Latest enhancements in the Spanish DBpedia*

Sara Sanz-Lucio, Oussama Tahiri-Alaoui, and
Mariano Rico[0000−0001−5878−8521]

Ontology Engineering Group,
Universidad Politécnica de Madrid, Spain
{sara.sanz.lucio@alumnos, oussama.talaoui@alumnos, mariano.rico@}upm.es

Abstract. The Spanish DBpedia is a data source used initially to support the Spanish community. However, our logs show that the Spanish language goes beyond Spanish speakers and many non-Spanish speakers use the Spanish DBpedia on a daily basis. In the last months we have made two important enhancements to the Spanish DBpedia: (1) we publish a nonstandard dataset containing the type of resources that in the standard distribution have no type, and (2) we update automatically our data every week by using the DBpedia databus. In this way, we satisfy a frequent request made by companies and we foster the usage of the Spanish language, the second mother language by the number of speakers (after Chinese), and the second in scientific papers (after English).

Keywords: Spanish DBpedia · Resource type · DBpedia data bus.

1 Introduction

1.1 The rising of the Spanish language

The data published by the Cervantes Institute in its 2020 report [4] are overwhelming: Spanish speakers have increased by 30% in the last decade, and the number of foreigners who study it has grown by 60%. More than 585 million people speak Spanish. Of these, almost 489 million are native Spanish speakers. Furthermore, Spanish is the second mother tongue by number of speakers after Mandarin Chinese, and the third language in the global count of users after English and Mandarin Chinese. On the Internet, it is the third most used and is the second language, behind English, publishing scientific texts.

The DBpedia project has long generated semantic information from English Wikipedia. Since June 2011, the information generation process has extracted information from Wikipedia in 111 of its languages, but only 18 languages have a DBpedia chapter with a website. One of them is Spanish. The DBpedia Internationalization Committee has assigned a website and a SPARQL [7] endpoint for

* Partially supported by HcommonK (RTC2019-007134-7) and Datos4.0 (TIN2013-46238-C4-3-R) projects.

each of these languages[1]. In the case of Spanish (with website es.dbpedia.org), the extraction process produces more than 100 million RDF triples from the Spanish Wikipedia. All these triples are available on the SPARQL endpoint `es.dbpedia.org/sparql` using Semantic Web [2] and Linked Data [1] technologies.

1.2 The DBpedia datasets

As we have mentioned previously, DBpedia extracts data from 111 different language editions of Wikipedia. Then, for each language we have a knowledge base (a "Knowledge Graph" in modern terminology, abbreviated as KG). The largest DBpedia KG is extracted from the English edition of Wikipedia, with around 400 million facts (triples) that describe 3.7 million resources (Wikipedia entries). The DBpedia knowledge graphs that are extracted from the other 110 Wikipedia editions together consist of 1.46 billion facts and describe 10 million additional resources. Therefore, two-thirds of the information in DBpedia comes from non-English Wikipedias.

From a technical perspective, the DBpedia project maps Wikipedia infoboxes [12] from 27 different language editions into the DBpedia ontology, a single shared ontology consisting of 320 classes and 1,650 properties. The mappings are created via a worldwide crowd sourcing effort and enable knowledge from the different Wikipedia editions to be combined. The DBpedia project publishes regular releases of all DBpedia knowledge bases for download and provides SPARQL query access to 18 out of the 111 language editions via a global network of local DBpedia chapters.

In addition to the regular releases, the project maintains a live knowledge base which is updated whenever a page in Wikipedia changes. DBpedia sets 27 million links pointing to many external data sources (e.g. Wikidata, Yago, Freebase) and thus enables data from these sources to be used together with DBpedia data. Several hundred data sets on the Web publish RDF links pointing to DBpedia and thus make DBpedia the central interlinking hub in the Linked Open Data (LOD) cloud[2].

1.3 The Spanish DBpedia datasets

Around 40% of the resources (entries) in the Spanish Wikipedia are not pointed (do not have links) by the English Wikipedia [5], which means they are "non-canonical" datasets, that is, 40% of the information stored in the Spanish DBpedia is exclusively stored in the Spanish DBpedia and is not available in the English DBpedia. This fact places the Spanish DBpedia as a valuable and exclusive source of local information. Additionally, we have to remark that the English DBpedia does not contain all the information stored in local DBpedias, but only a minimum part comprising labels and abstracts.

[1] See `https://www.dbpedia.org/members/chapter-overview/`

[2] See `https://lod-cloud.net`

2 New contributions to the Spanish DBpedia

The contributions made to the Spanish DBpedia during the last year are described in the next sections.

2.1 On new DBpedia types

Each DBpedia resource usually has more than one type. For example, the resource `Cervantes` has types `Agent`, `Person` and `Artist`, this is because the DBpedia ontology defines a well-known hierarchy of classes.

A previous study [8] showed that a large number of DBpedia resources, around 16%, do not have any type. DBpedia, in its 3.9 (English) version, has more than four million resources, but only around 50% of them have a type beyond level 1 (we consider `Thing` as level 0). Having correct types is important when working with semantic information, because it allows for better data queriability and discoverability. Thus, the more types we have and the more precise they are, the better data quality.

In the study mentioned previously, made by members of our research group, it was presented a new technique that improved the overall quality of the English DBpedia dataset by (1) providing type(s) to those resources lacking a type, and (2) adding more specialized types to already typed resources. Following the previous example about `Cervantes`, this means to infer that it also has the type `Writer`, a more specific type in the DBpedia hierarchy, although this fact was not in the original dataset. These inferred types are stored in a new dataset so that now it is made publicly available in Spanish and English.

This method surpasses the so-called SDTypes dataset as shown in figure 1. The small circle represents the number of types predicted by the SDType approach (1.38M). The big circle represents our approach, which produces 56.7% more types (2.15M). Our approach predicts most of the types predicted by the SDTypes approach[6], specifically 96.5% of them. Therefore, we conclude that our approach predicts most of the types predicted by the SDTypes approach.

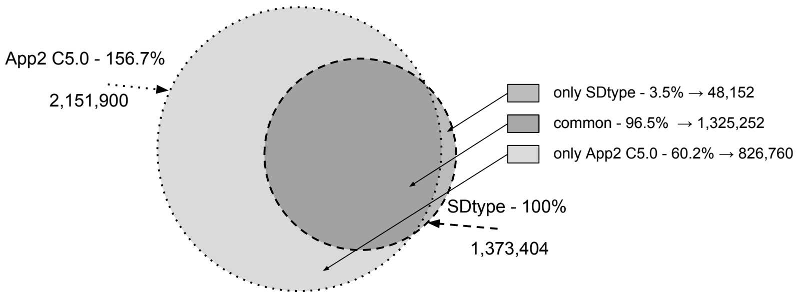

Fig. 1. Overlapping between the types predicted by our approach (light gray) and the ones predicted by the SDTypes approach (dark gray).

The contribution of the present work is to notify that we have overcome a technical issue that limited the number of resources that could be analyzed. This limitation restricted our study to old versions of DBpedia, in which we had less than 2.4 million resources. The master thesis [9] of the first author of this paper, Sara, was able to solve a limitation of the C5.0 library (a popular multiclass classifier). Now we can analyze any DBpedia and provide a new dataset to the DBpedia community, but we are focused mainly on the English and Spanish datasets. The publication mechanism is the DBpedia databus, mentioned in the next section.

2.2 On DBpedia automatic update process

In the past, the update process was based on a manual process in which we had to download the Spanish datasets provided periodically by DBpedia. This periodicity was the refresh rate of the DBpedia extraction process, and was in the range of months.

However, now we have an automated process and we get fresh data every week. This is possible by using the DBpedia Databus [3], which aims at providing an end-to-end pipeline easing auto-extraction, metadata-generation and publishing of Linked Data Knowledge Graphs at scale. The Databus platform provides two tools to connect consumers and producers, one for consumers (`https://databus.dbpedia.org`) and the SPARQL API (`https://databus.dbpedia.org/repo/sparql`) serves as a user interface to configure data set retrieval and combination in catalogs and the other for providers, the Databus Maven plugin (`http://dev.dbpedia.org/Databus/Derive/Maven/Integration`) which enables systematic upload and release of datasets on the bus.

The idea behind using Databus to get our data is the fact that it is able to generate metadata about datasets and then upload this metadata, so that anybody can query, download, derive, and build applications with this data via the Databus. Since the integration of data is easy with the Databus, many additional datasets have been integrated and loaded alongside DBpedia for the world to query. The process to load datasets on the bus comprises these phases:

- Acquisition: data is downloaded from the source and logged in.
- Conversion: data is converted to N-Triples and cleaned (Syntax parsing, datatype validation and SHACL).
- Mapping: the vocabulary is mapped on the DBpedia Ontology and converted (this was being done for Wikipedia's Infoboxes and Wikidata [11], but now it is done for other datasets as well).
- Linking: Links are mainly collected from the sources, cleaned and enriched.
- IDying: All entities found are given a new Databus ID for tracking.
- Clustering: IDs are merged onto clusters using one of the Databus IDs as cluster representative.
- Data Comparison: Each dataset is compared with all other datasets. We have an algorithm that decides on the best value, but the main goal here is transparency, i.e. to see which data value was chosen and how it compares to the other sources.

- A main knowledge graph fused from all the sources, i.e. a transparent aggregate.
- For each source, a local fused version called the "Databus Complement" is being produced. This is a major feedback mechanism for all data providers, where they can see what data they are missing, what data differs in other sources and what links are available for their IDs.
- The possibility to compare all data via a web service.

3 Conclusions and future work

We share with the community the enhancements made to the Spanish DBpedia in the last year. This has been the biggest change since its creation in 2011. The changes include a new user interface, automatic weekly updates, and the creation of a high-quality dataset on resource's types. All of this is publicly available (code, datasets and instructions) on the Internet [9, 10].

References

1. Auer, S., Lehmann, J., Ngonga Ngomo, A.C.: Introduction to linked data and its lifecycle on the web. In: Reasoning on the Web in the Big Data Era. vol. 6848, pp. 1–75 (01 2011)
2. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Scientific American **284**(5), 34–43 (2001), `http://www.jstor.org/stable/26059207`
3. Frey, J., Hofer, M., Obraczka, D., Lehmann, J., Hellmann, S.: DBpedia FlexiFusion the best of wikipedia > wikidata > your data. In: ISWC 2019
4. García-Montero, L.: Spanish: a living language. Report 2020. Tech. rep., Instituto Cervantes (2020), `https://cvc.cervantes.es/lengua/espanol_lengua_viva/pdf/espanol_lengua_viva_2020.pdf`
5. Mihindukulasooriya, N., Rico, M., García Castro, R., Gómez-Pérez, A.: An analysis of the quality issues of the properties available in the spanish dbpedia. In: AEPIA conference. pp. 198–209 (2015)
6. Paulheim, H., Bizer, C.: Improving the quality of linked data using statistical distributions. IJSWIS **10**(2), 63–86 (2014)
7. Prud'hommeaux, E., Seaborne, A.: SPARQL query language for RDF, W3C recommendation (2008), `http://www.w3.org/TR/rdf-sparql-query/`
8. Rico, M., Santana-Pérez, I., Pozo-Jiménez, P., Gómez-Pérez, A.: Inferring types on large datasets applying ontology class hierarchy classifiers: The dbpedia case. In: EKAW 2018. pp. 322–337 (2018)
9. Sanz-Lucio, S.: Detección de tipos en DBpedia. Master's thesis, Universidad Politécnica de Madrid (2021), `http://oa.upm.es/`
10. Tahiri-Alaoui, O.: An approach to automatically update the Spanish DBpedia using DBpedia Databus. Master's thesis, Universidad Politécnica de Madrid (2020), `http://oa.upm.es/63646/`
11. Vrandečić, D., Krötzsch, M.: Wikidata: A free collaborative knowledgebase. Communications of the ACM **57**(10), 78–85 (2014). https://doi.org/10.1145/2629489
12. Wu, F., Weld, D.: Automatically refining the wikipedia infobox ontology. In: Proceeding of the 17th International Conference on World Wide Web 2008, WWW'08. pp. 635–644 (04 2008)