WikInfoboxer: A Tool to Create Wikipedia Infoboxes Using DBpedia

Ismael Rodriguez-Hernandez, Raquel Trillo-Lado, and Roberto Yus

University of Zaragoza, Zaragoza, Spain
{587429, raqueltl, ryus}@unizar.es

Abstract. Wikipedia infoboxes present a summary, in a semistructured format, of the articles they are associated to. Therefore, they have become the main information source used by projects to leverage the knowledge in Wikipedia, such as DBpedia. However, creating quality infoboxes is complicated as current mechanisms are based on simple templates which, for example, do not check whether the information provided is semantically correct.
In this paper, we present *WikInfoboxer*, a tool to help Wikipedia editors to create rich and accurate infoboxes. WikInfoboxer computes attributes that might be interesting for an article and suggests possible values for them after analyzing similar articles from DBpedia. To make the process easier for editors, WikInfoboxer presents this information in a friendly user interface.

Keywords: Infoboxes, Wikipedia, DBpedia, Semantic Web

1 Introduction

Nowadays, Wikipedia[1] is the biggest free and collaborative encyclopedia thanks to the collaboration of thousands of anonymous editors along the world. Thus, it is one of the most visited websites on the Web. Moreover, Wikipedia is also the main information source for other projects, such as DBpedia [1] or Google Knowledge Graph [3], which aim to automatically process data in order to speed up research studies and analytics. Indeed, these projects benefit from the semistructured information in the *infoboxes* in some Wikipedia articles. Infoboxes are a fixed-format table with partially established parameters that provides standardized information across related articles and summarizes the most relevant facts in them. So, the higher the quality of Wikipedia articles and their infoboxes, the more reliable the results obtained in the studies based on them.

Currently, if a Wikipedia editor wants to include infoboxes in an article, firstly, she must select the infobox templates (which are also Wikipedia entries) appropriate for that article. After the selection process, the templates must be copied in the article and the editor has to fill in the values of the attributes in the template (e.g., birth place, name) by taking into account the textual

[1] https://www.wikipedia.org

hints provided in the articles describing the templates. This procedure is prone-to errors as templates are free form in nature and there is not any check of the type of input data either recommendation of values to fill in the attributes. Moreover, the same attribute coming from different templates can be interpreted in different ways and be filled with different inconsistent values. Therefore, only a quarter of the articles in Wikipedia have infoboxes approximately and a lot of infoboxes contain inaccurate data[2].

In this paper, we present *WikInfoboxer*[3], a tool based on Infoboxer (a previous prototype which has been updated [5]), whose main goal is becoming part o MediaWiki[4] to help Wikipedia editors to create rich and accurate infoboxes. Thus, our tool: 1) provides editors with suggestions about the templates to be used to create the infoboxes associated to their articles in Wikipedia, 2) ranks and aggregates the attributes of those templates, 3) suggests values to fill them whenever possible, and 4) links them to existing entities in Wikipedia. These functionalities are supported by statistic and semantic information extracted from DBpedia and Wikipedia templates. The rest of the paper is structured as follows. In Section 2, the architecture of WikInfoboxer and the technologies used to implement it are presented. Moreover, how to use WikInfoboxer and how to integrate it in MediaWiki and other WikiMedia Foundation[5] projects is depicted. Finally, the scenarios that will be presented in the demonstration session and the future work are described in Section 3.

2 Technical Overview

WikInfoboxer is a Web Information System composed of several modules (see Figure 1) that uses both statistical and semantic knowledge from a Knowledge Base to identify: 1) the most relevant attributes for a given Wikipedia entity and 2) the most relevant values and types for those attributes. Moreover, it also guides editors on the process of the selection of infoboxes templates for a Wikipedia entity, avoiding users to fill inconsistent data when redundant attributes appear in several templates selected.

The *User Interaction Module* manages user interactions and visualizes data previously obtained from the web server of the back-end of the system through HTTP calls. It can work in two different modes according to the expertise of the users: 1) *Expert Mode*, displaying statistical and semantic information about the attributes and their use (see Figure 2.a), and 2) *Basic Mode*, hiding details about the features of the attributes and their values for non-experts (see Figure 2.b). This module has been developed by using AngularJS[6] and Bootstrap[7].

[2] Data obtained from https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Infoboxes/Statistics accessed 25th April 2016.

[3] http://sid.cps.unizar.es/Infoboxer

[4] Software used to create and maintain Wikipedia (https://www.mediawiki.org).

[5] A non-profit organization that supports and operates Wikipedia and other free knowledge projects (https://wikimediafoundation.org).

[6] A framework to build Single Page Applications (https://angularjs.org).

[7] A framework to develop Web Responsive Applications (http://getbootstrap.com).

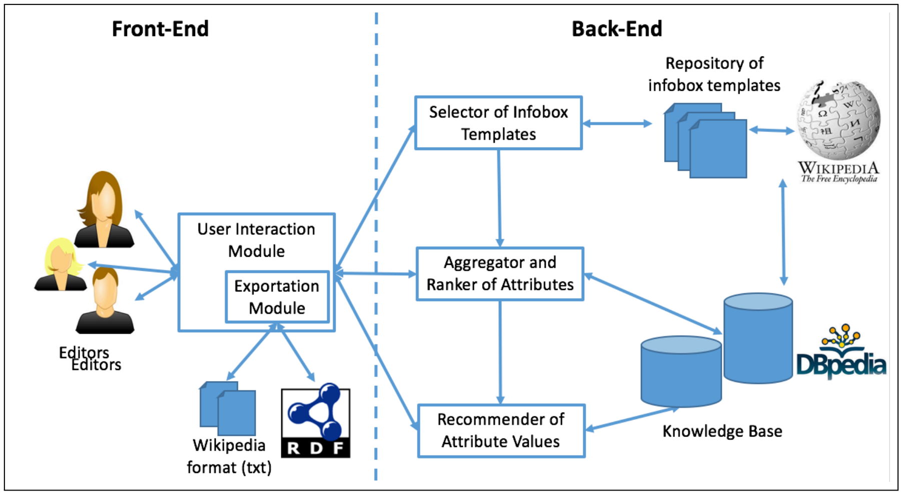


Fig. 1. High-level architecture of WikInfoboxer.

WikInfoboxer exports the created infoboxes in different formats. As the main goal is to help editors to create infoboxes for Wikipedia articles, the *Exportation Module* generates Wikipedia code in order to be directly copied and pasted in Wikipedia articles. Nevertheless, infoboxes are also translated into RDF to be loaded in the Knowlege Base. Moreover, the system can be easily adapted to export the information in the Wikidata format and publish it in Wikidata [4] by means of the Wikidata Toolkit[8].

The Back-End of WikInfoboxer is composed of three main modules (*Selector of Infobox Templates*, *Aggregator and Ranker of Attributes*, and *Recommender of Attribute Values*) in charge of selecting and computing statistical and semantics data, such as the domain and the range of properties, from Wikipedia and the *Knowledge Base (KB)* to guide editors in the process of creation of their infoboxes. It has been developed using the following technologies: Spring[9], OWL API[10] along with the HermiT reasoner[11], and a relational database management system (MySQL) to store statistics about users actions: filled properties, required timed, etc. and to maintain a cache of the results.

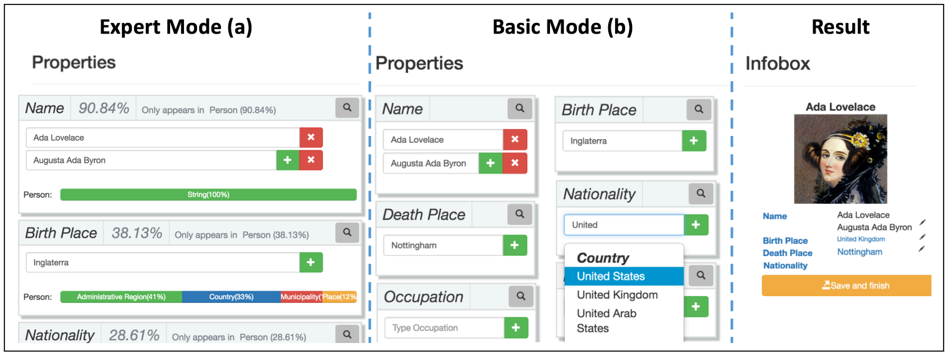
WikInfoboxer uses the DBpedia dump released in August 2015 as Knowledge Base. Nevertheless, other Open Knowledge Base or ontologies, such as Freebase or Wikidata, could be used. Performing efficient access to the Knowledge Base is essential to compute statistics for the different templates, attributes and values (and combining them) in real-time. So, the dump is managed by means of the

[8] An open source Java Library to interact with Wikidata (`https://www.mediawiki.org/wiki/Wikidata_Toolkit`).

[9] A framework to develop Java-based applications (`https://spring.io`).

[10] A Java API for OWL ontologies (`http://owlapi.sourceforge.net`).

[11] An OWL 2 reasoner (`http://www.hermit-reasoner.com`).



Name 90.84% Only appears in Person (90.84%)

Ada Lovelace

Augusta Ada Byron

Person: String(100%)

Birth Place 38.13% Only appears in Person (38.13%)

Inglaterra

Person: Administrative Region(41%) Country(33%) Municipality(Place(12%

Nationality 28.61% Only appears in Person (28.61%)

Name: Ada Lovelace; Augusta Ada Byron

Death Place: Nottingham

Occupation: Type Occupation

Birth Place: Inglaterra

Nationality: United

Ada Lovelace

Name: Ada Lovelace
Augusta Ada Byron

Birth Place: United Kingdom

Death Place: Nottingham

Nationality

Fig. 2. WikInfoboxer Graphical User Interface.

SPARQL server Apache Jena Fuseki using HDT [2] (a compressed RDF format) to speed up the select operations.

3 Demonstration Highlights and Future Work

The demonstration will consist of two phases. Firstly, we will ask attendances to create an infobox by using the current interface of Wikipedia for that purpose. After that, they could try to create an analogous infobox by using WikInfoboxer. Moreover, technical details, and data and graphics about the performance of WikInfoboxer will be shown.

As future work, we plan to incorporate WikInfoboxer as plugin of MediaWiki and integrate it on Wikipedia (see the proposal available on `https://meta.wikimedia.org/wiki/Grants:IEG/WikInfoboxer`). Moreover, we would like to explore the possibilities of WikInfoboxer as tool to populate domain-ontologies in different contexts such as public administrations and research environments.

Acknowledgments. This research was supported by the CICYT project TIN-2013-46238-C4-4-R, DGA FSE, and Keystone Action COST IC1302.

References

1. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: DBpedia - a crystallization point for the web of data. Web Semantics: Science, Services and Agents on the World Wide Web 7(3), 154–165 (2009)
2. Fernández, J.D., Martínez-Prieto, M.A., et al.: Binary RDF representation for publication and exchange (HDT). Journal of Web Semantics 19, 22–41 (2013)
3. Singhal, A.: Introducing the knowledge graph: Things, not strings. In: Official Blog of Google (2012)
4. Vrandecic, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Communications of the ACM 57, 78–85 (2014)
5. Yus, R., Mulwad, V., Finin, T., Mena, E.: Infoboxer: Using statistical and semantic knowledge to help create Wikipedia infoboxes. In: 13th ISWC. pp. 405–408 (2014)