Using Bibliodata LODification to Create Metadata-Enriched Literary Corpora in Line with FAIR Principles

Agnieszka Karlińska[1], Cezary Rosiński[2], Marek Kubis[3], Patryk Hubar[4], Jan Wieczorek[5]
[1]NASK National Research Institute, Warsaw, Poland
[2]Institute of Literary Research of the Polish Academy of Sciences, Poznań, Poland
[3]Adam Mickiewicz University, Poznań, Poland
[4]University of Warsaw, Warsaw, Poland
[5]Wroclaw University of Science and Technology, Wroclaw, Poland
agnieszka.karlinska@nask.pl, cezary.rosinski@ibl.waw.pl, marek.kubis@amu.edu.pl,
p.hubar@uw.edu.pl, jan.wieczorek@pwr.edu.pl

Abstract

This paper discusses the design principles and procedures for creating a balanced corpus for research in computational literary studies, building on the experience of computational linguistics but adapting it to the specificities of the digital humanities. It showcases the development of the Metadata-enriched Polish Novel Corpus from the 19th and 20th centuries (19/20MetaPNC), consisting of 1,000 novels from 1854–1939, as an illustrative case and proposes a comprehensive workflow for the creation and reuse of literary corpora. What sets 19/20MetaPNC apart is its approach to balance, which considers the spatial dimension, the inclusion of non-canonical texts previously overlooked by other corpora, and the use of a complex, multi-stage metadata enrichment and verification process. Emphasis is placed on research-oriented metadata design, efficient data collection and data sharing according to the FAIR principles as well as 5- and 7-star data standards to increase the visibility and reusability of the corpus. A knowledge graph-based solution for the creation of exchangeable and machine-readable metadata describing corpora has been developed. For this purpose, metadata from bibliographic catalogs and other sources were transformed into Linked Data following the bibliodata LODification approach.

Keywords: corpus, metadata, linked open data, FAIR, computational literary studies, digital humanities

1. Introduction

The practice of building corpora for literary studies is relatively nascent, and standardized procedures for curating collections of literary texts have not yet been fully developed (Gius et al., 2019). The direct application of corpus creation principles established in the field of linguistics to literary research is not an easy task, due to the significant differences between linguistic and literary corpora. The compilation of the latter is subordinated to the needs that result from the nature of linguistic hypotheses that concern phenomena from various levels of language (e.g. norms, orthography, lexis, syntax). Researchers' expectations of literary corpora concern how language (its different layers) is used to create literary content. Also, the textual material included in the literary corpus differs from the linguistic use and is usually more homogeneous by containing literary works or texts authored by literary experts. Another difference refers to the key of metadata categories selection used to describe the samples. In the case of linguistic corpora, the selection of categories is determined by the specifics of linguistic research (e.g. age of speakers, information on the periodization of language evolution, the standard of orthography or transcription used), while literary corpora require consideration of the specifics of literary research (e.g. place of publication, author's place of origin, number of issues, literary genre). The analysis of the aforementioned differences allows us to express the conclusion that both linguistic and literary corpora represent complementary resources.

Leveraging the expertise of bibliographers, we aim to move from a linguistic to a literary corpus, thereby bridging the gap between computational linguistics and digital humanities, esp. computational literary studies (Frank and Ivanovic, 2018). Based on the experience of creating Metadata-enriched Polish Novel Corpus from the 19th and 20th centuries (19/20MetaPNC), we will propose a corpus creation procedure that simultaneously provides data tailored to specific research needs and ensures reusability.

This paper makes several notable contributions. The first is the corpus itself, one of the first publicly available resources of its kind for Polish. It was created with very strict balancing criteria, taking into account the cultural context and drawing on the expertise of linguists on the one hand, and bibliographers and literary scholars on the other. Another significant contribution is the introduction of a comprehensive workflow for enriching and linking the metadata of a corpus of literary texts, which includes the implementation of the FAIR principles[1] developed in

[1] `https://www.go-fair.org/`

LREC-COLING 2024, pages 17271–17284
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

the context of scientific data management in corpus work. Additionally, we demonstrate how to present and share data following 7-star Linked Data Service standards[2], which extends the 5-star Open Data model proposed by Tim Berners-Lee[3].

We provide a Semantic Web-ready solution for creating exchangeable and machine-readable metadata for describing corpora, specifically adapted to the needs of a diverse scholarly community. We recommend employing the Resource Description Framework (RDF) as the metadata format for knowledge representation within a linked data environment, alongside the use of the Neo4J graph database management system. This software facilitates the storage and dissemination of data in a manner that is aligned with Semantic Web principles. We draw on the practice of bibliodata LODification, i.e. the conversion of publication metadata from bibliographic catalogs and other sources into Linked Open Data (LOD) (Lindemann, 2022; Lindemann et al., 2023). However, unlike the proponents of this approach, we recognize LODified data not only as publication metadata, citation relations, content-describing subject headings, or keywords but also as content-extracted data, both automatically and manually.

2. Related work

The linguistic literature demonstrates that when designing a corpus, its purpose plays a pivotal role—specifically, its adaptation to address particular research inquiries—as well as the methodology used for data selection. Apart from the corpus size, which depends on factors such as the linguistic aspect under investigation, the diversity within the studied language variety, and the frequency of text repetition within a given type or genre (Kennedy, 1998; Biber, 1990), there are also questions of representativeness and balance (Francis, 1982; Leech, 2007). Both of these are considered desirable but challenging to attain in practical corpus construction (Biber, 1993; Hunston, 2008; Baker, 2010). Building a corpus of historical material is deemed to be particularly difficult (Schlagenhauf, 2004; Jenset and McGillivray, 2017). It places very high demands on the metadata description (Depuydt and Brugman, 2019), and the implementation of the postulate of balanced source selection faces many more theoretical and practical obstacles than in the case of corpora of contemporary texts (Egan, 2019). The fundamental and insoluble problem is the limited knowledge of the writing of a given period. While it is possible to capture the overarching trends and identify the dominant types and genres of texts, the knowledge of the structures of the documents used in that era will always remain incomplete (Gruszczyński et al., 2020). The lack of reliable population data precludes the creation of a randomly sampled, statistically representative corpus and can lead to arbitrary text selection criteria, decisions based on speculation and theoretical assumptions, and favoritism towards canonical texts (Underwood, 2016).

A number of previous diachronic literary corpus projects have sought to reduce speculation and ahistoricity by using data on the production and reception of texts from bibliographical records, or by attempting to reconstruct reading practices in a given period. Others have relied solely on contemporary expertise. In the European Literary Text Collection (ELTeC), in order to capture the diversity of literary production while ensuring comparability across subcorpora, texts were categorized based on time span, canonicity (number of reprints), author gender, length, and authorship within the collection (ELTeC, 2021). For this purpose, the creators primarily used the catalogs of the national libraries, pointing out gaps in data availability (Schöch et al., 2021). In the FRANTEXT corpus, on the other hand, works were selected according to the "principle of authority," consulting recognized syntheses of 19th and 20th century French literature and compiling lists of works mentioned (Grieve-Smith, 2010). In the case of KOLIMO, bibliographic data were considered in addition to literature on the subject. Experts in the field evaluated the relevance of individual texts (Herrmann and Lauer, 2020). In the dProse corpus based on KOLIMO, metadata were verified and enhanced using repositories and literary encyclopedias (Gius et al., 2021).

Currently, there is a lack of a comprehensive and balanced corpus of Polish novels. Most Polish literary corpora are designed to meet specific research needs, often do not adhere to standard composition criteria, and do not provide thorough metadata descriptions. An exception is the Polish ELTeC subcorpus of 100 novels published between 1840 and 1920, according to the project's assumptions (Frontini et al., 2020). However, there are concerns about the inclusion of texts published after the specified period and the absence of a well-defined procedure for the acquisition of metadata, which is essential for assessing the quality of the corpus.

The authors' evaluation of the existing corpora reveals a common problem: insufficient metadata validation. This stems from an over-reliance on institutional sources for metadata and a presumption of their accuracy, coupled with limited opportunities for cross-referencing and validating metadata against other sources, primarily due to the underutilization of LOD, resulting in source isolation.

The resource gaps identified above, as well as

fair-principles

[2] https://www.ldf.fi

[3] https://5stardata.info/en

new initiatives in the area of metadata enhancement within textual corpora through the incorporation of external data, influenced our decision on the form and features of the 19/20MetaPNC. The initiative of Workset Creation for Scholarly Analysis (WCSA) directed by HathiTrust shares both methodological and technological similarities with our project (Jett, 2015). A distinctive aspect of WCSA is its use of HTRC worksets—user-compiled collections of volumes from the HathiTrust digital library, designated for data analysis via a diverse collection of HTRC tools and services. These worksets are a fundamental component for all analytical activities within HTRC Analytics, offering a platform for collaborative and referenced research, thereby advancing reproducibility. Similarly, the present project of 19/20MetaPNC creation also commences on the task of expanding string-based metadata with Uniform Resource Identifiers (URIs), a step towards refined data discovery and interoperability with external services (Jett et al., 2016).

3. Corpus design

3.1. Design principles

In designing the corpus, we adopted assumptions developed in the field of linguistics and benefited from the experience of creators of other literary corpora. The overall research goal was to trace the impact of historical and spatial factors on the dynamics of literary processes. We formulated specific research questions, focusing on the transformation of the urban-rural dichotomy in Polish fiction (Karlińska et al., 2022).

We assumed that the corpus would contain 1,000 texts. This number is large enough to enable distinguishing subcorpora and conducting computational literary studies. Like the authors of ELTeC (Schöch et al., 2021), we decided to maintain genre and language homogeneity. We included in the corpus only novels originally written in Polish (thus rejecting translated texts) and first published in book form between 1864 and 1939. In this way, we took into account the three distinct periods of Polish literary history—Positivism (1864–1890), Young Poland (1890–1918), and the interwar period (1918–1939)—and made it possible to carry out comparative analyses.

We aimed for representativeness and balance. We relied on texts available in digital form, drawing from a variety of sources and assessing the quality of the data. Due to the lack of complete bibliographies and information on the literary production and reception of the period, we could not define the population precisely, including both all published novels and their authors. This limitation made it impossible for us to assess the representativeness of the data. Therefore, we focused on ensuring maximum balance by relying on a broad set of metadata. We used ELTeC as a model and also drew inspiration from KorBa (Gruszczyński et al., 2022), a corpus of Polish texts from the 17th and 18th centuries. KorBa employed a sophisticated selection procedure and categorized texts based on chronology, genre, subject, and, unusually for corpora, geographical origin (Gruszczyński et al., 2020).

The distinctive geopolitical and socio-cultural context of the Polish territories during the second half of the 19th century and the first half of the 20th century played a central role in shaping the metadata structure, content, and criteria for balancing the corpus. In the late 18th century, Poland ceased to exist as a sovereign state and was partitioned, falling under the rule of the Habsburg Monarchy, the Kingdom of Prussia, and the Russian Empire until 1918. These partitions showed significant differences in terms of socio-economic development, urbanization, and civil liberties (Kaczynska, 1970). An additional criterion for text selection was the time frame of the narrative, which was set no earlier than 1815, the year of the Congress of Vienna, which established national borders that remained largely unchanged for over a century.

Following a methodology similar to ELTeC, we decided to include in our corpus both novels considered part of the contemporary canon and those that have fallen into relative obscurity. To measure their reception, we took into account the number of reprints of a given publication.

Given the challenge of maintaining a balance between classes, we defined each text class's minimum and maximum representation in the corpus. We established the following criteria:

1. **Date/literary period:**

   Positivism (1864–1890) >= 20%

   Young Poland (1890–1918) >= 20%

   the Interwar Period (1918–1939) >= 20%

2. **Gender:**

   female author 10%–50%

3. **Place of publication:**

   Austrian partition >=15%

   Prussian partition >=15%

   Russian partition >=15%

4. **Level of reception:**

   no more than 2 reprints >= 30%

   more than 2 reprints >= 30%

3.2. Data collection

We examined four open sources of 19th and 20th-century Polish prose with the goal of collecting texts for our corpus:

1. **ELTeC corpus** (ELTeC, 2021) which contains novels encoded in the TEI format.
2. **Wolne Lektury** (Modern Poland Foundation, 2022), an online repository that is mainly oriented towards collecting school readings and offering them in reader-friendly data formats.
3. **Polish edition of the Wikisource** (Wikimedia Foundation, 2022) project which includes transcriptions of printed books that have fallen into the public domain encoded in the MediaWiki format.
4. **Polona**, a digital library maintained by the National Library of Poland (2022) that provides scans of printed books.

Our initial dataset consisted of 100 Polish novels from ELTeC, 193 literary works from Wolne Lektury, 225 texts from Wikisource, and ca. 6,000 volumes from Polona. It has to be noted that the collected texts vary greatly in quality. Novels published by Wolne Lektury were thoroughly edited and contemporized. ELTeC texts retained original (historic) spelling and punctuation, but the hyphenated words were merged in transcription. In the case of Wikisource, not only spelling and punctuation is preserved, but also hyphenation is kept in the original form. Digital copies of physical books provided by Polona contain OCR-derived textual layers only. Hence, they contain errors introduced in the process of optical character recognition and retain spelling, punctuation, and hyphenation of the physical copies. In order to make the texts from all the sources more uniform we cleaned OCR-related errors and normalized punctuation and hyphenation with the use of custom scripts adapted from (Kubis, 2021). We also utilized a diachronic normalizer (Jassem et al., 2017; Dudzic et al., 2024) to modernize spelling.

3.3. Metadata description

The metadata that we have used to describe the corpus defy the traditional model of archiving and sharing collections, familiar from domain bibliographies or library catalogs. None of the common bibliographic data formats are comprehensive enough to include information crucial to our research, such as the geographic coordinates of the places described in the novels or the attribution of the partitions in which they were published. We use the information in the catalogs for research, the original purpose of which is to collect and preserve the textual production. Our activity is what Foulonneau and Cole (2005) refer to as "the process of adapting metadata for another application than originally envisioned when the metadata records were created". The phenomenon we are dealing with is metadata repurposing, understood as the use of data in a new context not originally intended (Deng, 2010). Hence, the metadata we extracted from the National Library of Poland catalog was the author's name, the title of the book including subtitle, place and year of publication, and genre.

It is worth noting that a source of enriching information about individual texts beyond the content of the catalogs and an opportunity to increase corpus research potential is the use of NLP techniques. The 19/20MetaPNC benefited from Named Entity Recognition (NER), which made it possible to label all the places that appear in the texts of the novels as settings. For this purpose, we used the PolDeepNer2 system and its pre-trained model, learned from the KPWr corpus (Marcińczuk et al., 2018; Marcińczuk and Radom, 2021).

However, neither the arbitrary metadata notation we initially adopted, nor the available and widely used metadata formats facilitated the scientific reuse of the literature corpus due to a lack of interoperability. This was only made possible by the use of LOD structures, which allowed us to identify, harmonize, and enrich the original metadata. This type of intervention places corpus-building activities close to the achievements of the FAIR principles and the 7-star Linked Data Service. Thus, the metadata obtained through LODification were the author's persistent identifier (PID), place of birth and gender, and place of publication PID. For both types of places we acquired geographic coordinates.

Conducting NLP-based enrichment alongside PID and LOD enrichment, we developed a four-stage toponym disambiguation workflow to identify and standardize geographic entities (geo-entities). The workflow utilized leading approaches in Geographic Information Retrieval (Buscaldi, 2011; Derungs and Purves, 2014) and primarily relied on the Geonames database. The first stage involved identifying historical place name variants through knowledge-based methods with historical registers, directories, and dictionaries used as data sources. In the second stage, records from the Geonames database were assigned to the identified geo-entities using a list of historical name variants. When multiple records were found (e.g., Paris as the capital of France or Paris as a Polish village) we filtered search results based on the population. The third stage aimed to determine whether the identified names referred to cities or villages. The final stage focused on determining the partition in which a village or city was located. Historical maps of Polish territories after the post-1815 partitions were used, along with georeferencing through software like QGIS and OpenStreetMap resources. This stage involved plotting polygons on the maps corresponding to each partition to determine precise border coordinates and subsequently assign geo-entity affiliations to specific partitions.

We then performed a semi-automated metadata verification to ensure that the collected texts met the project's criteria. For genres, manual verification was required, as the bibliographic information was not always accurate. At the same time, as a result of automatic and manual work, we have added the following attributes to the description of the corpus: literary period, assignment to the partition on the basis of the place of publication, and the time of the novel's action (before or after 1815). For the latter, we considered information from the titles and subtitles obtained from the National Library, details found in the opening pages of the relevant book regarding the time of the narrative, as well as information from subsequent pages, including descriptions of historical events or technological advancements not feasible before 1815. As a result, we obtained a collection of 2,927 novels. After thorough deduplication, which included author names, titles, and years of first printing, we obtained 1,707 novels, from which we randomly selected 1,000 based on the adopted balancing criteria.

In conclusion, two types of metadata have been used to describe the corpus: the first type is produced from the perspective of the needs of the information systems so that it is primarily used for knowledge retrieval and only secondarily can be used for research purposes. The other type—which we call research question-based metadata—is produced in a specific research process and is originally used to answer the problems posed by the researchers, but can be used to improve information retrieval. Additionally, the second layer of metadata understanding has been produced and includes metadata obtained from library catalogs, manual completions, NLP-based extractions, and reconciliations through LODification.

3.4. Data publication

One of the primary objectives of publishing corpora is to facilitate their long-term utility and reusability for diverse scholarly applications. In line with this goal, the 19/20MetaPNC corpus has been shaped by the principles of 7-star data and FAIR. Rather than employing conventional storage as static files in a data repository, the corpus is made available through direct download links. This ensures that users always have access to the most current dataset, as files are retrieved directly from the publisher. To further enhance accessibility, Python code is provided to simplify the downloading process. Complementing these measures, texts in the 19/20MetaPNC corpus have been purposefully selected to fall under public domain licenses, permitting unrestricted access, use, and distribution. Each step of corpus development, including design decisions, balancing criteria, statistical metrics, and Python code, has been comprehensively documented and made publicly available in an open GitHub repository [4]. These decisions, taken collectively, result in a corpus designed to optimize accessibility, interoperability, and long-term reusability.

Providing such a structured corpus allows detailed exploration of the data, mainly through the application of graph visualizations, advanced and precise filtering of results, and the use of complex queries in the SPARQL language. This approach not only enriches the data retrieval process, but also highlights the semantic relationships between the data. Most importantly, this form of knowledge sharing facilitates the formulation of new research questions, promoting a continuous cycle of inquiry and discovery within the academic community.

3.5. Literary corpus creation workflow

Based on the experience of creating a 19/20MetaPNC corpus, we propose a workflow for the creation and reuse of a meta-corpus based on the two perspectives of metadata generation and metadata use, as well as the guidelines for sharing text collections as described by the FAIR and 7-star standards. This workflow consists of nine consecutive stages.

1. **Research question-based design:** Metadata design should encompass both general-purpose metadata derived from existing cataloging information and metadata tailored to address specific research questions. In corpus design, it is imperative to define the target population and establish criteria for text inclusion, as well as balancing criteria.
2. **Data collection and reuse:** To avoid redundant work, the corpus creators should make use of existing textual resources and metadata databases, while complementing them with any essential elements that may be unavailable in the digital environment.
3. **Data evaluation and preprocessing:** All collected resources should undergo quality assessment and subsequently be standardized in terms of metadata consistency, data types, and formats. This may involve tasks such as OCR error correction and diachronic normalization.
4. **NLP-based enrichment:** At this stage, techniques such as NER or topic modeling can be used to enhance the text description. This newly acquired data, together with metadata, can then be used to balance the corpus.
5. **PIDs and LOD enrichment:** Unique identifiers should be assigned to the data to ensure

[4] https://github.com/CHC-Computations/19-20MetaPNC

accessibility and interoperability with authoritative databases. Adherence to LOD structures can improve the comprehensibility and reusability of the corpus.

6. **Metadata verification and completion:** Metadata should be rigorously validated to ensure both accuracy and completeness. The evaluation process includes quantitative and qualitative analysis, preferably conducted by domain experts. In cases where gaps are identified, it may be necessary to supplement the metadata.

7. **Balancing:** Based on the adopted balancing criteria and the collected metadata, along with the data extracted from the texts, the corpus should be balanced. In cases where there are significant discrepancies between different text classes, sampling may be a necessary step.

8. **Semantic environment and ontology building:** During this stage, relationships between various (meta)data elements should be defined, establishing a structured framework for analysis. It is recommended to store the information in a graph-based database, allowing for efficient querying and analysis while preserving the semantic relationships. This structured approach facilitates a comprehensive analysis and the potential for uncovering new insights.

9. **Publishing:** The final stage involves publishing the corpus in a way conducive to computational literary studies, making it accessible through services and repositories that facilitate resource reuse. The published corpus should be accompanied by comprehensive documentation describing aspects such as licenses, creators, and standards employed.

4. Dataset in numbers

The proportions between text classes were consistent with the assumed balancing scheme described in 3.1. Women were the authors of 29.4% of the novels in the corpus, well above the assumed minimum of 10%. The collection is dominated by texts of low reception with no more than 2 reprints (63.3%). Among the partitions, the Russian partition is predominant (58.5%), while the Austrian (19.7%) and Prussian (15.4%) partitions are less represented. Novels published in foreign centers account for 4.4%. It should also be noted that 20 novels (2%) were published simultaneously for the first time in two places belonging to different partitions. In terms of dates of publication, the least represented group are texts from the period of Positivism (20.7%). 39.1% of the works were published in Young Poland and 40.2% in the Interwar period. The number of titles for each balance criterion is shown in Fig.1.

The novels in the corpus were written by 390 different authors, including 111 women and 279 men. On average, there were 2.56 novels per author, with a median of 1 and a std=3.63. A total of 14 authors, including 5 women, exceeded the number of 10 novels, and as many as 227 authors are represented by a single text in the corpus.

19/20MetaPNC comprises a total of 64,313,110 tokens, distributed across 4,255,570 sentences. On average, each novel contains about 64,313.11 tokens (median 55,571.50) and 4,255.57 sentences (median 3,593.50). The relatively high standard deviation indicates significant variations in the length of novels. Notably, there is a substantial difference in length between novels of high and low reception (the former being much longer). Additionally, variations in length exist among novels from different literary periods, with positivist novels being the longest and those from the interwar period being the shortest. Complete statistics for the number of tokens and sentences based on the adopted balancing criteria can be found in Tab. 1 and 2.

5. Discussion

In constructing 19/20MetaPNC, we aimed to reconcile the expectations of computational linguists regarding the structure and text selection criteria of the corpus with the needs of literary scholars and digital humanists. The novelty of our resource lies in an advanced approach to balancing that takes into account the spatial dimension, the inclusion of non-canonical texts not previously covered by other corpora, and a complex and multi-stage procedure of metadata enrichment and verification. Creating the corpus involved several challenges, primarily related to the quality of the data and metadata on which we relied.

The lack of complete population data makes it challenging to assess the representativeness of the corpus. Since it is based on texts already available in digital form, 19/20MetaPNC can be categorized as an opportunistic corpus. However, we have drawn from a variety of sources and conducted a rigorous selection of texts, achieving a good balance across four different criteria. Despite the strict requirements, we successfully managed to include the works of 390 authors, creating the largest open corpus of its kind for Polish. We also considered the geographic-political criterion—the territory of Poland in the 19th century and at the beginning of the 20th century was divided into three partitions—Prussian, Russian and Austrian. Polish literature was produced in all the partitions, but the circumstances in which the literary circuits were formed differed significantly. We have consequently

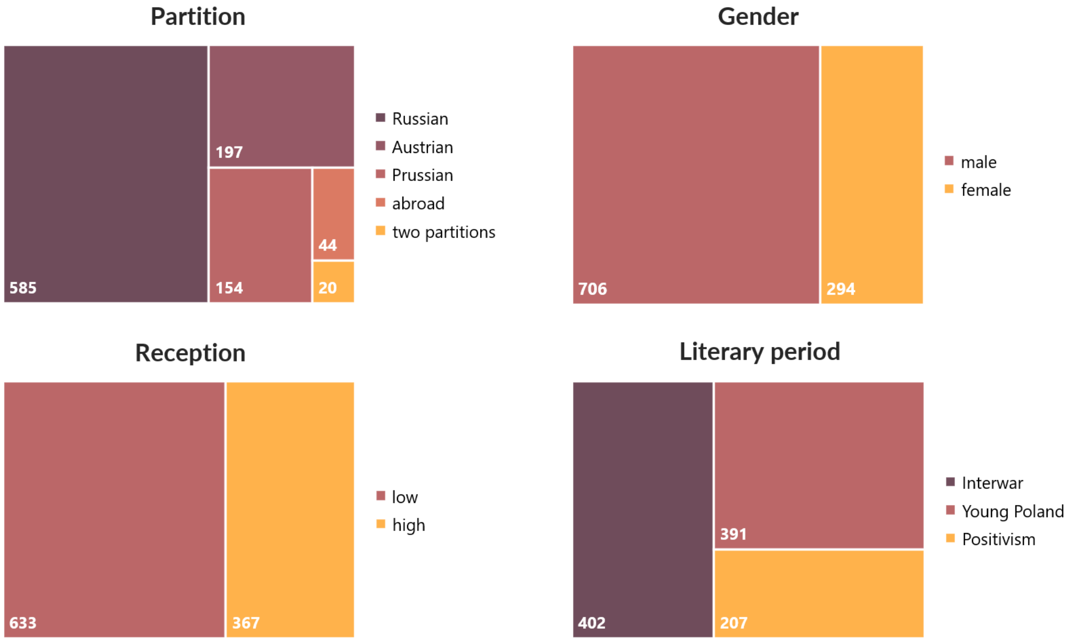

Figure 1: Treemaps of the balance criteria in 19/20MetaPNC

Table 1: Corpus statistics (number of tokens)

feature	mean	std	25%	50%	75%
In total	64313.11	42433.66	37416.25	55571.50	82112.50
Gender					
Female	61760.83	39234.87	35540.00	55868.00	78374.25
Male	65375.95	43679.45	37574.75	55548.50	83191.00
Partition					
Austrian	64563.74	34456.49	38966.00	57163.00	82349.00
Prussian	56628.18	31590.98	36142.50	49925.50	73404.50
Russian	64809.38	45321.22	35957.00	54909.00	83223.00
Abroad	75226.57	59361.35	42487.75	63759.00	84335.25
Two partitions	82492.70	46596.78	49451.00	78812.50	100881.75
Period					
Positivism	73408.68	45077.16	42589.50	63683.00	89686.00
Young Poland	66979.41	47471.30	38870.50	57330.00	83121.00
Interwar	57036.23	33875.48	35258.50	48930.00	74560.00
Reception					
High	80777.91	51066.59	46927.00	70239.00	99687.50
Low	54767.16	32974.61	32825.00	48966.00	69789.00

attempted to balance the corpus in terms of where novels were written. We have also covered works created in exile—outside the territory of the former Poland (Fig.1). At this point, the balance does not extend to the length of the texts. To prevent longer texts (e.g., multi-volume novels) from exerting a disproportionate influence on the results, it is worthwhile to use sampling. Due to the diversity of sources, we had to address issues related to different formats and text quality. A significant portion of the material was generated by the OCR process, which is prone to errors. This required extensive pre-processing and data harmonization.

A significant challenge in building a literary corpus is the need to rely on library catalogs, which are produced with remarkable regularity, but are full of omissions and inconsistent data. An additional difficulty is the inability to compare data with other sources due to the limited interoperability of library catalogs, although it should be noted that libraries' Linked Data Services are increasingly emerging. Another problem is the limitation of metadata related to the original purpose of library resources so that the richness of the description of the texts and entities present is insufficient for at least corpus balancing.

Table 2: Corpus statistics (number of sentences)

feature	mean	std	25%	50%	75%
In total	4255.57	2920.47	2340.00	3593.50	5452.50
Gender					
Female	4120.91	2849.00	2126.00	3406.50	5514.50
Male	4311.65	2949.90	2397.75	3632.50	5388.00
Partition					
Austrian	3990.22	2204.64	2439.00	3658.00	5100.0
Prussian	3867.04	2409.96	2170.00	3306.00	5107.25
Russian	4332.10	3062.80	2295.00	3546.00	5480.00
Abroad	5087.36	4387.72	2599.0	4172.50	5528.25
Two partitions	5792.50	3663.12	3247.50	5340.50	6608.75
Period					
Positivism	4020.37	2495.06	2256.50	3477.00	5274.00
Young Poland	4512.17	3402.11	2395.00	3711.00	5484.00
Interwar	4127.10	2585.39	2320.00	3475.00	5387.50
Reception					
High	5394.23	3434.85	3061.50	4613.00	6958.50
Low	3595.40	2336.69	2008.00	3101.00	4686.00

Connecting metadata to LOD repositories such as VIAF, Wikidata, Geonames, and Library of Congress Subject Headings enables cross-checking of information between sources. Discrepancies in spelling of author names across different editions of the same book can be either resolved by voting or at least detected automatically. The overlap of metadata across different LOD sources can be utilized to align book titles that have undergone diachronic changes and missing metadata required for corpus balancing can be supplied from the linked repositories. Without connections to LOD repositories, all these problems have to be resolved by hand.

19/20MetaPNC was created with the intention of not only building a textual resource but also with the conviction that a literature corpus is a data set that must be built according to FAIR principles and 7-star data. Every action taken was not only research-motivated but also aimed at reuse by other researchers, so in order to increase the visibility of 19/20MetaPNC and its reusability potential, we implemented bibliodata LODification from the beginning. Our goal was to develop a procedure that could be easily adapted to other types of literary research. In order to achieve this, it was necessary to provide a universal, open, and flexible way to add more categories of metadata according to current scholarly needs and to propose such methods of constructing them that they would be intuitive for subsequent researchers, including those who had not previously worked with the resource. Therefore, the target form of presentation of the corpus is to present the entire collection in the form of a knowledge graph.

In the process of enriching the (meta)data and converting it to formats compatible with the triple data model and semantic web, we published its metadata in two different structures: a Neo4J graph database using the Labelled Property Graph (LPG) model and RDF. The availability of 19/20MetaPNC as the knowledge graph enables data exploration through complex queries and visualizations that would otherwise be impossible. This not only allows us to seek answers to well-established research questions but also to pose new questions and uncover non-obvious connections in the data, ultimately contributing to the development of computational literary studies.

Particularly useful for our research application of the corpus was the visualization showing books published in a partition other than the author's birth partition (Fig. 2). Partitions are marked in green, authors in yellow, and literary texts in blue. Two types of relationships were included in the query: "published in" and "born in". The latter is shown in purple and bold. The query highlighted the clear disparity found in the data between publishing practices in the different partitions. Both the Russian Partition (top right) and the Austrian Partition (bottom) show a large proportion of books published by authors born in the other partitions. At the same time, a large number of authors willingly published elsewhere were born in both partitions. In the Prussian Partition, on the other hand, a significant disproportion can be seen; many books by authors born in other partitions were published here, but the number of authors originating from this partition and published in the others is much smaller. This is a valuable observation for literary scholars and historians, illustrating the different strategies of Polish-language literature under external authority

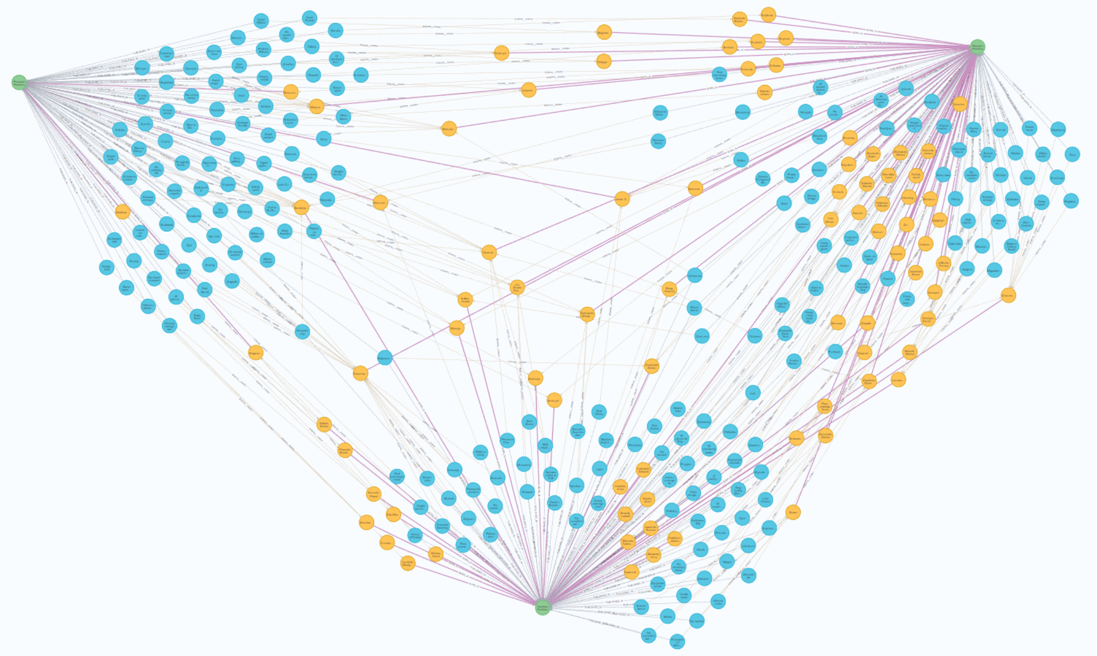

Figure 2: Query in the Neo4J graph database showing books published in a different partition than the author's partition of birth

and the possibilities for the transfer of literary texts between the partitions.

6. Conclusion and future work

The creation of a literary corpus is a partly separate endeavor from the creation of a linguistic corpus. Clear guidelines are lacking. The approach we propose, based on bibliodata LODification, will allow the design of resources that are tailored to specific usage scenarios, while at the same time enabling a much wider range of applications and stimulating new research questions. The 19/20MetaPNC project demonstrates that metadata-enriched corpora have great potential and can be an important step towards computational literary studies. Their main advantage is the ease of further extension and adaptation to different research contexts and disciplinary knowledge. Combined with an extensive balancing procedure, this will overcome the criticisms of ahistoricism and disregard of sociocultural determinants directed at the computational analysis of literary texts, while ensuring the relevance and comparability of the results.

The released resource is the first phase of a larger project. In the next iterations, we will expand the corpus to include historical novels (with a plot time before 1815). We will rebalance the corpus according to the length of the novels. We also plan to extend the metadata, in particular through NLP-based enrichment. First, we will use a combination of supervised and unsupervised approaches to determine the thematic content of each text. For this purpose, we have proposed a solution that uses a crowdsourced dataset consisting of texts labeled with literary motifs by volunteers, and enhances data quality through expert validation. By exploring how different results meet the needs of both researchers and bibliographers, we will select the optimal model.

The biggest task, however, will be the development of the Text Corpora Ontology (TCO), which has been initiated in parallel with the 19/20MetaPNC corpus. TCO will be tailored for the publication of text corpora within a Semantic Web environment, identifying objects and their bibliographic relations across written documents such as books, journal articles, and conference papers. TCO will also help to represent crucial corpus creation attributes like balance, representativeness, and relevance. TCO will integrate existing ontologies such as schema.org, FOAF[5], BiRO[6] , FaBiO[7]. Future work on the Text Corpora Ontology will focus on diversifying the representation of text corpora metadata: structurally, content-oriented, and encompassing research questions related to the corpora. This aims to standardize metadata representation, encapsulate document content, and integrate research inquiries within the ontology, facilitating a more comprehensive analysis of text corpora in a web environment.

[5] http://xmlns.com/foaf/0.1/

[6] https://sparontologies.github.io/biro/current/biro.html

[7] https://sparontologies.github.io/fabio/current/fabio.html

7. Bibliographical References

Paul Baker. 2010. *Research Methods in Linguistics*, chapter Corpus Linguistics. Continuum, London.

Douglas Biber. 1990. Methodological Issues Regarding Corpus-based Analyses of Linguistic Variation. *Literary and Linguistic Computing*, 5(4):257–269.

Douglas Biber. 1993. Representativeness in Corpus Design. *Literary and Linguistic Computing*, 8(4):243–257.

Davide Buscaldi. 2011. Approaches to disambiguating toponyms. *SIGSPATIAL Special*, 3(2):16–19.

Sai Deng. 2010. Optimizing workflow through metadata repurposing and batch processing. *Journal of Library Metadata*, 10(4):219–237.

Katrien Depuydt and Hennie Brugman. 2019. Turning digitised material into a diachronic corpus: Metadata challenges in the nederlab project. In *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage*, DATeCH2019, pages 169–173, New York, NY, USA. Association for Computing Machinery.

Curdin Derungs and Ross S. Purves. 2014. From text to landscape: locating, identifying and mapping the use of landscape features in a Swiss Alpine corpus. *International Journal of Geographical Information Science*, 28(6):1272–1293.

Kacper Dudzic, Filip Gralinski, Krzysztof Jassem, Marek Kubis, and Piotr Wierzchon. 2024. Two Approaches to Diachronic Normalization of Polish Texts. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 207–212, St. Julians, Malta. Association for Computational Linguistics.

Thomas Egan. 2019. Non-representativeness in corpora: perils, pitfalls and challenges. *CogniTextes*, 19. http://journals.openedition.org/cognitextes/1772. Accessed: 2023-10-17.

Muriel Foulonneau and Timothy W. Cole. 2005. Strategies for reprocessing aggregated metadata. In *International Conference on Theory and Practice of Digital Libraries*, pages 290–301, Berlin, Heidelberg. Springer.

Nelson W. Francis. 1982. Problems of assembling and computerizing large corpora. In S. Johansson, editor, *Computer corpora in English language research*, pages 7–24. Norwegian Computing Centre for the Humanities.

Andrew Frank and Christine Ivanovic. 2018. Building Literary Corpora for Computational Literary Analysis – A Prototype to Bridge the Gap between CL and DH. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Francesca Frontini, Carmen Brando, Joanna Byszuk, Ioana Galleron, Diana Santos, and Ranka Stanković. 2020. Named Entity Recognition for Distant Reading in ELTeC. In *CLARIN Annual Conference 2020*, Virtual Event, France.

Evelyn Gius, Svenja Guhr, and Inna Uglanova. 2021. "d-prose 1870–1920" a collection of german prose texts from 1870 to 1920. *Journal of Open Humanities Data*, 7:11.

Evelyn Gius, Katharina Krüger, and Carla Sökefeld. 2019. Korpuserstellung als literaturwissenschaftliche aufgabe. In *DHd 2019 Digital Humanities: multimedial & multimodal Konferenzabstracts*, pages 164–166, Frankfurt am Main und Mainz.

Robert L. Górski. 2008. Charakterystyka chronologiczna i stylistyczna korpusu dla „wielkiego słownika języka polskiego". In P. Żmigrodzki and R. Przybylska, editors, *Nowe studia leksykograficzne*, pages 117–127. Lexis.

Stefan Th. Gries. 2009. *Quantitative corpus linguistics with R. A practical introduction*. Routledge, New York.

Angus Grieve-Smith. 2010. *Building a Representative Theater Corpus. A Broader View of Nineteenth-Century French*, chapter FRANTEXT's Corpus of Nineteenth-Century French. Palgrave Pivot, Cham.

Włodzimierz Gruszczyński, Dorota Adamiec, Renata Bronikowska, Witold Kieraś, Emanuel Modrzejewski, Aleksandra Wieczorek, and Marcin Woliński. 2022. The Electronic Corpus of 17th- and 18th-century Polish Texts. *Language Resources and Evaluation*, 56(1):309–332.

Włodzimierz Gruszczyński, Dorota Adamiec, Renata Bronikowska, and Aleksandra Wieczorek. 2020. Elektroniczny korpus tekstów polskich z XVII i XVIII w.– problemy teoretyczne i warsztatowe. *Poradnik Językowy*, 777(8):32–51.

Susan Hunston. 2008. Collection strategies and design decisions. In A. Ludeling and M. Kyto, editors, *Corpus Linguistics: an international handbook*, volume 1, pages 154–168. De Gruyter.

Krzysztof Jassem, Filip Graliński, and Tomasz Obrębski. 2017. Pros and Cons of Normalizing Text with Thrax. In *Proceedings of 8th Language & Technology Conference*, pages 230–235.

Gard B. Jenset and Barbara McGillivray. 2017. Methodological challenges in historical linguistics. In *Quantitative Historical Linguistics: A Corpus Framework*, pages 1–35. Oxford University Press, Oxford.

Jacob Jett. 2015. Modeling worksets in the hathitrust research center. CIRSS Technical Report WCSA0715. University of Illinois at Urbana-Champaign.

Jacob Jett, Timothy W. Cole, Christopher Maden, and J. Stephen Downie. 2016. The hathitrust research center workset ontology: A descriptive framework for non-consumptive research collections. *Journal of Open Humanities Data*, 2:e1.

Elżbieta Kaczynska. 1970. *Dzieje robotników przemysłowych w Polsce pod zaborami*. Warszawa.

Agnieszka Karlińska, Cezary Rosiński, Jan Wieczorek, Patryk Hubar, Jan Kocoń, Marek Kubis, Stanisław Woźniak, Arkadiusz Margraf, and Wiktor Walentynowicz. 2022. Towards a contextualised spatial-diachronic history of literature: mapping emotional representations of the city and the country in Polish fiction from 1864 to 1939. In *Proceedings of the 6th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 115–125, Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Graeme Kennedy. 1998. *An Introduction to Corpus Linguistics*. Longman, London.

Marek Kubis. 2021. Quantitative analysis of character networks in Polish 19th- and 20th-century novels. *Digital Scholarship in the Humanities*, 36(Supplement 2):ii175–ii181.

Geoffrey Leech. 2007. *Corpus Linguistics and the Web*, chapter New resources, or just better old ones? The Holy Grail of representativeness. Brill, Leiden, The Netherlands.

David Lindemann. 2022. LOD-ification of bibliographical data using free software: CLB-LOD wikibase. Mutual Learning Workshop for Improving Cultural Heritage Bibliographical Data, Prague.

David Lindemann, Christiane Klaes, and Penny Labropoulou. 2023. Bibliodata LODification using free software. DH2023 Collaboration as Opportunity, Graz.

Michał Marcińczuk and Jarema Radom. 2021. A single-run recognition of nested named entities with transformers. *Procedia Computer Science*, 192:291–297.

Michał Marcińczuk, Jan Kocoń, and Michał Gawor. 2018. Recognition of named entities for polish-comparison of deep learning and conditional random fields approaches. In *Proceedings of the PolEval 2018 Workshop*, pages 77–92. Institute of Computer Science, Polish Academy of Science.

Lukas Schlagenhauf. 2004. Challenges in modelling a richly annotated diachronic corpus of german. In *Workshop on XML-based Richly Annotated Corpora*.

Christof Schöch, Roxana Patras, Tomaž Erjavec, and Diana Santos. 2021. Creating the European Literary Text Collection (ELTeC): Challenges and Perspectives. *Modern Languages Open*, (1):25.

Ted Underwood. 2016. The real problem with distant reading (blog). https://tedunderwood.com/2016/05/29/the-real-problem-with-distant-reading/. Accessed: 2023-10-17.

8. Language Resource References

ELTeC. 2021. Polish novel collection (ELTeC-pol). Ed. by Joanna Byszuk. COST Action Distant Reading for European Literary History.

Berenike Herrmann and Gerhard Lauer. 2020. Kolimo. a corpus of literary modernism for comparative analysis. `https://kolimo.uni-goettingen.de/about`. Accessed: 2023-10-10.

Modern Poland Foundation. 2022. About the Project. `https://wolnelektury.pl/info/o-projekcie/`. Accessed: 2023-10-10.

National Library of Poland. 2022. About Polona Website. `https://polona.pl/page/about-polona/`. Accessed: 2023-10-10.

Wikimedia Foundation. 2022. About Wikisource. `https://wikisource.org/wiki/Wikisource:About_Wikisource`. Accessed: 2023-10-10.

A. Text Corpora Ontology

As a result of preliminary work on the Text Corpora Ontology (TCO), we have distinguished the following classes:

1. a single text **(tco:Text)**
2. the entities responsible for the single text, in particular persons **(tco:Person)**
3. places of publication and other locations relevant to the balance of the corpus **(tco:Place and tco:Partition)**
4. time periods, in particular literary epochs **(tco:Epoch)**
5. a general Corpus class **(tco:Corpus)**, representing a single set of texts

The attributes of the individual classes comprising the ontology not only enable the description of basic bibliographic information, such as title, authorship, and place of publication, but also include attributes important to the process of corpus construction. These latter attributes have been identified based on an extensive analysis of existing text collections and their respective construction principles. For individual classes, these attributes might encompass:

1. **foaf:gender** and **schema:birthPlace** for **tco:Person class**
2. **tco:numberOfReissues**, **tco:numberOfTokens** and location of the place of publication in a particular partition **(tco:inPartition)** for **tco:Text class**
3. information on the literary period within which the text included in the corpus was written **(tco:inEpoch)**

Each class instance can be further extended with external identifiers **(owl:sameAs)**, and individual documents can contain direct references to full-text files **(schema:contentUrl)**.

B. Implementation of literary corpus creation workflow

The proposed workflow for the creation and reuse of a meta-corpus represents a generalization of our experience in creating the 19/20MetaPNC corpus. To illustrate this connection, an overview of the subsequent stages of the workflow and their implementation in 19/20MetaPNC is presented in Tab. 3.

Table 3: Literary corpus creation workflow implementation in 19/20MetaPNC

Stage	**19/20MetaPNC Implementation**
Research question-based design	• The overall research goal was to trace the impact of historical and spatial factors on the dynamics of literary processes. • Specific research questions were formulated, focusing on the emotional polarization of literary images of the city and the country in Polish prose of the turn of the 20th century. • In order to achieve the research goal and answer the questions posed, the target population was defined, along with criteria for text inclusion (i.e., novels originally written in Polish and published between 1864 and 1939, with the time of the plot later than 1815), metadata schema, and balancing criteria.
Data collection and reuse	• Four open collections of 19th and 20th-century Polish prose are reused: – **ELTeC corpus** (ELTeC, 2021), – **Wolne Lektury** (Modern Poland Foundation, 2022), – **Wikisource** (Wikimedia Foundation, 2022), – **Polona** (National Library of Poland, 2022).
Data evaluation and preprocessing	• The most frequent categories of OCR errors found in **Polona** texts are automatically fixed. • The spelling of texts from **Wikisource** and **Polona** is modernized with the use of a diachronic normalizer. • Punctuation and hyphenation are normalized. • A common data format is introduced.
NLP-based enrichment	• The locations in the novel were tagged using PolDeepNer2, one of the latest and most effective NER tools for the Polish language (Marcińczuk et al., 2018; Marcińczuk and Radom, 2021). • NLP techniques were applied as part of a four-stage toponym disambiguation workflow to identify and standardize geographical entities. • A combination of supervised and unsupervised approaches will be used to determine the subject matter of the novel.
PIDs and LOD enrichment	• Automatic metadata enrichment was performed using the services of the National Library of Poland, VIAF, Wikidata, and Geonames. The following metadata were obtained: author's PID, place of birth, gender, and place of publication PID. Coordinates were determined for both types of places. • A four-stage toponym disambiguation workflow assigned records from the Geonames database to geo-entities identified as settings in the texts of the novels.

Stage	19/20MetaPNC Implementation
Metadata verification and completion	• Semi-automatic metadata verification was conducted to ensure that the collected texts met the corpus design criteria, particularly: – First edition dates were verified. – Original language was confirmed. – Genre was verified. – Authors' names and gender were verified, taking into account the use of pseudonyms. • As a result of automatic and manual work, information on the literary period, the assignment to the partition on the basis of the place of publication, and the time of the novel's action (before or after 1815) was completed.
Balancing	• The corpus was balanced historically and geographically based on the date and place of publication. • Additionally, balancing considered the gender of the authors and the level of reception, determined by the number of reprints of each publication. • The minimum and maximum share of a particular text class in the corpus were determined. • Further rebalancing of the corpus will be conducted based on the length of the novels.
Semantic environment and ontology building	• The entire metadata collection is stored in the form of a knowledge graph. • RDF format is employed for interoperability and to provide a structured framework for conducting computational literary studies. • Neo4J graph database is used for exploration, visualization and answering complex queries.
Publishing	• The meta corpus is published in a public GitHub repository. • The source texts are referenced in the knowledge graph by RDF triples that point to direct download links. • Python scripts for simplifying 19/20MetaPNC processing are provided in the GitHub repository.