The French Enlightenment Novel as a Graph? Potentials and Challenges in the Construction of a Knowledge Network

Maria Hinzmann[1], Julia Röttgermann[1], Anne Klee[1], Moritz Steffes[1], and Christof Schöch[1]

[1]Trier Center for Digital Humanities, University of Trier

1 Introduction

Our contribution aims at presenting different dimensions and lessons learned in the ongoing construction of a knowledge network for literary history based on the Linked Open Data paradigm (Berners-Lee, 2006; Chiarcos and Pareja-Lora, 2020). Our specific subdomain in the current project phase is the French novel of the second half of the 18th century (Delon and Malandain, 1996; Mylne, 1981) which we investigate and represent using three different types of information sources: 1. bibliographic metadata, 2. textual features from primary sources, 3. statements from scholarly publications. In order to illustrate the potentials, but also to discuss the challenges of this approach as concretely as possible, we focus on one particular issue, namely the mining and modeling of narrative locations.[1] Specifically, we show: (a) how we extract data on narrative locations from two of the three information sources, (b) how we make them queryable via the infrastructure of a Wikibase instance using a SPARQL endpoint, and (c) how such a graph-based approach to literary history opens up new perspectives on the French Enlightenment novel.

2 Approaches to mining and modeling 'narrative locations'

The identification of place names in literary texts poses several challenges. In light of our experience, the challenges discussed in previous research can be divided into four main – partly overlapping – dimensions:

1. Ambiguity: a) terms like 'Providence' or 'Hope' can represent place names, but also non-spatial concepts; b) the existence of homonymous locations, like Athens in Greece and Athens in the United States; c) places change their names throughout history (see 4) (Heuser et al., 2016; Jockers, 2016; Nielsen, 2016).

2. (Geo-)Localization and fictionality: Nielsen (2016) points out that a) named geographical places do "not necessarily associate with a geographical position" and b) fictional places are "not necessarily not geolocatable". In addition, there are c) generic places that serve as narrative locations (e.g. ‚monastery', ‚island').

[1] See https://mimotext.uni-trier.de/ for details on "Mining and Modeling Text".

3. Extraction of ‚place names' and modeling ‚narrative locations': In our context, there is a further need for differentiation, which is specific to fictional narrative texts: a) Not all place names identified by named entity recognition describe the narrative location as the "the setting of the ongoing action" or "primary/first level narration" Pier (2014), but they can also concern flashbacks and flashforwards; b) place names can function as elements contributing to a discourse of a place as Heuser et al. (2016) point out. We would add that c) further narrative functions could be differentiated – also with regard to 'generic places' and 'cultural reference points'.

4. Historicity: Furthermore, special challenges arise with regard to spatiotemporal entanglements and vague or shifting spatial references. In this respect, we underline that our epistemological interest is not a historiographical one.

Our experiences and approaches largely correspond to the findings of previous research, but are also specific in some points. This is what we would like to outline in the following regarding firstly our corpus of novels and textual features of these primary sources (Klee and Röttgermann, 2020; Röttgermann, 2021) as well as secondly rich bibliographic metadata as two different sources of information. For our domain, the Bibliographie du genre romanesque français 1751-1800 (BGRF), published in 1977 by Mylne, Martin, and Frautschi, is central, as it defines the population of about 2000 French novels published in the second half of the 18th century.

We thus have the advantage of disposing of the bibliographic metadata that contains information about narrative locations and helps us contextualize the results of the named entity recognition on the full texts. This concerns an important point of our handling of ambiguity; we illustrate our approaches and strategies with regard to the further challenges differentiated above in the next section.

3 'Narrative locations' in the French Enlightenment Novel

We discuss our approach to integrating statements about 'narrative locations' into a knowledge network – starting with mining aspects, followed by an exposition of our modeling decisions, and concluding with information about the infrastructure.

3.1 Mining

A corpus of eighteenth-century French novels (Röttgermann, 2021) is analysed via SpaCy's named entity recognition pipeline using the French language package "fr core news lg". Named entity recognition (NER) is a popular information retrieval technique "to identify and segment named entities and classify or categorize them under various predefined classes" (Sarkar, 2019). Within the French language package of SpaCy one can distinguish the following types of named entities: LOC, PER, MISC and ORG entities.[2]

The five most common "LOC" (location) entities within each novel and their numerical occurences per text are extracted. We use the tool OpenRefine[3] to further process the results, in particular data cleaning and matching the extracted strings to Wikidata objects (see figure 1). The matching process is partly automated but requires human intervention to disambiguate between the possible Wikidata matches.

[2] SpaCy: https://spacy.io/; French language model: https://spacy.io/models/fr, package we used: https://github.com/explosion/spacy-models/releases/tag/fr_core_news_lg-3.1.0 (21.10.2021)

[3] See https://github.com/OpenRefine/OpenRefine

Our second relevant source, the BGRF, has been extensively analysed and modeled according to current bibliographic standards (including full-text digitization and semi-automatic encoding with a machine learning classifier – Conditional Random Fields) by Andreas Lüschow (Lüschow, 2020). This can then be followed by a more detailed semantic modeling of the statements derived from the different keywords, starting with 'thematic statements' (Schöch et al., 2022) and currently focusing on 'narrative locations'. As in the case of the NER results, we were able to partially automate the matching process of the keywords using OpenRefine. In both cases, a controlled spatial vocabulary was being used and expanded step by step.[4]

All	Column 1	bgrf	LOC1	Wikidata Label fr	URL	Quantity LOC1
1.	Abbes_Voyage	58.5	Palais	palais	https://www.wikidata.org/wiki/Q16560	3
2.	Anonym_Suzon	83.9	Couvent	couvent	https://www.wikidata.org/wiki/Q1128397	21
3.	Anonyme_Zoloe	00.37				
4.	Arnaud_Epoux	83.15	Paris	Paris	https://www.wikidata.org/wiki/Q90	27
5.	Arnaud_Matinees	99.43				
6.	Arnaud_Sentiment	70.21	Nancy	Nancy	https://www.wikidata.org/wiki/Q40898	111
7.	Barthelemy_Voyage	88.27	Grèce	Grèce	https://www.wikidata.org/wiki/Q41	440

Figure 1: We use OpenRefine to reconcile our entities with Wikidata.

3.2 Modeling

A fundamental aspect of our approach involves importing both the original text strings as well as (if possible) matched items (i.e. Wikibase:Items linked to Wikidata identifiers). This, in combination with referencing all sources via the property "stated in", enables a clear traceability of the origin of each individual statement, which seems to us to be particularly important due to our semi-automatic approach (see figure 2).

Our spatial vocabulary was built up incrementally. Each concept is represented multilingual (FR, EN, DE) as well as – if available – by the geodata that we were able to add via matching in OpenRefine using the property 'coordinate location'.[5] With respect to the georeferencing, we have chosen a way that does not overstate the problems outlined in 2. In this sense, we included 'generic spatial concepts' in our vocabulary which allows matching with non-georeferenceable spatial concepts having Wikidata-Identifiers and we also decided to reuse the Wikidata property 'narrative location' as directly as possible.[6] We consider Wikidata as a kind of hub and focus currently on linking to Wikidata items. In relation to the need for differentiation outlined in point 3, we have, as mentioned above, opted for a very broad notion of 'narrative location'. With regard to the challenges outlined under point 4, it can be noted that in the case of established historically defined spatial concepts (e.g. Constantinople), we have resorted to matching them with Wikidata identifiers.

[4] See https://github.com/MiMoText/vocabularies/blob/main/Raumvokabular.tsv.

[5] See https://www.wikidata.org/wiki/Property:P625.

[6] See (Nielsen, 2016), https://www.wikidata.org/wiki/Property:P840 and https://www.wikidata.org/wiki/Q44613 for 'monastery'.

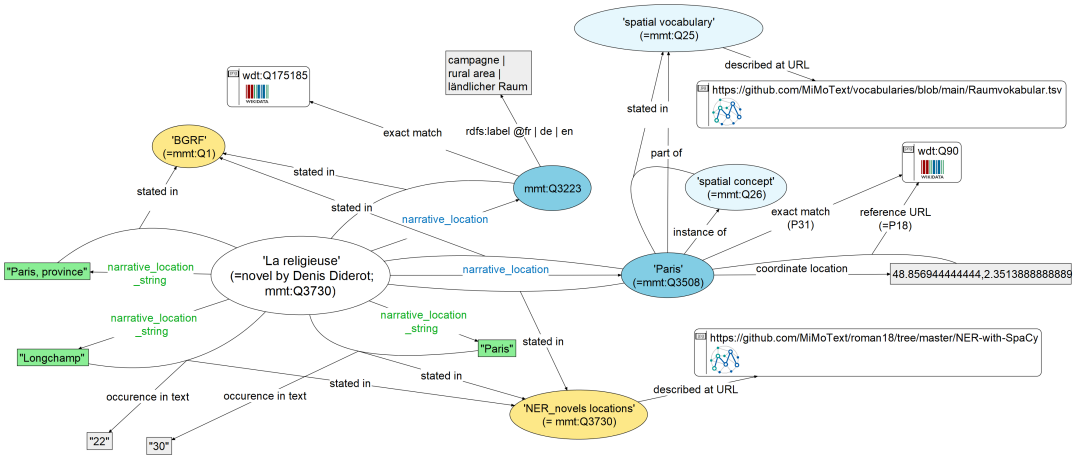

Figure 2: Our data model for 'narrative locations' from NER (novels) and bibliographic data.

3.3 Infrastructure

For the provision of data as well as the infrastructure in the project, Open Science principles are fundamental. This concerns, among other things, the publication of FAIR data (Röttgermann and Schöch, 2020; Schöch, 2021) as well as the use of open source tools and software – in particular Wikibase.

We use the Python library Pywikibot for inclusion of the RDF (Resource Description Framework) triples in our Wikibase instance. Pywikibot is a tool for automating work on a MediaWiki.[7] For this process, we have developed an individual bot that allows us to easily import and update our data via TSV files. In order for this script to work, each TSV file has a self defined header. This header makes the import expandable. It is possible to add new properties, items and statements without further expenditure. The script can be found on our GitHub Repository Wikibase-Bot.[8]

4 Knowledge Graph 'in action'

Having all our data triples on French Enlightenment novels and authors stored in our Wikibase instance enables us to query it as a graph. The DockerWikibaseQueryService interface provides numerous visualization options that allow exploration and analysis of data at different levels of granularity. Earlier results of an analysis of patterns of thematic clusters via Topic Modeling (Klee and Röttgermann, 2020) and metadata on publication date and narrative forms (Lüschow, 2020) are already stored in this triplestore. An overview of the narrative locations of the bibliographic metadata of approximately 2000 Enlightenment novels included in our database shows that they are often set in Paris and/or France as well as in rural settings (see figure 3).

In our knowledge network, we can use various queries to explore the narrative locations in more detail. Thus, we can for instance analyze the emergence, decrease, or continued relevance of certain locations over the course of 50 years and relate it to previous research.

---

[7] See concerning Wikibase https://wikiba.se/ and with regard to the bot https://www.mediawiki.org/wiki/Manual:Pywikibot/de.

[8] See https://github.com/MiMoText/Wikibase-Bot.

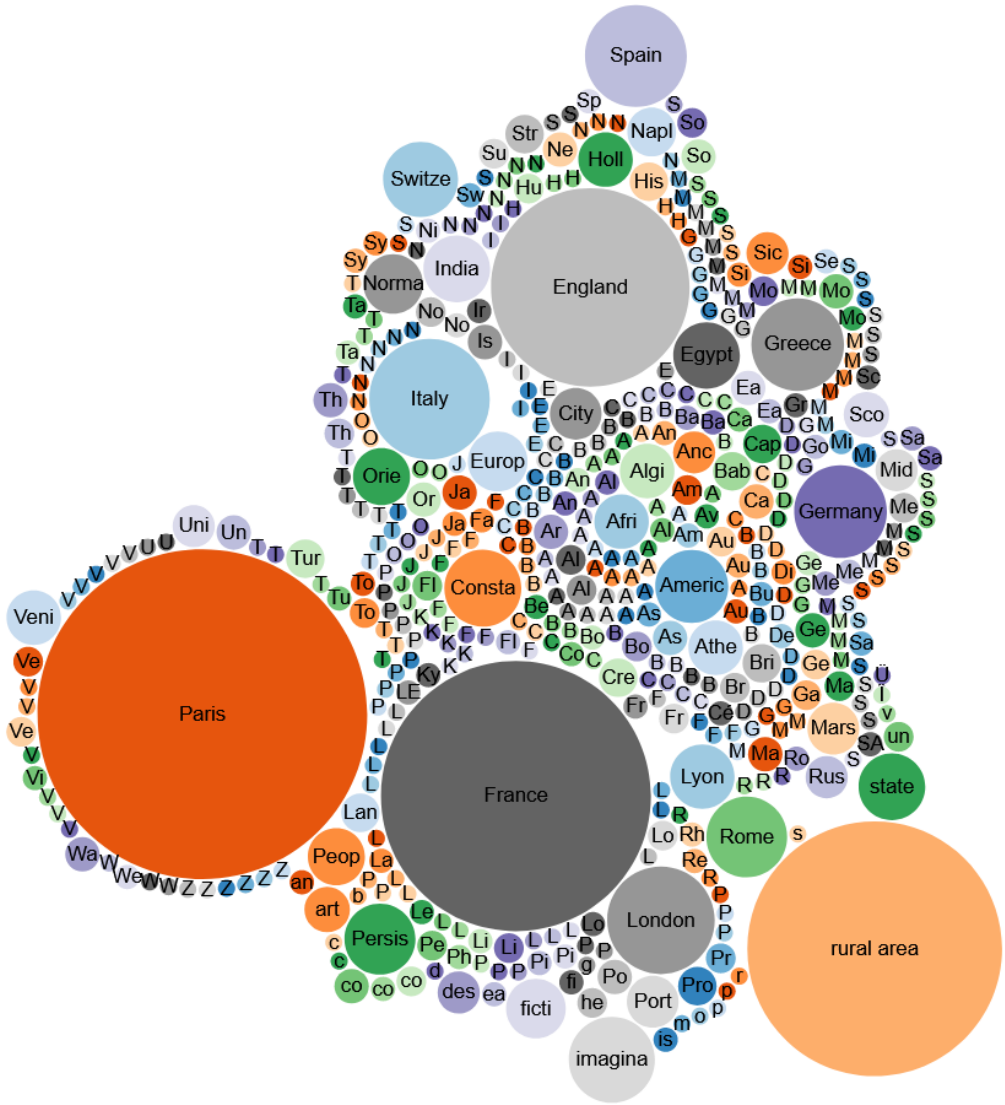

Figure 3: Overview of the narrative locations of about 1700 French Enlightenment Novels.

In the following, we will consider exemplary linkages of specific narrative locations and themes in two directions – on the one hand, starting from a specific location that is linked to diverse themes and, vice versa, using a theme as a starting point. If we ask for example, whether the narrative location "rural area" is linked to certain thematic concepts, we see that this is the case for sentimentalism, gallantry, sentiment and unhappiness (see figure 4).

Conversely, we can gain an overview of which narrative locations are linked to a specific thematic concept. In this way, we can further explore the narrative locations of those novels in our corpus that are about 'miracles' (see figure 5). Strikingly, these places of action are located in spatial (Orient, India, China, Constantinople) or temporal distance (e.g. Roman Empire, Ancient Greece, Persis) to eighteenth-century France or are explicitly named as 'fictional spaces' (imaginary place, fictional island, fictional country).[9]

[9] See Annex / SPARQL-Query 2.

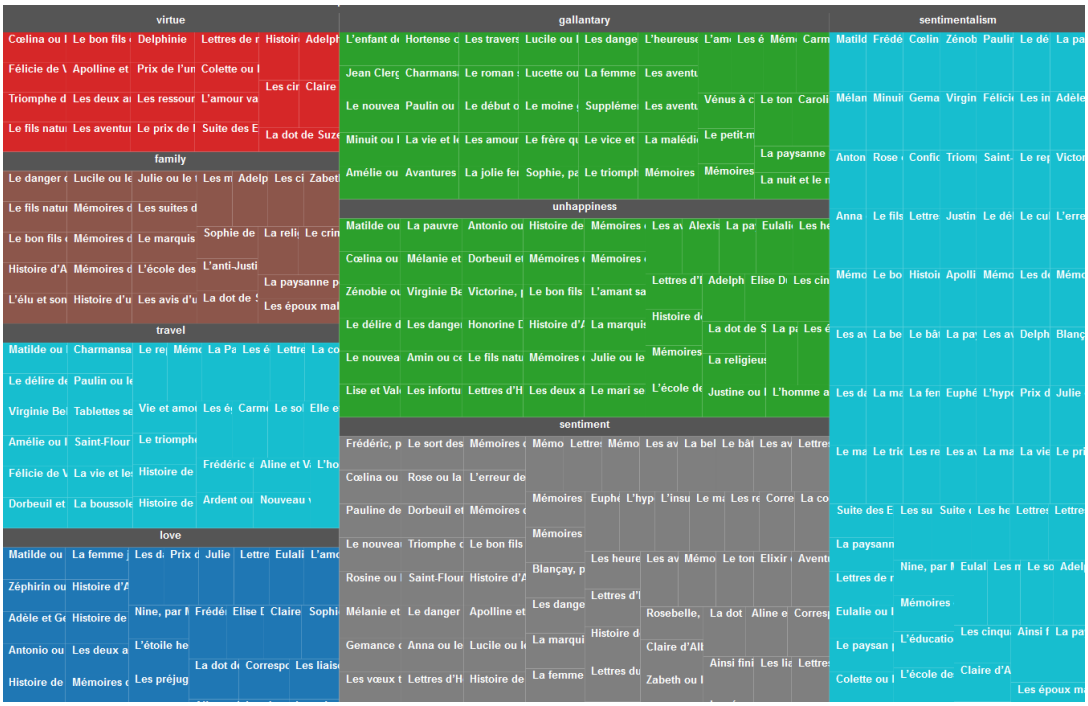

Figure 4: Thematic concepts linked to the narrative location “rural area” [extract].

5 Conclusion

The ’spatial turn’, which has been discussed in the social and cultural sciences for quite some time (Günzel, 2017), has been playing an increasingly important role in literary studies for about two decades. This concerns the categories ’space’ and partly also ’movement’ in literary texts (Hallet and Neumann, 2009) and especially also in the context of a ’spatial narratology’ (Dennerlein, 2009). We believe that our approach of providing statements about narrative location within the framework of a data-based literary history opens up new perspectives (Schöch, 2021).

With regard to the challenges and differentiation options worthy of discussion (see 2), we have accepted an intentional loss of information and complexity reduction at the level of the individual information sources, which however results in a gain of useful information and thus also potentially new knowledge in the use of our MiMoText graph. In the sense of the Linked Open Data paradigm, the project-specific knowledge graph does not stand on its own, but will be linked to the Wikidata graph (Schelstraete and Van Remoortel, 2019). After the publication of our prototype (public local Wikibase instance), it is currently planned to add all work and novel items that have not yet been defined in Wikidata and then link them to our local instance via a MiMoText ID.

Ultimately, we see the potential of our knowledge network of the French Enlightenment novel on three levels: (a) We hope that the community of literary scholars will gain new perspectives on the specific subject matter through our integration and aggregation of a relatively large amount of data for our domain. (b) Our approach is transferable to other sub-disciplines of literary history (e.g., German-language modern drama), but also to other disciplines in the humanities (philosophy, history, or cultural studies). (c) Due to our Wikibase infrastructure and the attention to connectivity and

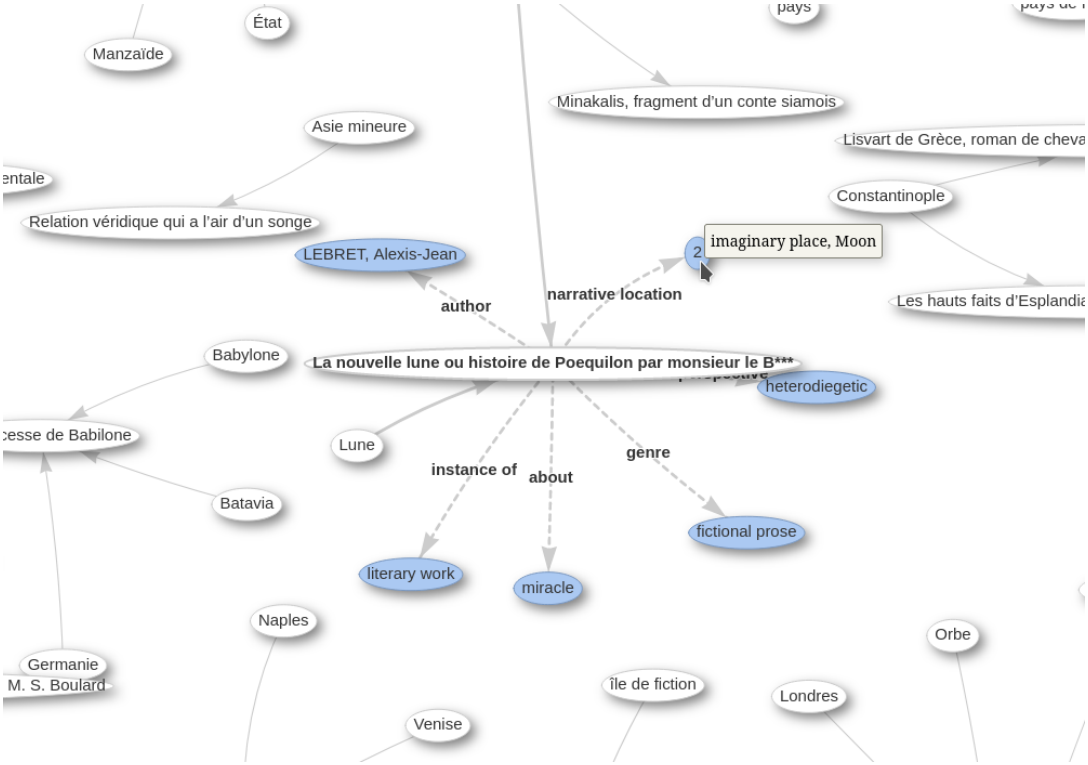

Figure 5: Excerpt from the graph (narrative locations linked to the thematic concept "miracle").

reuse of Wikidata, our data is also usable outside the literary history community and can be linked to items and knowledge particles in other projects.

6 Annex: SPARQL Queries

6.1 SPARQL-Query 1 - thematic concepts per location ('rural area')

```
SELECT
   ?value ?valueLabel ?item ?itemLabel
WHERE
{
 ?item wdt:P25 ?value. # theme
?item wdt:P52 wd:Q3223 # items with narrative location: rural area
#P52 represents 'narrative location'

  SERVICE wikibase:label { bd:serviceParam wikibase:language "en". }
  #defaultView:TreeMap
}
```

6.2 SPARQL-Query 2 - narrative locations per thematic concept ('miracle')

```
SELECT
   ?value ?valueLabel ?item ?itemLabel
WHERE
```

```
{
 ?item wdt:P52 ?value. #exploring narrative location of works
?item wdt:P25 wd:Q2970 #works with 'miracle' as theme

  SERVICE wikibase:label { bd:serviceParam wikibase:language "en". }
}
```

References

Berners-Lee, T. (2006). Linked Data - Design Issues. w3.org/DesignIssues/LinkedData.html.

Chiarcos, C. and A. Pareja-Lora (2020). Open Data—Linked Data—Linked Open Data—Linguistic Linked Open Data (LLOD): A General Introduction. In Development of Linguistic Linked Open Data Resources for Collaborative Data-Intensive Research in the Language Sciences, pp. 1–17. Boston: MIT Press.

Delon, M. and P. Malandain (1996). Littérature française du XVIIIe siècle. Paris: PUF.

Dennerlein, K. (2009). Narratologie des Raumes. Number 22 in Narratologia. Berlin ; New York: de Gruyter.

Günzel, S. (2017). Raum: eine kulturwissenschaftliche Einführung. Number Band 143 in Edition Kulturwissenschaft. Bielefeld: Transcript. OCLC: on1004222139.

Hallet, W. and B. Neumann (Eds.) (2009). Raum und Bewegung in der Literatur: die Literaturwissenschaften und der Spatial Turn. Bielefeld: Transcript.

Heuser, R., M. Algee-Hewitt, and A. Lockhart (2016). Mapping the Emotions of London in Fiction, 1700–1900: A Crowdsourcing Experiment. In Literary Mapping in the Digital Age. London: Routledge.

Jockers, M. L. (2016). The Ancient World in Nineteenth-Century Fiction; or, Correlating Theme, Geography, and Sentiment in the Nineteenth Century Literary Imagination. Digital Humanities Quarterly 10(2), §1–§32.

Klee, A. and J. Röttgermann (2020). Doing topic modeling on French 18th century novels in the context of MiMoText project [Data set: github.com/mimotext/topicmodeling].

Lüschow, A. (2020). Automatische Extraktion und semantische Modellierung der Einträge einer Bibliographie französischsprachiger Romane. In Spielräume: Digital Humanities zwischen Modellierung und Interpretation. Konferenzabstracts, Paderborn, pp. 80–84. DHd-Verband.

Mylne, V. (1981). The Eighteenth-Century French Novel: Techniques of Illusion (2 ed.). Cambridge ; New York: Cambridge University Press.

Nielsen, F. (2016). Literature, Geolocation and Wikidata. In Wiki@ICWSM.

Pier, J. (2014). Narrative Levels (revised version; uploaded 23 April 2014). lhn.uni-hamburg.de/node/32.html.

Röttgermann, J. (2021). Collection de romans français du dix-huitième siècle (1750-1800) / Eighteenth-Century French Novels (1750-1800) [Data set: github.com/mimotext/roman18].

Röttgermann, J. and C. Schöch (2020). FAIRe Daten in den Literaturwissenschaften? Das Beispiel Mining and Modeling Text und der französische Roman des 18. Jahrhunderts. blog.fid-romanistik.de.

Sarkar, D. (2019). Named Entity Recognition. In Text analytics with Python: a practitioner's guide to natural language processing (Second edition ed.)., pp. 536–557. New York, NY: Apress.

Schelstraete, J. and M. Van Remoortel (2019). Towards a Sustainable and Collaborative Data Model for Periodical Studies. Media History 25(3), 336–354.

Schöch, C. (2021). Open Access für die Maschinen. In M. Effinger and H. Kohle (Eds.), Die Zukunft des kunsthistorischen Publizierens, pp. 79–94. Heidelberg: arthistoricum.net.

Schöch, C., M. Hinzmann, J. Röttgermann, A. Klee, and K. Dietz (2022). Smart Modeling for Digital Literary History. International Journal of Humanities and Arts Computing (IJHAC) Special issue on Linked Open Data. In press.