Taking DBpedia Across Borders: Building the Serbian Chapter

Uroš Milošević, Vuk Mijović, Sanja Vraneš*
* Mihajlo Pupin Institute, Belgrade, Serbia
{uros.milosevic, vuk.mijovic, sanja.vranes}@pupin.rs

***Abstract*—With the emergence of Linked Data, DBpedia has steadily grown to become one of the largest and most important structured knowledge sources we know of today. Adopting Wikipedia's practice of entrusting the community with most of the work, the DBpedia internationalization committee has made a major step towards the move from unstructured to structured knowledge on the Web. Still, with new languages come new challenges. In this paper, we inspect some common obstacles that need to be tackled in order to add one more language to this popular data hub, but also some that haven't been encountered before in this domain. More specifically, we explore the digraphic nature of the Serbian language, analyze the state of the DBpedia Extraction Framework with respect to its support for languages that use multiple scripts, and propose solutions towards overcoming this problem. Finally, we deploy the first digraphic DBpedia edition, taking the leading position amongst all DBpedia versions in the percentage of all covered Wikipedia templates, and all template occurrences in Wikipedia that are mapped, while adding a valuable new chapter to the DBpedia project and enriching the Linked Open Data Cloud even further.**

I. INTRODUCTION

The rise of the Linked Data Web has brought a paradigm shift to the world of information retrieval. We're no longer interested in the short answers to *who*, *what*, *where* or *when*, but would also like to know about the relations between and the background behind those answers. However, these connections imply structure, and structured knowledge is often hard to come by.

DBpedia leverages the existing efforts by the ever-growing Wikipedia communities worldwide by extracting the crowdsourced data and transforming them to RDF. Not only has this approach helped create a bridge between unstructured and structured data on the Web; it has pushed DBpedia towards becoming the most important data hub in the Linked Open Data Cloud (Figure 1). It is also one of the largest multilingual datasets on the Web of Data and, as such, expected to be able to cope with all the difficulties that go along with multilingualism.

In this paper, we give a detailed survey of the internationalization challenges standing in the way of enriching DBpedia with more languages, an analysis of the state of the DBpedia framework itself with respect to such challenges, and a report on the completed and ongoing efforts being put into the Serbian edition of this dataset.

In Section 2, we take a look at the current state of the DBpedia project, with a focus on its main components. In Section 3, we try to understand the issues related to the move from the Serbian Wikipedia to the very first Serbian DBpedia. Section 4 outlines the achieved results, and Section 5 concludes our findings and proposes future work.

II. DBPEDIA

Creating and maintaining a multilingual knowledge base can require an enormous (and, thus, expensive) amount of work. Crowdsourcing unstructured knowledge, on the other hand is free, and has given birth to Wikipedia – one of the most important knowledge sources mankind knows of today. Being maintained by thousands of contributors from all over the world, it is bound to grow even larger.

Realizing the potential of this global effort, the Linked Data community has gathered around a joint project of their own to breathe life into the Linked Data Cloud by enriching it with crowdsourced knowledge straight from Wikipedia, in RDF form.

English DBpedia alone describes 4.0 million things[1]: 832,000 persons, 639,000 places, 372,000 creative works (music albums, movies, video games etc.), 209,000 organizations, 226,000 species, 5,600 diseases etc. All versions of DBpedia together contain descriptions of 24.9 million things, (out of which 16.8 million are interlinked with the English DBpedia concepts), 12.6 million labels and abstracts, 24.6 million image links, 27.6 million links to external web pages. Moreover, they contain 45.0 million links to other, external, LOD datasets, and 67.0 million Wikipedia, and 41.2 million YAGO category links. Together, they make up a dataset of 2.46 billion RDF triples, out of which 1.98 billion were extracted from non-English language editions.

A. DBpedia Ontology

The DBpedia Ontology is a manually created directed-acyclic graph based on the most commonly used infoboxes within Wikipedia. It is a shallow, cross-domain ontology that covers 529 classes which form a subsumption hierarchy and are described by 2,333 different properties[2].

The research presented in this paper is partly financed by the European Union (FP7 LOD2 project, Pr. No: 257943), and in part by the Ministry of Science and Technological Development of Republic of Serbia (SOFIA project, Pr. No: TR-32010).

[1] http://wiki.dbpedia.org/About
[2] http://wiki.dbpedia.org/Ontology

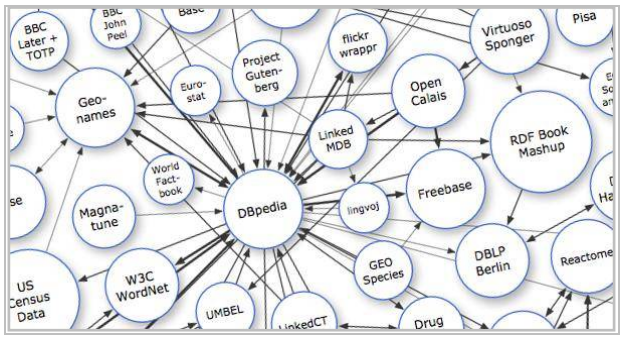

Figure 1. DBpedia at the heart of the LOD Cloud

A downside to any mass community effort is often lack of standardization and quality control. As such, the DBpedia Ontology, as detailed as it is, appears to be suffering from many issues caused by the *ad hoc* nature of the solutions contributed by its creators for their specific purposes. Some of these problems are often reflected in the missing classes, properties, as well as the inadequate names, ranges and domains, defined for some of those properties. For example, the *constellation* property is limited to the *Galaxy* domain. However, other *CelestialBody* instances often also have references to constellations, yet no other property exists for this purpose.

An even detailed look using an ontology evaluation methodology, such as OntoClean [1], would reveal other suboptimal engineering choices. For instance, the subsumption hierarchy has many typical role-concepts [2] (e.g. *Celebrity*, *Engineer*, *Criminal*, *Athlete* etc.) subsumed by a basic-concept - *Person*, which might not be appropriate for an ontology with such wide coverage (or any reusable ontology, for that matter). This, however, could be well outside the scope of this paper and will not be covered here.

B. Information Extraction Framework

The Wikipedia information is extracted and triplified using a flexible and extensible framework (DBpedia Information Extraction Framework – DIEF), written in Scala and structured into different modules, where the *core module* contains the essential components of the framework, and the *dump extraction module* is aimed at Wikipedia data extraction. The framework also relies on a growing set of *extractors* – mappings from page nodes to graphs of statements describing those nodes. The DIEF provides a total of 16 extractors, 7 of which are language (mostly English) specific. Two of them are aimed at the richest sources of structured knowledge on Wikipedia: infobox templates.

The *Generic Infobox Extractor* is tuned to create triples using the article URI as the subject, the infobox property name (in camel case form, appended to the `http://dbpedia.org/property` namespace) as the predicate, and the attribute value as the object.

With the introduction of the internationalization (i18n) filters, the *Mapping-based Infobox Extractor* has become the single most important extractor for any internationalization effort. It relies on manually created Wikipedia infobox property to DBpedia Ontology property mappings (using a relatively simple syntax) to extract triples, binding any mapped properties to the `http://dbpedia.org/ontology` namespace.

TABLE I.
MAPPING THE *INFOKUTIJA BIOGRAFIJA* INFOBOX

Template mapping	
Map to class	*Person*
Property mapping	
Template property	*datum_rođenja*
Ontology property	*birthDate*

The template parameter values are parsed according to the data types specified in the DBpedia Ontology and the Wikipedia resources are classified based on the used infobox template and additional classification rules specified in the mapping configurations. TABLE I shows a mapping example with a specified class and a single property representing the date of birth of a person.

C. Mappings Wiki

Although it may appear that the fact that the mappings need to be created manually takes the entire idea of exploiting the crowdsourcing efforts of the Wikipedia community one step back, the actual amount (and nature) of work is trivial in comparison with that of maintaining Wikipedia. Thanks to the Mappings Wiki, infobox templates can be collaboratively mapped to the corresponding ontology classes and properties, across different Wikipedia language editions [4]. Moreover, the Wiki can validate a single mapping without starting up the entire Extraction Framework, and also let you preview the mapping result by triplifying the infoboxes of a small number of test articles on the fly.

There are currently 26 language specific mappings in DBpedia's Mapping Wiki[1].

III. BUILDING A SERBIAN DBPEDIA

The case of the Greek DBpedia brought attention to the many issues related to internationalization and paved the way for other languages that use non-US-ASCII encoded scripts.

The Serbian version of Wikipedia is by no means the largest, or the richest Wikipedia amongst the 287 versions currently available[2]. It is 26th in article count, with approximately 242,000 articles. However, it is one of very few editions supporting multiple writing systems, due to Serbian being the only European language with active *synchronic digraphia* (using two scripts for the same language).

A. Wikipedia in Serbian

We have mentioned earlier the problems that often go along with many mass community projects. Unfortunately, Wikipedia is no different, and the lack of a coordinated approach is reflected in data errors, noise and redundancy.

Take, for example, one of the most frequently occurring infobox properties defining the Web page of a resource (*foaf:homepage* in RDF). A total of 36 variants of this property exist in the Serbian Wikipedia, the most common of which, *веб-страна*, alone is used in 108 infobox

[1] http://mappings.dbpedia.org/

[2] http://meta.wikimedia.org/wiki/List_of_Wikipedias

templates. *веб* is found on 28 occasions, *вебсајт* on 16, *страница* on 10, *website* on 7 etc.

The digraphic nature of the language only doubles the potential for error. The same often happens with entire infobox templates. For instance, there are three different infoboxes used to describe an actor: *Глумац*, *Кутијица за глумце*, and *Glumac-lat.*

It is clear that Wikipedia itself is in need of a collective effort towards standardization and a common vocabulary.

B. Coping with Digraphia

The issue of digraphic Wikipedia is best illustrated in the case of information retrieval. Most of the Serbian online communities rely on the Latin alphabet for communication/interaction on the Web. That means a large portion of the information available online is (and, often, expected to be) encoded in ISO 8859-2 (i.e. Latin-2). And, yet, most of the information in the Serbian Wikipedia dumps is encoded in Cyrillic. So, unless the information retrieval software performs transliteration (romanization or cyrillization) on-the-fly (at retrieval time, as in the case of Wikipedia), many attempts at information extraction will be doomed to fail. This directly affects common tasks such as keyword search, label-based SPARQL querying, named entity recognition, etc.

As it may be unrealistic to impose this requirement on the software developers, the only reasonable, yet, perhaps, not so elegant workaround is to have the knowledge base keep the information encoded in both character sets. Although such approach would double the space requirements needed for storing any Cyrillic or Latin string literal, there is also the matter of perspective - one could argue that although the information being stored is essentially the same, the very fact that different character sequences are needed to describe the same piece of knowledge makes this problem fall into the domain of multilingualism.

In such a case, a single IRI would still be used, but two separate triples would be stored for any string literal in Serbian. For example:

```
http://sr.dbpedia.org/resource/парсер
                        rdfs:label "Парсер"@sr ;
                        rdfs:label "Parser"@sr .
```

It is worth noting that the IANA Language Subtag Registry[1] contains separate tags for Serbian Latin and Cyrillic (sr-Latn and sr-Cyrl, respectively), but lists them as redundant.

As the current version of the DIEF doesn't provide the means for transliteration, let alone duplicating literals, we developed a small post-processor that first transliterates the strings in the DIEF dumps, and then merges them back with the original dump (Figure 2). It should be noted that, although it is safe to assume that all Cyrillic strings can be transliterated directly to their Latin counterparts, vice-versa is not always straight forward. For instance, many resources on the Serbian Wikipedia have labels that reflect their original names that are not meant to be transliterated (e.g. *ASCII* would never be transliterated to *АСЦИИ*). The Serbian edition of Wikipedia has a number of syntactical

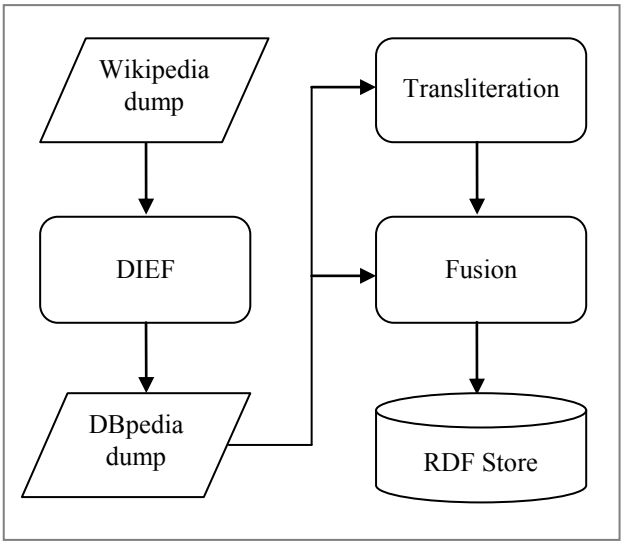

Figure 2. DBpedia Extraction process

constructs that can be used to keep the original encoding of a string. However, there are no mechanisms available for accomplishing the same based on a generated DBpedia RDF dump. Moreover, in cases where transliteration is possible, doing so is not always as easy as with romanization.

TABLE II shows the nature of Serbian *digrams* (pairs of characters, each used to identify a single phoneme). The rule says that any *њ* character is always transliterated to *nj*. Ideally, the same should hold for the other way around, but there are exceptions. For instance, *konjunkcija* (en: *conjunction*) is not transliterated to *коњункција*, but *конјункција*, as that is the original form of the word (*н* and *j* are treated as separate characters).

Therefore, we perform only romanization of Cyrillic string literals in our post-processing module.

C. Serializing the DBpedia Dumps

Serializing multilingual data in RDF is not a straightforward process either [5]. Below we take a look at the serialization formats supported by the DIEF, with respect to their support for internationalization.

N-Triples, a subset of Notation 3, lack shortcuts such as CURIEs, which is why they are less readable and more difficult to create manually. What's more important for our task, is that they support only the 7-bit US-ASCII character encoding instead of UTF-8, meaning there's no support for IRIs either .

TABLE II.
SERBIAN DIGRAMS

Serbian Latin alphabet	Serbian Cyrillic alphabet	International Phonetic Alphabet (IPA) value
Lj lj	Љ љ	/ʎ/
Nj nj	Њ њ	/ɲ/
Dž dž	Џ џ	/dʒ/

[1] http://www.iana.org/assignments/language-subtag-registry/language-subtag-registry

Turtle (Terse RDF Triple Language) is a subset of, and compatible with, Notation 3 and a superset of the minimal N-Triples format. It's compact, human-readable, and UTF-8 based and, therefore, makes a great solution for internationalization [7]. Turtle is also part of the SPARQL query language for expressing graph patterns.

TriX is an experimental format that aims to provide a highly normalized, consistent XML representation for RDF graphs, allowing the effective use of generic XML tools such as XSLT, XQuery, etc. Its form helps it retain expressivity, while providing a compact and readable alternative to RDF/XML [8].

RDF/JSON (JavaScript Object Notation), requires less overhead with respect to parsing and serializing than XML, and encodes text in Unicode, thereby making the use of IRIs possible. The percent character doesn't need special treatment; the only characters that need escaping are quotation marks, reverse solidus and the control character (U+0000 through U+001F). RDF serialization in JSON follows a non-standardized specification, but can be considered a good overall solution for internationalization [9]. This serialization, however, is experimental and available only for the DBpedia *Live* module.

As it provides the most compact (non-experimental) solution with full internationalization support, we serialize the Extraction Framework output in Turtle.

IV. Results

As mentioned earlier, finding an appropriate mapping for every single Wikipedia infobox/property is not possible (nor needed). Still, using the Mapping wiki, we have successfully mapped:

- 38.90 % of all templates in Wikipedia (440 of 1131).
- 20.52 % of all properties in Wikipedia (6042 of 29438).
- 96.92 % of all template occurrences in Wikipedia (174594 of 180146).
- 67.52 % of all property occurrences in Wikipedia (1485832 of 2200536).

The mapping statistics show that the Serbian DBpedia is on par with the best covered Wikipedias, taking the leading position in the percentage of all Wikipedia templates that are mapped at 38.90%, and all template occurrences in Wikipedia that are mapped, with total coverage of 96.92% (Figure 3).

Configured to use only the label and the mapping extractor, the Extraction Framework produces a dataset of 3,051,772 triples (TABLE III). This number can be further boosted by including other extractors, such as the *Page Links*, *Disambiguation*, *Wiki Page* and, especially, the *Infobox* extractor. As previously mentioned, the last module increases the triple count at the expense of data quality, by extracting all properties from all infoboxes. As such property types are not part of a subsumption hierarchy and there is no consistent ontology for the infobox dataset, the use of this extractor is advised only if an application requires complete coverage of all Wikipedia properties (and noisy data is not an issue). The Generic Infobox Extractor alone produces 2,795,427 triples.

TABLE III.
Extraction results

Resource type	Number of triples
Instance types	979,022
Labels	553,368
Mappings	1,519,382
Transliterated literals	415,406
Total	3,467,178

The mapped infobox properties hold 1,016,180 string literals (Cyrillic, Latin, purely numeric and other), 415,406 of which are encoded in Cyrillic. As previously described, the same number is transliterated to the Serbian Latin alphabet using our post-processor, and then fused back with the original dataset.

V. Conclusions and Future Work

Thanks to the achieved results, we're certain the produced dataset will prove to be an invaluable resource for the Serbian Linked Data community.

The future updates should involve less manual work, as most of the mappings are in place, and some of the post-processing work is already being transferred to the DBpedia Extraction Framework. The Digraphic Extractor will be able to automatically transliterate both Cyrillic and Latin string literals, while skipping those that are not meant to be transformed (by detecting the MediaWiki syntax constructs and magic words[1] that prevent them from being transliterated).

Furthermore, the findings concerning the ontology itself will be reported back to the DBpedia community with suggestions for improvements, such as using an evaluation methodology to make it theoretically sound and complete, and, most of all, reusable. In order to give back to Wikipedia, the results will be announced to the Serbian Wikipedia community. To help better coordinate and standardize their efforts, the DBpedia ontology will be recommended as a common vocabulary for Wikipedia infoboxes and properties.

Finally, we're safe to assume the collective experience and findings we've come across on our way to producing the largest Serbian Linked Data knowledge source should not only lead to better future versions of this particular dataset, but to a better DBpedia, in general.

References

[1] J. Brank, M. Grobelnik, and D. Mladenic, "A Survey of Ontology Evaluation Techniques", *Proceedings of Conference on Data Mining and Data Warehouses*, 2005

[2] K. Kozaki, Y. Kitamura, M. Ikeda, and R. Mizoguchi, "Hozo: An Environment for Building/Using Ontologies based on a Fundamental Consideration of Role and Relationship", *Proceedings of the 13th International Conference Knowledge Engineering and Knowledge Management*, 2002, pp. 213-218

[3] D. Kontokostas, C. Bratsas, S. Auer, S. Hellmann, I. Antoniou, G. Metakides, "Internationalization of Linked Data: The case of the Greek DBpedia edition", *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. XV, pp. 51-61, September 2012

[4] S. Hellmann, C. Stadler, J. Lehmann, and S. Auer, "DBpedia live extraction", *Proceedings. of 8th International Conference on*

[1] https://www.mediawiki.org/wiki/Help:Magic_words

Ontologies, *DataBases, and Applications of Semantics*, vol. 5871 of Lecture Notes in Computer Science, 2009, pp. 1209 - 1223.

[5] U. Milošević, "I18n of Linked Data Tools with respect to Western Balkan Languages", *Proceedings of the 3rd International Conference on Information Society Technology*, 2013

[6] W3C, *N-Triples*, Retrieved from http://www.w3.org/2001/sw/RDFCore/ntriples/

[7] D. Beckett, T. Berners-Lee, *Turtle - Terse RDF Triple Language*, 2011, Retrieved from http://www.w3.org/TeamSubmission/turtle/

[8] J. J. Carroll, P. Stickler. *RDF Triples in XML*, 2004. Retrieved from http://www.hpl.hp.com/techreports/2003/HPL-2003-268.pdf

[9] I. Davis, T. Steiner, A. J. Le Hors, *RDF 1.1 JSON Alternate Serialization (RDF/JSON)*, 2013, Retrieved from https://dvcs.w3.org/hg/rdf/raw-file/default/rdf-json/index.html

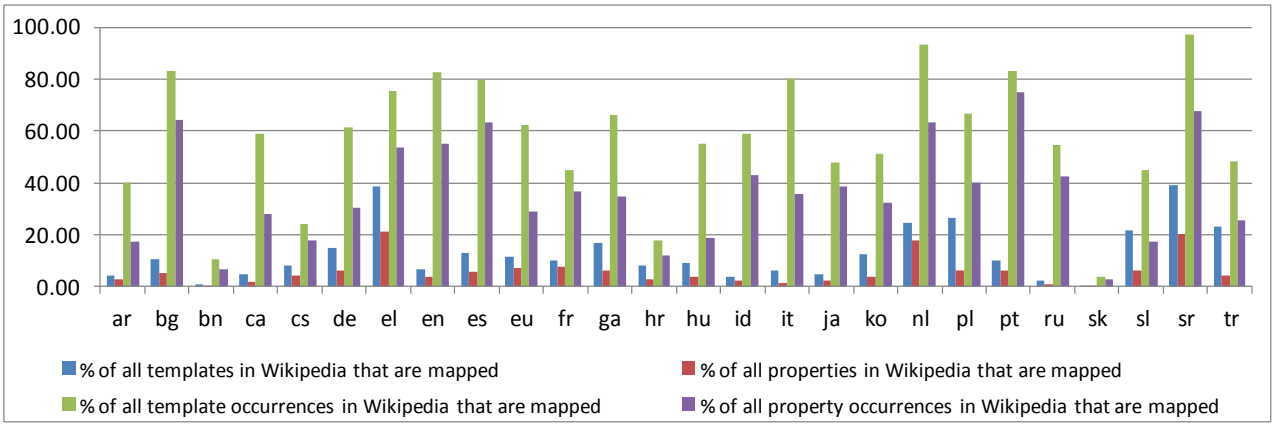

Figure 3. Mapping statistics