# Infotheca (Q25460443) in Wikidata

Ranka Stanković, Lazar Davidović



Infotheca (Q25460443) in Wikidata | Ranka Stanković, Lazar Davidović | Infotheca | 2021 | |

10.18485/infotheca.2021.21.1.5

http://dr.rgf.bg.ac.rs/s/repo/item/0005151

# Infotheca (Q25460443) in Wikidata

UDC 004.62: [030:004.738.5
DOI 10.18485/infotheca.2021.21.1.5

**ABSTRACT:** Wikidata 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. In this paper, we will present 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 the articles published in *Infotheca*, *Journal for Digital Humanities* as an example.

**KEYWORDS:** Semantic Web, Open Linked Data, Wikidata, Infotheca, journal metadata.

**PAPER SUBMITTED:** 24 June 2021
**PAPER ACCEPTED:** 16 July 2021

Ranka Stanković
ranka.stankovic@rgf.bg.ac.rs
*University of Belgrade*
*Faculty of Mining and Geology*
*Belgrade, Serbia*

Lazar Davidović
lazarmdavidovic@gmail.com
*University of Belgrade*
*Belgrade, Serbia*

## 1 Introduction

Wikidata[1] is a Wikimedia Foundation knowledge base, a common source of various kinds of data, both concrete and abstract. The stored data can be used by other Wikimedia projects, such as Wikipedia and the wider community alike, for different purposes. This contributes to extending the boundary from machine readable to machine comprehensible data on the web. In this paper, we will present 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 the articles published in *Infotheca*, *Journal for Digital Humanities* as an example.

Semantic web is an extension of the existing web where information is given precise meaning allowing better collaboration between computers and their users. The open and partially structured nature of the resources whose development was organized by Wikimedia provided the basis for the creation

1. Wikidata

of many machine readable resources, like DBPedia,[2] for example, relying on the standardized languages of the semantic web. The concept of the semantic web and open linked data technologies expand the traditional web by using a standard markup language and similar processing tools, where RDF (Resource Description Framework) plays a significant role and makes more efficient information retrieval solutions possible (Shah et al. 2002). In order for the semantic web to act the part, computers should have access to structured collections of information and be able to set out defined rules of automated management. Wikidata actually fits the trends of information technology development that extend the boundary from machine readable to machine comprehensible data on the web.

The Scholia project[3] (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. The results of this particular project and the availability of the Infotheca articles in different digital formats provided the inspiration for the "Wikification" of the articles published in the Infotheca journal. Having seen the content of the web pages scholiaEvent[4] and scholiaTopic,[5] a similar project was launched with the goal of creating linked (RDF) data about authors and scholarly articles based metadata and adding links to Wikidata to the journal articles, as well as showing the co-authorship graph as an interactive page on the website of the Biblisha digital library[6] (Stanković et al. 2015). The implementation is a case study that can be further extended to other use cases, such as conferences and digital libraries. The Scholia tool is being developed as part of a lager initiative, WikiCite,[7] aiming to index bibliographic data in Wikidata on the resources that can be used to corroborate the claims made in Wikidata, Wikipedia or elsewhere. At the time when we are inundated by false information on the web, proper corroboration of information by relevant sources certainly plays a key role. Since we wanted to automatize the process of preparing and entering information, we looked

2. DBPedia

3. Scholia, Scholia at Wikidata

4. An overview of past and present conferences with an organizer containing information about article submission deadlines ScholiaEvent

5. An overview of scholarly and professional articles, as well as their authors and topics appearing together, grouped by thematic entities: Wikipedia, machine learning, biology, food and the like. scholiaTopic

6. Biblisha

7. WikiCite

into different solutions and ended up using two of them, namely, OpenRefine[8] and QuickStatements[9] that will be discussed further in the sections to follow.

The collaboration of Wikimedia Serbia[10] with the University of Belgrade is a longstanding one (Stakić 2009). The University Library Svetozar Marković together with the Faculty of Mathematics of the University of Belgrade and Wikimedia Serbia launched the (Wiki-Librarian) project in 2015 with the idea of making as much quality content as possible available on Wikipedia (Popović, Ševkušić, and Stakić 2015). Wikiadata, as an open data network was used by Andonovski (Андоновски 2020) to describe language resources, namely, novels forming part of the Serbian-German literary corpus (Andonovski, Šandrih, and Kitanović 2019). For a number of years now, students at the Faculty of Mining and Geology have been undergoing training to enter data into and use the Wikidata[11] database, while the Intelligent Systems doctoral course features the subjects Knowledge Representation and Semantic Web that explore the potential for application of open data. As part of the "Distant Reading for European Literary History"[12] се ради на уносу метаподатака о српским романима из корпуса *srpELTeC*[13] COST Action CA16204 (2017-2021) metadata about Serbian novels included in the srpELTEC corpus is being entered into the knowledge base (Krstev et al. 2019) and Wikidata linked to various applications, one of which is Aurora.[14] Members of JeRTeh Language Resources and Technologies Society[15] too contributed to the results presented in this article.

8. OpenRefine (formerly Google Refine) is a tool for working with messy data: cleaning it; transforming it from one format into another, together with extending it with external data via web services. OpenRefine

9. The tool for editing Wikidata items: adding and removing statements, labels, descriptions, etc. QuickStatements

10. Wikimedia

11. Input data to Wikidata and their use

12. One of the most important aims of this action is preparing a multilingual corpus (titled European Literary Text Collection - ELTeC) which, when fully complete, will feature a hundred novels from each participating country first published in the period 1840-1920.

13. srpELTeC

14. Aurora

15. JePTex

## 2 Wikidata

Wikidata is a knowledge base whose purpose is to be a common source of certain kinds of data (for example, the population of a country, place of birth, date of founding) used by other Wikimedia projects, such as Wikipedia. In that sense, it is similar to Wikimedia storage where media files accessed from other Wikimedia projects are stored. Wikidata is oriented towards documents, focused on items representing topics, concepts or objects. Every item has been assigned a persistent identifier, a positive integer with an upper-case Q as a prefix, known as *QID*. This makes translation of the basic information necessary for recognizing a topic covered by an item without favouring any language whatsoever, the aim being to ensure the uniqueness of meaning of a particular concept.

These are some of the examples of items: places (Novi Sad: Q55630, London: Q84, Zvezdara (Belgrade): Q12645852), people (Đorđe Balašević: Q342045, Tim Berners-Lee: Q80, Hedy Lamarr: Q49034), events (First Serbian Uprising: Q368689, concert: Q182832, marathon: Q40244), objects (chair: Q15026, glass: Q81727, frying pan: Q127666)), concepts (joy: Q935526, fear: Q44619, concept: Q151885), literary works (*Gorski vijenac*: Q1192476, *Don Quixote*: Q480, *Game of Thrones* (books): Q1751870), films (*Lepota poroka*: Q4239792, *Hair* (film): Q757156), TV series (*Game of Thrones*: Q23572, *'Allo 'Allo!*: Q425628), ballet (*Don Quixote* (ballet): Q1239463)... The concepts behind items should be unique, but as it happens, there can exist two items under the same name, Nikola Tesla (Q9036) refers to the famous scientist, while Nikola Tesla (Q2732597) refers to a housing project (Q486972) in Niška Banja (Q954986) named after him. It is recommended that in the case of polysemous entities, like the above-mentioned *Don Quixote* (ballet) or *Game of Thrones* (books) an additional explanation should be given in the parenthesis. An item is, thus, linked to a unique identifier (QID), the identifier is, in turn, linked to the item's corresponding title and description, so as to remove any ambiguity.

An identifier of a data item (QID) can, in addition to being linked to a title and a description, have a number of aliases and statements (claims, expressions) representing its properties and values. A statement is an ordered triple (item, property, value), where item (Q) is any topic (person, object, place, concept), item (P) is property . The relation[16] or a characteristic relevant to an item can be, for example: hair colour (P1884) for people,

16. In mathematics, if an ordered pair $(x, y)$ is the relation $\rho$ then the element $x$ is has established a relation to the element y and it is written as a triple: $x\rho y$.

| | |
|---|---|
| The University of Belgrade is a public university and a member of the European University Association. | |
| The University of Belgrade was established in 1808 by Dositej Obradović. | |
| Q240631 P31 Q875538.<br>Q240631 P463 Q868940.<br>Q240631 P112 Q347659.<br>Q240631 P571 “1808”. | Q240631 P31 Q875538;<br>P463 Q868940;<br>P112 Q347659;<br>P571 “1808”. |
| Dositej Obradović was born on February 17th, 1742 in Čakovo and died on April 7th, 1811. He was a linguist, poet, writer and philosopher. | |
| Q347659 P569 “17 February 1742”.<br>Q347659 P19 Q325736.<br>Q347659 P570 “7 April 1811”.<br>Q347659 P106 Q14467526.<br>Q347659 P106 Q49757.<br>Q347659 P106 Q36180.<br>Q347659 P106 Q4964182. | Q347659 P569 „17 February 1742“;<br>P19 Q325736;<br>P570 „7 April 1811“;<br>P106 Q14467526;<br>P106 Q49757;<br>P106 Q36180;<br>P106 Q36180. |

**Table 1.** Examples of Wikidata items

publisher (P123) for published works, founding (P571) for organizations and the like. The value of an item cam be a literal itself, that is, a character string (for instance: the length of the Danube is 2860 km) or a reference to some other item (the capital of Serbia is Belgrade, for example). An item can be described by a string of statements, each of which provides a fact or a piece of data about the item. Table 1 shows several examples of natural language sentences and the encoding of this information in Wikipedia, represented as triples of subject, predicate and object (left), and in shortened notation (right).

In the above example, the second column of the table features triples i.e. sentences following the subject-predicate-object pattern. More precisely, we could say that these are RDF triples, where RDF is an abbreviation standing for web Resource Description Framework (Q54872). The sentences end in a period. The third column shows abridged notation doing away with the repetition of the subject, so that the punctuation mark “.” indicates that the predicate that follows refers to the same subject.

Similarly, the items in Wikidata represent relations as triples, so that the relation Tesla-way of life-vegetarianism is encoded as Q9036 P1576 Q83364.

An important characteristic of Wikidata is that it has two facets. One is intended for people and the other one for machines, which enables numerous applications in the domain of natural language processing. Let's mention some of them: text classification, indexing, text analysis, summarizing, normalizing, linking, etc. Another important feature is multilingual support, making it possible to link each item to a label in any language registered in Wikimedia resources, which, in turn, opens up the possibilities for numerous applications, from automatic translation and classification of multilingual documents, to analysing web and social media content.

## 3 Automated Wikidata entry

The entry of individual pieces of data is often a time-consuming task, but it can be sped up in the situations where data already exists and is stored in different digital formats. With proper prior preparation, it can be entered in Wikidata semiautomatically. Therefore, the basic idea was to speed up the entry of data about the results of the research in the domain of digital humanities in Serbia, as well as about old Serbian novels, so as to increase the visibility of both the Serbian language, our cultural heritage and the results of the research in Serbia and certainly pave the way for many other data sets.

In order for data entry automatization to be possible, it was necessary to make the first step consisting of collecting and preparing data. The second step refers to the choice of Wikidata labels that will be used to identify the predicate and creating a data entry outline. The outline defines the linking of a value to an item, namely, a subject via predicate as a mediator.

Although the ultimate aim was the entry of data about articles, entering information about their authors was an indispensable step and a prerequisite for further work. After data entry has been completed, SPARQL [17] queries for different views were created, using Wikidata integrated technologies too for visualizing the results.

Every Infotheca journal article from the Biblisha bilingual digital library has been linked to the corresponding Wikidata entry, so that not only each individual, particular Wikidata view can be accessed directly, but also that some of the useful visual representations can be integrated within the application itself. Figure 1 features an example of a pattern in Biblisha showing the three top-ranked articles in the collection and illustrating the fact that behind every article title there is a link to a Wikidata resource path.

17. [18]

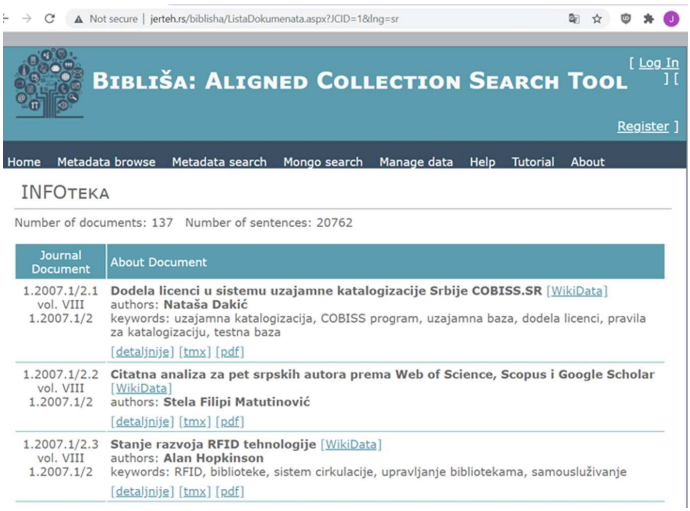


**Figure 1.** Biblisha digital library panel showing an overview of the metadata about Infotheca.

The information about the first article on the list would be translated into the language of Wikidata in the following way:

„Додела лиценци у систему узајамне каталогизације Србије COBISS.SR“ (Q98785010)
instance of (P21) academic journal article (Q18918145);
author P(50) Наташа Дакић (Q99281474).

We can see that the article is represented by the identifier Q98785010, that it is an instance (P21) of the class of scholarly articles (Q18918145) and that it has the author Q99281474.

The above-mentioned tool OpenRefine initially developed by Google, as well as the QuickStatements tool developed by the Wikidata team member, Magnus Manske are often implemented together and can be said to complement each other. QuickStatements uses textual TSV or CSV formats efficiently generated by the OpenRefine tool. The difference between these two tools is in the granularity of transactions, since OpenRefine inputs the changes in a single step, so resolving data entry errors can lead to data duplicates, while QuickStatements inputs each item individually, allowing better monitoring of the entire process. The examples of good practice indicate that OpenRefine is used for preparing data entry in the Wikidata database, while

the actual entry of RDF triples is performed by using the QuickStatements tool.

Preparing data in the form of a CSV file is certainly the first step, followed by creating an OpenRefine project and loading the prepared data. What comes next is recognition of the existing Wikidata items – an indispensable step enabling the linking of file content to identifiers (QID) of the existing items and entry of new ones, if they do not already exist. In this phase, manual checking and possibly making changes are necessary. Creating a dataset schema defines the predicates that will link subjects to objects in RDF triples and it is a very important step. Here are some characteristic examples with comments:

- Title (P1476), in English and Serbian (both scripts, Cyrillic and Latin, for the sake of search);
- Main subject of the creative work (P921), key words, where the existing ones are linked and new ones are added as instances with labels in Serbian and English;
- Publisher (P123);
- Language of the work or name (P407);
- Publication date (P577), represented by year only;
- Published in (P1433) Infotheca (Q25460443);
- Licence (P275);
- Full text available at (P953).

In view of the fact that properties are added more and more rarely, it is recommended to perform a search of similar properties and properties of similar resources before the decision for them to be added is made. Moreover, special attention should be paid to the limitations in the domain of properties that can be seen in the suggestions provided when entering data. Joint work of distributed users unavoidably sometimes leads to Wikidata duplicates. The solution to this situation is making use of the option to merge or eliminate duplicates. Additional information about it is available at this page.

After the initial data entry, the input of data into the database continued after each newly published issue of Infotheca. As a result, 38 Infotheca articles are now made available. An HTML integrating Wikidata Query Service with Biblisha was created. The queries retrieving tables of the latest published articles, frequency of the keywords in articles, pictures of authors, author profile table, co-authorship graph, distribution of authors by sex, etc. were written.The information about authors consists of basic data that

should definitely be enriched with new content in the forthcoming period, including data about the institutions where they work, research interests, references to authoritative research databases and the like. By way of example, here is a simple query, available at this site, showing a list of Infotheca articles, issue, volume and publication date.

```
SELECT ?paper ?paperLabel ?vol ?publication ?publication_date
WHERE {
?paper wdt:P1433 wd:Q25460443;
       wdt:P577 ?publication_date;
       wdt:P478 ?vol;
SERVICE wikibase:label {bd:serviceParam
                        wikibase:language "en,sr".}
}
ORDER BY DESC(?vol )
```

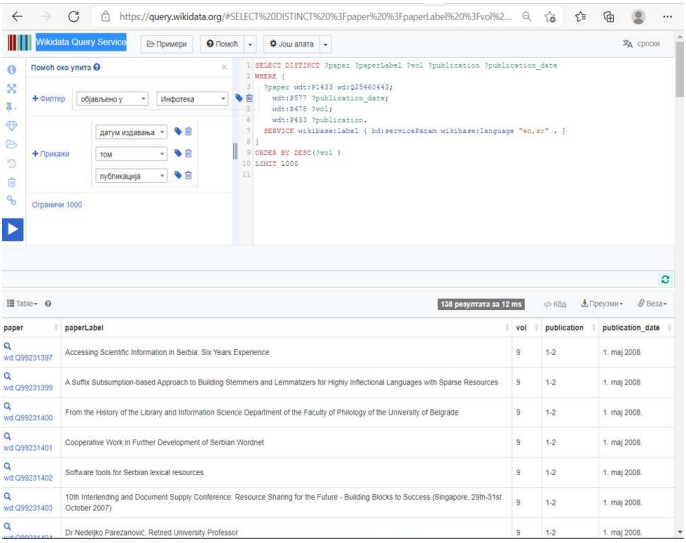


**Figure 2.** *Wikidata Query Service* query interface

Figure 2 provides an illustration of the above query in the Wikidata Query Service user interface, where the upper part of the panel is used for

making queries, while the results are shown in the lower part. The view (table, graphical representation, grid, timeline, chart, map etc.) can be chosen depending on query type.

Figure 1 features part of a co-authorship graph[19] pulling data from Wikidata via Wikidata Query Service.

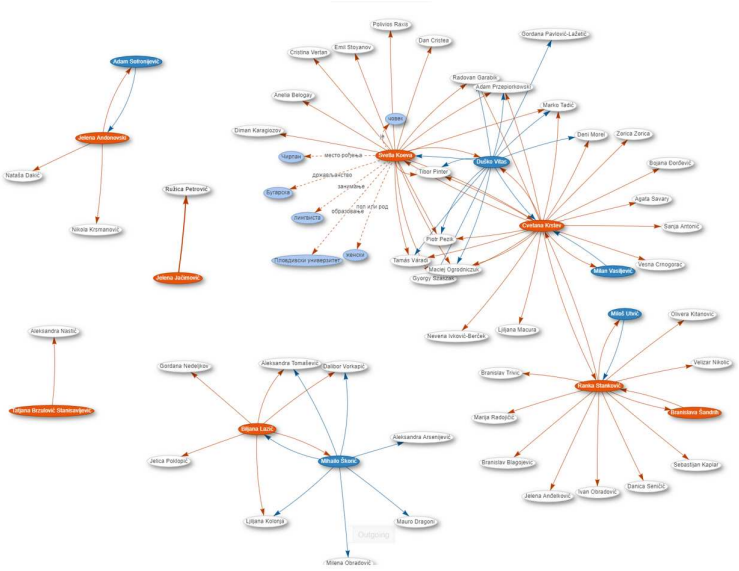


**Figure 3.** *Infotheca* articles co-authorship graph

## 4 Conclusion and future plans

Positive experiences of working with Infotheca Wikidata were drawn upon when entering in Wikidata the data on the novels and their authors belonging to the ELTeC multilingual collection (European Literary Text Collection) one subcollection of which will consist of a hundred Serbian novels from the period 1840-1920 developed as part of the Distant Reading for European Literary History Cost Action: CA16204 by members of the JeRTeh society,

19. The co-authorship graph is available, where the SPARQL query itself can be accessed or the view changed to table, timeline, graph and the like.

led by Cvetana Krstev and Ranka Stanković. A set of metadata about the novels digitized up to the present and prepared to fit the requirements of the action was built. The work on Wikidata is seen as a continued activity where special attention will be paid to linked open data in the domain of linguistics – LLOD and its application. We must certainly be aware of the problems and limitations related to Wikidata and other kinds of linked open data, so as to be able to look into the ways of overcoming or at least mitigating them.

Wikidata is a truly immense knowledge base that is 1) available to everyone – for reading information, making queries, editing and improving it; 2) Open – multiple use is available under the Creative Commons CC0 licence granting complete freedom to use data; 3) multilingual – entities can be named and described in any natural language. These three key features are the main driving forces for the many applications, which, we believe, will inspire the wiki community further to devote more attention to this resource. The so-called small languages, including Serbian, should make use of all the possibilities to find their place in the digital space. Thus, the activities carried out as part of this and similar projects and initiatives are a humble attempt to contribute to the preservation of the Serbian language in the digital age.

## References

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

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." In *Book of abstracts of DH_ BUDAPEST_ 2019,* 36–37.

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

Popović, Aleksandra, Milica Ševkušić, and Đorđe Stakić. 2015. "Biblioteke i Vikipedija zajedno na vebu: slobodno znanje za sve." *Digitalna humanistika: tematski zbornik u dve knjige, knj. 1,* 151–161.

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.

Stakić, Đorđe. 2009. "Wiki Technology - Origin - Development and Importance." *INFOtheca-Journal of Informatics & Librarianship* 10 (1-2): 69–78.

Stanković, Ranka, Cvetana Krstev, Biljana Lazić, and Dalibor Vorkapić. 2015. "A bilingual digital library for academic and entrepreneurial knowledge management." In *Proceeding of 10th International Forum on Knowledge Asset Dynamics-IFKAD,* 1764–1777.

Андоновски, Јелена. 2020. "Мрежа отворених података и језички ресурси у процесу изградње српско-немачког литерарног корпуса." PhD diss., Универзитет у Београду, Филолошки факултет, јануар.