

Towards Linked Data Internationalization - Realizing the Greek DBpedia

Dimitris Kontokostas
Web Science Program
Mathematical Department
Aristotle University of
Thessaloniki
<http://webscience.auth.gr>
jimkont@math.auth.gr

Sebastian Hellmann
University Leipzig
Institut für Informatik
<http://aksw.org>
hellmann@informatik.uni-leipzig.de

Charalampos Bratsas
Web Science Program
Mathematical Department
Aristotle University of
Thessaloniki
<http://webscience.auth.gr>
cbratsas@math.auth.gr

Ioannis Antoniou
Web Science Program
Mathematical Department
Aristotle University of
Thessaloniki
<http://webscience.auth.gr>
iantonio@math.auth.gr

Sören Auer
University Leipzig
Institut für Informatik
<http://aksw.org>
auer@informatik.uni-leipzig.de

George Metakides
Web Science Program
Mathematical Department
Aristotle University of
Thessaloniki
<http://webscience.auth.gr>
george@metakides.net

ABSTRACT

This paper describes the realization of the Greek DBpedia as part of the DBpedia Internationalization process. “I18n filters” are proposed as pluggable components of the *DBpedia Information Extraction Framework*, in order to address issues concerning covering more knowledge from non-English Wikipedia’s and International Resource Identifier (IRI) support. Moreover, a new extractor is introduced that uses the *Wikipedia Interlanguage Links* to connect international DBpedia’s and transitively to the LOD Cloud. Finally, the paper illustrates the first international project which provides Transparent Content Negotiation (TCN) rules for International Resource Identifier’s (IRI’s) for de-referencing purposes. This work could serve as a guide not only for other multilingual DBpedias, but for publishing linked data in languages based on non-Latin character sets as well.

Categories and Subject Descriptors

I.2.4 [Artificial Intelligence]: Knowledge Representation Formalisms and Methods; E.2 [Data]: Data Storage Representations—*Linked Representations*

General Terms

Management, Languages

Keywords

DBpedia, multi-lingual, internationalization, I18n, IRI, URI, TCN, Linked Data

1. INTRODUCTION

This paper presents the recent DBpedia Internationalization effort starting with the realization of the Greek DBpedia. The early versions of the DBpedia framework were restricted to the English Wikipedia as its sole source. Since the start of DBpedia in 2007 [6], however, the focus has shifted. DBpedia is now becoming an increasingly fused version, which integrates information from many different Wikipedia editions. The emphasis of this fused DBpedia is still on the English Wikipedia, as it is the most abundant language edition. During the fusion process, however, country specific information is lost or ignored. Using the Greek Wikipedia as a base, the aim of the research described in this article is to create a Greek DBpedia and establish best practices (complemented by software) to allow the DBpedia community¹ to easily generate, maintain and properly interlink a language specific DBpedia edition.

The Greek Wikipedia is, when compared to other Wikipedia language editions, still relatively small - 47th in article count - with around 60.000 articles. Furthermore, it is not as well-organized as the English one regarding infobox usage, one of the preliminaries to extract high-quality data with the DBpedia approach. As only few Greek data sources are published as Linked Open Data (LOD) yet, the Greek DBpedia could not only serve as the core where all these datasets could be interlinked, but also provide a guideline on how they could be published, how non-latin characters could be handled and how the Transparent Content Negotiations (TCN) rules (RFC 2295) [5] can be implemented.

2. CURRENT DATA TOPOLOGY OF DBPEDIA

¹The authors established a DBpedia Internationalization Committee to collect other interested community members to create a network of internationalized DBpedias (<http://dbpedia.org/internationalization>)

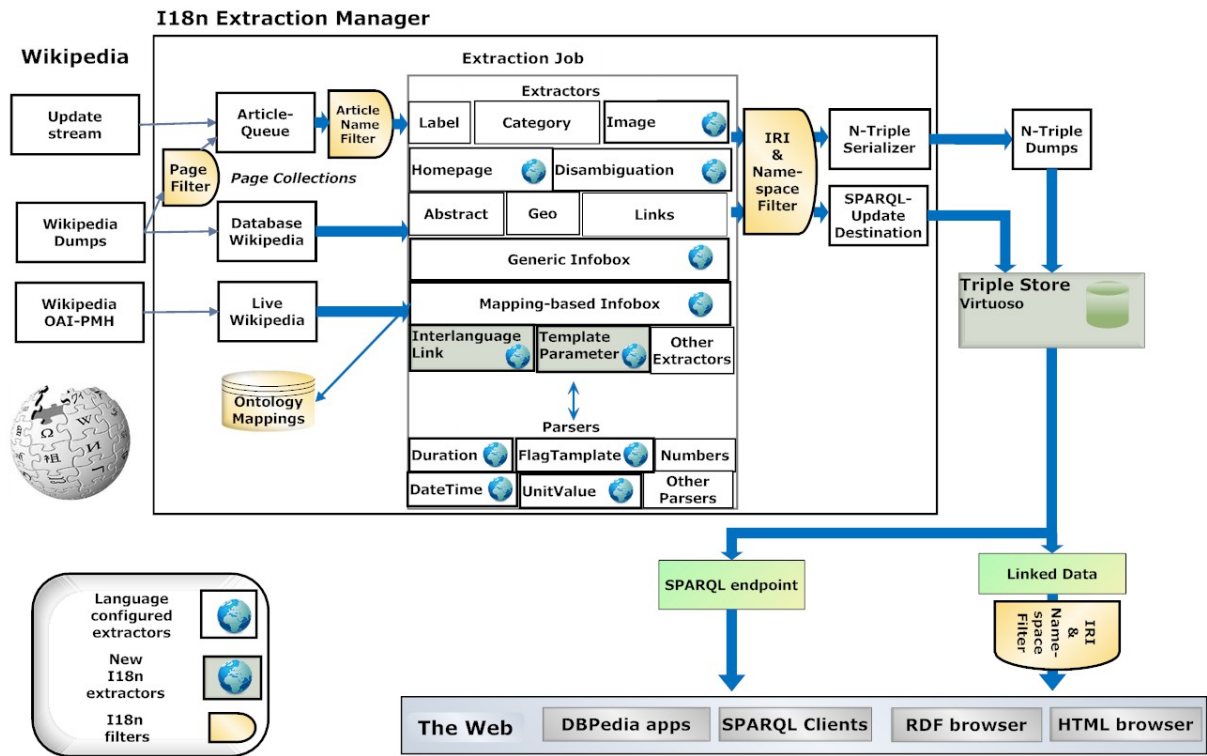


Figure 1: The Internationalized DBpedia Information Extraction Framework including the I18n filters and the two new extractors.

The data sets extracted by the DIEF are made available either via the DBpedia SPARQL Endpoint or as downloadable dumps². The focus of these data sets is clearly on the English part of Wikipedia. Although other language editions are used as sources, neither the DIEF nor the pluggable extractors nor the parsers have been tailored to deal with specialties of other language edition. With minor exception, the extractors which have been designed for the English version have been applied *as-is* on other sources and lack *customizability*. Currently, around 95 language editions are used and basic information such as abstracts, labels, titles, links, page links and geographic information is provided. In order to provide a uniform resource identifier (URI) for all languages, DBpedia currently uses the Wikipedia inter-language links³, assigning non-English articles the corresponding English resource identifier. For instance, there exist articles about the Greek city of Thessaloniki in Greek, which are translated to other languages, all translations using the same (English) resource name <http://dbpedia.org/resource/Thessaloniki>.

2.1 Current State of the Internationalization Effort

The introduction of the Mapping-Based Extractor in [6] alongside crowd-sourcing approaches in [4] allowed the International DBpedia community to easily define infobox-to-

²Endpoint: [urlhttp://dbpedia.org/sparql](http://dbpedia.org/sparql)

Downloads: <http://wiki.dbpedia.org/Downloads>

³http://en.wikipedia.org/wiki/Help:Interlanguage_links

ontology mappings using a very simple syntax⁴. As a result of this development, there are presently mappings defined in 14 languages⁵ in addition to English.

At the time of writing, three official DBpedia chapters, apart from the English one, exist : the German⁶, the Korean⁷ and the Greek⁸. While the first two provide URI-based datasets and SPARQL Endpoints, the Korean chapter has also made an effort to export their datasets with localized IRIs. In the Greek DBpedia, we managed to define Transparent Content Negotiation rules (RFC 2295) [5] for IRI de-referencing.

3. EXTENSION OF THE DBPEDIA INFORMATION EXTRACTION FRAMEWORK

In this contribution the (new) *Internationalized DBpedia Information Extraction framework* (I18n-DIEF) is outlined as illustrated in Figure 1. In particular with regard to the infobox extraction process, solutions are presented so as the Greek Wikipedia can be improved and be aligned with the English version. The main approach to facilitate internationalization was to implement “internationalization filters”, where necessary, that are plugged into the DBpedia extraction framework and provide required I18n functionality.

The first priority was to improve the extractors and cus-

⁴<http://mappings.dbpedia.org>

⁵en, de, fr, pl, it, es, nl, pt, ca, hu, sl, hr, el, ga, ru

⁶<http://de.dbpedia.org>

⁷<http://ko.dbpedia.org>

⁸<http://el.dbpedia.org>

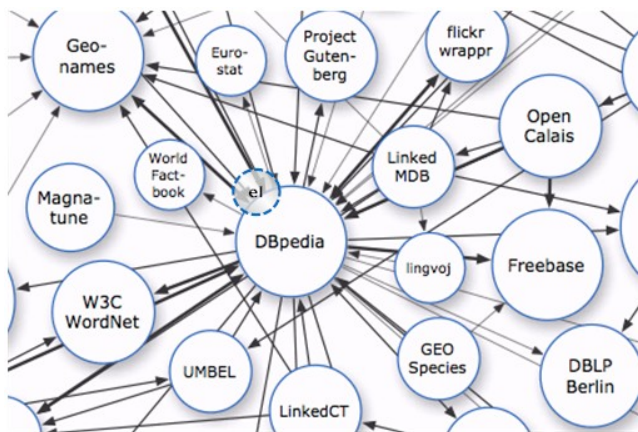


Figure 2: The Greek DBpedia in the i18n LOD Cloud (modification Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch <http://lod-cloud.net/>)

tomize/tune them - where appropriate - for language specific aspects in order to increase the amount and quality of the produced triples. Five extractors and four parsers could be enhanced: the Homepage, Image, Disambiguation, Generic Infobox and the Mapping-Based Extractor and the Date-Time, Duration, Flag-Template and Unit-Value Parsers were extended with more I18n options.

We argue that the DBpedia naming scheme has to be improved in order to properly incorporate knowledge from other Wikipedia language editions. Before our effort, information extracted from different language editions was merged into one single namespace (i.e. <http://dbpedia.org/resource/>). However, it is more appropriate that language specific namespaces are introduced, such as <http://el.dbpedia.org/resource/> for Greek. And as new languages start publishing their data, the English DBpedia could be transferred in <http://en.dbpedia.org> sub-domain and the default domain could be used solely for the “Cross-language knowledge fusion” [6]

A new extractor was developed for inter-DBpedia linking. Using the country specific resource domain approach, DBpedia can use the Wikipedia interlanguage links to link international DBpedia resources using the `owl:sameAs` predicate. Using these links, the smaller DBpedia language editions will be linked to the English DBpedia, and thus transitively to all other resources. To accomplish this, the *Interlanguage-Links Extractor* (ILL in Figure 1) was implemented, that extracts such links, e.g. `dbp:Thessaloniki owl:sameAs dbp-el:Θεσσαλονίκη`.

Additionally to the inter-DBpedia linking, another tool was developed that generates automatically the links between a non-English DBpedia and all the external LOD datasets that are linked to the English DBpedia. To achieve this, the inter-DBpedia linking (`owl:sameAs`) datasets are used and we join them with the external LOD datasets that DBpedia offers as downloads. The created datasets discard the RDF triples which do not link with the non-English DBpedia. With this tool, a non-English DBpedia can share all LOD links relevant with the English DBpedia. As a result of our

About: **Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης**
 An Entity of Type: **organisation**, from Named Graph: <http://el.dbpedia.org>, within Data Space: el.dbpedia.org

Το Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης (Α. Π. Θ.) είναι πανεπιστημιακό ίδρυμα της Ελλάδας, με έδρα την Θεσσαλονίκη. Ιδρύθηκε το 1925 κατόπιν ευγενικής του τότε πρωθυπουργού Αλέξανδρου Παπαναστασίου κατά την περίοδο της πρώτης Ελληνικής Δημοκρατίας, και αρχικά ονομάζονταν Πανεπιστήμιο Θεσσαλονίκης. Μετονομάστηκε το 1954 και σήμερα λειτουργεί με ονομασία 42 τμήματα. Τα οποία οργανώνονται σε 12 σχολές, καλύπτοντας ένα ευρύ φάσμα επιστημονικών πεδίων. Το πανεπιστήμιο συστάθηκε Αριστοτέλεια προς τιμήν του φιλόσοφου Αριστοτέλη, ενώ το όνομα του πανεπιστημίου αποκαλύπτει τον Άγιο Δημήτριο, ο οποίος είναι προστάτης της πόλης της Θεσσαλονίκης. Η πανεπιστημιακή (campus) του ΑΠΘ εκτείνεται σε μία έκταση 430.000 τετραγωνικών μέτρων περίπου και βρίσκεται κοντά στο κέντρο της Θεσσαλονίκης. Άδρια πικρή όχθη του campus, αλλά και για διάφορα ιστορικά μνημεία, μερικά από τα εγκαταστάσεις του πανεπιστημίου βρίσκονται εκτός της πανεπιστημιακής (ή ακόμη και εκτός του παλαιότερου συγκροτήματος της Θεσσαλονίκης). Το Α. Π. Θ. είναι σε αριθμό φοιτητών το μεγαλύτερο εκπαιδευτικό ίδρυμα της Ελλάδας. Στην έκθεση Academic Ranking of World Universities για το 2006, κατατάχθηκε στις θέσεις 303-401, ανάμεσα στα 500 καλύτερα πανεπιστήμια στον κόσμο.

Property	Value
<code>dbpedia-owl:abstract</code>	<ul style="list-style-type: none"> Το Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης (Α. Π. Θ.) είναι πανεπιστημιακό ίδρυμα της Ελλάδας, με έδρα την Θεσσαλονίκη. Ιδρύθηκε το 1925 κατόπιν ευγενικής του τότε πρωθυπουργού Αλέξανδρου Παπαναστασίου κατά την περίοδο της πρώτης Ελληνικής Δημοκρατίας και αρχικά ονομαζόταν Πανεπιστήμιο Θεσσαλονίκης. Μετονομάστηκε το 1954 και σήμερα λειτουργεί με ονομασία 42 τμήματα. Τα οποία οργανώνονται σε 12 σχολές, καλύπτοντας ένα ευρύ φάσμα επιστημονικών πεδίων. Το πανεπιστήμιο συστάθηκε Αριστοτέλεια προς τιμήν του φιλόσοφου Αριστοτέλη, ενώ το όνομα του πανεπιστημίου αποκαλύπτει τον Άγιο Δημήτριο, ο οποίος είναι προστάτης της πόλης της Θεσσαλονίκης. Η πανεπιστημιακή (campus) του ΑΠΘ εκτείνεται σε μία έκταση 430.000 τετραγωνικών μέτρων περίπου και βρίσκεται κοντά στο κέντρο της Θεσσαλονίκης. Άδρια πικρή όχθη του campus, αλλά και για διάφορα ιστορικά μνημεία, μερικά από τα εγκαταστάσεις του πανεπιστημίου βρίσκονται εκτός της πανεπιστημιακής (ή ακόμη και εκτός του παλαιότερου συγκροτήματος της Θεσσαλονίκης). Το Α. Π. Θ. είναι σε αριθμό φοιτητών το μεγαλύτερο εκπαιδευτικό ίδρυμα της Ελλάδας. Στην έκθεση Academic Ranking of World Universities για το 2006, κατατάχθηκε στις θέσεις 303-401, ανάμεσα στα 500 καλύτερα πανεπιστήμια στον κόσμο.
<code>dbpedia-owl:campus</code>	<ul style="list-style-type: none"> <code>dbpedia-el:Θεσσαλονίκη</code> <code>dbpedia-el:Ελλάδα</code> 150px(Λογότυπο ΑΠΘ)
<code>dbpedia-owl:motto</code>	
<code>dbpedia-owl:numberOfPostgraduateStudents</code>	9000 (xsd:integer)
<code>dbpedia-owl:numberOfStudents</code>	90000 (xsd:integer)
<code>dbpedia-owl:numberOfUndergraduateStudents</code>	86000 (xsd:integer)
<code>dbpedia-owl:staff</code>	2330 (xsd:integer)
<code>dbpedia-owl:thumbnail</code>	<ul style="list-style-type: none"> http://upload.wikimedia.org/wikipedia/commons/thumb/5/5c/Flag_of_Greece.svg/200px-Flag_of_Greece.svg.png http://www.apth.gr/ http://www.apth.gr/services/personal/accounts http://www.apth.gr
<code>dbprop:hasPhotoCollection</code>	<ul style="list-style-type: none"> http://www2.welcs.fu-berlin.de/ckw/wap/wikipedia/infobox/University_of_Thessaloniki http://www.w3.org/2006/03/wmw/20/instances/synse/university-noua-2
<code>dbprop:wordnet_type</code>	
<code>dbprop:wikiPageUsesTemplate</code>	<ul style="list-style-type: none"> <code>dbpedia:Πρότυπο:Πανεπιστήμιο</code>

Figure 3: HTML representation using TCN rules

inter-DBpedia linking in the Greek DBpedia, 112,000 more RDF triples were created, linking the Greek DBpedia, with 20 external LOD datasets⁹ (cf. Figure 2).

In order to facilitate non-Latin DBpedia language editions we migrated the DBpedia extraction framework from URIs [1] to Internationalized Resource Identifiers (IRIs) [2] thus avoiding unreadable identifiers due to the %-encoding of non-Latin characters (cf. Figure 3).

Problems encountered in this process are discussed, especially concerning the definition of Transparent Content Negotiation (TCN). The peculiarity in defining TCN rules for IRI's, lies in the HTTP protocol (rfc 2616) [3], which accepts only URI's. The followed approach was to store data in the IRI form but encoding and decoding IRIs into URIs for de-referencability purposes. A general solution is also proposed that allows to deal not only with de-referencing, but with data serialization issues concerning IRI's.

4. CONCLUSIONS

As a result of the DBpedia I18n effort, there is an increase by 62.6% in total triples, compared to the standard DBpedia approach. Also, the usability for non-Latin characters was substantially improved. The Greek DBpedia is a step towards LOD Internationalization. To our knowledge, this is the first successful attempt to serve Linked Data with de-referencable IRI's for LOD publishing in non-latin languages. Since more than half of the Web users nowadays are speakers of a language using a non-Latin-based script, for LOD to be successful on the Web in a global scale the support for non-Latin character sets has to be improved.

5. ACKNOWLEDGMENTS

This project would not have been completed without the continuous support of the DBpedia team, the students and the staff of the Webscience MSc, Mr.Konstantino Stampoulis¹⁰, Greek Wikipedia administrator and the Webscience MSc program of Aristotle University of Thessaloniki that facilitated this effort. The administrative and financial support of the municipality of Veria is gratefully acknowledged.

⁹<http://el.dbpedia.org/en/datasets>
¹⁰<http://el.wikipedia.org/wiki/User:Geraki>

A part of this work was supported by a grant from the European Union's 7th Framework Programme provided for the project LOD2 (GA no. 257943).

6. REFERENCES

- [1] T. Berners-Lee, R. Fielding, and L. Masinter. Rfc 3986, uniform resource identifier (uri): Generic syntax. Request For Comments (RFC), 2005.
- [2] M. Duerst and M. Suignard. Internationalized Resource Identifiers (IRIs). RFC 3987 (Proposed Standard), January 2005.
- [3] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, and T. Berners-Lee. Hypertext transfer protocol – http/1.1 (rfc 2616). RFC 2616 (Proposed Standard), 1999.
- [4] S. Hellmann, C. Stadler, J. Lehmann, and S. Auer. DBpedia live extraction. In *Proc. of 8th International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE)*, volume 5871 of *Lecture Notes in Computer Science*, pages 1209–1223, 2009.
- [5] K. Holtman and A. Mutz. Transparent Content Negotiation in HTTP. RFC 2295 (Experimental), March 1998.
- [6] J. Lehmann, C. Bizer, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann. DBpedia - a crystallization point for the web of data. *Journal of Web Semantics*, 7(3):154–165, 2009.