From ELTeC Text Collection Metadata and Named Entities to Linked-data (and Back)

Milica Ikonić Nešić, Ranka Stanković, Christof Schöch and Mihailo Škorić

Proceedings of the Linked Data in Linguistics Workshop @ LREC2022, pages 7–16
Marseille, 24 June 2022

From ELTeC Text Collection Metadata and Named Entities to Linked-data (and Back)

Milica Ikonić Nešić*, Ranka Stanković†, Christof Schöch‡, Mihailo Škorić†
*University of Belgrade, Faculty of Philology, Serbia
milica.ikonic.nesic@fil.bg.ac.rs, †University of Belgrade, Faculty of Mining and Geology, Serbia
{ranka.stankovic, mihailo.skoric}@rgf.bg.ac.rs
‡University of Trier, Germany; schoech@uni-trier.de

Abstract

In this paper we present the wikification of the ELTeC (European Literary Text Collection), developed within the COST Action "Distant Reading for European Literary History" (CA16204). ELTeC is a multilingual corpus of novels written in the time period 1840—1920, built to apply distant reading methods and tools to explore the European literary history. We present the pipeline that led to the production of the linked dataset, the novels' metadata retrieval and named entity recognition, transformation, mapping and Wikidata population, followed by named entity linking and export to NIF (NLP Interchange Format). The speeding up of the process of data preparation and import to Wikidata is presented on the use case of seven sub-collections of ELTeC (English, Portuguese, French, Slovenian, German, Hungarian and Serbian). Our goal was to automate the process of preparing and importing information, so OpenRefine and QuickStatements were chosen as the best options. The paper also includes examples of SPARQL queries for retrieval of authors, novel titles, publication places and other metadata with different visualisation options as well as statistical overviews.

Keywords: Wikidata, linked data, SPARQL, distant reading, literary corpus, named entity linking, ELTeC

1. Introduction

The COST Action "Distant Reading for European Literary History"[1] ran from 2017 to 2022 and aimed to use computational methods for the analysis of large collections of literary texts. The main goal of this networking project was to compile and analyse a multilingual open-source collection of novels, named European Literary Text Collection (ELTeC). ELTeC contains corpora of 100 novels per language written between 1840 and 1920 that are encoded in XML, are linguistically annotated and contain detailed metadata (Schöch et al., 2021).

The term *distant reading* (Moretti, 2000) describes an alternative or a complement to *close reading*: Instead of detailed, qualitative interpretations of selected literary texts, the idea is to analyse large collections of literary texts using quantitative methods of text analysis and machine learning. Formal and quantifiable textual features are used as indicators for relevant literary phenomena, with their patterns of occurrence then being related to categories such as authors, genres, or literary periods (Schöch et al., 2020).

This paper presents an approach for publishing the metadata and named entities (NE) from the sub-collections of ELTeC as linked open data. More precisely, the paper presents results for 700 novels from the first seven languages (English, Portuguese, French, Slovenian, German, Hungarian and Serbian) that are morpho–syntactically tagged (Stanković et al., 2022b) and partially annotated with named entities (Stanković et al., 2019; Frontini et al., 2020), as well as the case study on Named Entity Linking (NEL) for the Serbian ELTeC sub-collection.

Linked open data for literary texts is slowly gaining traction, as evidenced by resources such as BookSampo (Mäkelä et al., 2013) or projects like POSTDATA (Bermúdez-Sabel et al., 2021) and Mining and Modeling Text (Schöch et al., 2022). The motivation for the presented activity was to increase the visibility of the ELTeC collection, to connect it to open knowledge bases, as well as to allow searching and analyzing texts using linked open data. The incentive for the presented activity was the successful initial implementation for Serbian (Ikonić Nešić et al., 2021) that was further applied to other six languages with support of the sub-collection coordinators.

We use the term *wikification* not only for entity linking with Wikidata as the target Knowledge base, but also for creating and populating Wikidata items related to novels which will be further used for entity linking.

The crucial point for automation of wikification was the synergy of the powerful open source tools OpenRefine (Huynh, 2012) and QuickStatements (Manske, 2019). This enabled 700 novels from the core collections and 20 from extended sub-collections of ELTeC to be described in Wikidata, including associated items for their first editions, print editions, digital editions and the ELTeC (electronic) editions. This resulted in approximately 20,900 automatically added statements. To the best of our knowledge, this work is the first example of data about literary corpora for seven languages being automatically imported into Wikidata using different open source tools.

Section 2 is dedicated to the ELTeC: in Subsection 2.1 an overview of the text collection is given, in Subsec-

[1]Distant Reading for European Literary History (CA16204), `https://www.distant-reading.net`.

tion 2.2 the XML/TEI encoding of novels is explained, while in Subsection 5.1 the NER approach applied to novels is introduced.
The ELTeC Linked data model is presented in Section 3: in Subsection 3.1, the main data model, automation and the management of ELTeC Wikidata are presented, while the pipeline, from data preparation to Wikidata linking, is presented in Subsection 3.2.
The process of automation of ELTeC Wikidata population is presented in Section 4. The entity linking is described in Section 5.2: entity recognition and linking with Wikidata identifiers.
The development of a user friendly interface with predefined SPARQL queries with visualization is presented in Section 6. A set of web pages was developed with integrated results of SPARQL queries to help literary scholars that are not familiar with SPARQL. Several different visualisation options, based on Wikidata Query Service should allow new aspects of distant reading of the literary data. Section 7 concludes and summarizes our entire research and outlines several possibilities of extensions to this research.

2. ELTeC Text Collection

2.1. Overview of ELTeC Collection

Within the COST Action "Distant Reading for European Literary History", a research network of more than 200 researchers from more than 30 countries was built to foster digital, cross-lingual research into the history of the European novel. The envisaged activities were to build a multilingual corpus of European novels and develop appropriate, digital methods of analysis. Its main objective was the production of a unified, uniform, multilingual, digital novel collection dubbed the "European Literary Text Collection", or ELTeC for short (Odebrecht et al., 2021), containing novels first published between 1840 and 1920 in Europe.
ELTeC is a multilingual resource that provides learning opportunities regarding collaborative research for the European, multilingual community of researchers in (computational) literary studies. It is also a foundation for the development of cross-lingual methods and a first step towards a history of European literature that would be truly digital, multilingual and diverse (Schöch, 2022).
The novels are selected from the time period 1840-1920 and currently, 10 corpora are complete while seven more are in progress, in addition to several extension collections. The latest release (v1.1.0) was published in April 2021, containing 14 sub-collections and 1,200 novels. Its key characteristics are that each corpus represents the variety of production, that texts are encoded in XML-TEI, that they are linguistically-annotated (morpho-syntactically, NE) and that everything is published under open licences (Schöch et al., 2021; Burnard et al., 2021).
ELTeC is designed to support a wide range of distant reading methods. Such methods cover various computational approaches to literary text analysis, regarding authorship and textuality, time and space, theme and style, or character and plot (more in (Schreibman and Siemens, 2008; Eve, 2022)). Many of them have already been applied to ELTeC, among them stylometric authorship attribution (Škorić et al., 2022; Cinkova and Rybicki, 2020), stylistic analysis (Stanković et al., 2022a; Patras et al., 2021; Krstev, 2021b) or direct speech detection (Byszuk et al., 2020). Linguistic annotation and detailed metadata support many of these methods.

2.2. XML/TEI Encoding of Novel's Metadata

The ELTeC coding scheme was produced with no intention to present the original documents in all their original structure or layout complexity, but to make it easier to access the texts that are encoded in a predictable manner. The relevant COST Action working group agreed that the ELTeC should be delivered in a TEI–encoded format, using a schema developed specifically for the project (Burnard et al., 2021).
In order to be compliant with the TEI guidelines, a documents needs to provide metadata in the `<teiHeader>`. Each novel from the ELTeC collection at level-1 (text with structural and layout annotations) is prepared as an XML/TEI document and contains a TEI header with the following required XML elements:

- `<fileDesc>`: description of the electronic edition, which includes the title of the work and the name of the author, as well as the statements of responsibility (scanning, correction, annotation), date of publication, size (measured by the number of words). Identifiers can be assigned to authors and their work, such as VIAF and Wikidata.
- `<sourceDesc>`: brief bibliographic description of the first edition and the edition used as the source for ELTeC (if different from the first edition).
- `<profileDesc>`: description of the text in terms of meeting criteria used for the selection of novels (e.g. author's gender, novel's size, time slot of the first edition, number of recent reprints,...).
- `<revisionDesc>`: review of all changes to the digital edition since its first publication.

An opportunity for speeding up the process of data preparation for Wikidata was seen in using information already encoded in the header of each novel (Krstev, 2021a; Ikonić Nešić et al., 2021). This approach will be elaborated in Section 3.

3. ELTeC Linked Data Model

3.1. Wikidata Class Selection

Wikidata is an open source knowledge base where the underlying structure in RDF is a collection of triples,

each consisting of a subject (Wikidata item to which the claim refers), a predicate (Wikidata property), and the object (value). A value can be another item, a string, a time, a period, a location, an URL, or a quantity, depending on the property type. Statements can use qualifiers that show the contexts of the validity of the statement and they can include references. Qualifiers and references are also represented in the form of triples, where the subject is the claim.
The items and properties in Wikidata that are used to structure the ontology are:

- classes: class (Q16889133), entity (Q35120) and Wikidata metaclass (Q19361238),
- properties: instance of (P31) and subclass of (P279)

Classes conceptually group together similar items.

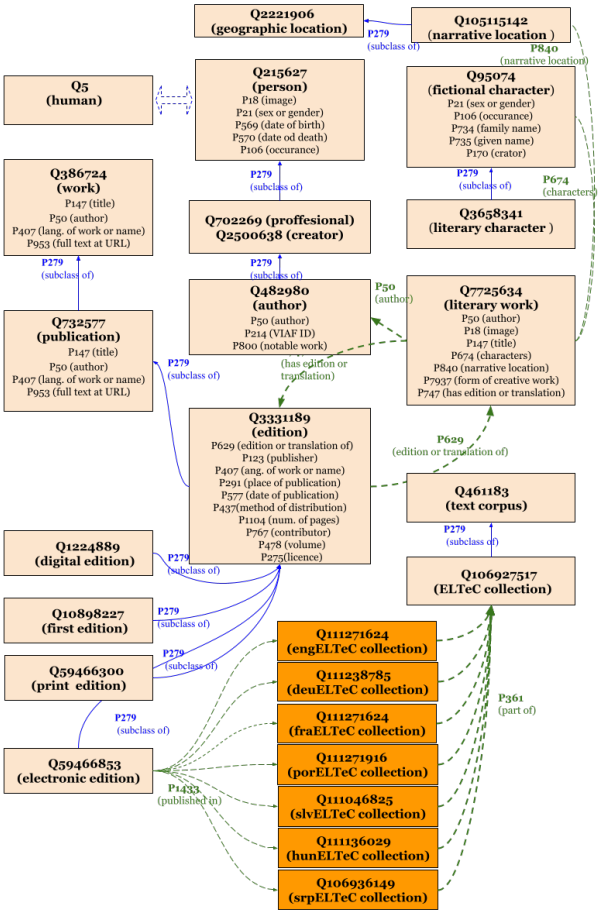

Figure 1: The class diagram of Wikidata used for novels and editions in ELTeC text collection.

Figure 1 presents the class/instance relation of all classes and relations that are used in this research. The blue lines represent "subclass of" relations between classes, while green lines presents other properties. The class person (Q215627) is used as "instance of" (P31) class humans (Q5), as recommended in the Wikidata documentation (Class person). Each item for a novel is connected with an appropriate item that is an instance of electronic edition (Q59466853), first edition (Q10898227), print edition (Q59466300) and digital edition (Q1224889) using property (P747) (has edition or translation), and every item of edition must be connected with a corresponding item for a novel with inverse property (P629) (edition or translation of). Orange boxes represent items for each of the seven corpora of ELTeC that are (P279) subclasses of electronic edition (Q59466853). All seven are published in ELTeC Collection (Q106927517) which is "subclass of" text collection (Q461183). A list of all properties that are used for authors, novels and editions is presented as a part of Wikidata: WikiProject ELTeC (Property overview). It is necessary to emphasize that for now only items for novels in the Serbian part of ELTeC are connected with appropriate items for main characters and narrative places. All items for main characters are created manually and all of them are instances of literary character (Q3658341). Narrative places are instances of class city (Q515).

3.2. ELTeC Data Model Aligning with Wikidata Classes

Having consistent TEI headers enabled extraction of metadata and linking with Wikidata. Data extraction was a necessary step to automate the process of importing novels and editions into Wikidata. After careful selection of classes and properties, it was necessary first to find exact mappings between them and elements of the novels' XML documents. Figure 2 shows an example of the mapping for the French novel *Lucingole* (Q111366753) written by Catulle Mendès (Q971215).
A set of metadata of the ELTeC novels was extracted from the element `<teiHeader>`, the part of which is presented in Table 1. The first column of Table 1 represents the TEI XPath to and element or attribute for ELTeC edition (the upper cell) and for different types of editions, where *type* can be first, print or digital (the lower cell). The second column contains information about the class of the instantiated data that is used for mapping. More about mapping and chosen classes can be seen in (Ikonić Nešić et al., 2021).

4. Automatization of ELTeC Wikidata Population

An opportunity for speeding up the process of item creation was seen in using the information encoded in the header of each novel, as explained in Subsection 2.2. The main aim of this research was to build Wikidata entities by using the model and mapping presented in Subsection 3.2 for the novels belonging to those ELTeC corpora that already provide a so-called level-2 encoding with morpho-syntactic and NE annotation: English, Portuguese, French, Slovenian, German, Hungarian and Serbian.
As the guideline model for the automation activities,

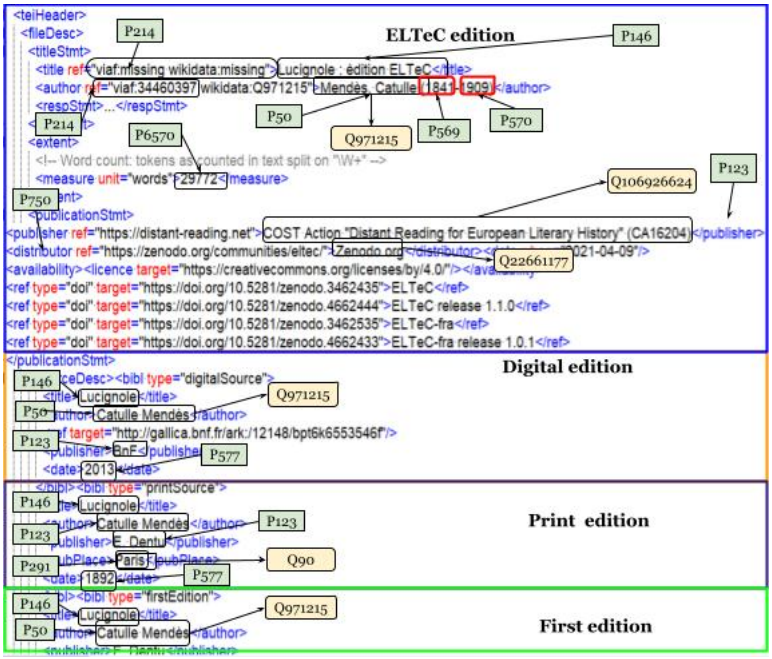

Figure 2: Mapping between metadata header and Wikidata (the novel *Lucignole* (Q111366753))

TEI XPath to element or attribute for ELTeC edition data	element is instance of
/titleStmt/title	Q783521 (title)
/titleStmt/author	Q482980 (author)
/extent/measure[unit="words"]	Q8034324 (word count)
/publicationStmt/publisher	Q105044823 (publisher)
/publicationStmt/distributed	Q12540664 (distributor)
/publicationStmt/availability/licence@target	Q79719 (licence)
/profileDesc/langUsage/language[ident="fr"]	Q34770 (language)
TEI XPath to element or attribute for different type edition data	**element is instance of**
/sourceDesc/bibl[type=@typeSource]/title	Q783521 (title)
/sourceDesc/bibl[type=@typeSource]/author	Q482980 (author)
/sourceDesc/bibl[type=@typeSource]/publisher	Q105044823 (publisher)
/sourceDesc/bibl[type=@typeSource]/pubPlace	Q1361759 (place of pub.)
/sourceDesc/bibl[type=@typeSource]/data	Q1361758 (date of pub.)

Table 1: Mapping between metadata to Wikidata for editions

the use case of SrpELTeC at Wikidata (Ikonić Nešić et al., 2021) was employed.

Data preparation and the import process were done via the synergy of OpenRefine (Verborgh and Wilde, 2013) – a tool for working with messy data, like cleaning, converting from one format to another, with the addition of external data via a web service – and QuickStatements, a Wikidata editor for adding and removing statements, tags, properties, labels and descriptions.

The following processing steps were performed on all novels with level-2 annotations:

- preparation of metadata of ELTeC sub-collections for import into Wikidata,
- import of data into OpenRefine and reconcile data with external source (Wikidata),
- importing data into Wikidata using QuickStatments,
- analysis of imported dataset using a set of SPARQL queries.

The procedure for the extraction of all metadata from the headers into one *CSV* (comma separated values) file, appropriate for further transformations and exploitation of text collections in OpenRefine, was integrated in the already existing tool for creation, management and exploitation of lexical resources *Leximir* (Stanković and Krstev, 2012).

After mapping metadata to Wikidata, OpenRefine was used to automate the data preparation, check existence and perform disambiguation. A process of manually checking of extracted metadata was required to solve some uncertainties. Namely, several instances of wrong date of birth or death of authors or missing VIAF IDs etc. were found and solved in collaboration with members of other teams of the working group for different languages.

Since author-related entries are a precondition for the automatic item creation, the OpenRefine *reconciling*

process was used to check if each entry existed. Reconciliation is the process of matching our dataset with that of an external source – in this case we use this process to identify existing items in Wikidata – a necessary step that enables linking of the file contents to the identifiers (QID) of existing Wikidata items and the creation of new ones for those that do not exist.

For missing authors, items as instances of authors (Q482980), were automatically created, with labels, description and properties such as dates of birth and death and the author's gender, which were extracted from the element `<author>` from metadata `<teiHeader>`, if the information was available.The process of entering authors in Wikidata will not be described, and we will focus on entries for novels and editions in Wikidata. The main entities involved in these tasks are: text collection (Q461183), novels (Q7725634) and version, edition, or translation (Q3331189).

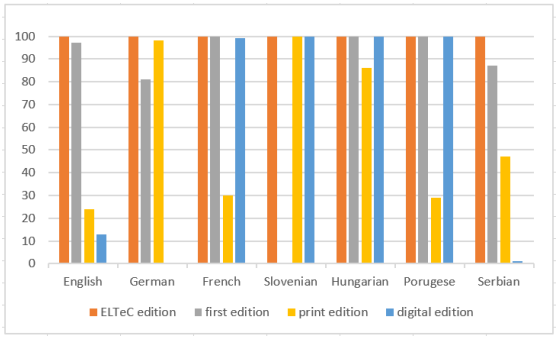

Figure 3: Statistical overview of edition items.

The next step was editing the Wikidata schema using OpenRefine. Creating a Wikidata input set schema defines subjects (items which we create), predicates (properties) that will connect subjects, and objects in RDF triples which are values of extracted metadata. The subject of the statement one or more properties whose value can be a Wikidata item, external URL, or literal (string). The subject of the statement one or more properties whose value can be a Wikidata item, external URL, or literal (string). After editing and saving the Wikidata schema, we exported it as a *QuickStatements* file and automatically added it to Wikidata.

700 novels of the ELTeC level-2 collection with 700 ELTeC (electronic) editions, 565 first editions, 414 print editions and 413 electronic editions were automatically added (totaling in approximately 20,900 statements).The statistical overview of quantities automatically added to Wikidata for each language is presented in Figure 3. More information about the metadata mapping can be found in (Ikonić Nešić et al., 2021).

5. NER for ELTeC

5.1. Literary Characters and Narrative Locations in Novels

The main goal of named entity recognition, in general, is to indicate in a text names of persons, their roles, locations, organizations, and other entities relevant for specific purposes. The NER team agreed that seven categories of entities should be indicated in the novels: PERS, ROLE, DEMO, ORG, LOC, WORK, and EVENT, which were assessed as being of the greatest importance for further literary studies (Stanković et al., 2019). Developing the NE layer of the ELTeC, testing the automatic NER for Distant Reading in ELTeC and fostering NER results and analysis are presented in (Frontini et al., 2020).

Entities belonging to one of the following NE classes were represented in Wikidata in this phase: PERS entities which correspond to main characters of a novel, ROLE entities used for their titles, professions or positions and LOC entities that designate places where the action of a novel takes place (geopolitical locations). This research was focused on two categories, PERS and LOC. The main characters of the novel can be found in the list of the extracted PERS entities, while in the LOC entity list one expects to find where the narrative of the novel is set. All entities in both categories were sorted by frequency of occurrence in each novel, and the most frequent entities are taken as literary characters (Q3658341) and narrative places, i.e. geographic location (Q2221906). This task cannot be fully automated, since the names of same characters can be mentioned in a text in a number of different ways, such as: *Čedomir Ilić*, *Čedomir*, *Ilić*, *Čeda*, and it is not clear enough if places mentioned in novel are narrative places or places that are mentioned by some characters. Using these extracted named entities we were able to manually add 123 narrative locations and 904 main characters for 69 novels to Wikidata.

The main characters were described with a set of properties: gender, profession, whether the character is fictional or not, relations between characters (husband, wife, parent, child, etc.) and professions of characters related to the main ones. Since the basic information for each novel and its author is already in Wikidata, e.g. the birthplace of an author, his/her residence at the time of writing, the place of novel's first publication, it is now possible to relate the ELTeC geodata the (place of publication and places of narrative) to other time/space coordinates, and consider more detailed mapping visualizations as presented in Section 6.

Using SPARQL query https://w.wiki/5BX7, we produce graph (Figure 4) with the number of novels that are mentioning particular locations (places) based on Wikidata. *Srbija (Serbia)* is mentioned in 39 novels and *Beograd (Belgrade)* is mentioned in 19 novels. The graph with number of characters in novels, generated using https://w.wiki/5BX9, is presented in Figure 5. It can be seen that *Djuradj Branković : istoričeskih ro-*

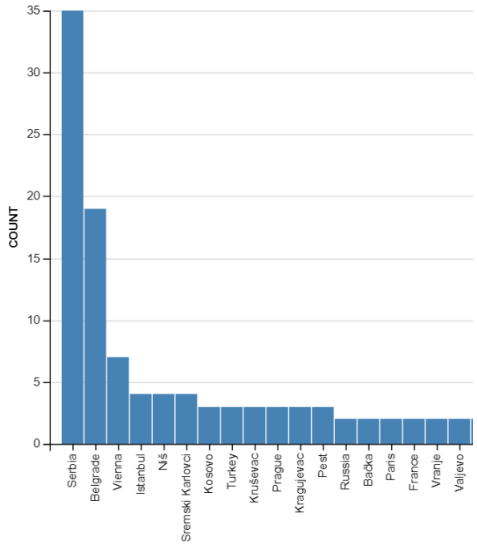

Figure 4: Number of novels mentioning the locations.

man (Djuradj Branković : a historical novel) has the largest number of characters (41). Currently, only the

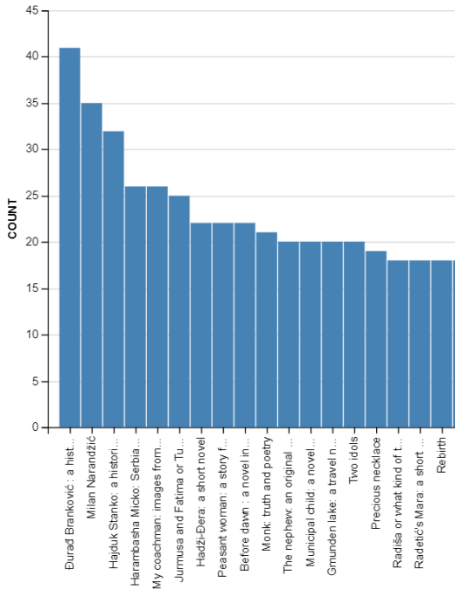

Figure 5: Main characters mentioned in novels.

narrative locations and literary characters in the Serbian part of ELTeC collection are populated, the other languages were not covered with this research.

5.2. From NE Extraction to Wikidata across Inception to NIF

After the main characters and narrative places were manually added, in order to validate the viability of our approach in a realistic scenario, we used the tool INCEpTION (Klie et al., 2018) for the Wikidata named entity linking on a subset of SrpELTeC collection. INCEpTION is a web-based environment for interactive text annotation and knowledge management with integrated machine-learning based assistance features and entity linking with Wikidata. The user identifies entity mentions and links them to Wikidata. To link text to an item (a class or instance), the user selects a span of text and searches for the linking item using an autocomplete text with items from Wikidata. (Castilho et al., 2018)

For the purpose of our research, two Serbian novels Ivkova slava : pripovetka (Ivko's patron saint's day: a short story) and Nečista Krv (Impure blood) were imported into INCEptION and linked with main characters and locations. We present the main characters and locations for the novel Impure blood in Table 2.

Main characters	**Narrative locations**
Sofka (Q109693861)	Vranje (Q211645)
Magda (Q10974671)	Srbija (Q403)
Marko (Q109747266)	Beograd (Q3711)
Arsa (Q109747507)	Turska (Q43)
Mita (Q109747662)	Carigrad (Q16869)
Simka (Q109748862)	Solun (Q210176)
Todora (Q109748881)	Morava (Q211328)
Tone (Q109748906)	
Ahmet (Q109748924)	
Milenija (Q109748942)	
Tomča (Q109748839)	
Stana (Q110283369)	
baba-Simka (Q110826779)	

Table 2: Characters and locations in *Impure blood*.

Figure 6 presents an example of linking character *Sofka* from the novel *Impure blood* with Wikidata item *Sofka* (Q109693861).

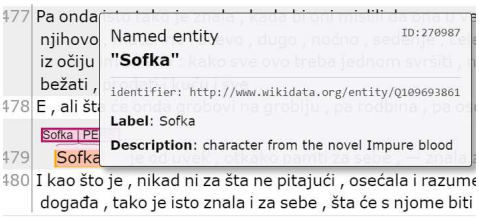

Figure 6: Inception & Wikidata NEL in *Impure blood*.

The workflow of linking characters and narrative places is presented in Figure 7.

The full process of linking entities with knowledge bases using the INCEpTION annotation platform is described in (Klie et al., 2020).

After linking annotations in INCEpTION to the knowledge base, we were able to write queries to find occurrences of all linked entities (e.g. specific persons) or find verbs that precede specific places. First steps towards RDF editions of the ELTeC corpus are publishing two Serbian novels Ivkova slava : pripovetka

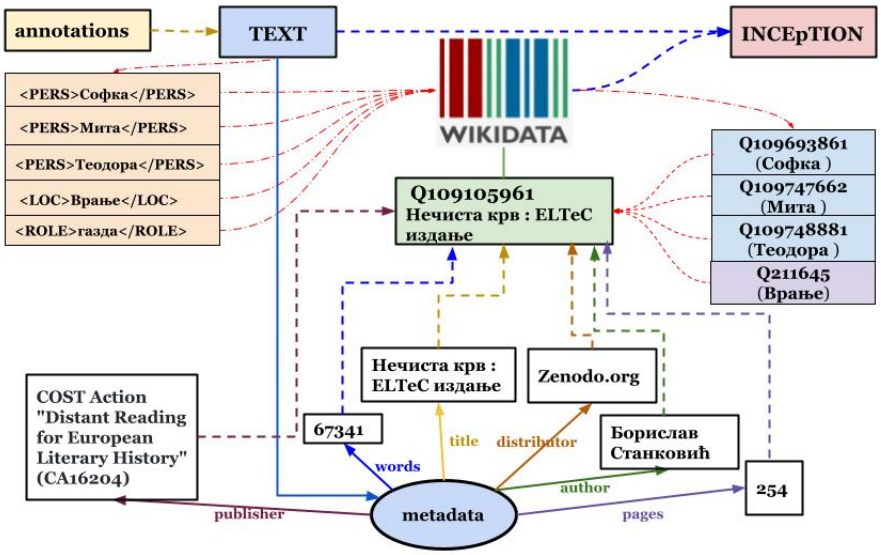

Figure 7: The Workflow: mapping metadata with Wikidata and Inception on the novel *Impure blood.*

(Ivko's patron saint's day: a short story) and Nečista Krv (Impure blood), POS-tagged, lemmatized, with NER and NEL with Wikidata, available in NIF (Ikonić Nešić and Stanković, 2022b). An example of an NIF excerpt of the novel *Nečista krv (Impure blood)* is presented in Figure 8.

6. The Overview of ELTeC@Wikidata by SPARQL Queries

In order to facilitate the use of Wikidata about ELTeC, we created a website with a set of predefined SPARQL queries that enable retrieval of authors, novel titles, publication places, characters, family relations of characters, their roles and others, and offer different visualization options (Ikonić Nešić and Stanković, 2022a). Different queries were written that supplied the tables: the title of the novel, the name of the author, the author's pictures, the year of publication, the main characters, and for those with imported narrative places and main characters also the relations between them, as the number of places mentioned by authors, and etc.

Figure 9 represents the timeline visualization of all authors in seven sub-collections URL. Figure 10 represents the map of first publication places.

The query presented below produces https://w.wiki/5BpU, a map of places of birth of authors, colour-coded by the time span.

```
#defaultView:Graph
SELECT DISTINCT ?person ?name ?bplace
?byear ?coord ?layer
WHERE {
?novel  wdt:P747 ?edition;
        wdt:P50  ?person.
?edition wdt:P1433 ?coll.
?coll wdt:P361 wd:Q106927517}
?person wdt:P570 ?dod;
        wdt:P19 ?place
?place wdt:P625 ?coord.
OPTIONAL{?person wdt:P569 ?dob.}
OPTIONAL{?person wdt:P18 ?image.}
BIND(YEAR(?dob) AS ?byear)
BIND(IF(byear<1851,"-1850",
IF(byear<1901,"1851-1900",
IF(byear<1951,"1901-1950",
"after-1950")))) AS ?layer)
?person rdfs:label ?name.
FILTER((LANG(?name)) = "en") ?place
rdfs:label ?bplace.
FILTER((LANG(?bplace))="en") }
ORDER BY (?byear)
```

In Figure 11, blue points represent time spans before 1700, orange between 1751-1800, green between 1801-1850, and red between 1851-1900.

The list of all novels, authors and editions for English, German, French, Portuguese, Slovenian, Hungarian and Serbian collection is presented in WikiProject ELTeC.

7. Conclusion and future work

In this paper we presented our recently finished activity of populating Wikidata with 720 novels from the ELTeC for seven languages (English, Portuguese, French, Slovenian, German, Hungarian and Serbian). The presented approach is language independent, so we hope that this can be an inspiration for other ELTeC corpora to expand their visibility using open linked data. The research in the digital humanities has increasingly advanced the importance of linked (open) data and with this activity we try to contribute to the distant reading methods using linked data.

Current activities include manual Named Entity Linking with Wikidata using INCEPTION platform, but future activities will be focused on training a model for automatic Named Entity Linking and exploring the formal data structures for tabular formats in language technology: CoNLL-RDF and CoNLL-RDF ontology (Chiarcos et al., 2021).

The second type of future activities will concern publishing entire annotated corpora as Linguistic Linked

```
<file:/srv/inception/repository/project/34/document/359/source/SRP19101_1.tsv#offset_91555_91560>
        a                       nif:EntityOccurrence , nif:OffsetBasedString , nif:Word ;
        nif:anchorOf            "Sofku" ;
        nif:beginIndex          "91555"^^xsd:nonNegativeInteger ;
        nif:endIndex            "91560"^^xsd:nonNegativeInteger ;
        nif:lemma               "Sofka" ;
        nif:nextWord
                <file:/srv/inception/repository/project/34/document/359/source/SRP19101_1.tsv#offset_91561_91562> ;
        nif:posTag              "PROPN" ;
        nif:previousWord
                <file:/srv/inception/repository/project/34/document/359/source/SRP19101_1.tsv#offset_91542_91554> ;
        nif:referenceContext
                <file:/srv/inception/repository/project/34/document/359/source/SRP19101_1.tsv#offset_0_96737> ;
        nif:sentence
                <file:/srv/inception/repository/project/34/document/359/source/SRP19101_1.tsv#offset_91497_91628> ;
        itsrdf:taClassRef       <PERS> ;
        itsrdf:taIdentRef       <http://www.wikidata.org/entity/Q109693861> .
```

Figure 8: SrpELTeC NIF sample

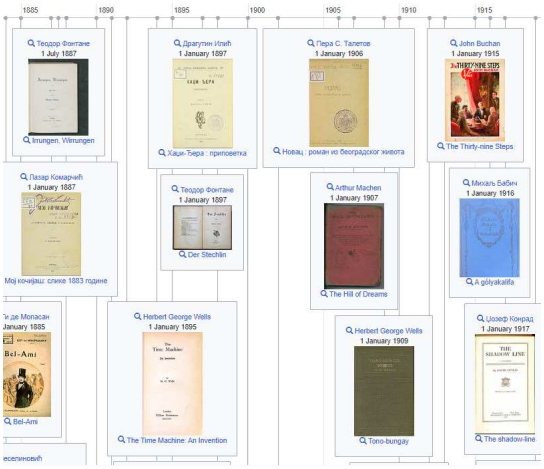

Figure 9: ELTeC sub-collections timeline

Figure 10: Map of first publication places

Open Data. Using NIF or Web Annotation / Open Annotation, the export of all level-2 novels additionally supplied with NEL layer could be published in the RDF store to be available via the SPARQL endpoint.

Following the current and future trends and challenges,

Figure 11: Birthplaces of authors, time span coloured.

as well as a proposal of the use of design patterns (Khan et al., 2021) we will apply the OntoLex-FrAC: Frequency, Attestations, Corpus Information module for complementing dictionary of lesser known, archaic words extracted from the old novels.

8. Acknowledgements

The preparation of the text corpora and a virtual mobility were supported by the COST Action "Distant Reading for European Literary History" (CA16204). Linked data development was done in the scope of the project "WikiELTeC–Wikidata about old Serbian novels from collection ELTeC" and supported by the COST Action "NexusLinguarum, European network for Web-centred linguistic data science" (CA18209). Both Actions are funded by COST (European Cooperation in Science and Technology, see www.cost.eu. The authors would like to thank Prof. dr Cvetana Krstev for her valuable comments which helped to improve the manuscript.

9. Bibliographical References

Bermúdez-Sabel, H., Díez Platas, M. L., Ros, S., and González-Blanco, E. (2021). Towards a common model for European Poetry: Challenges and solutions. *Digital Scholarship in the Humanities*.

Burnard, L., Schöch, C., and Odebrecht, C. (2021). In search of comity: TEI for distant reading. *Journal of the Text Encoding Initiative*, (14).

Byszuk, J., Woźniak, M., Kestemont, M., Leśniak, A., Łukasik, W., Šeļa, A., and Eder, M. (2020). Detecting direct speech in multilingual collection of 19th-century novels. In *Proceedings of LT4HALA 2020-1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 100–104.

Castilho, R. E. D., Klie, J.-C., Kumar, N., Boullosa, B., and Gurevych, I. (2018). Linking Text and Knowledge Using the INCEpTION Annotation Platform. In *2018 IEEE 14th International Conference on e-Science (e-Science)*, pages 327–328.

Chiarcos, C., Ionov, M., Glaser, L., and Fäth, C. (2021). Formal Data Structures for Tabular Formats in Language Technology.

Cinkova, S. and Rybicki, J. (2020). Stylometry in a Bilingual Setup. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 977–984, Marseille, France, May. European Language Resources Association.

Eve, M. P. (2022). *The Digital Humanities and Literary Studies*. The Literary Agenda. Oxford University Press, Oxford, New York, February.

Frontini, F., Brando, C., Byszuk, J., Galleron, I., Santos, D., and Stanković, R. (2020). Named Entity Recognition for Distant Reading in ELTeC. In *CLARIN Annual Conference 2020*.

Ikonić Nešić, M., Stanković, R., and Rujević, B. (2021). Serbian ELTeC Sub-Collection in Wikidata. *Infotheca – Journal for Digital Humanities*, 21(2):60–87.

Khan, A. F., Chiarcos, C., Declerck, T., Gifu, D., García, E. G.-B., Gracia, J., Ionov, M., Labropoulou, P., Mambrini, F., McCrae, J. P., et al. (2021). When Linguistics Meets Web Technologies. Recent advances in Modelling Linguistic Linked Open Data. *Semantic Web journal*.

Klie, J.-C., Eckart de Castilho, R., and Gurevych, I. (2020). From Zero to Hero: Human-In-The-Loop Entity Linking in Low Resource Domains. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6982–6993, Online, July. Association for Computational Linguistics.

Krstev, C. (2021a). The Serbian Part of the ELTeC Collection through the Magnifying Glass of Metadata. *Infotheca – Journal for Digital Humanities*, 21(2):26–42.

Krstev, C. (2021b). White as Snow, Black as Night – Similes in Old Serbian Literary Texts. *Infotheca – Journal for Digital Humanities*, 21(2):119–136.

Mäkelä, E., Hypén, K., and Hyvönen, E. (2013). Fiction literature as linked open data—The BookSampo dataset. *Semantic Web*, 4(3):299–306.

Moretti, F. (2000). Conjectures on World Literature. *New Left Review*, 1 (February):54–68.

Patras, R., Odebrecht, C., Galleron, I., Arias, R., Herrmann, B. J., Krstev, C., Poniž, K. M., and Yesypenko, D. (2021). Thresholds to the "Great Unread": Titling Practices in Eleven ELTeC Collections. *Interférences littéraires/Literaire interferenties*, 25:163–187, October.

Schreibman, S. and Siemens, R. (2008). *Companion to Digital Literary Studies*. Blackwell Companions to Literature and Culture. Blackwell Publishing Professional, Oxford, hardcover edition, December.

Schöch, C., Eder, M., Arias, R., and Pieter Francois, A. P. (2020). Foundations of Distant Reading: Historical Roots, Conceptual Development and Theoretical Assumptions around Computational Approaches to Literary Texts. In *Digital Humanities 2020*.

Schöch, C., Patras, R., Erjavec, T., and Santos, D. (2021). Creating the European Literary Text Collection (ELTeC): Challenges and Perspectives. *Modern Languages Open*.

Schöch, C., Hinzmann, M., Röttgermann, J., Dietz, K., and Klee, A. (2022). Smart Modelling for Literary History. *International Journal of Humanities and Arts Computing*, 16(1):78–93.

Schöch, C. (2022). What is ELTeC all about? In *Belgrade Training School 2022: Exploring ELTeC: Use-Cases for Information Extraction and Analysis. Belgrade, March 21-23, 2022*.

Stanković, R., Santos, D., Frontini, F., Erjavec, T., and Brando, C. (2019). Named Entity Recognition for Distant Reading in Several European Literatures. In *DH Budapest 2019*.

Stanković, R., Krstev, C., Šandrih Todorović, B., Vitas, D., Škorić, M., and Nešić, M. I. (2022a). Distant Reading in Digital Humanities: Case Study on the Serbian Part of the ELTeC Collection. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'22)*, Marseille, France, june. European Language Resource Association (ELRA).

Stanković, R., Škorić, M., and Šandrih Todorović, B. (2022b). Parallel bidirectionally pretrained taggers as feature generators. *Applied Sciences*, 12(10).

Verborgh, R. and Wilde, M. D. (2013). *Using OpenRefine*. Packt Publishing, 1st edition.

Škorić, M., Stanković, R., Ikonić Nešić, M., Byszuk, J., and Eder, M. (2022). Parallel stylometric document embeddings with deep learning based language models in literary authorship attribution. *Mathematics*, 10(5).

10. Language Resource References

David Huynh. (2012). *OpenRefine*. `https://openrefine.org/`, 3.5.

Milica Ikonić Nešić and Ranka Stanković. (2022a). *SparqlELTeC*. `http://jerteh.rs/resursi/WIKIDATA-SPARL/`.

Milica Ikonić Nešić and Ranka Stanković. (2022b). *srpNIF*. `http://llod.jerteh.rs/ELTEC/srp/NIF/`.

Klie, Jan-Christoph and Bugert, Michael and Boullosa, Beto and Eckart de Castilho, Richard and Gurevych, Iryna. (2018). *The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation*. https://inception-project.github.io.

Magnus Manske. (2019). *QuickStatements*. https://quickstatements.toolforge.org/, 2.0.

Carolin Odebrecht and Lou Burnard and Christof Schöch. (2021). *European Literary Text Collection (ELTeC): April 2021 release with 14 collections of at least 50 novels.* Zenodo, https://github.com/COST-ELTeC.

Ranka Stanković and Cvetana Krstev. (2012). *LeXimir - Tool for lexical resources management and query expansion*. 1.0.