Serbian ELTeC Sub-Collection in Wikidata

Milica Ikonić Nešić, Ranka Stanković, Biljana Rujević

Serbian ELTeC Sub-Collection in Wikidata | Milica Ikonić Nešić, Ranka Stanković, Biljana Rujević | Infotheca | 2021 | |

10.18485/infotheca.2021.21.2.4

http://dr.rgf.bg.ac.rs/s/repo/item/0006192

Serbian ELTeC Sub-collection in Wikidata

UDC 004.62: [030:004.738.5
DOI 10.18485/infotheca.2021.21.2.4

Milica Ikonić Nešić
milica.ikonic.nesic@fil.bg.ac.rs
University of Belgrade
Faculty of Philology
Belgrade, Serbia

Ranka Stanković
ranka.stankovic@rgf.bg.ac.rs

Biljana Rujević
biljana.rujevic@rgf.bg.ac.rs
University of Belgrade
Faculty of Mining and Geology
Belgrade, Serbia

ABSTRACT: This paper presents an example of integration of Wikidata with digital libraries and external systems, as well as some best practices for speeding up the process of data preparation and import to Wikidata, on the use case of SrpELTeC, Serbian subcollection of the ELTeC multilingual collection (European Literary Text Collection). After preliminary work on the manual Wikidata population with SrpELTeC novels, the goal was to automate the process of preparing and importing information, so different solutions were analysed and finally synergy of two, OpenRefine and QuickStatements, was chosen as the best option. The paper also brings examples of SPARQL queries for retrieval of authors, novel titles, publication places and other metadata with different visualisation options.
KEYWORDS: Wikidata, distant reading, literary corpus, named entity linking, ELTeC, SrpELTeC.

PAPER SUBMITTED: 05 November 2021
PAPER ACCEPTED: 12 November 2021

1 Introduction

Wikidata[1] is a Wikimedia Foundation knowledge base, a common source of various kinds of data, used not only by other Wikimedia projects, but also increasingly by numerous semantic web applications. Integration of Wikidata with digital libraries and external systems is envisaged as a useful task for various applications. Wikidata has grown significantly since its launch in October 2012. It has also become the most edited Wikimedia project,

1. Wikidata

supporting 150–500 edits per minute, or half a million per day— about three times as many as the English Wikipedia. About 90% of these edits are made by bots created by contributors to automate tasks, yet almost one million edits per month are still made by humans (Vrandečić and Krötzsch 2014). It supports over 350 languages, especially English, Dutch, French and German, contains more than 200 million statements on about 56 million items and has a higher edit frequency than Wikipedia,[2] and in this work it is used for Serbian.

Wikidata, as an open data network, was used by Andonovski (Андоновски 2019) to describe language resources, namely, novels from the Serbian-German literary corpus (Andonovski, Šandrih, and Kitanović 2019). Stanković and Davidović (2021) presented an example of integration of Wikidata with digital libraries and external systems, as well as the potential for speeding up the process of data preparation and entry, using articles published in the journal for digital humanities *Infotheca*, as an example. Wikidata's popularity in medicine and bioinformatics is also growing very fast. The potential use of Wikidata as a useful resource for biomedical data integration and semantic interoperability between biomedical computer systems is rising. Different knowledge resources can be automatically processed by users as well as by computer methods and programs, and it was shown how that can be useful for various medical purposes such as clinical decision support (Turki et al. 2019). The Scholia[3] project (Nielsen, Mietchen, and Willighagen 2017) is one of the first comprehensive endeavours of its kind aimed at representing bibliographical data, scholarly profiles of authors and institutions using Wikidata. To the best of our knowledge, this work is the first example of automatically imported data about literary text corpus in Wikidata using different open source tools. The main concepts and usage of Wikidata in our research are presented in Section 2.

The opportunity for speeding up the process of data preparation was seen in using information already encoded in the header of each novel (Krstev 2021) in the SrpELTeC, a subcollection of ELTeC – European Literary Text Collection[4] of novels from the period 1840-1920, developed as part of the "Distant Reading for European Literary History Cost Action"[5] (COST Ac-

2. Language Statistica for Items
3. Scholia, Scholia in Wikidata
4. ELTeC (Distant Reading for European Literary History)
5. D-reading home page

tion CA16204) by members of the JeRTeh society, led by Cvetana Krstev and Ranka Stanković (Stanković et al. 2019; Frontini et al. 2020).

Cooperation of Wikimedia Serbia[6] and their education program[7] with the University of Belgrade has a long tradition. Work on entering metadata about Serbian novels from the SrpELTeC corpus (Krstev et al. 2019) and linking Wikidata to various applications, one of which is *Aurora*,[8] has been going on for many years. Students of the Faculty of Mining and Geology are trained to populate and use Wikidata, and the application possibilities of open data are studied within the subject Presentation of Knowledge and the Semantic Web at the University of Belgrade multidisciplinary studies PhD program Intelligent Systems. Before the activities described in this paper, entry of ELTeC metadata was manual.

A set of metadata of the SrpELTeC novels, which will be presented in the third section of this paper, is extracted from the `<TEIHeader>` element, to fit the requirements of the ELTeC action schema.[9] Mapping between dataset selected from metadata defined by DR WG1 (Distant Reading working group 1) and Wikidata will be presented, as well as possibilities related to some further, optional data, such as novel's main characters, important places etc. In Section 3 Wikidata concepts will be presented and illustrated by SrpELTeC novel entities and their properties.

Since the automation of the data preparation and import process was envisaged, different solutions were analysed and finally synergy of OpenRefine[10] and QuickStatements[11] tools was chosen as the best option, similar to the approach presented in (Stanković and Davidović 2021). Elaboration of the automation process is given in Section 4.

After completing the SrpELTeC novels Wikidata, a set of web pages integrated results from different queries with different visualisation options, based on Wikidata Query Service, with the Aurora, but further integration with other systems is envisaged. Queries were written that supplied the tables: the title of the novel, the name of the author, the author's pictures,

6. Wikimedia Serbia
7. Wikimedia Educational Program
8. Aurora
9. ELTeC XML Schemas
10. OpenRefine, is a tool for working with messy data: cleaning, converting from one format to another, with the addition of external data via a web service, web page
11. QuickStatements, Wikidata editor: add and remove statements, tags, descriptions, etc., web page

the year of publication, the main characters, the author's distributions by gender, etc. The experience with using SPARQL[12] for data validation will be shared in Section 5. Data on the main characters include elementary data, which should be supplemented with new content in the future: whether the characters are fictional or not, and if they are not, their short biography.

2 Wikidata

Tim Berners Lee believed that the web will enable machines to comprehend semantic documents and data, but not human speech and writings. Properly designed, the semantic web can assist the evolution of human knowledge as a whole (Berners-Lee, Hendler, and Lassila 2001). Nowadays, the semantic web is an extension of the existing web, where information is given a precisely defined meaning, and which enables better cooperation between computers and users. The concept of the semantic web and open related data technologies extend the traditional web using standard markup language and supporting processing tools, where the RDF (Resource Description Framework),[13] a framework for describing resources on the web, plays a major role and provides more efficient solutions for finding information (Shah et al. 2002). For the semantic web operability, computers need to have access to structured collections of information and establish defined rules for automated management. Wikidata is fitting into these trends in information technology development, which are pushing the boundaries from machine readability to machine comprehensibility (understanding) of data on the web, namely from web of documents to web of data. The underlying structure of any expression (statement) in RDF is a collection of triples, each consisting of a subject, a predicate, and an object.

Wikidata is document-oriented, item-centered, representing topics, concepts, or objects and consist of two types of entities: items (e.g. https://www.wikidata.org/wiki/Q107648205) and properties (e.g. https://www.wikidata.org/wiki/Property:P1433). Each item is assigned a unique, permanent identifier "QID" or Q number, which is the unique identifier of a data item on Wikidata, comprising the letter "Q" followed by one or more digits. It is used to help people and machines understand the difference between items with the same or similar names, but different meanings. This number appears next to the name at the top of each Wikidata item. Properties

12. SPARQL
13. RDF

cannot be directly created by regular users, to prevent duplication and disorganization of Wikidata properties (e.g. Risk factor property proposal).[14]

The subject of the triple is the Wikidata item to which the claim refers, the predicate is a Wikidata property, and the object is a 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 be made more precise using qualifiers. These qualifiers show the contexts of the validity of the statement. Statements can be annotated by including references. Qualifiers and references are also represented in the form of triples, where the subject is the claim. A claim and its references are considered a statement (Turki et al. 2019). Statements are how any information known about an item is recorded in Wikidata.

The items and properties in Wikidata that are used to structure the ontology are class (Q16889133), entity (Q35120), Wikidata metaclass (Q19361238), instance of (P31), and subclass of (P279). Classes are items that conceptually group together similar items, as human (Q5) groups together humans.

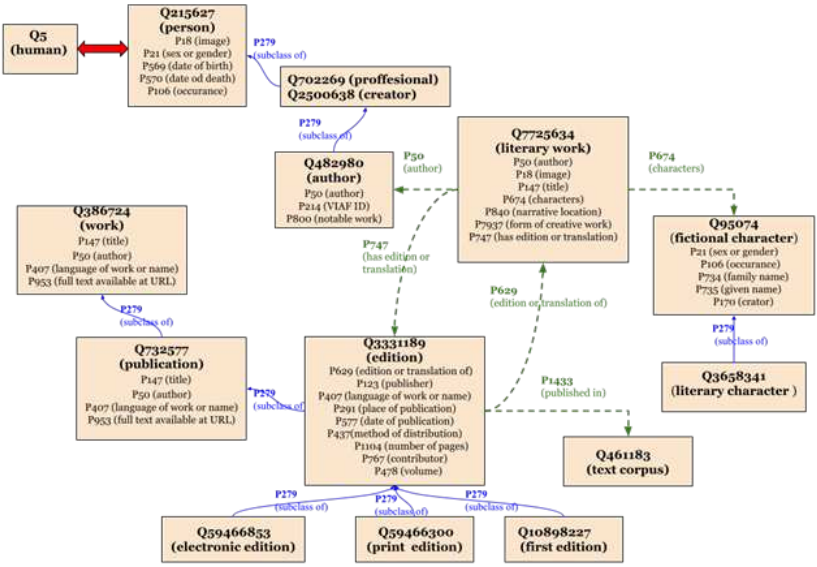

Figure 1. The class diagram of Wikidata used for novels in SrpELTeC.

14. Property proposal/Creative work

Figure 1 presents the class/instance relation of all classes and relations that are used in this paper and that will be presented in more detail in Section 3. Blue lines represent "instance of" relations between classes, while green lines present properties that connect two classes by other relations. It is very important to notice that, for the purpose of our work, the class person (Q215627), which represents a "being that has certain capacities or attributes constituting personhood" is used as class humans (Q5) with property "instance of" (P31) as recommended in Wikidata documentation.[15] The class properties are used to describe items in our work, and Figure 1 emphasizes their usage. A very important aspect of the presented work is that each novel is represented with three different items in Wikidata, which will be associated with the appropriate properties, as explained in Subsection 3.2. We will just mention here that some properties like P18 (image), P1104 (number of words), P214 (VIAF[16] ID), occurrence (P106) are general Wikidata properties and P478 (volume) is a qualifier of edition.

3 SrpELTeC Data Structure in Wikidata

To automate the entry of items in Wikidata, the first step was to retrieve and prepare the data, as presented in Subsection 3.1. The second step involved selecting XML elements and attributes in the data to be used to identify predicates and create an input schema. The schema defines the relationship of the value to the item, i.e., the subject, using predicates as intermediaries. Wikidata concepts exemplified by new SrpELTeC entities and their properties are introduced in Section 3.2.

3.1 Data Preparation and Mapping with Wikidata

The work on the ELTeC corpus envisages a basic annotation for all subcollections, the so-called level 1 annotation. The annotation consists of marking the basic structural elements of the text (chapters and other units) and some basic textual elements. The annotation was performed in accordance with the TEI recommendations (TEI Consortium 2021), where from a rich set of elements defined by these recommendations, only a subset was selected as mandatory or allowed.[17]

15. Class person

16. The Virtual International Authority File – an international reference file for authors and books that includes bibliographic and subject metadata (Loesch 2011).

17. Encoding Guidelines for the ELTeC: level 1

A metadata header `<TeiHeader>` is required for each text annotated in accordance with TEI recommendations, so it is also the case with all ELTeC corpus novels. The mandatory header elements are uniform for all collections and they must contain:

- 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). The author and the work can be joined by identifiers, such as viaf and Wikidata.
- A brief catalog description of the first edition and the edition used as the source for ELTeC (if different from the first edition).
- Description of the text in terms of meeting the balance criteria (e.g. author gender, size, time...) (see (Trtovac, Milnović, and Krstev 2021) from this issue).
- Review of all changes to the digital edition since its first publication.

The first column of Table 1 represents the properties of Wikidata that are used in statements. The XPath[18] expressions used to retrieve the metadata from the TEI header are in the second column. The extracted elements from the TEI header are the values of the property in the statement, where value type can be item, URL or string. All extracted data were labeled, the same properties labeled in the same column. The name in the third column is used in further processing and automation steps. The last column contains information about the class of the instantiated data that is used for mapping, which will be explained later.[19]

In the first step of automation, Wikidata items were added for all novels that are in the SrpELTeC collection (more in Section 4), where each novel was created as an instance of literary work (Q7725634), and related with its editions. The editions of the novel, using property P747 (has edition or translation), are connected with a novel with property P629 (edition or translation of). As shown in Figure 2, the data from the first edition and from the ELTeC edition are extracted from the TEI header and mapped to appropriate Wikidata properties, entities and values. The properties that are used to create new items for novels are some of those presented in Table 1, such as P50 (author), P146 (title), P407 (language of work or name) and also new ones such as P674 (characters) and P840 (narrative location), which will be explained later. The data for authors are extracted from TEI header and

18. XML Path Language (XPath) 3.1
19. Properties table

Currently exists on Wikidata as a property	**TEI XPath to element (attribute)**	**Name of a column in prepared data**	**Instance of**
	//fileDesc		
	/titleStmt		
P1476 (title)	/title	Title	Q783521 (title)
P214 (viaf id)	/title@ref	_ViafID	Q19832964 (VIAF ID)
P50 (author)	/author	Author	Q482980 (author)
P214 (viaf id)	/author@ref	_ViafID	Q19832964 (VIAF ID)
	/extent		
P657 (number of words)	/measure@unit	Words	Q8034324 (word count)
P1104 (number of pages)	/measure@unit	Pages	Q1069725 (page)
	/publicationStmt		
P123 (publisher)	/publisher	Publisher	Q105044823 (publisher)
P750 (distributed by)	/distributor	Distributor	Q60614978 (distributor)

mapped to Wikidata properties, such as sex or gender (P21), date of birth (P569), date of death (P570), and VIAF ID (P214). One should note that the author's date of birth (P569) and the date of death (P570) are extracted from the author element in the header (framed red in Figure 2).

Figure 2 shows an example of mapping between the metadata header of a novel *Ivkova slava* (Ivko's feast) (SRP18950) and Wikidata. Green boxes represent properties that are prefixed by P (e.g. P214 (VIAF ID), P146 (title)) and that are pointing to xml elements or attributes. The content that is framed represent values in the Wikidata statement, and if they have their own QID, they are associated with an appropriate Q identifier (e.g. Stevan Sremac (Q559989), Beograd (Q3711)). The contents that are framed but are not associated with a Q identifier (e.g. 185 (number of words), Ivkova slava: pripovetka: ELTeC izdanje (title), 1895 (date of publication)) are literals. The blue box displays information used to create item ELTeC edition (subclass of edition (Q3331189)) of a novel in Wikidata (e.g. "Ivkova slava: pripovetka: ELTeC izdanje" (Q107648205)). The orange contains information used to create the first edition of a novel in Wikidata (e.g. "Ivkova slava: pripovetka" (Q109336719)).

The narrative characters from a literary work and places where the action takes place can be found in Wikidata for well described novels (e.g. Romeo and Juliet (Q83186); Don Quixote (Q480)). This information is not a part of the metadata header and other extraction methods are required. The SrpELTeC is published in the so-called level-2[20] as well, which supplies more detailed information by annotating all words in the text with their part-of-

20. Encoding Guidelines for the ELTeC: level 2

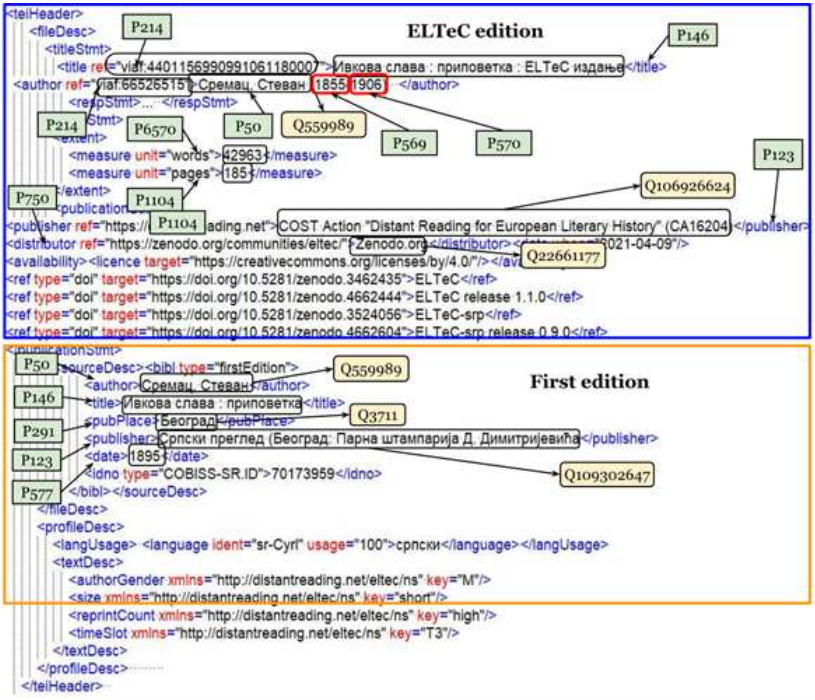

Figure 2. Mapping between metadata header and Wikidata.

Currently exists on Wikidata as a property	**TEI XPath to element (attribute)**	**Name of a column in prepared data**	**Instance of**
P6216 (copyright status)	/availability /licence @target	Licence	Q20007257 (CC BY 4.0)
	SourceDesc/bibl	FirstEdition	
P50(author)	/author	_author	Q482980 (author)
P146 (title)	/title	_title	Q783521 (title)
P291 (place of publication)	/pubPlace	_pubPlace	Q1361759 (place of publication)
P123 (publisher)	/publisher	_publisher	Q105044823 (publisher)
P577 (publication date)	/date	_date	Q1361758 (date of publication)
	//profileDesc		
P407 (language of work or name)	langUsage /language	Language	Q34770 (language)
P21 (sex or gender)	/textDesc authorGender@key	authorGender	Q290 (sex)

Table 1. Extraction of information from metadata header.

speech, lemma (word's vocabulary headword form), and optionally other morphosyntactic descriptions, as well as by annotating named entities.

The main goal of named entity recognition in general, is to indicate in a text names of persons, their roles, locations, organizations, and other relevant entities for specific purposes. The first system for recognizing named entities for Serbian is based on manually created rules, which rely on comprehensive Serbian lexical resources (Krstev et al. 2014).

At the level of the whole action, it was agreed that only 7 categories of entities should be indicated in the novels: PERS, ROLE, DEMO, ORG, LOC, WORK, EVENT, which were assessed as being of the greatest importance for further literary studies (Frontini et al. 2020). For the purpose of the work presented here, only two categories PERS and LOC are used. In the list of the extracted PERS entities the main characters of the novel can be found, 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 frequently entities in the PERS category are taken as literary characters (Q3658341), while the most frequently entities in the LOC category are taken as narrative places, i.e. geographic location (Q2221906). This task cannot be fully automated, since names of the same character can be mentioned in a text in a number of different ways, such as: *Ivko*, *Ivka*, *Ivku*, *Ivko Mijalković*.

3.2 Structure of SrpELTeC Wikidata Items

The structure of Wikidata statements allows encoding of the basic information needed to identify the topic covered by an item, without favoring any language, in order to ensure the uniqueness of the meaning of a particular term. Some examples of items used are Beograd (Q3711), Srbija (Q47561), Ivo Andrić (Q47561), Ivkova slava (Q107648205), ELTeC collection (Q106927517). It happens sometimes that there are two items under the same name, e.g. Ivkova Slava (Q107648205), which represents the novel by Stevan Sremac (Q559989), and Ivkova Slava (Q12752161), which represents the movie based on the novel and directed by Zdravko Šotra (Q1253494). It is recommended that in the case of ambiguous entities additional clarification is given in parentheses, such as Ivkova slava (movie). Thus, the item is associated with a unique identifier (QID), while the identifier is associated with a pair: title and description, in order to remove any ambiguity. In our work we used several properties as explained in Subsection 3.1 and some of these properties are used both for novels (literary work) and authors as instance of (P31) and VIAF ID (P214).

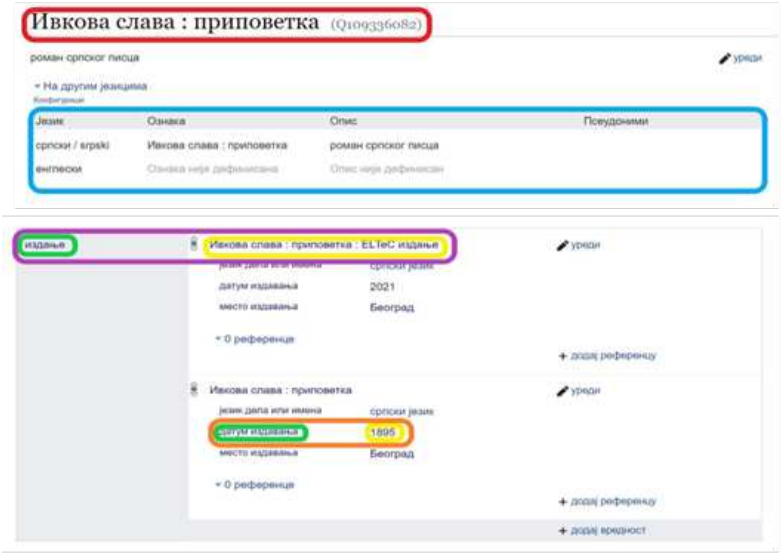

Figure 3. Wikidata data structure: the case of a novel.

Figure 4. Wikidata data structure: the case of an ELTeC edition.

Figure 3 and Figure 4 present examples of ELTeC items in Wikidata. Items include an identifier (framed by red boxes), a list of labels, a description and aliases in different languages (blue boxes), and a list of statements composed of claims (purple), qualifiers (orange) and references (gray). Claims, references and qualifiers are triples, where the predicate (green) is a Wikidata property and the object (yellow) is a value, external URL, date, string or another Wikidata item. Some items have been cropped from the image for clarity.

Every item can be described, as shown in Table 2, by using the Wikidata's unique identifier of a data item QID (e.g. Q107648205, Q559989) as the subject, properties (e.g. P50, P577, P135) and objects, which can be either a literal like "1899" or another item like Q36180, by a series of statements, each providing one fact or information about the item. Table 2 gives several examples of sentences in natural language, their annotation with Wikidata IDs and finally their encoding as Wikidata statements, using reifying for RDF triples (Hernández, Hogan, and Krötzsch 2015) where the same subject is not repeated. Statements with the same subject are separated by a semicolon ";" and the last one is finished by a dot ".".

Ivkova slava is a literary work written by Stevan Sremac.	
Ivkova Slava was published in Belgrade in 1899.	
Ivkova slava (Q107648205) is (P31) a literary work (Q7725634) written by (P50) Stevan Sremac (Q559989).	Q107648205 P31 Q7725634; P50 Q559989; P291 Q3711; P577 "1899".
Ivkova Slava (Q107648205) was published in (P291) Belgrade (Q3711) in (P577) "1899".	
Stevan Sremac was born on 23rd November 1855 in Senta, and he died on 26th august 1906. He was a writer and belonged to realism. He's VIAF ID is 66526515.	
Stevan Sremac (Q559989) was born on (P569) "23rd November 1855". in (P19) Senta (Q571136). and he died on (P570) "26th August 1906". He (Q559989) was (P106) a writer (Q36180) and belonged to (P135) realism (Q667661). He (Q559989) has VIAF ID (P214) "66526515".	Q559989 P569 "23rd november 1855"; P19 Q571136; P570 "26th august 1906"; P106 Q36180: P135 Q667661; P214 "66526515".

Table 2. Transforming natural language into Wikidata

4 SrpELTeC Wikidata Entry and Enrichment Automation

Manual population of Wikidata with individual data is often a time-consuming task. As mentioned in Section 1, the initial population of Wikidata with ELTeC editions of novels was done manually, through a user-friendly interface (Figure 3 and Figure 4). In this way 54 novels from the SrpELTeC sub-collection were described in Wikidata by University of Belgrade students through different activities. The control of manual entries revealed that some of the entries were incomplete or contained incorrect information, such as incomplete novel title, an author's VIAF ID entered as a novel's VIAF ID , connecting wrong persons as authors (e.g. football player Dušan Đurić (Q116994) instead of the writer with the same name (Q108986248)) or wrong year of the first edition. For these reasons, a systematic validation of manual entries was performed: each dataset retrieved from the headers was compared with the corresponding dataset retrieved by a SPARQL query, in

order to identify properties for which statements are missing, yielding results presented in Figure 5. Half of all missing statements were publication places (only 4 items had this information). In some statements the author was missing, which was caused by the fact that these authors were not represented in Wikidata by an QID, so the students used a string label with the author's name instead of the proper author QID. Such problems and shortcomings motivated us to start automating the whole process.

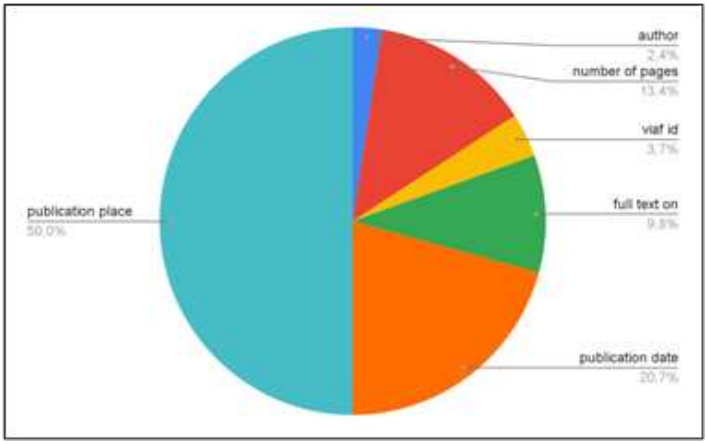

Figure 5. The overview of missing statement per property.

Successful automation of the population of Wikidata for the Infotheca journal (Stanković and Davidović 2021), showed that the population of SrpELTeC Wikidata can also be enhanced by using various procedures and tools, presented in (Turki et al. 2019). The advantage of the ELTeC collection was that the required metadata were available in the header of each novel, as described in Subsection 3.1 The procedure for extraction of all metadata from headers into one *CSV* (comma separated) file in tabular format, appropriate for further transformations, as well as transformation and exploitation of text collections, was integrated in the existing tool for creation, management and exploitation of lexical resources *Leximir* (Ranka et al. 2011).

Before automatically creating items that were missing, it was necessary to fix some problems. According to the WikiProject Books,[21] every book must

21. Wikidata:WikiProject_Books

have a property of either edition (Q3331189) or written work (Q47461344) as (P31). Collaborators working on this project manually added for 48 novels that an instance of (P31) literary work (Q7725634) is also an edition (Q3331189), which was incorrect and we had to remove that from the statement using *QuickStatement* as shown in Table 3 (blue and bold).

Statement	**Remove statement: literary work (Q7725634)**
Glava šećera: ELTeC izdanje (Q106936423) is (P31) literary work (Q7725634)	-Q106936423 P31 **Q7725634**
Bez oca i majke: ELTeC izdanje (Q108838098) is (P31) literary work (Q7725634)	-Q108838098 P31 **Q7725634**

Table 3. Removing a statement from a Wikidata item

Another problem was that students used the property page/s (P304), which represents the location of the claim, for the number of pages, and this had to be replaced by the property *number of pages* (P1104). Also, instead of the property edition (Q3331189) students used the property edition (Q397239), which represents the process of making a version of a work (usually a book). For the number of words a qualifier *determination method* (P459) had to be added in the statement, as illustrated in Table 4 (blue and bold).

Statement	**Add determination method (P459): word count (Q8034324)**
Glava šećera: ELTeC izdanje (Q106936423) number of words (P6570) 11035.	Q106936423 P6570 11035 **P459 Q8034324**
Bez oca i majke: ELTeC izdanje (Q108838098) number of words (P6570) 29321	Q108838098 P6570 29321 **P459 Q8034324**

Table 4. Adding the qualifier in a Wikidata item

For successful entry of novels, the entry of their authors is an indispensable step and a precondition, because the authors have to be separate items in the Wikidata. After checking whether the authors exist in Wikidata by using SPARQL queries, which will be described in Section 5, it was found that

37 out of 68 different authors from the ELTeC collection already existed in Wikidata. For some existing authors some missing properties were detected and added, e.g. P214 (VIAF ID), P21 (gender) and P106 (occupation). For missing authors, items were automatically created, with description and properties, such as gender of the author, date of birth and date of death, which were extracted from the column author. It should be noted that for some authors particular information was missing in metadata (as unknown), for example, there was no VIAF ID, date of birth and/or date of death. For the sake of simplicity, the process of entering authors in Wikidata will not be described, and we will focus on the entry of novels in Wikidata.

For the purpose of our work, it was necessary to add two instances for each novel: the first is a novel as a literary work (Q7725634), which is the subclass of written work (Q47461344), and the second is an edition as instance of (Q3331189). For the edition, two instances are recorded: the first edition (Q10898227) and the electronic, i.e. ELTeC edition (Q59466853). Since we wanted to automate the process of preparing and entering information, we tried different solutions and ended up using two of them, namely, *OpenRefine* and *QuickStatements*. In order to successfully automate the process, several mandatory steps were required. For the first step, the actual preparation of input data, a custom procedure was written that extracts metadata from the TEI file header in tabular form, suitable for further automation, as explained in the Subsection 3.1. Some information for a few novels were missing in TEI headers, because the author was unknown, or the year or the place of the first edition were unknown, while the majority of novels did not have their VIAF ID.

For the novels that were already in the Wikidata and for which some statements were missing, we had to fix all the missing fields. The data were labeled, so that each column represented one statement (predicate) of the novel. It was also necessary to select labels to identify predicates and to create a Wikidata schema. The schema allowed the item to be automatically linked to Wikipedia. Before creating the schema, reconciling each column was necessary. During column reconciliation, a very important process was identification of 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. At this stage, manual verification of data was possible and their correction, if necessary. Each column contained information that was extracted as was presented in Table 1. Some examples of reconciling cells for ELTeC edition of novels are the following:

1. **Title** (P1476) to an entity of type *edition*, *version*, or *translation* (Q3331189)
2. **Author** (P50) to an entity of type *human* (Q5) and then search for match
3. **Language of work or name** (P407) to an entity of type *language* (Q34770)
4. **Number of pages** (P1104) to an entity of type *natural number* (Q21199)
5. **Number of words** (P6570) to an entity of type *natural number* (Q21199)
6. **Published in** (P1433) to an entity of type *text corpus* (Q461183)
7. **VIAF ID** (P214) to an entity of type *VIAF ID* (Q19832964)
8. **Full work available at URL** (P953) to an entity of type *URL* (Q42253)
9. **Publication date** (P577) to an entity of type *calendar year* (Q3186692)
10. **Place of publication** (P291) to an entity of type *city* (Q515)
11. **Volume** (ID of novel) (P478) to an entity of type *volume* (Q1238720)

The next step was editing the Wikidata schema by using OpenRefine. Creating a Wikidata input set schema defines predicates (properties) that will connect subjects and objects in RDF triples. Each statement for a subject has a property and value that can be a Wikidata item, external URL, or literal (string). As presented in Table 1, the property from the first column is related to content (values: items or literals) in the third column. After editing and saving the Wikidata schema it was exported as a *QuickStatements* file. A few lines from this file are given in Figure 6. In the final stage the prepared file was exported in the *QuickStatements* tool and Wikidata items were automatically created.

5 The Overview of SrpELTeC@Wikidata by SPARQL Queries

In this section, we will present a statistical overview of the status of srpELTeC collection in Wikidata, illustrated by characteristic SPARQL queries and their results. We created SPARQL queries for various views, using the integrated technologies in Wikidata to visualize the results. We wrote queries that retrieved the tables with columns for: the title of the novel, the name of the author, the author's pictures, the year of publication, the authors distribution by gender, etc.

```
CREATE
LAST    Lsr "Ђул-Марикина прикажња : приповетка : ELTeC издање"
LAST    Dsr "ELTeC издање романа српског писца"
LAST    P31 Q3331189
LAST    P1433   Q106927517
LAST    P1433   Q106936149
LAST    P1476   sr:"Ђул-Марикина прикажња : приповетка : ELTeC издање"
LAST    P50 Q3625974
LAST    P407    Q9299
LAST    P577    +2021-00-00T00:00:00Z/9
LAST    P291    Q3711
LAST    P1104   107
LAST    P6570   20244
LAST    P953    "https://distantreading.github.io/ELTeC/srp/SRP19012.html"
LAST    P478    "SRP19012"
```

Figure 6. *QuickStatements* file for creating the statements of a novel.

The first validation using SPARQL, retrieved authors that already existed in the Wikidata and it was later used for statistical overview. Before adding Wikidata items, the number of authors in the srpELTeC collection in Wikidata was only 38, while now there are 69 authors, with more than 300 statements as illustrated in Figure 7.

The following query lists authors and novel titles with default view as tree:

```
#defaultView:Tree
SELECT DISTINCT ?author ?authorLabel ?novel ?novelLabel
WHERE {
   # novel published in (P1433) ELTeC collection (Q106927517)
   ?novel wdt:P1433 wd:Q106927517;
          # novel instance of (P31) literary work (Q7725634)
          wdt:P31 wd:Q7725634.
   # show the author (P50) of the novel if there is one
   OPTIONAL {?novel wdt:P50 ?author}
SERVICE wikibase:label
{bd:serviceParam wikibase:language "sr,[AUTO_LANGUAGE],en".}}
```

The statement `?novel P1433 Q106927517` in `WHERE` clause retrieves all novels (?novel) that are published (P1433) in ELTeC: European Literary Text Collection (ELTeC) 1850-1920 (Q106927517). The rows that starts with "#" are comments, introduced to help understand the query. The prefix *wdt:* stands for namespace http://www.wikidata.org/prop/ used for properties and prefix *wd:* is used for objects (QIDs) for namespace https://www.wikidata.org/wiki/. The result of the previous query is given in Fig-

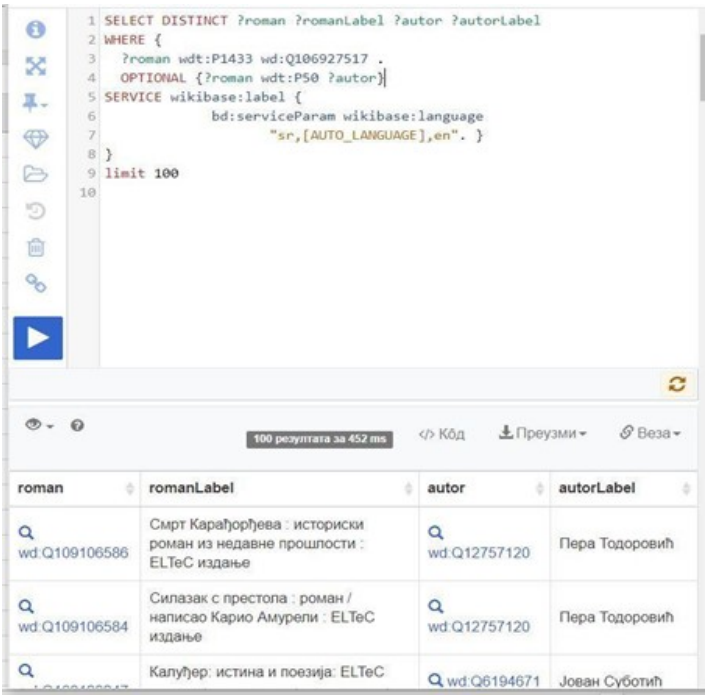

Figure 7. *Wikidata Query Service* with an example.

ure 8; the whole query and results can be retrieved by Wikidata query service at the link https://w.wiki/4Lja.

The process of entering novels into Wikidata using OpenRefine and QuickStatment was very successful. As a result, there are now 100 novels in Wikidata that are part of the Serbian ELTeC sub-collection and also 10 novels that are in the Serbian extended ELTeC sub-collection, with more than 700 statements.

Using Wikidata Query Service we can display, for example, all novels in the ELTeC collection that have a VIAF ID, number of pages and number of words, with the following query:

```
# defaultView:BubbleChart
SELECT DISTINCT ?novel ?novelLabel ?num_pages ?num_words ?viaf
WHERE {
  # novel published in (P1433) SrpELTeC coll. (Q106936149)
  ?novel wdt:P1433 wd:Q106936149;
```

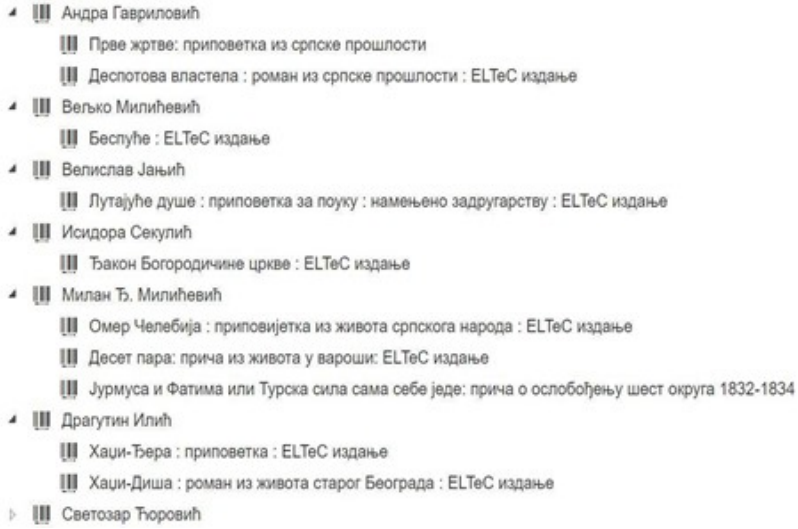

Figure 8. The graph of authors and their works.

```
        # number of pages (P1104)
        wdt:P1104 ?num_pages;
        # number of words (P6570)
        wdt:P6570 ?num_words;
        # viaf id (P214)
        wdt:P214 ?viaf.
SERVICE wikibase:label
{bd:serviceParam wikibase:language "sr,[AUTO_LANGUAGE],en".}}
```

The result of this query is represented in Figure 9, where the size of the circle reflects the number of pages in a novel. Full query results can be retrieved by Wikidata query service on the following link: https://w.wiki/4i9k.

For some novels we imported pictures of cover pages using the Wikimedia commons[22] repository; presently, we are preparing pictures of cover pages for the remaining novels. Figure 10 represents the timeline visualization of novels, sorted by year of their first publication, which was obtained with the following query (https://w.wiki/4LjP):

```
#defaultView:Timeline
SELECT DISTINCT ?novel ?novelLabel ?image ?date ?author
```

22. Upload Wizard

Figure 9. Bubble chart visualization.

```
?authorLabel
WHERE {
  # novel published in (P1433) ELTeC collection (Q106927517)
  ?novel wdt:P1433 wd:Q106927517;
         # has edition or translation (P747)
         wdt:P747 ?edition.
         # edition instance of (P31) first edition (Q10898227)
         ?edition wdt:P31 wd:Q10898227;
         # image (P18)
         wdt:P18 ?image.
  # optional date of publication (P577)
  OPTIONAL { ?edition wdt:P577 ?date. }
  # optional author (P50)
  OPTIONAL { ?novel wdt:P50 ?author. }
SERVICE wikibase:label
{bd:serviceParam wikibase:language "sr,[AUTO_LANGUAGE],en".}}
```

One of the possible view options is a map preview for queries that have coordinates in the output list. Figure 11 presents a map with places of novel publication for the following query (https://w.wiki/4hSR):

Figure 10. The timeline visualization.

```
# defaultView:Map
# names of the publication places for the first editions
SELECT DISTINCT ?place ?coor
WHERE {
  # edition instance of (P31) first edition (Q10898227)
  ?edition wdt:P31 wd:Q10898227;
         # published in (P143)
         # ELTeC collection (Q106927517)
         wdt:P1433 wd:Q106927517;
         # publication place (P291)
         wdt:P291 ?place.
         # place coordinate location (P625)
         ?place wdt:P625 ?coor.
SERVICE wikibase:label
{bd:serviceParam wikibase:language"sr,[AUTO_LANGUAGE],en".}}
```

It is also possible to visualize data as interactive graphs of authors and ELTeC editions (https://w.wiki/4j6D), where the click on an item reveals a set of its properties and related items (Figures 12 and 13).

```
#defaultView:Graph
SELECT DISTINCT ?author ?authorLabel ?edition
?editionLabel
WHERE {
  # published in (P1433) ELTeC collection (Q106927517)
```

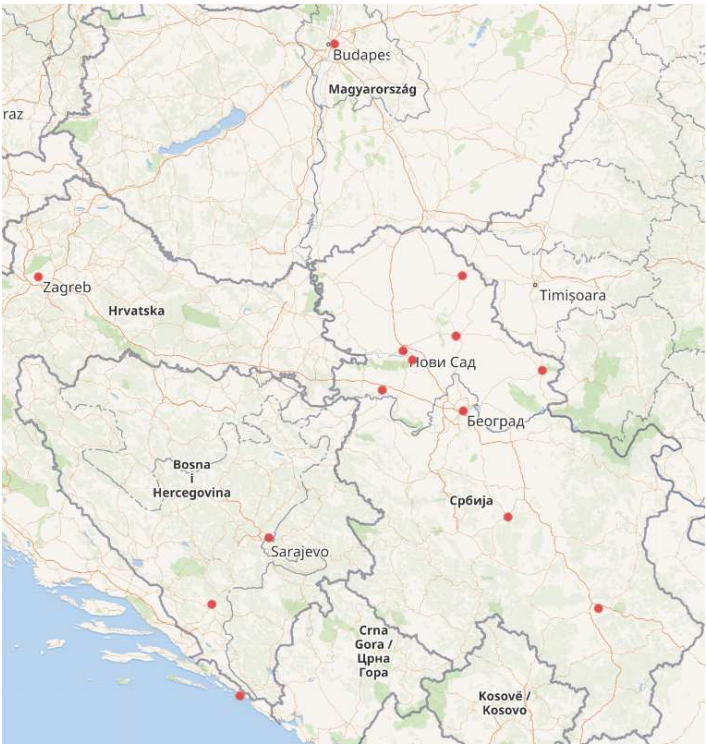

Figure 11. The map with places in which novels were first published.

```
  ?edition wdt:P1433 wd:Q106927517;
         # instance of (P31)
         # version, edition, or translation (Q3331189)
         wdt:P31 wd:Q3331189.
         # publisher (P123)
         # COST action "Distant Reading for European"
         # Literary History" (CA16204) (Q106926624)
         wdt:P123 wd:Q106926624.
  # optional author (P50)
  OPTIONAL {?edition wdt:P50 ?author}
SERVICE wikibase:label
{bd:serviceParam wikibase:language"sr,[AUTO_LANGUAGE],en".}}
```

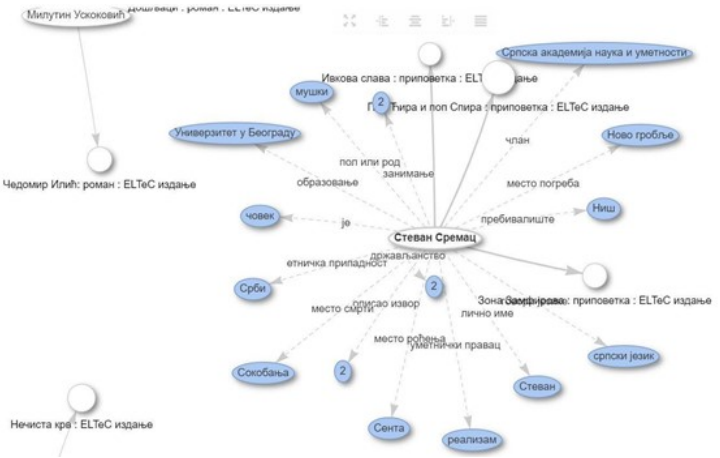

Figure 12. Visualization of an author (Stevan Sremac) and the set of his properties.

As we have already mentioned, the result of this work is that now each novel in Wikidata has items for two editions, the first edition and the electronic (ELTeC edition). Currently, there are 110 ELTeC novels in Wikidata – 100 from SrpELTeC and 10 from SrpELTeC-extended, and since each novel and its associated editions have at least 20 statements, the results is that there are more than 2500 statements for the whole SrpELTeC collection.

6 Conclusions and Future work

This paper presented the automation of the preparation and import of data to Wikidata, illustrated by SrpELTeC, the Serbian sub-collection in the ELTeC multilingual collection (European Literary Text Collection). After the extraction of metadata from TEI headers, mapping with Wikidata schema was defined and the synergy of OpenRefine and QuickStatements tools was used for import. As a result of this work, there are now 110 novels from the ELTeC collection in Wikidata, with associated items for the first edition and the electronic ELTeC editions. That means that approximately 2500 statements were automatically added. Future research will use a list of locations, associated with different texts in the corpus, to explore ways to enrich and relate this data to knowledge bases and build a larger context around it. Also, we plan to add data on the main characters in Wikidata, which will

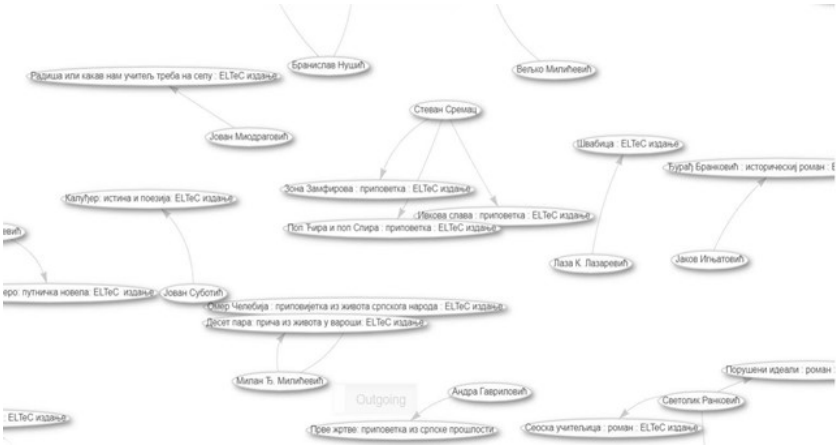

Figure 13. The graphs of ELTeC novels and their authors.

include some basic data: gender, profession, whether the character is fictional or not, and if the character is real, a short biography will be entered. With the basic information for each novel (birthplace of author, residence at time of writing, place of publication), one can begin to relate the ELTeC geodata (place of publication and places of narrative) to other time/space coordinates, and consider more detailed mapping visualizations. The analysis of available data about other editions will be explored, as well as other data related to the novel. The research presented is language independent, and the same approach can be used for automation of data import for other ELTeC collections.

Acknowledgment

The text collection preparation is supported by the COST Action 16204 – Distant Reading for European Literary History support. Linked data development is supported by the Wikimedia Serbia "WikiELTeC – Wikidata about old Serbian novels from collection ELTeC". The authors thank the master students of the Department of Library and Information Sciences of the Faculty of Philology, University of Belgrade for their help in filling in the Wikidata with data on the characters from the novels and the places where the action of the novels takes place.

References

Andonovski, Jelena, Branislava Šandrih, and Olivera Kitanović. 2019. "Bilingual lexical extraction based on word alignment for improving corpus search." *The Electronic Library.*

Berners-Lee, Tim, James Hendler, and Ora Lassila. 2001. "The semantic web." *Scientific american* 284 (5): 34–43.

Frontini, Francesca, 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.*

Hernández, Daniel, Aidan Hogan, and Markus Krötzsch. 2015. "Reifying RDF: What works well with wikidata?" *SSWS@ ISWC* 1457:32–47.

Krstev, Cvetana. 2021. "The Serbian Part of the ELTeC Collection through the Magnifying Glass of Metadata." *Infotheca - Journal for Digital Humanities* 21 (2): 26–42. ISSN: 2217-9461. https://doi.org/10.18485/infotheca.2021.21.2.2.

Krstev, Cvetana, Jelena Jaćimović, Branislava Šandrih, and Ranka Stanković. 2019. "Analysis of the first Serbian Literature Corpus of the Late 19th and Early 20th century with the TXM platform." *DH Budapest 2019,* http://elte-dh.hu/wp-content/uploads/%202019/09/DH_BP_2019-Abstract-Booklet.pdf.

Krstev, Cvetana, Ivan Obradović, Miloš Utvić, and Duško Vitas. 2014. "A system for named entity recognition based on local grammars." *Journal of Logic and Computation* 24 (2): 473–489.

Loesch, Martha Fallahay. 2011. "VIAF (The Virtual International Authority File)–http://viaf. org." *Technical Services Quarterly* 28 (2): 255–256.

Nielsen, Finn Årup, Daniel Mietchen, and Egon Willighagen. 2017. "Scholia, scientometrics and Wikidata." In *European Semantic Web Conference,* 237–259. Springer.

Ranka, Stanković, Obradović Ivan, Krstev Cvetana, and Vitas Duško. 2011. "Production of morphological dictionaries of multi-word units using a multipurpose tool." In *Proceedings of the Computational Linguistics-Applications Conference, October 2011, Jachranka, Poland,* 77–84.

Shah, Urvi, Tim Finin, Anupam Joshi, R Scott Cost, and James Matfield. 2002. "Information retrieval on the semantic web." In *Proceedings of the eleventh international conference on Information and knowledge management,* 461–468.

Stanković, Ranka, and Lazar Davidović. 2021. "Infotheca (Q25460443) in Wikidata." *Infotheca - Journal for Digital Humanities* 21 (1): 87–98. https://doi.org/10.18485/infotheca.2021.21.1.5.

Stanković, Ranka, Diana Santos, Francesca Frontini, Tomaž Erjavec, and Carmen Brando. 2019. "Named Entity Recognition for Distant Reading in Several European Literatures." *DH Budapest 2019.*

TEI Consortium, ed. 2021. *TEI P5: Guidelines for Electronic Text Encoding and Interchange. 4.3.0(31-08-2021).* TEI Consortium. http://www.tei-c.org/Guidelines/P5/.

Trtovac, Aleksandra, Vasilije Milnović, and Cvetana Krstev. 2021. "The Serbian Part of the ELTeC Collection – from the Empty List to the 100 Novels Collection." *Infotheca - Journal for Digital Humanities* 21 (2): 7–25. ISSN: 2217-9461. https://doi.org/10.18485/infotheca.2021.21.2.1.

Turki, Houcemeddine, Thomas Shafee, Mohamed Ali Hadj Taieb, Mohamed Ben Aouicha, Denny Vrandečić, Diptanshu Das, and Helmi Hamdi. 2019. "Wikidata: A large-scale collaborative ontological medical database." *Journal of biomedical informatics* 99:103292.

Vrandečić, Denny, and Markus Krötzsch. 2014. "Wikidata: a free collaborative knowledgebase." *Communications of the ACM* 57 (10): 78–85.

Андоновски, Јелена. 2019. "Мрежа отворених података и језички ресурси у процесу изградње српско-немачког литерарног корпуса." PhD diss., Универзитет у Београду, Филолошки факултет.