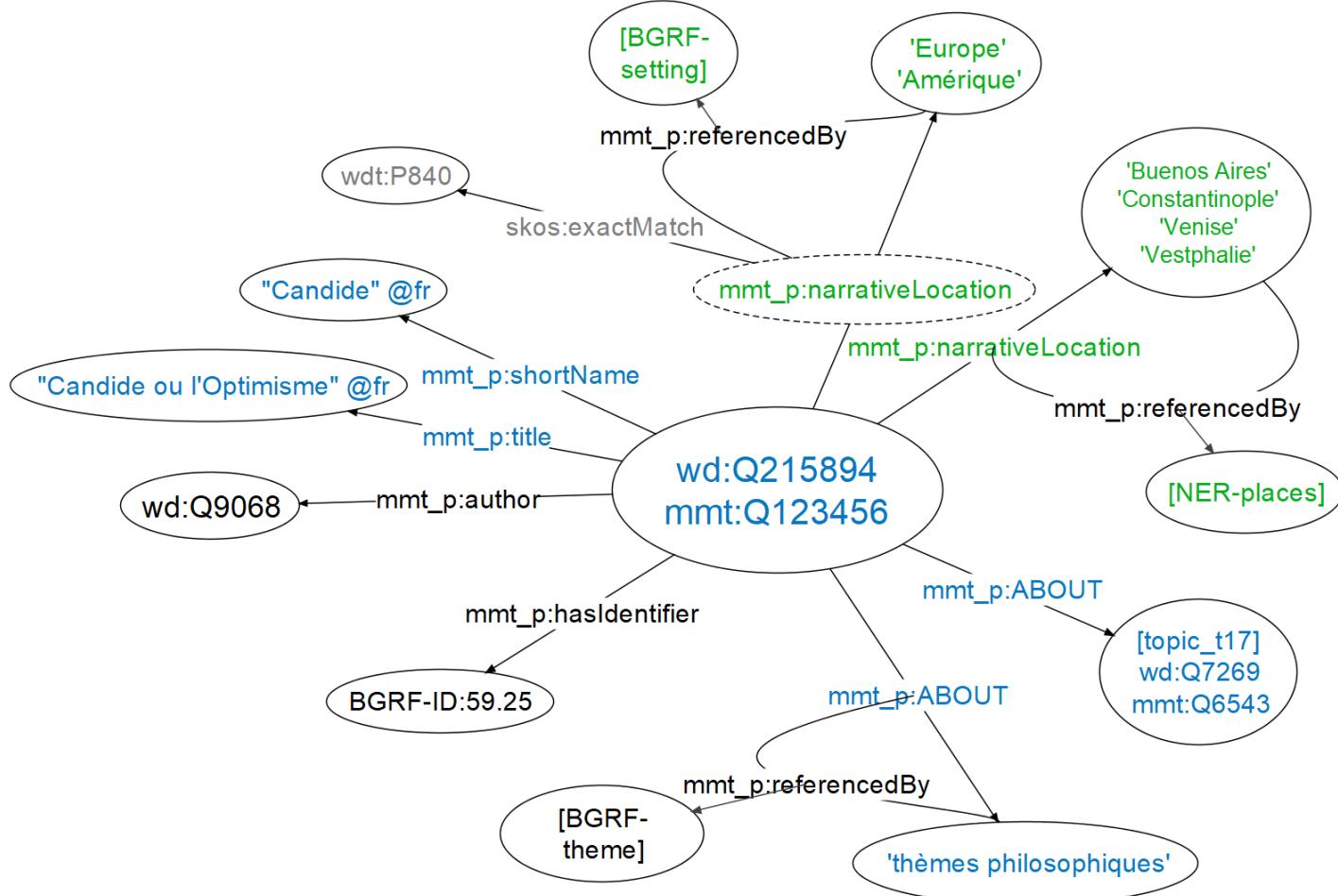
Candide ist das meistgelesene Werk Voltaires und war es wohl schon zu Lebzeiten des Autors. Als es 1759 in Genf erstmals im Druck erschien, wurde es zwar sofort verboten, aber doch nur mit dem Ergebnis, daß es im gleichen Jahr noch dreizehn Neuauflagen erlebte.
(Erich Köhler, Aufklärung II, 1984)

Säule 3: Aussagen aus Fachliteratur

- Aussagen in zitiertem Beispiel
 - Candide PUBLICATION_DATE 1759
 - Candide LEGAL_STATUS censored
 - Candide RECEPTION_INTENSITY high
- weitere Aussagen:
 - Candide GENRE novel; satire; utopia
 - Voltaire TOPIC_INTEREST Gesellschaftskritik
 - Voltaire INFLUENCED_BY Leibniz

(5) Fazit: LOD für die Literaturgeschichtsschreibung

Wissensnetzwerk ('Knowledge Graph')



Herausforderungen

- Modellierung von "Meta-Tripeln" ('reification'; 'qualifier' etc.)
 - {Voltaire AUTHOR_OF Candide} SOURCE Köhler_1984
 - {Candide LEGAL_STATUS censored} TEMPORAL_SCOPE 1759-1765
 - {[Subject] PREDICATE [Object]} RELIABILITY (low | middle | high)
- wenig Konsens in/über Literaturgeschichte; tendenziell Entkoppelung von 'Theorie' und 'Praxis'
- Standardisierung am Anfang (kein "literaturhistorisches CIDOC-CRM")
- Implementierung (über verschiedene Tools hinweg)
- Konkretisierung von Nutzungsszenarien
- Mehrsprachigkeit (z.B. NER Fachliteratur)
- Entwicklung kontrollierter Vokabulare (z.B. 'Themen'-Werte in Pilotprojekt)

Chancen LOD für die Literaturgeschichtsschreibung

- (1) Linked Open Data-Paradigma als Ansatz
 - Pluralität von Perspektiven, Heterogenität der Quellen
 - Referenzierung der Quellen ist möglich; gezieltere Suchoptionen in Abhängigkeit von Qualifiern insgesamt
 - unsichere Informationen können berücksichtigt werden (Zuverlässigkeitsgrad "qualifizierbar")
 - widersprüchliche, komplementäre Informationen können nebeneinander bestehen
- (2) Standardisierungsprozesse als Reflexionsanlass & Dialogpotential
 - viele Fragen: Was sind in einer Disziplin / Community die relevanten Entitäten und Relationen? Welche Aussagetypen sind (jenseits von Metadaten) notwendig? Welche könnten nützlich sein?
 - Metaperspektive auf disziplinären Diskurs

Ziele

- Unser Ziel: "Wikidata für die Literaturgeschichte"
 - literaturhistorisches Informationssystem
 - LOD-Basis, SPARQL-Endpoint, Suchmaske
- Aber mit:
 - viel spezifischerem Fokus (Romanliteratur 1750-1800)
 - stark erweiterter Abdeckung (Autoren, Werke)
 - stark erhöhter Aussagendichte
 - systematischer Ontologie von Aussagentypen
 - vielfältigen Anwendungsszenarien für die Literaturgeschichtsschreibung

Vielen Dank!

Fragen oder Kommentare?

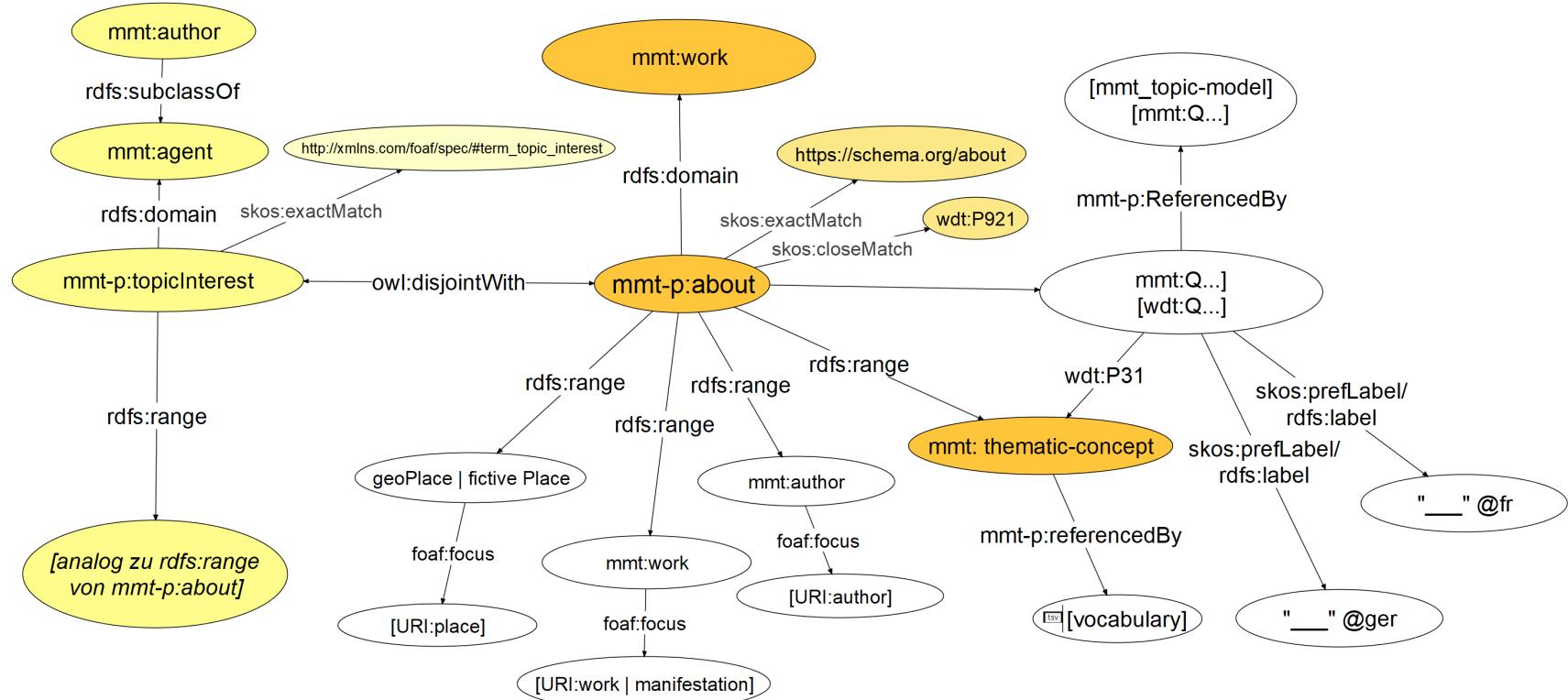
Folien: <https://mimotext.github.io/lod-listhist>

Projekt: <https://mimotext.uni-trier.de>

Lizenz: Creative Commons Attribution (CC BY), 2020

Bonus-Folien

Ontologie-Auszug: Pilotprojekt ("thematische Aussagen")



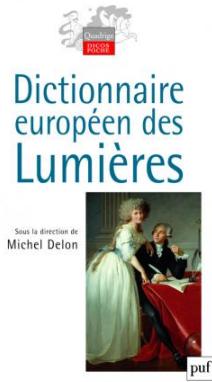
Wikidata query

The image shows the Wikidata Query Service interface. On the left, there is a vertical toolbar with various icons: a red bar, a blue bar, a magnifying glass, a double arrow, a pin, a diamond, a folder, a circular arrow, a trash can, a link, and a play button. The main area contains the following SPARQL query:

```
1 SELECT DISTINCT ?book ?bookLabel
2 WHERE {
3     ?book wdt:P31 wd:Q7725634 ; # books that are literary works
4         wdt:P407 wd:Q150 ;      # books written in French
5         wdt:P921 wd:Q5891; #main subject: philosophy
6     SERVICE wikibase:label {
7         bd:serviceParam wikibase:language "[AUTO_LANGUAGE],fr" .
8     }
9 }
```

The interface includes a navigation bar at the top with "Wikidata Query Service", "Examples", "Help", "More tools", and language selection ("English"). A refresh icon is located in the bottom right corner.

Kontrolliertes Vokabular



Passions (Michel DELON)	→ L'opéra
Passion (Représentation des) (Madeleine PINAULT SORENSEN)	→ Mémoires
Patriotisme (Gonthier Louis FINK)	→ émotion
Paysage (Sophie LE MENAHÈZE-LEFAY)	→ affection
Paysans, paysannerie (Gérard BEAUR)	→ société
Pédagogie → Éducation, instruction et pédagogie	
Peinture (Christian MICHEL)	→ artiste
Perfectibilité → Progrès	
Persiflage (Michel DELON)	→ sophistry
Perspective (Madeleine PINAULT SORENSEN)	→ logique
Pessimisme → Optimisme, pessimisme	
Peuple (Jochen SCHLOBACH)	→ citoyen
Philosophe (Jochen SCHLOBACH)	→ philosophe
Philosophie allemande (Werner SCHNEIDERS)	→ philosophie
Philogistique (Bernadette BENSAUDE-VINCENT)	→ scepticisme
Physiocratie (René LE MÉE)	→ sociologie
Physiognomonie (Philippe KAENEL)	→ vision
Physiologie (Alain TOUWAIDE)	→ biologie
Physique expérimentale (Robert LOCQUEINEUX)	→ physique
Piétisme (Udo STRATER)	→ orthodoxie
Pittoresque (Madeleine PINAULT SORENSEN)	→ esthétique
Poésie en Europe (Cariona SETH)	→ lyrique
Poésie en France (Édouard GUITTON)	→ poème
Rococo (Jean WEISGERBER)	→ art
Roman (Érik LEBORGNE)	→ roman
Roman noir (Michael BERNSEN)	→ fiction
Ruines (Philippe JUNOD)	→ artefact
Rumeur (Arlette FARGE)	→ affaire
Russie (Georges DULAC)	→ nation
Salon de peinture (William HAUPTM)	→ exposition
Salons (Jean-Noël PASCAL)	→ exposition
Satire (Roland MORTIER)	→ caricature
Sauvage → Barbare, sauvage	
Scepticisme → Doute, scepticisme, Sciences	
Sciences (Diffusion et vulgarisation de COHEN)	→ science
Sculpture (Bent SØRENSEN)	→ art
Sensibilité (Gerhard SAUDER)	→ émotion
Sensualisme (Sylvain AUROUX)	→ sens
Sexualité (Représentation de la)	→ sexualité
ABRAMOVICI	→ artiste
Siècle (Jochen SCHLOBACH)	→ époque
Silhouette, découpage (Madeleine PINAULT SORENSEN)	→ art
Sociabilité (Catherine LARRÈRE)	→ société
Socrate (Raymond TROUSSON)	→ philosophe

- Kern bzw. Basis: domänenspezifische Ressource für themat. Konzepte (*Dictionnaire européen des Lumières*. Hrsg. von Michel Delon, PUF, Paris, 1997.)
- Erweiterung
 - temporäre Dynamik: Ergänzung von Konzepten (alle 3 Infoquellen)
 - Prozess der Konsolidierung (u.a. über Identifier/Normdaten)
 - work in progress: <https://github.com/MiMoText/vocabularies/>

Publikationen als Daten (prospektiv)

- Digital und Open Access
- Publikationen als (maschinenlesbare) Daten
 - Reichhaltige Metadaten
 - Explizite, semantisch kodierte Textstruktur
 - Auszeichnung und Identifikation von Entitäten (Normdaten)
 - Kernaussagen als LOD-Statements
 - Alles in offenen Standardformaten

"Ziele der Literaturgeschichtsschreibung" (Borkowski/Heine 2013)

- zwei Konzeptionen
 - Fokus auf Historisierung
 - Fokus auf Herstellung von Gegenwartsbezügen
- Ziele der historisierenden Konzeption
 - Rekonstruktion: Etablierung wahrer oder wahrscheinlicher Aussagen über Literatur
 - Konsolidierung: Sammeln und Verbreiten des fundierten Wissensbestands

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

Romananalyse: 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

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

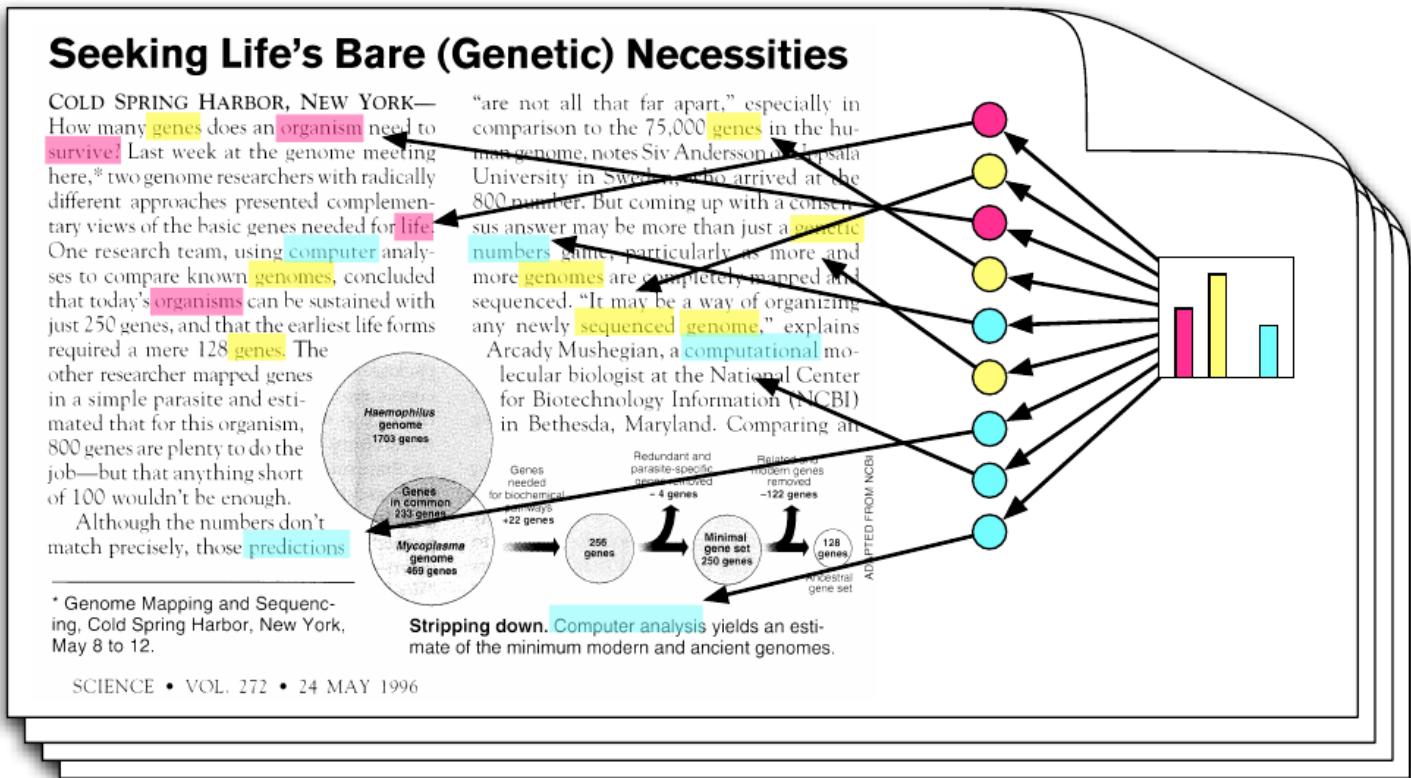
"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes 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

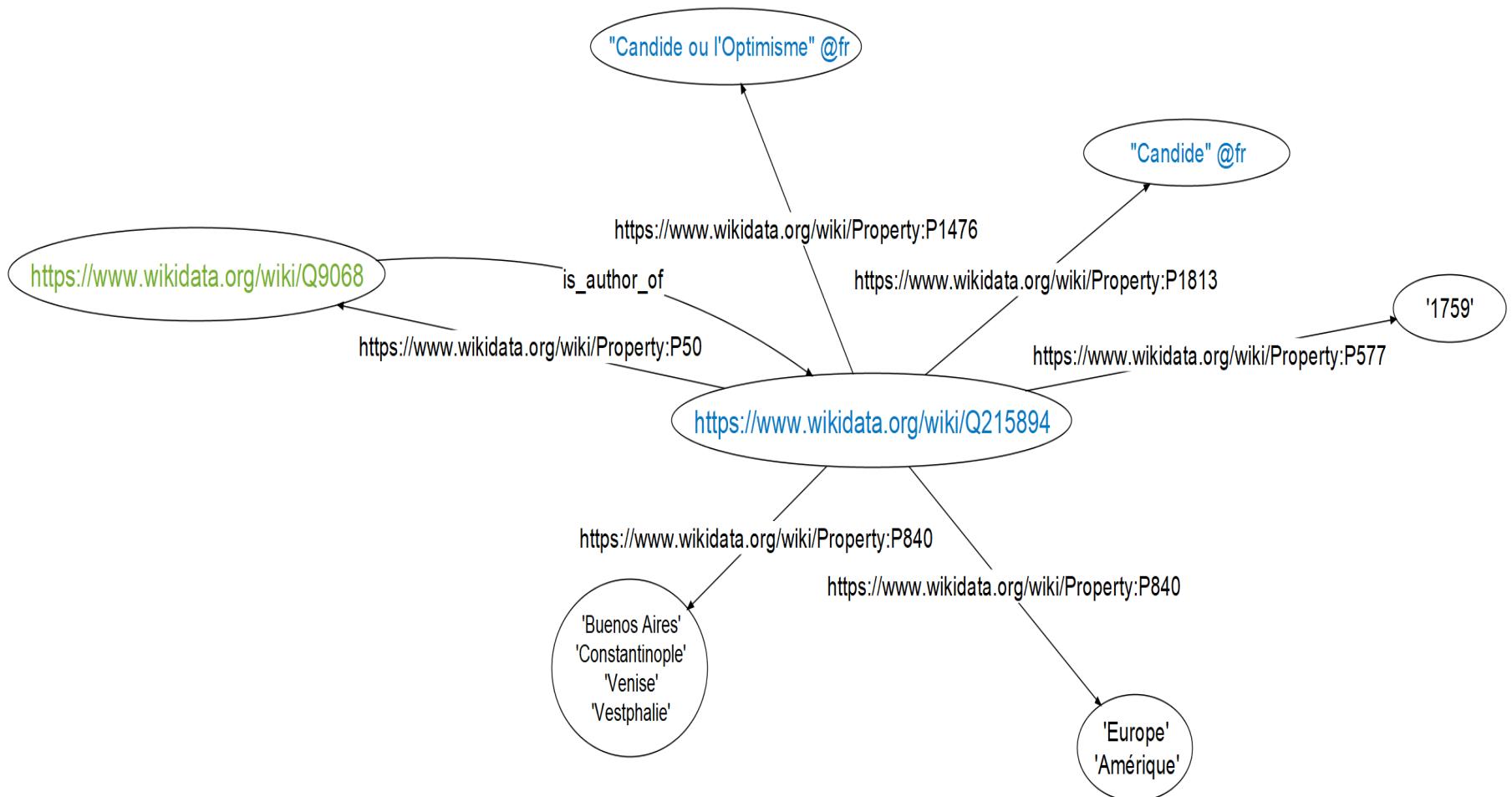
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

Topic proportions and assignments



LOD-Prinzipien: Eindeutige Identifier (mit URLs)



Aussagen in Triple-Struktur

