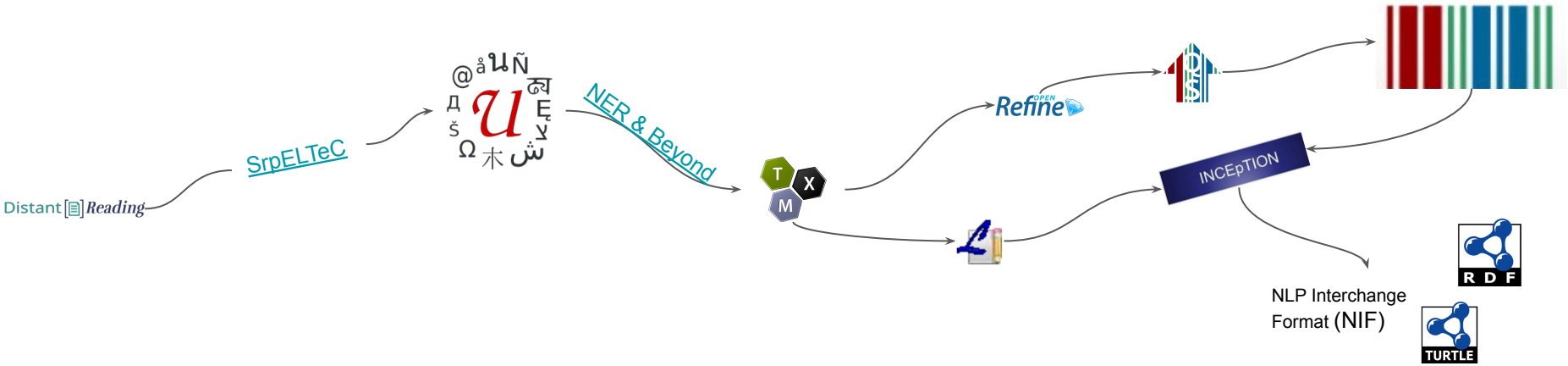
Named Entities' Journey from Unitex to Wikidata

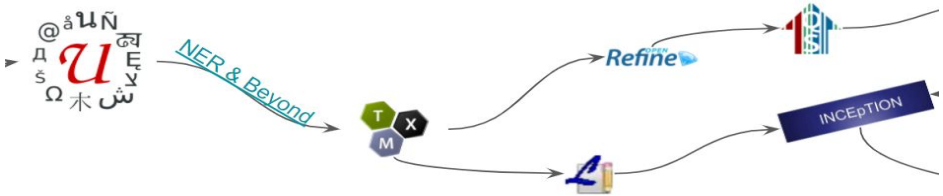


Ranka Stanković, University of Belgrade

Journées « Unitex », 15-16 Décembre 2021

Key topics

- Named Entities: recognition, annotation, extraction, linking with Wikidata
- Speeding up the process of data preparation and entry
- Tools to support the process:



- Use case:
  - Serbian ELTeC collection of novels from period 1840-1920
  - Wikidata data model and examples
  - SPARQL queries of SrpELTeC Wikidata
  - NEL (with Wikidata) in INCEpTION

The Serbian ELTeC collection of novels from period 1840-1920

- The COST action CA16204 began in 2017 and will end at the beginning of 2022.
- One of the most important goals of this action is the preparation of a multilingual corpus (called European Literary Text Collection - ELTeC) https://www.distant-reading.net/eltec/
- When fully completed, ELTeC will contain 100 novels first published in the period 1840-1920. for a large number of European languages
    - {cze, deu, eng, fra, hun, pol por, rom, slv, srp} subcollections with 100+ novels
    - 7 subcollections are under development for level-1
- SrpELTeC collection contains
    - 100 novels that are part of ELTeC and
    - 13 novels from SrpELTeC-ext extended collection
- TEI-conformant ODD specification https://github.com/COST-ELTeC/Schemas

Distant Reading

SrpELTeC text formats: TEI - level 1

- Text preparation
  - <teiHeader> with rich information about author, novel, publication and digital version
  - In <text> structural elements: , <head> , <pb>,<milestone/> , <quote>, <l>, <foreign>, <ref>, <note>, <hi>,...
  - more in Special Issue of Infotheca “The Serbian Part of the ELTeC Collection through the Magnifying Glass of Metadata” (Krstev, 2021)
- As TEI XML:
  https://github.com/COST-ELTeC/ELTeC-srp/tree/master/level1
- As HTML:
  https://distantreading.github.io/ELTeC/srp/index.html
- More on level-1 schema:
  https://distantreading.github.io/Schema/eltec-1.html

```
<quote>
<l>Мито море, Митанче,</l>
<l>Цркни ми, пукни, душманче </l>
<l>Мито море, Мито, Мито!</l>
</quote>
<p>И то „Мито“ све прихватише и почеше са толико чежње, страсти,
лудила да певају и изговарају, да и сама амамџика, раздрагана,
смешећи се, мораде да оде, још јаче да притвори спољна амамска
врата, да се тамо, у чаршији, не би чуло и око амама почели да
скупљају мушки. </p>
<pb n="158"/>
<p>Симка већ беше обучена и села у чело астала. Испред ње је већ
стајало стакло ракије, кисела грожђа, крушака и паприка. Око ње
почеше и остале да се ређају, освежавајући се воћем и осталим
слаткишима.</p>
<p>И Софка је грабила, да би била брзо готова. Сиђе и она, седе међ
њих. Почеше и њу да служе, и, на изненађење свих, не само што поче
јести, него испи и неколико чаша ракије, а једнако са оним задржаним
осмехом на устима, с очима мало стиснутим и обрвама једва приметно
набраним, као смејући се свима и све их сажаљевајући...</p>
</div>
<div type="chapter" xml:id="SRP19101_C16">
<head>XVI</head>
<p>У полумрак вратиле се из амама. И све док су пролазиле чаршијом,
прелазиле онај мост, увијене, погнуте и крijући се једна уз другу,
```

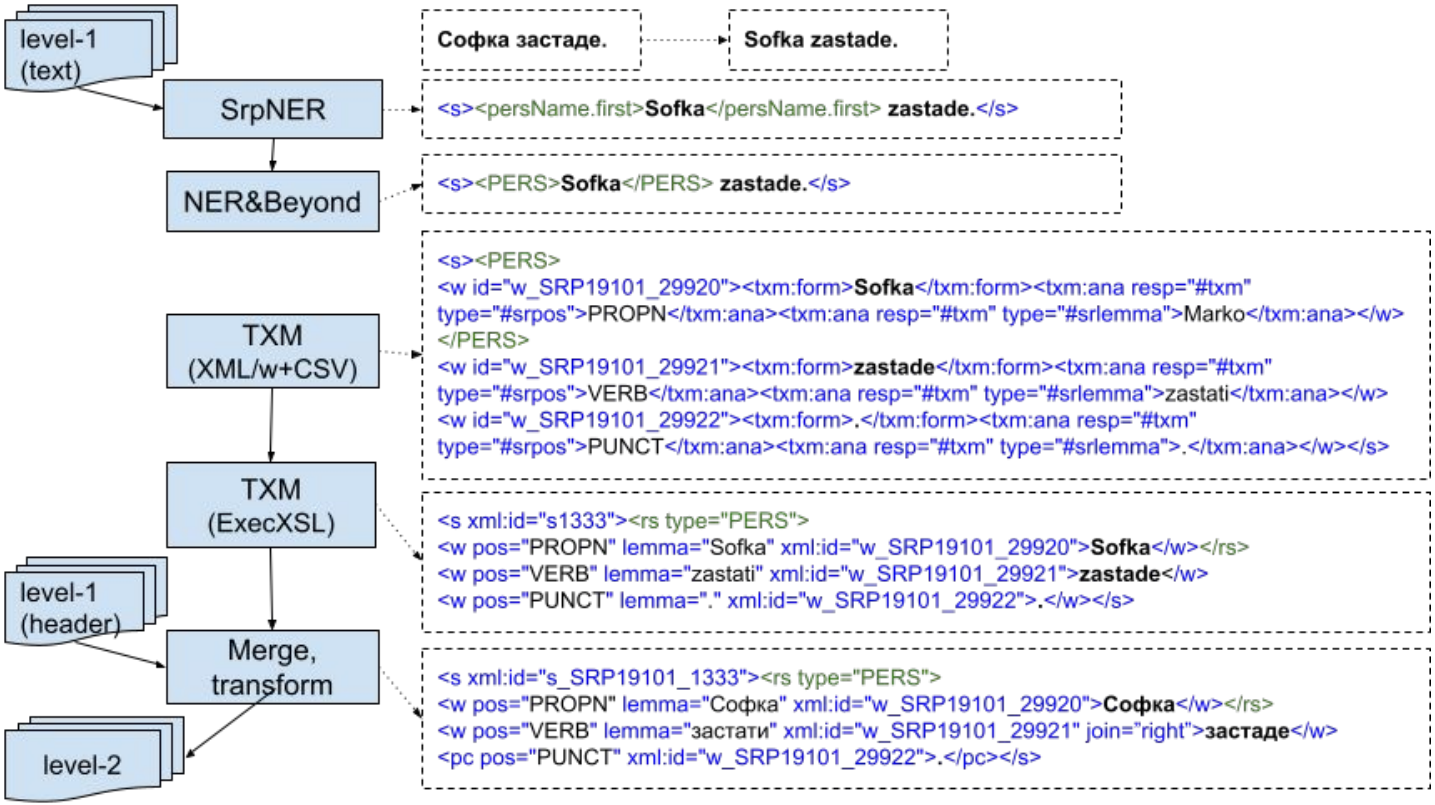
ELTeC level-2 specification

- ELTeC level-2 is level-1 augmented by: lemmatisation, POS-tagging (UD tagset), NER annotation, morphosyntactic annotation optional, direct speech and sentiment analyses envisaged
- ELTeC level-2 includes all elements existing in level-1 and introduces some new ones:
  - <s> as the sentence tag, used for segmentation of text into sentences, and
  - <w> and <pc> for tokenization
    - mandatory linguistic attributes @pos, @lemma, and @join,
    - optional @xml:id, @msd
  - <rs> (referring string) for NER with @type attribute: **PERS, LOC, ORG, DEMO, WORKS, EVENTS, ROLE**
- The element <s> can contain a sequence of <w> elements, either directly or in the sub-paragraph elements <corr>, <emph>, <foreign>, <hi>, <label>, <title>.
- More about level-2: https://distantreading.github.io/Schema/eltec-2.html

ELTeC level-2 production

- Two main problems encountered in producing Serbian ELTeC level-2 were similar to those encountered for other languages:
  - 1) majority of morphosyntactic taggers do not work well with XML format and
  - 2) harmonization of NER and morphosyntactic annotations, which are performed separately with different tools.
- A solution was found in the TXM tool (https://txm.gitpages.huma-num.fr/textometrie/),
  - an environment that enables tagging of XML files,
  - solved the problem of alignment of NER and morphosyntactic tags.

The Serbian SrpELTeC Level-2 pipeline



SrpNER in Unitex

- Sentence splitting performed by Unitex transducer (Krstev 2008; Paumier 2021)
  - adapted to consider level-1 tags
  - outputs the start tag <s> at the beginning of a sentence and the end tag </s> at its end
- NER performed by the rule-based system SrpNER (Krstev et al. 2014), based on large-scale lexical resources for Serbian (Krstev, 2008), coupled with local grammars in the form of finite-state transducers (Vitas and Krstev, 2012).
- Since SrpNER works on Latin texts, it is necessary to transliterate Cyrilic texts to Latin.
- SrpNER recognises 11 classes of NEs:
  - dates and time (moments and periods),
  - money and measurement expressions,
  - geopolitical names (countries, settlements, oronyms and hydronyms), and
  - personal names (last, first, nicknames, full, …).
- Semi-automatic removal of the embedded NER tags from the SrpNER output

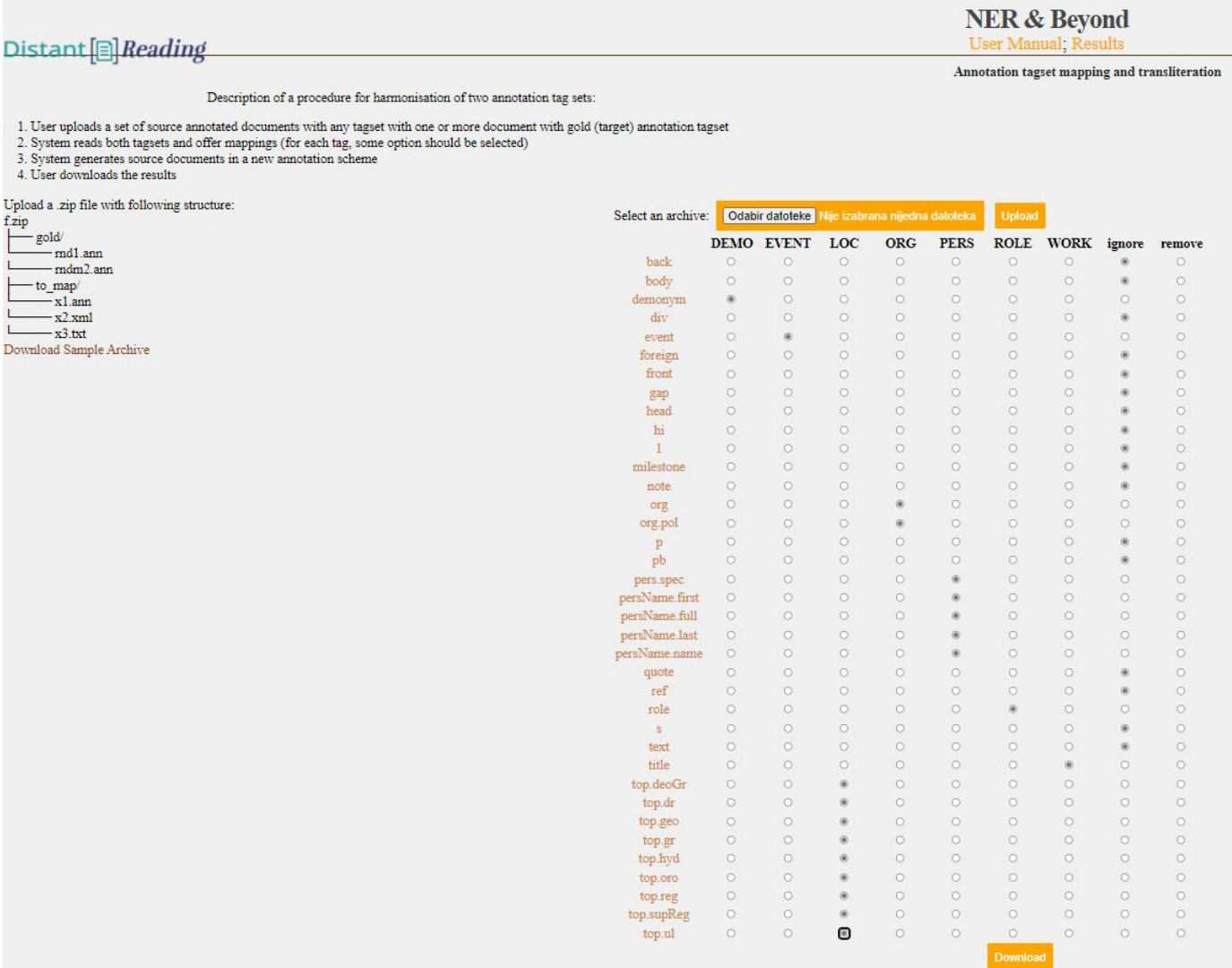
1. <pers.spec><role>carice</role> <persName.full>Marije Terezije</persName.full></pers.spec>
2. <pers><role>Sekretar</role> <persName.last>Živanović</persName.last></pers>
3. <org>Saborne crkve u <top.gr>Beogradu</top.gr></org>
4. <org>manastira <pers.spec>Sv. Marka</pers.spec></org>
5. <pers>Ali-<pers><persName.last>Milevu</persName.last></pers></pers>

1. <role>carice</role> <persName.full>Marije Terezije</persName.full>
2. <role>Sekretar</role> <persName.last>Živanović</persName.last>
3. <org>Saborne crkve u Beogradu</org>
4. <org>manastira Sv. Marka</org>
5. <pers>Ali-Milevu</pers>

The tagsets mapping in the NER&Beyond portal

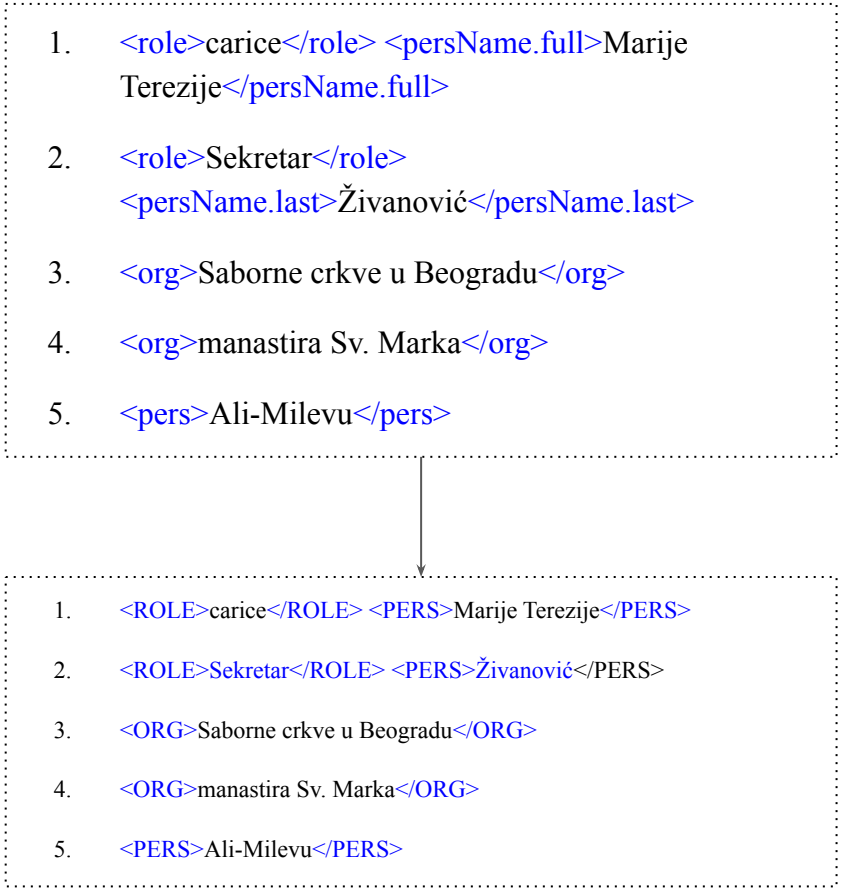
http://nerbeyond.jerteh.rs/

- Level-2 tagset: PERS, ROLE, LOC, ORG, DEMO, EVENT, WORK
- An automatic procedure implemented as part of the NER&Beyond portal (Stanković et al. 2019; Šandrih Todorović et al. 2021) was developed and used to map SrpNER tags into the 7-categories ELTeC NER schema.



The tagsets mapping

- The mapping procedure allows mapping, ignore or removal of XML elements.
- The following XML elements are ignored: back, body, div, foreign, front, gap, head, hi, l, milestone, note, p, pb, quote, ref, s, text
- Mapping is defined as follows:
  - pers.spec, persName.first, persName.full, persName.last, persName.name → **PERS**
  - top.deoGr, top.dr, top.geo, top.gr, top.hyd, top.oro, top.reg, top.supReg, top.ul → **LOC**
  - demonym → **DEMO**
  - event → **EVENT**
  - org, org.pol → **ORG**
  - role → **ROLE**
  - title → **WORK**



TXM: POS-tagging and lemmatisation

- TXM (Heiden, 2010) import option “XML/w+CSV”
  - CSV file with metadata
  - TreeTagger serbian model used for POS tagging and lemmatization (Stanković et al. 2020)
  - tokenization applied by a set of rules (not perfect)
- TreeTagger also requires a lexicon and Serbian morphological dictionaries were used as a lexicon for training
- POS tagset: Universal dependencies.

```
<s><PERS>
 <w id="w_SRP19101_29920"><txm:form>Sofka</txm:form>
   <txm:ana resp="#txm"   type="#srpos">PROPN</txm:ana>
   <txm:ana resp="#txm"
type="#srlemma">Sofka</txm:ana></w></PERS>
 <w id="w_SRP19101_29921"><txm:form>zastade</txm:form>
   <txm:ana resp="#txm" type="#srpos">VERB</txm:ana>
   <txm:ana resp="#txm" type="#srlemma">zastati</txm:ana></w>
 <w id="w_SRP19101_29922"><txm:form>.</txm:form>
   <txm:ana resp="#txm" type="#srpos">PUNCT</txm:ana>
   <txm:ana resp="#txm" type="#srlemma">.</txm:ana></w>
</s>
```

Additional transformations

execXSL macro:
txm-front-teitxm2xmlw.xsl
adapted for level-2

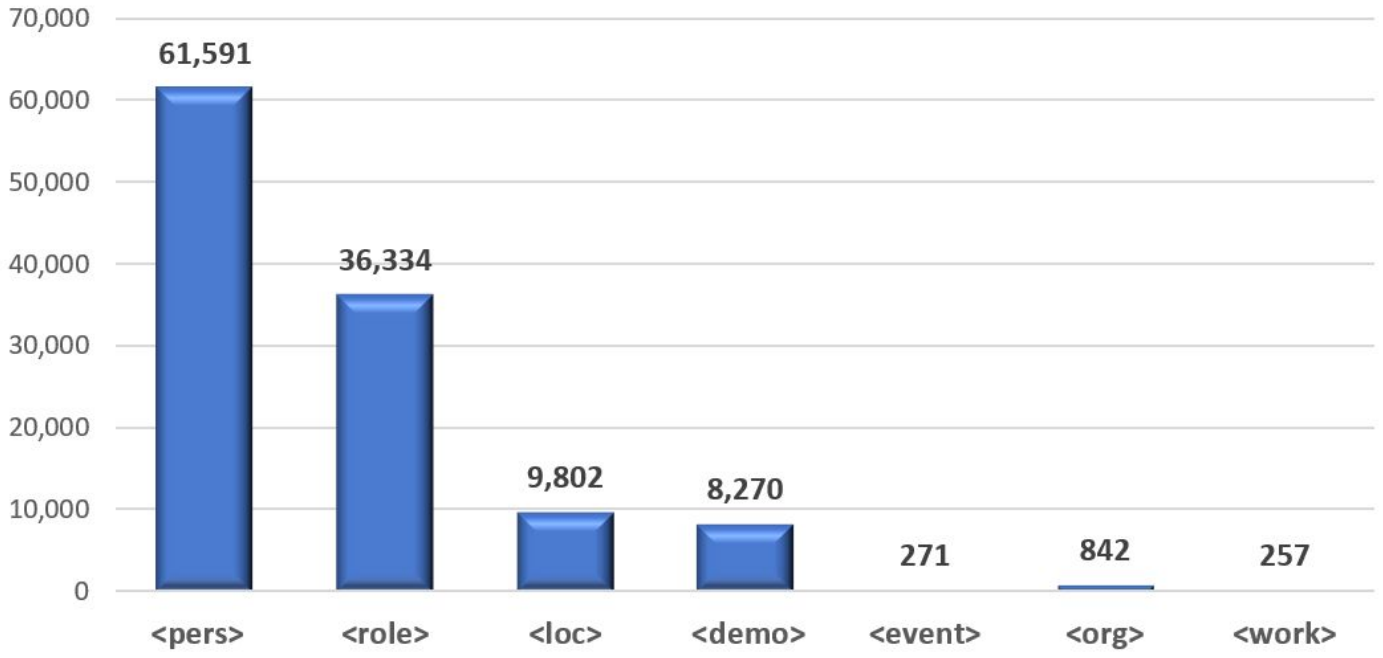
- sentence counting
- use of required namespaces for the attributes xml:id, xml:lang
- removing some attributes
- mapping XML elements for NER tags
  - **<PERS>, <LOC>, <ORG>, <DEMO>, <ROLE>, <WORK>, <EVENT>**
  - **<rs>, with @type from {PERS, LOC, ORG, DEMO, ROLE, WORK, EVENT}**

```
<xsl:template match=
"tei:PERS|tei:LOC|tei:ORG|tei:DEMO|tei:ROLE|tei:WORK|tei:EVENT">
    <!-- produce a referring string element -->
    <xsl:element name="rs" namespace="http://www.tei-c.org/ns/1.0">
        <xsl:attribute name="type"><xsl:value-of select="local-name()"/>
        </xsl:attribute>
            <xsl:apply-templates select="tei:w"/>
            <xsl:apply-templates select="tei:foreign"/>
    </xsl:element>
</xsl:template>
```

```
<s xml:id="s1333"><rs type="PERS">
<w pos="PROPN" lemma="Sofka"
xml:id="w_SRP19101_29920">Sofka</w></rs>
<w pos="VERB" lemma="zastati"
xml:id="w_SRP19101_29921">zastade</w>
<w pos="PUNCT" lemma="." xml:id="w_SRP19101_29922">.</w></s>
```

```
<s xml:id="s_SRP19101_1333"><rs type="PERS">
<w pos="PROPN" lemma="Софка"
xml:id="w_SRP19101_29920">Софка</w></rs>
<w pos="VERB" lemma="застати" xml:id="w_SRP19101_29921"
join="right">застаде</w>
<pc pos="PUNCT" xml:id="w_SRP19101_29922">.</pc></s>
```

Number of occurrences of NER elements by class (in 65 novels of SrpELTeC)



Frequencies of **PERS** AND **ROLE** categories in 65 novels of SrpELTeC

PERS	F	NumN	RelFKc
bog	2211	0.52	358.50
Boža	641	0.57	112.82
Milan	601	0.2	60.60
Pera	536	0.24	50.66
Miloš	1203	0.19	41.46
Mara	732	0.24	40.94
Milana	335	0.18	36.19
Stojan	572	0.16	34.57
Srba	257	0.32	31.59
Danica	672	0.14	31.07
Marka	382	0.31	31.00
Jov	428	0.21	30.25
Ljubica	671	0.15	28.08
Sima	634	0.19	26.97
Jelena	366	0.1	26.77
Nikola	356	0.21	23.12
Darinka	765	0.08	22.19
Milica	244	0.12	19.09
Ana	563	0.13	17.94
Steva	481	0.14	17.37
Mari	230	0.22	17.33
Sava	265	0.24	16.39
Petar	253	0.24	15.73
Jova	193	0.17	15.00
Ivan	362	0.15	14.72

srlemma	F	Num	RelFK
gospodin	2238	0.57	431.87
pop	1257	0.44	156.53
gospođa	1123	0.43	144.23
kapetan	1075	0.31	142.40
učitelj	996	0.45	135.87
gazda	655	0.51	96.91
seljak	558	0.52	95.79
g .	650	0.40	90.66
gospodar	639	0.41	70.64
kmet	810	0.30	63.31
ministar	453	0.24	59.29
doktor	407	0.35	58.37
đak	590	0.36	50.28
gospođica	426	0.33	47.37
trgovac	312	0.47	46.51
predsednik	433	0.27	45.07
činovnik	232	0.40	36.38
sluga	289	0.43	34.89
lekar	283	0.35	27.26
car	294	0.30	27.22
pandur	272	0.28	25.79
knez	416	0.24	25.16
pisar	386	0.27	23.49
poslanik	178	0.25	21.04
radnik	192	0.33	19.78

Frequencies of **DEMO** AND **LOC** categories in 65 novels of SrpELTeC

DEMO	F	NumNo	RelFKol
Turčin	1788	0.37	159.95
srpski	1047	0.52	148.41
turski	674	0.41	71.44
Srbin	273	0.33	33.20
Ciganin	139	0.26	12.54
beogradski	166	0.27	11.98
nemački	93	0.25	9.82
Srpkinja	83	0.19	8.43
ruski	128	0.21	6.13
grčki	95	0.2	5.29
francuski	67	0.17	3.82
Arnautin	137	0.07	3.76
Grkinja	158	0.04	3.14
Hrvat	58	0.06	2.75
Grk	74	0.14	2.20
Nemac	43	0.14	1.96
Madžar	67	0.1	1.94
Arapin	84	0.09	1.76
užički	35	0.09	1.72
Vlah	26	0.11	1.69
Rus	79	0.11	1.69
carigradski	35	0.11	1.31
Srba	24	0.11	1.11
Bugarin	37	0.11	1.10
Ciganka	31	0.11	1.06

LOC	Freq	Num	RelFK
Beograd	785	0.52	120.39
Srbija	594	0.42	57.25
Niš	118	0.20	7.44
Rusija	157	0.20	6.98
Dunavo	118	0.24	6.93
Kosovo	111	0.24	6.78
Sava	105	0.20	5.24
Beč	99	0.19	5.10
Carigrad	107	0.18	4.45
Morava	81	0.17	4.39
Turska	89	0.19	3.79
Pariz	66	0.16	3.22
Kragujevac	62	0.16	3.03
Pešta	102	0.11	2.75
Dunav	60	0.18	2.72
Stambol	50	0.13	2.33
Bosna	86	0.09	2.05
Užice	62	0.08	1.98
Kruševac	65	0.08	1.80
Zemun	30	0.14	1.65
Šumadija	30	0.15	1.59
Karlovci	74	0.07	1.51
Kalemegdan	43	0.07	1.39
Srem	24	0.15	1.37
Rudnik	90	0.07	1.31

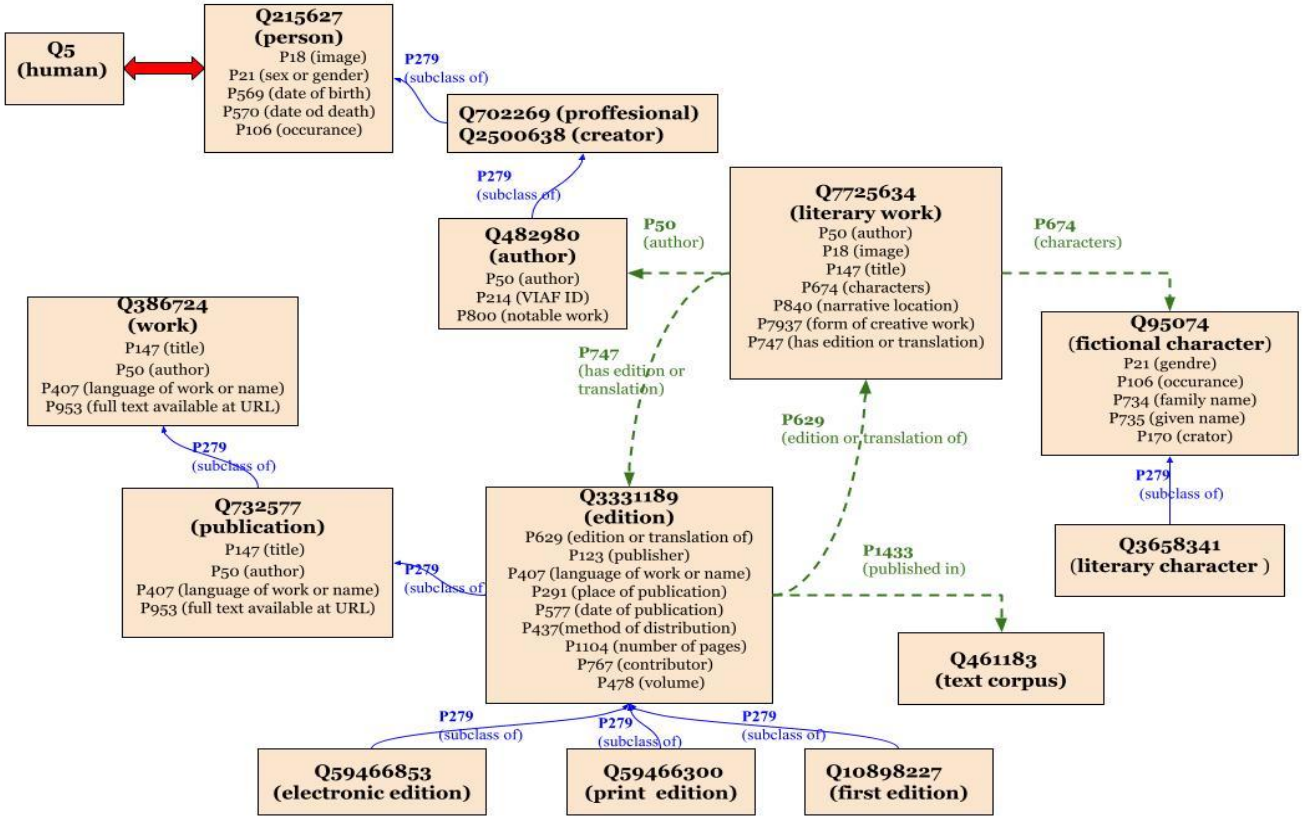
SrpELTeC data structure in Wikidata

Data preparation and mapping with Wikidata

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

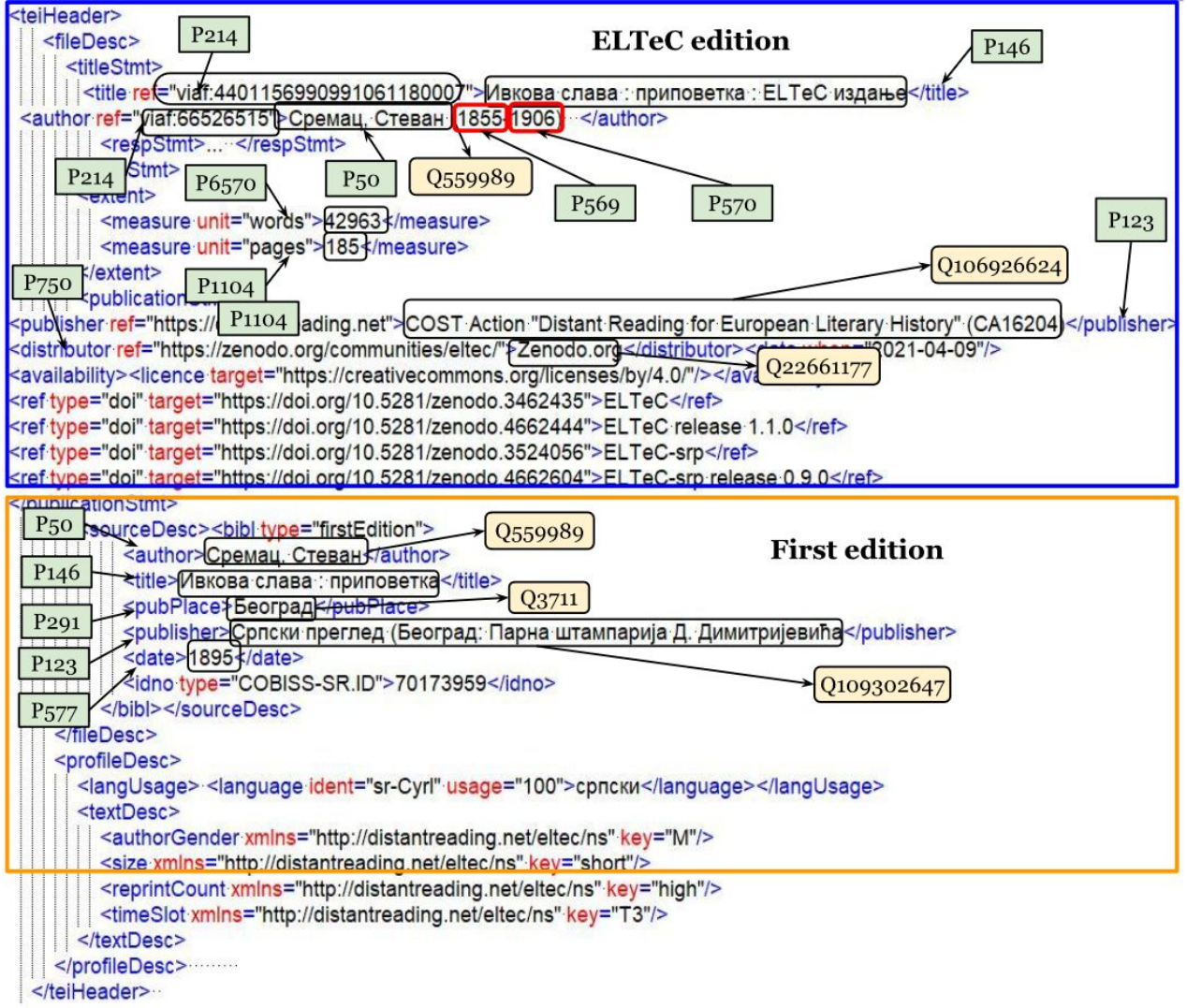
1. Description of the electronic edition, which includes the title of the work and the name of the author as well as the statements of responsibility (scanning, correction, annotation), date of publication, size (measured by the number of words). The author and the work can be joined by identifiers, such as viaf and wikidata.
2. A brief catalog description of the first edition and the edition used as the source for ELTeC (if different from the first edition).
3. Description of the text in terms of meeting the balance criteria (e.g. author's gender, size, date of publication,...).
4. Review of all changes to the digital edition since its first publication.

Wikidata - ELTeC data model: classes and instances

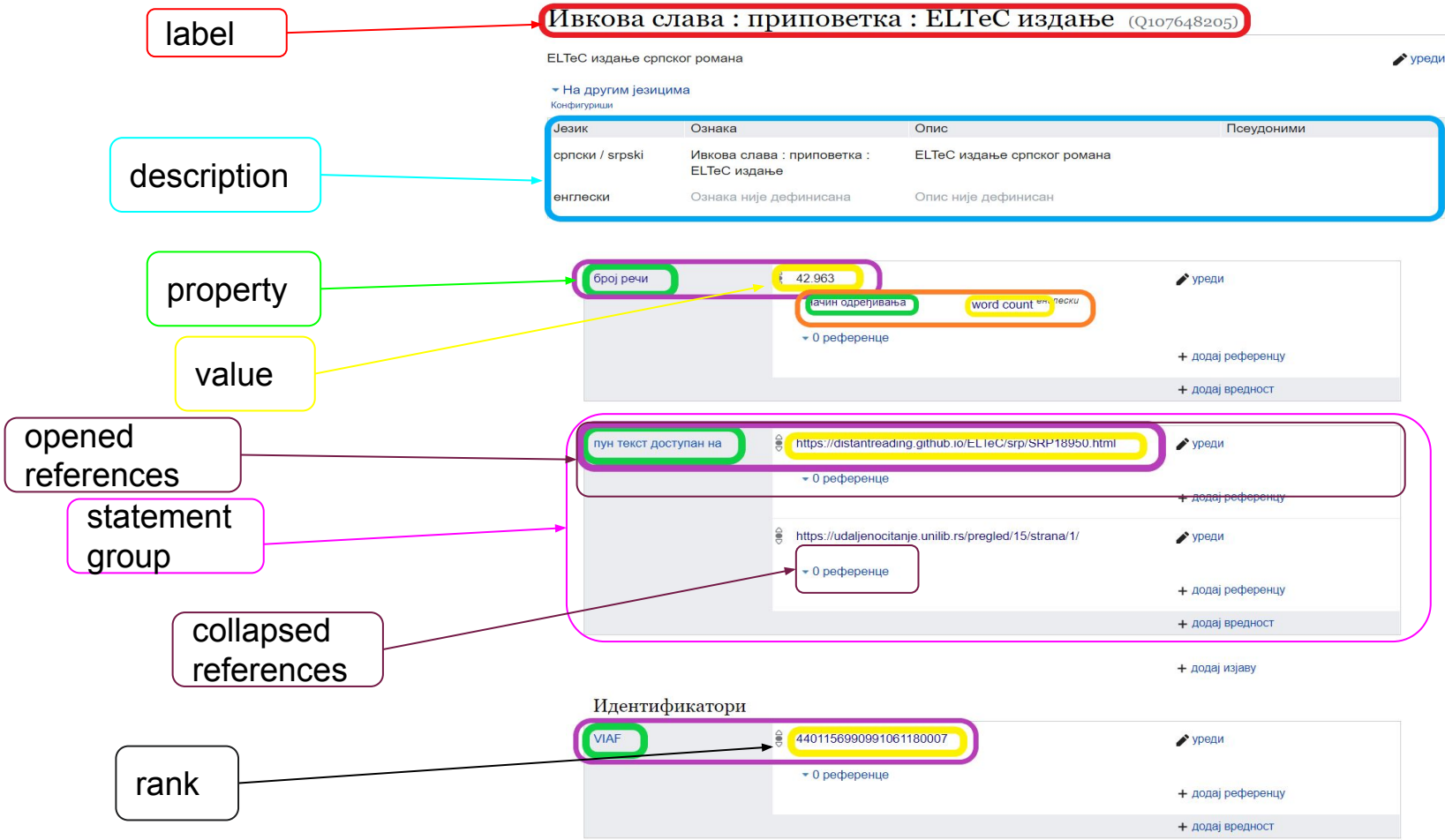


Wikidata properties	TEI XPath to element (attribute)	Name of column in prepared data	instance of
P214 (viaf id)	//fileDesc/titleSmt/title@ref	Title_ViafID	Q19832964 (VIAF ID)
P1476 (title)	//titleDesc.titleSmt/title	Title	Q783521 (title)
P50 (author)	//fileDesc/titleSmt/author	Author	Q482980 (author)
P214 (viaf id)	//fileDesc/titleSmt/author@ref	Author_ViafID	Q19832964 (VIAF ID)
P657 (number of words)	//fileDesc/extent/measure@unit	Words	Q8034324 (word count)
P1104 (number of pages)	//fileDesc/extent/measure@unit	Pages	Q1069725 (page)
P123 (publisher)	//fileDesc/publicationSmt/publisher	Publisher	Q105044823 (publisher)
P750 (distributed by)	//fileDesc/publicationSmt/distributor	Distibutor	Q60614978 (distributor)
P6216 (copyright status)	//fileDesc/publicationSmt/availability/licence@target	Licence	Q20007257 (CC BY 4.0)
P50(author)	//fileDesc/bible/author	FirstEdition_author	Q482980 (author)
P146 (title)	//fileDesc/bible/title	FirstEdition_title	Q783521 (title)
P291 (place of publication)	//fileDesc/bible/pubPlace	FirstEdition_pubPlace	Q1361759 (place of publication)
P123 (publisher)	//fileDesc/bible/publisher	FirstEdition_publisher	Q105044823 (publisher)
P577 (publication date)	//fileDesc/bible/date	FistEdition_date	Q1361758 (date of publication)
P407 (language of work or name)	//profileDesc/langUsage/language	Language	Q34770 (language)
P21 (sex or gender)	//textDesc/authorGender	authorGender	Q290 (sex)

Mapping between SrpELTeC metadata and Wikidata



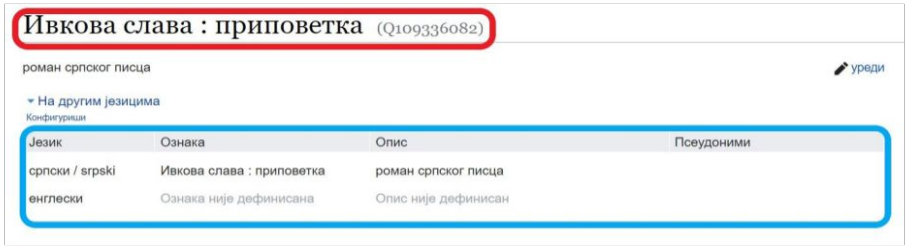
Structure of SrpELTeC Wikidata items (example of ELTeC edition)

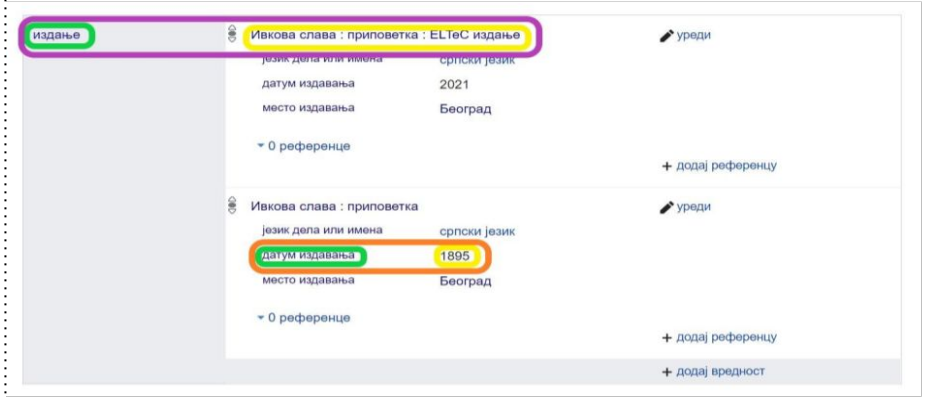


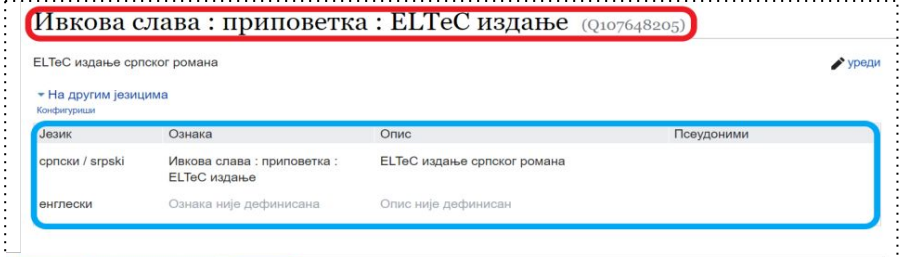
Wikidata data structure for ELTeC novels

the case of a novel:
as a literary work (Q7725634)

the case of edition:
as instance of edition (Q3331189)
1) first edition (Q10898227) and
2) electronic (ELTeC) edition (Q59466853)







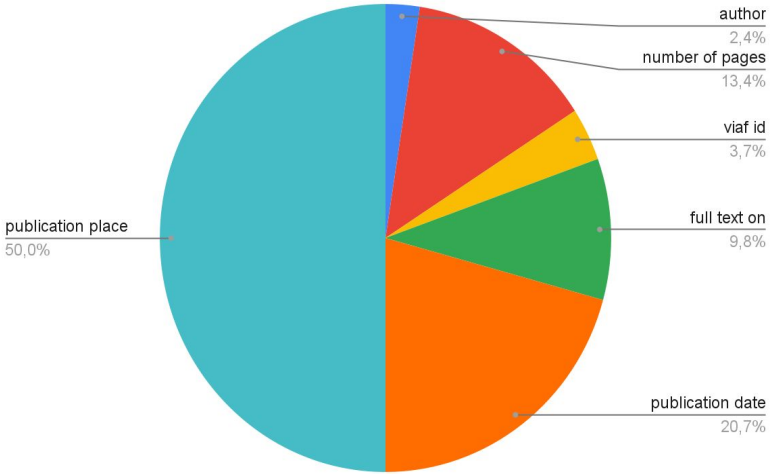
Transforming natural language into Wikidata

Ivkova slava is a literary work written by Stevan Sremac.	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 was published in Belgrade in 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".

SrpELTeC Wikidata entry and enrichment automation

The manual population of individual data (in our case ELTeC editions of novels) into Wikidata is often a time-consuming task. In this way 54 novels from the SrpELTeC sub-collection were described in Wikidata by students at the University of Belgrade 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 entered as a work' s VIAF , connecting wrong people as authors (e.g. a football player Dušan Đurić (Q116994) with the same name as a writer (Q108986248)) or wrong year of the first edition.

Is there a possibility for the input to be faster?



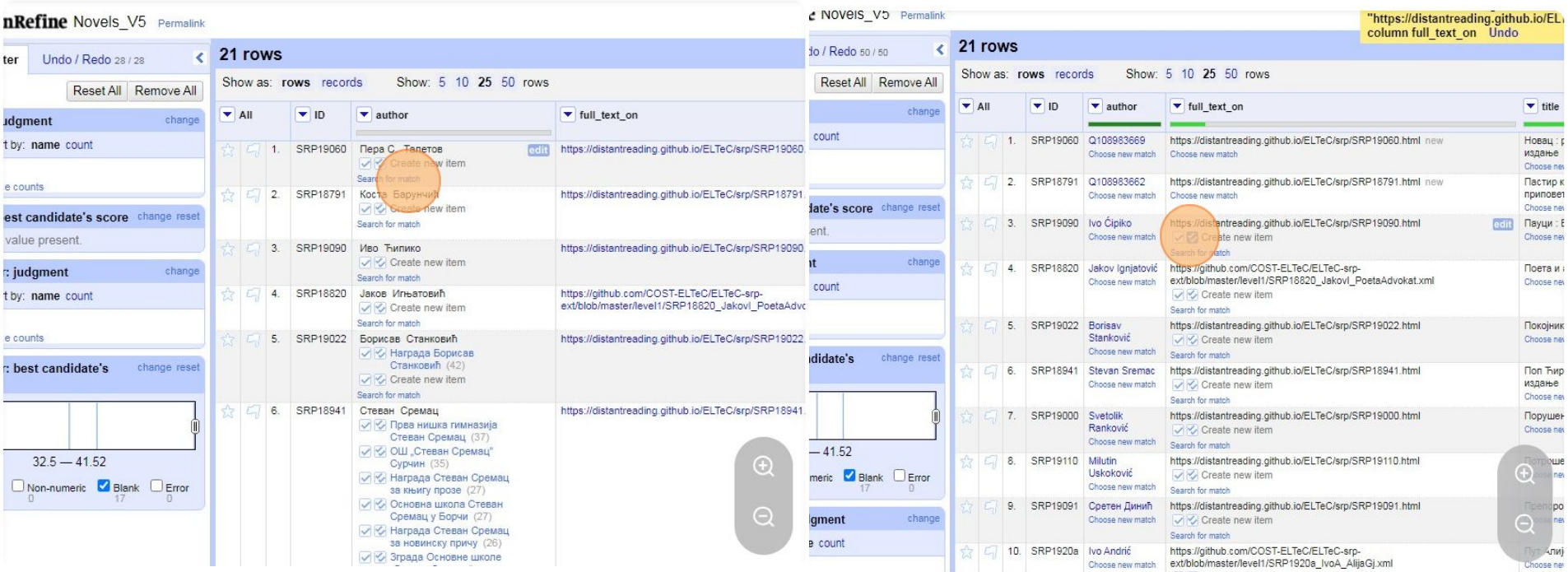
Automatization with web resources

Since the automation of the process of data preparation and import was envisaged (using <TeiHeader> data), the different solutions were analysed and finally synergy of OpenRefine and QuickStatements tools was chosen as the best choice.

[Tutorial for automatization using OpenRefine and QuickStatement]

Tutorial for automatization using OpenRefine and QuickStatement

… Step by step, ilustrated tutorial with SrpELTeC text collection



When creating new item "Create new item" manual option was chosen for each cell, to check the results.
There is an option to automatically create a new item, for each cell in the column.

Examples of reconciling cells for ELTeC edition of novels

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)
    For all novels their identification (ID) assigned as the property volume of a book. e.g. SRP19012

Finally… 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 on slide 18, the property from the first column is related to content (values: items or literals) in the third column.

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.

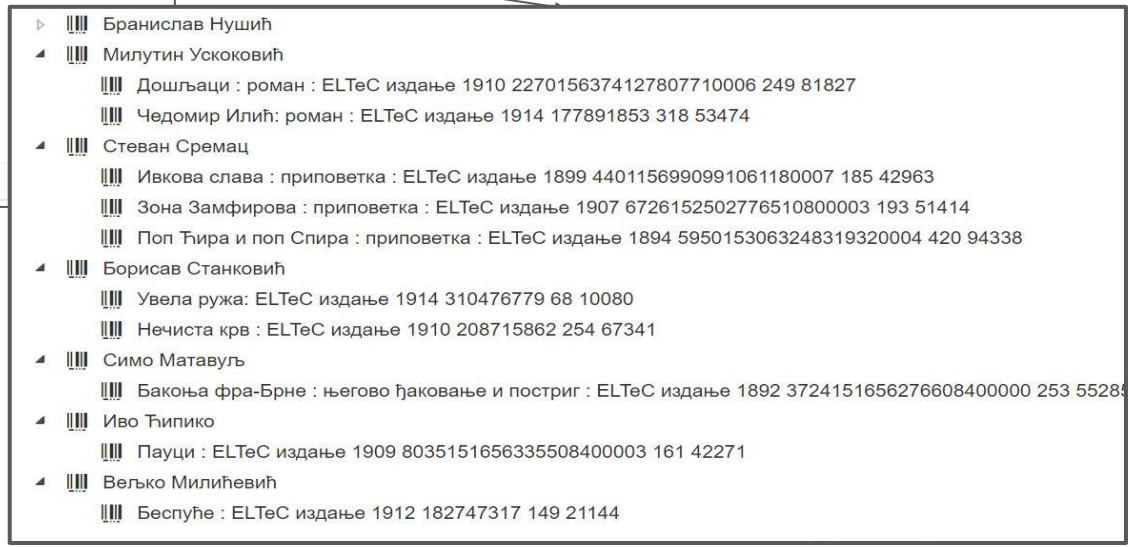
In the final stage the prepared file was exported in the QuickStatements tool and Wikidata items were automatically created.

Wikidata properties	TEI XPath to element (attribute)	Name of column in prepared data	instance
P214 (viaf id)	//fileDesc/titleSmt/title@ref	Title_ViafID	Q19832964 (VIAF
P1476 (title)	//titleDesc.titleSmt/title	Title	Q783521 (title)
P50 (author)	//fileDesc/titleSmt/author	Author	Q482980 (author)
P214 (viaf id)	//fileDesc/titleSmt/author@ref	Author_ViafID	Q19832964 (VIAF
P657 (number of words)	//fileDesc/extent/measure@unit	Words	Q8034324 (word c
P1104 (number of pages)	//fileDesc/extent/measure@unit	Pages	Q1069725 (page)
P123 (publisher)	//fileDesc/publicationSmt/publisher	Publisher	Q105044823 (publ
P750 (distributed by)	//fileDesc/publicationSmt/distributor	Distibutor	Q60614978 (distrib
P6216 (copyright status)	//fileDesc/publicationSmt/availability/licence@target	Licence	Q20007257 (CC BY

```
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"
```

Overview of SrpELTeC@Wikidata by SPARQL queries

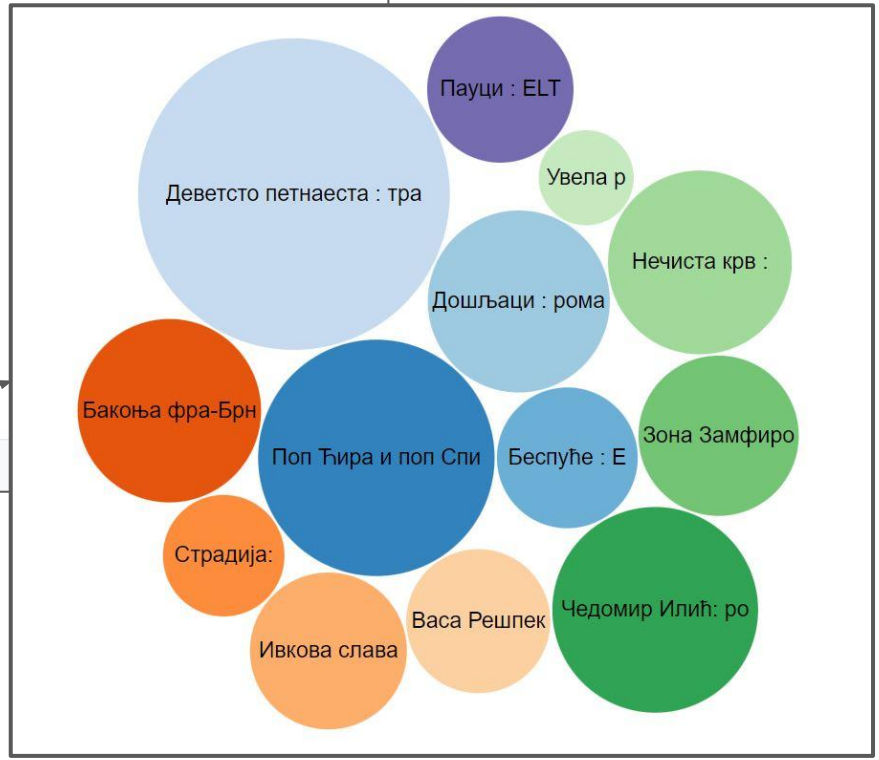
Visualization of authors of novels.
https://w.wiki/4Lja



Wikidata Query Service | Примери | Query Builder | Помоћ | Још алата

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

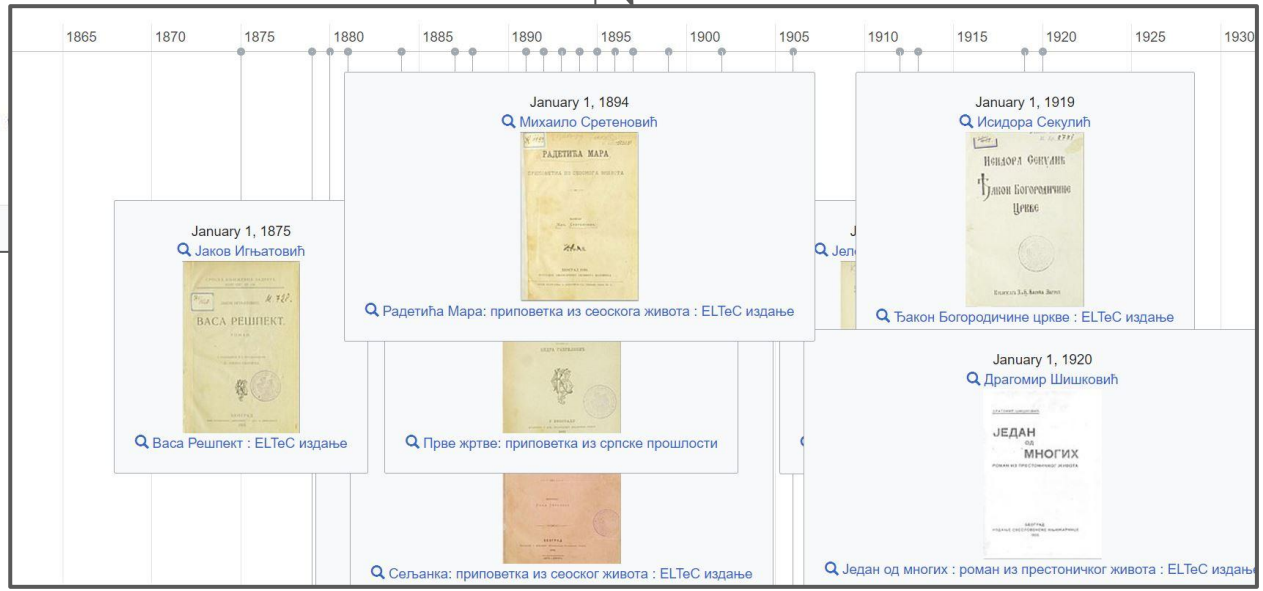
Bubble chart of novels with number of pages, number of words and VIAF ID.
https://w.wiki/4LjC



Wikidata Query Service | Примери | Query Builder | Помоћ | Још алата

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

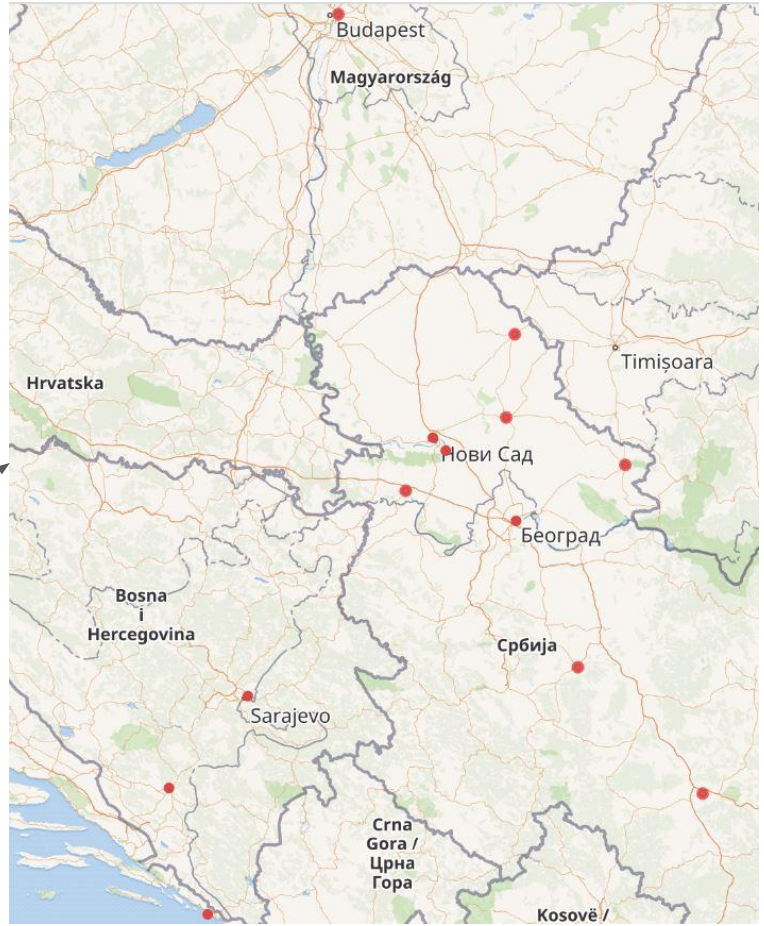
Timeline visualization of novels sorted by year of first publication
https://w.wiki/4LjZ



Wikidata Query Service | Примери | Query Builder | Помоћ | Још алата

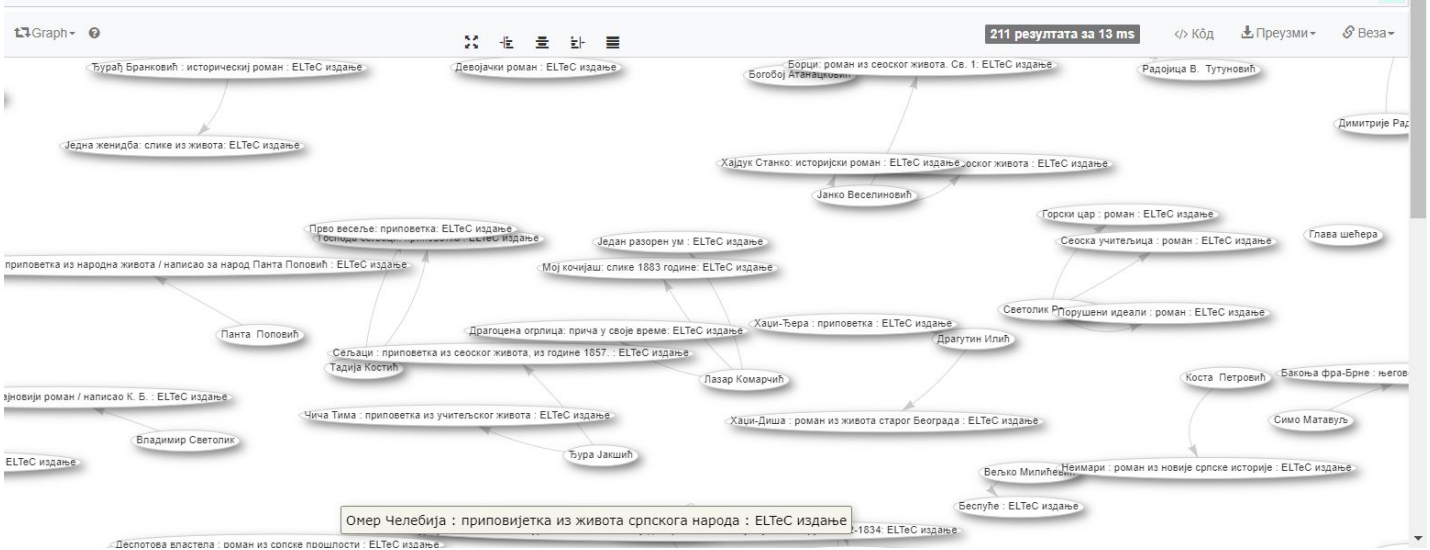
```
#defaultView:Map
# presenting place of first edition
SELECT DISTINCT ?place ?coor
WHERE {
  # edition instance of (P31) edition (Q3331189)
  ?edition wdt:P31 wd:Q3331189;
        # published in (P143) ELTeC collection (Q106927517)
         wdt:P1433 wd:Q106927517;
         # publication place (P291)
         wdt:P291 ?place.
         # place coordinate location (P625)
         ?place wdt:P625 ?coor.
  SERVICE wikibase:label { bd:serviceParam wikibase:language "sr,[AUTO_LANGUAGE],en". }
}
```

Map with places of novel publication https://w.wiki/4Lje

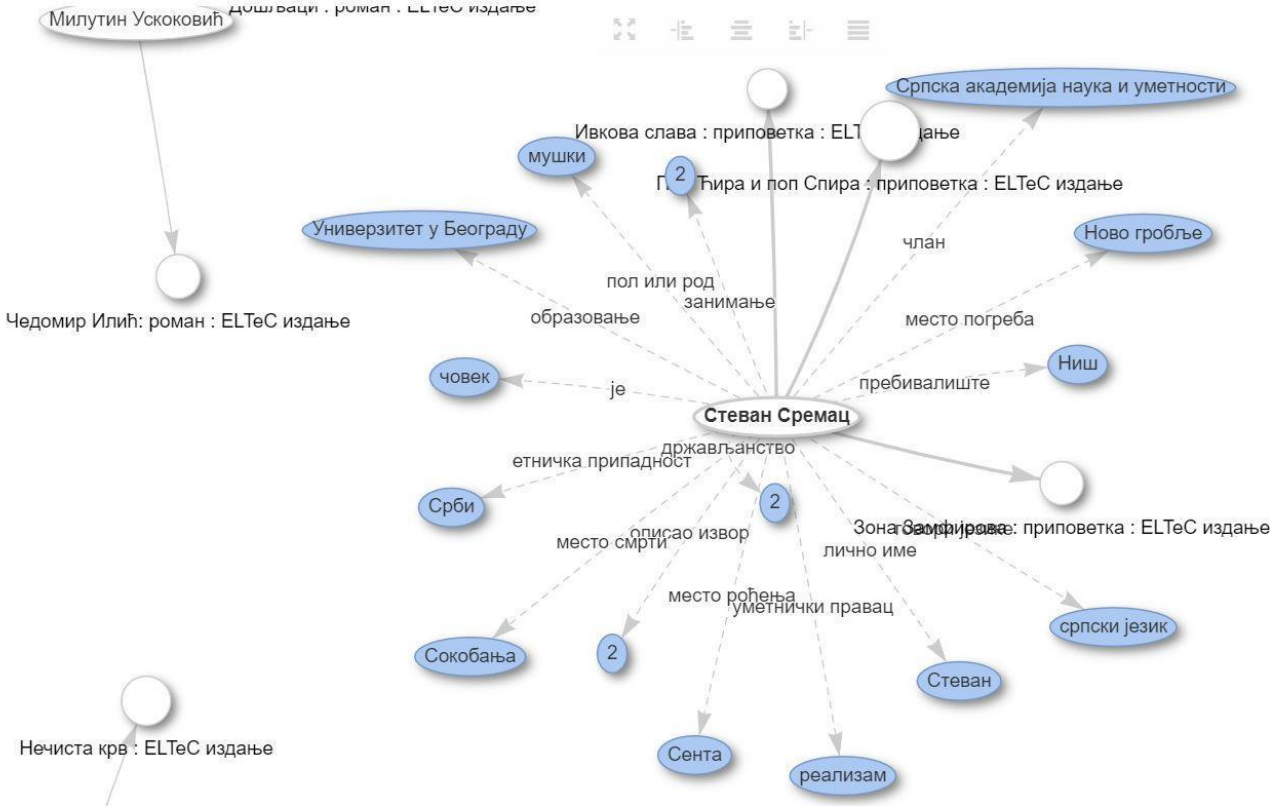



```
#defaultView:Graph
SELECT DISTINCT  ?author ?authorLabel ?edition ?editionLabel
WHERE {
 # published in (P1433) ELTeC collection (Q106927517)
  ?edition wdt:P1433 wd:Q106927517;
            #  instance of (P31)   version, edition, or translation (Q3331189)
             wdt:P31  wd:Q3331189.
  # optional author (P50)
  OPTIONAL {?edition wdt:P50 ?author}
SERVICE wikibase:label
{ bd:serviceParam wikibase:language "sr,[AUTO_LANGUAGE],en". }}
```

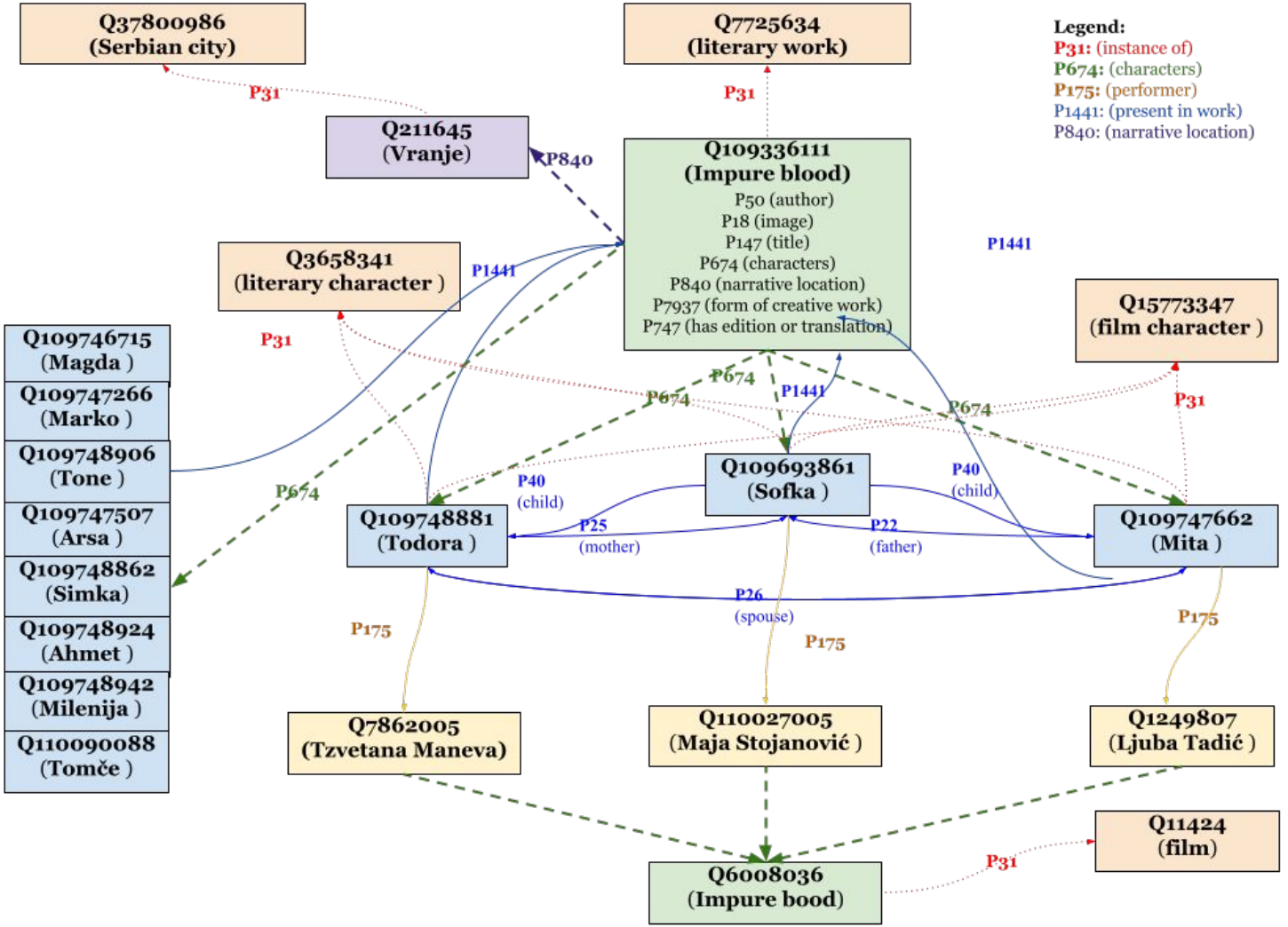
https://w.wiki/4LuC



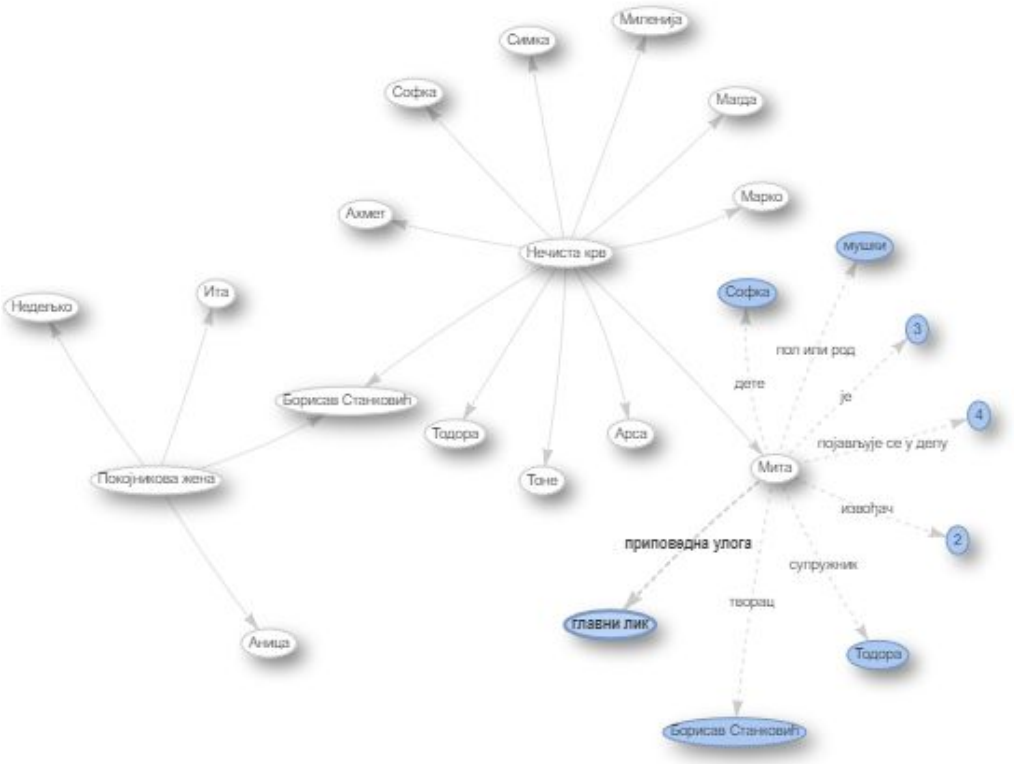
Visualization of an author (Stevan Sremac) and his properties



Characters in “Impure blood”: Sofka, Mita, Todora, Magda, Marko, … and narative location: Vranje

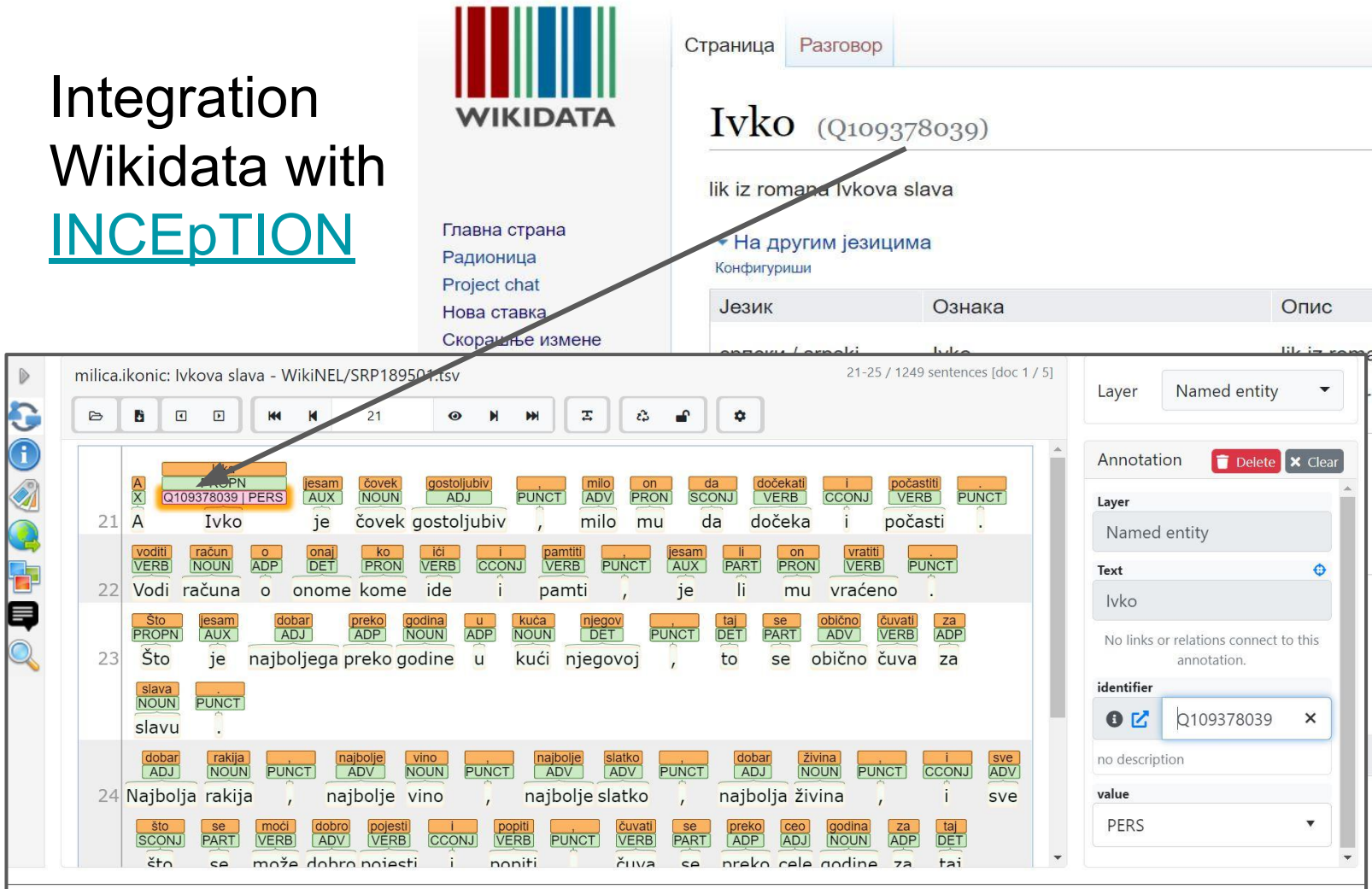


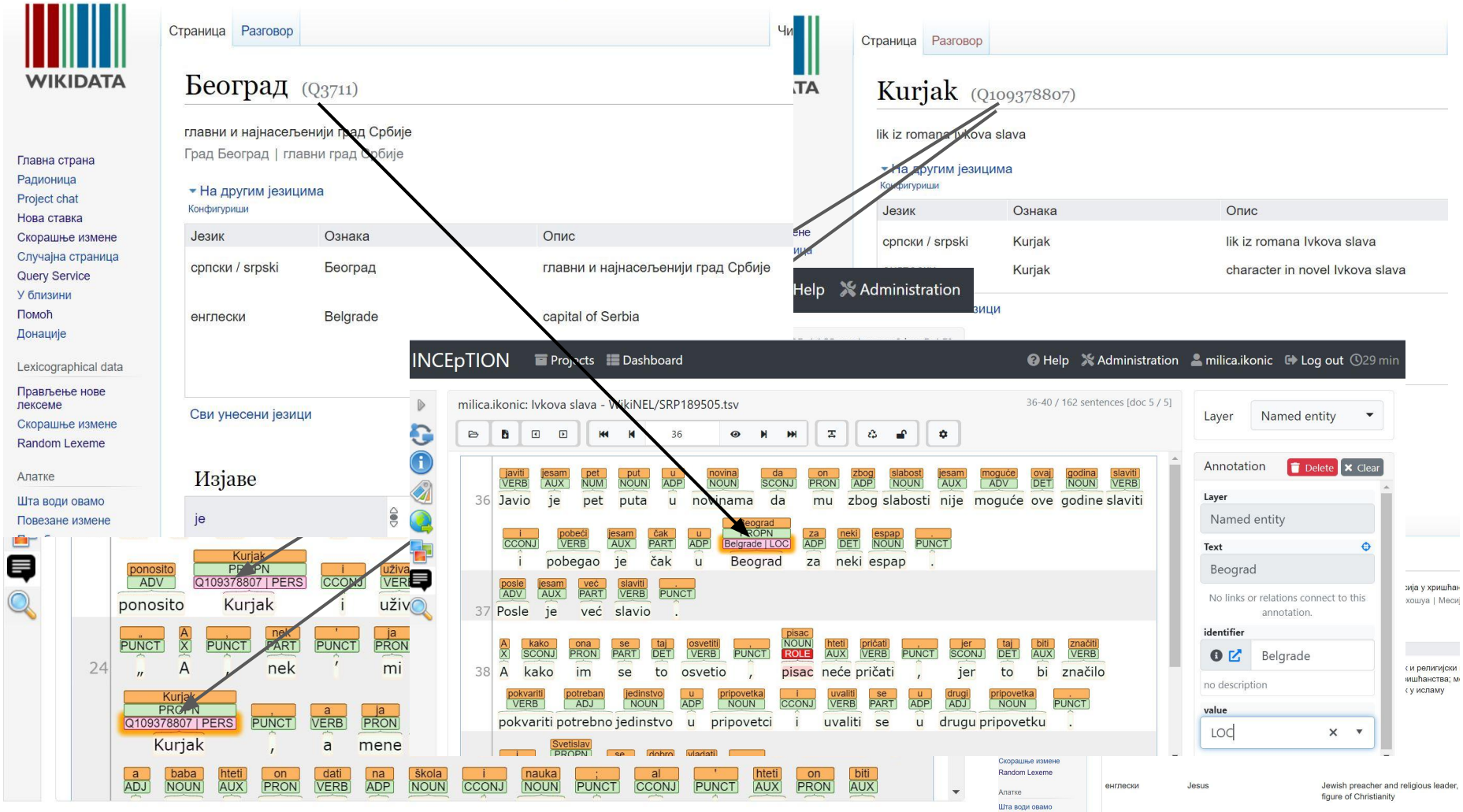
Novels and its characters

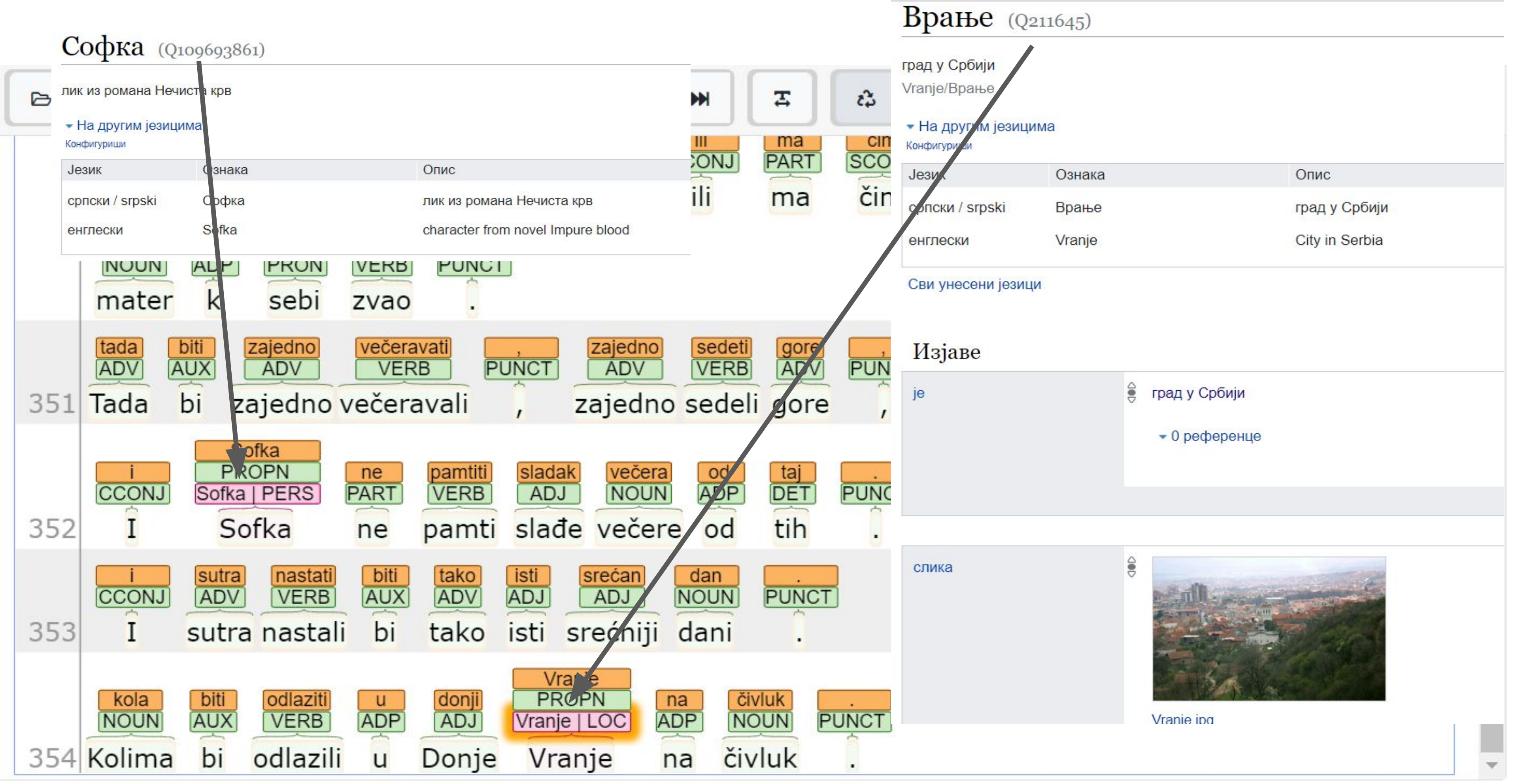


https://w.wiki/4YaE

Integration Wikidata with [INCEpTION]

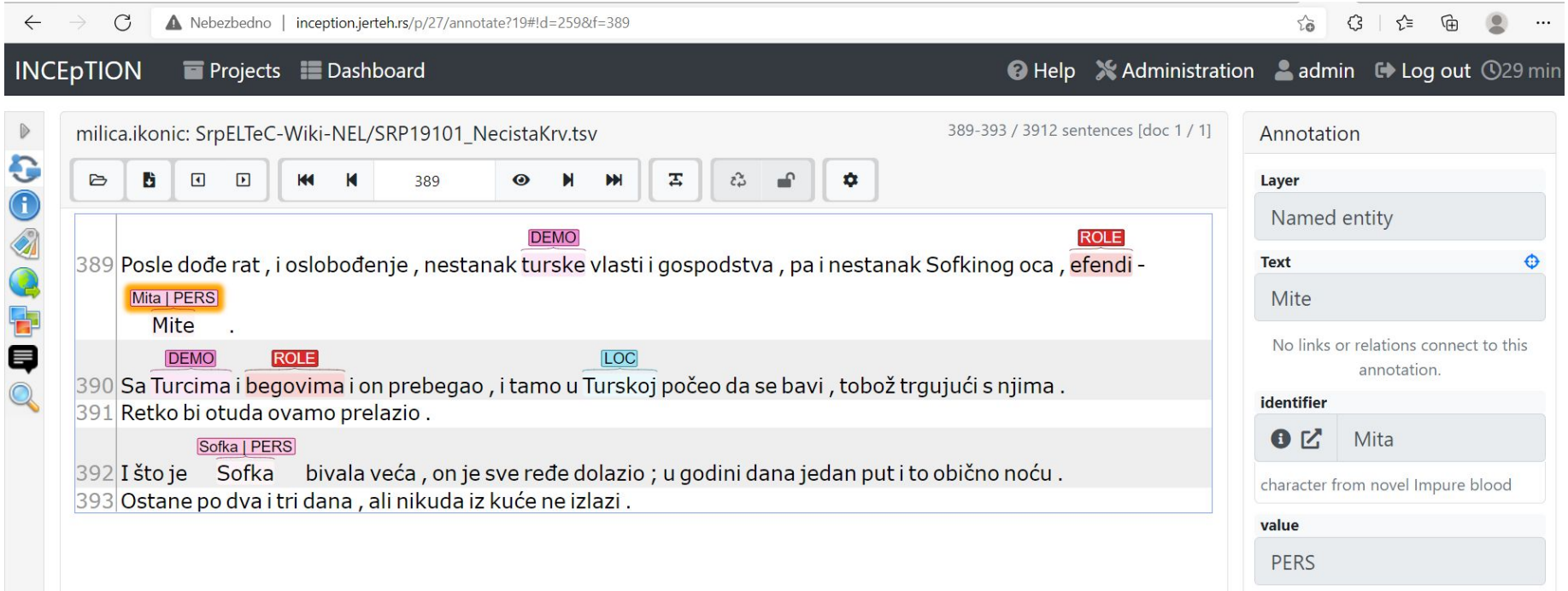



WIKIDATA
Страница
Разговор
Београд (Q3711)
главни и најнасељенији град Србије
Град Београд | главни град Србије
На другим језицима
Конфигуриши
Језик
Ознака
Опис
српски / srpski
Београд
главни и најнасељенији град Србије
енглески
Belgrade
capital of Serbia
Сви унесени језици
Изјаве
Главна страна
Радионица
Project chat
Нова ставка
Скорашње измене
Случајна страница
Query Service
У близини
Помоћ
Донације
Lexicographical data
Прављење нове лексеме
Скорашње измене
Random Lexeme
Алатке
Шта води овамо
Повезане измене
Kurjak (Q109378807)
lik iz romana Ivkova slava
character in novel Ivkova slava
Help
Administration
INCEpTION
Projects
Dashboard
milica.ikonic
Log out
29 min
milica.ikonic: Ivkova slava - WikiNEL/SRP189505.tsv
36-40 / 162 sentences [doc 5 / 5]
Layer
Named entity
Annotation
Delete
Clear
Text
Beograd
No links or relations connect to this annotation.
identifier
Belgrade
no description
value
LOC
Javio je pet puta u novinama da on mu zbog slabosti nije moguće ove godine slaviti
i pobegao je čak u Beograd za neki espap .
Posle je već slavio .
A kako im se to osvetio , pisac neće pričati , jer to bi značilo
pokvariti potrebno jedinstvo u pripovetci i uvaliti se u drugu pripovetku .
ponosito Kurjak i uživ
„ A , nek ' mi
Kurjak , a mene
a baba hteti on dati na škola i nauka ; al ' hteti on biti
Jesus
Jewish preacher and religious leader, figure of Christianity



Софка (Q109693861)
лик из романа Нечиста крв
На другим језицима
Конфигуриши
Језик | Ознака | Опис
српски / srpski | Софка | лик из романа Нечиста крв
енглески | Sofka | character from novel Impure blood
mater k sebi zvao .
351 Tada bi zajedno večeravali , zajedno sedeli gore ,
tada ADV | biti AUX | zajedno ADV | večeravati VERB | , PUNCT | zajedno ADV | sedeti VERB | gore ADV
352 I Sofka ne pamti slađe večere od tih .
i CCONJ | Sofka PROPN Sofka | PERS | ne PART | pamtiti VERB | sladak ADJ | večera NOUN | od ADP | taj DET | . PUNCT
353 I sutra nastali bi tako isti srećniji dani .
i CCONJ | sutra ADV | nastati VERB | biti AUX | tako ADV | isti ADJ | srećan ADJ | dan NOUN | . PUNCT
354 Kolima bi odlazili u Donje Vranje na čivluk .
kola NOUN | biti AUX | odlaziti VERB | u ADP | donji ADJ | Vranje PROPN Vranje | LOC | na ADP | čivluk NOUN | . PUNCT
Врање (Q211645)
град у Србији
Vranje/Врање
На другим језицима
Конфигуриши
Језик | Ознака | Опис
српски / srpski | Врање | град у Србији
енглески | Vranje | City in Serbia
Сви унесени језици
Изјаве
je
град у Србији
0 референце
слика
Vranje.jpg

Layers can be reduced (hiden)…



Export to NIF

“**NLP** Interchange Format (**NIF**):

- an RDF/OWL-based format that allows to combine and chain several **NLP** tasks in a flexible, light-weight way.
- The core of **NIF** consists of a vocabulary, which can represent Strings as RDF resources.
- A special URI Design is used to pinpoint annotations to a part of a document.”

```
<file:/srv/...#offset_39967_39972>
        a                       nif:Word , nif:EntityOccurrence , nif:OffsetBasedString ;
        nif:anchorOf            "Sofki" ;
        nif:beginIndex          "39967"^^xsd:nonNegativeInteger ;
        nif:endIndex            "39972"^^xsd:nonNegativeInteger ;
        nif:lemma               "Sofka" ;
        nif:nextWord            <file:/srv/...#offset_39973_39974> ;
        nif:posTag              "PROPN" ;
        nif:previousWord        <file:/srv/...#offset_39965_39966> ;
        nif:referenceContext    <file:/srv/...#offset_0_378682> ;
        nif:sentence            <file:/srv/...#offset_39885_40052> ;
        itsrdf:taClassRef       <PERS> ;
        itsrdf:taIdentRef       <http://www.wikidata.org/entity/Q109693861> .
```

SrpELTeC @Wikidata status

Wikidata has 110 novels from SrpELTeC collection with associated items for the first edition and electronic SrpELTeC editions.

That means that we automatically added more than **2500** statements.

Conclusion

- Presented examples can be seen as succesful show cases
- Similar approach can be used in many other similar cases
- Places and main characters from novels will be further added to Wikidata and linked in text
- Web annotation Data Model, W3C Recommendation (2017), implementation is planned

Step by step authors: https://scribehow.com/shared/OpenRefine__0K5TU3J9SmalqRBYePhdhQ

Step by step novels: https://scribehow.com/shared/OpenRefine__AOPgpu7dRqmxUkYBtVbGgQ