Marefa:
Turning Publishers Catalogs’ Data Into Linked Data
Ahmed Ktob, Beihang University, Haidian, China
Zhoujun Li, Beihang University, Haidian, China

ABSTRACT
This article describes how recently, many new technologies have been introduced to the web; linked data is probably the most important. Individuals and organizations started emerging and publishing their data on the web adhering to a set of best practices. This data is published mostly in English; hence, only English agents can consume it. Meanwhile, although the number of Arabic users on the web is immense, few Arabic datasets are published. Publication catalogs are one of the primary sources of Arabic data that is not being exploited. Arabic catalogs provide a significant amount of meaningful data and metadata that are commonly stored in excel sheets. In this article, an effort has been made to help publishers easily and efficiently share their catalogs’ data as linked data. Marefa is the first tool implemented that automatically extracts RDF triples from Arabic catalogs, aligns them to the BIBO ontology and links them with the Arabic chapter of DBpedia. An evaluation of the framework was conducted, and some statistical measures were generated during the different phases of the extraction process.

KEYWORDS
Arabic Catalogs, Extraction Framework, Linked Data, RDF, Semantic Web, Web of Data

1. INTRODUCTION
In recent years, several exciting technologies empowered the web and made it richer and dynamic. Soon after its introduction by Berners-Lee, linked data (Berners-Lee, 2006) has grown to be a major technology and led to the creation of the Web of data (Bizer, Heath, & Berners-Lee, 2009). Data providers are continuously adopting a set of best practices (Atemezing & Villazón-Terrazas, 2014; Berners-Lee, 2006; Schmachtenberg, Bizer, & Paulheim, 2014) to publish their data on the web. For instance, the Linking Open Data (LOD) project (Abele, McCrae, Buitelaar, Jentzsch, & Cyganiak, 2017) included in its core 1163 published datasets along with statistical analysis about them (Abele et al., 2017; Schmachtenberg et al., 2014). These datasets are both cross-domain, such as DBpedia (Christian Bizer et al., 2009; Lehmann et al., 2015), and domain-specific such as LinkedGeoData (Stadler, Lehmann, Höflinger, & Auer, 2012). Most of these datasets are available in English; thus, neither non-English speakers nor applications can benefit from them.

DOI: 10.4018/IJSWIS.2018070108

Copyright © 2018, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
According to Cisco, the number of internet users in the Middle East and Africa is expected to grow by 6-folds between 2016 and 2021 with a compound rate of 44% annually (Cisco, 2017). Therefore, the number of Arabic users and Arabic resources on the web is and going to be enormous. Only a few Arabic datasets are available as linked data, yet. In addition, Ktob and Li (2017) identified the challenges pertinent to the creation of the Arabic Knowledge Graph (AKG) in which the scarcity of Arabic linked datasets was a significant factor. Besides the proprietary factor of data, the inexistence of tools designed with Arabic support in scope affects significantly the number of Arabic datasets published. Therefore, Marefa, meaning “knowledge”, is a tool developed to automatically extract and ease the publication of Arabic catalogs’ data as Resource Description Framework (RDF) datasets. It, to the best of our knowledge, is the first tool implemented to accomplish the goal mentioned above. Marefa, built with a flexible extraction framework, is user-friendly and easy to use.

The main contributions of this work are:

1. Marefa\(^1\) is the first framework developed to convert Arabic catalogs’ data to linked data;
2. Compared to DBpedia and Makh tota+ (Al-Rajebah & Al-Khalifa, 2012), our solution has more coverage (cf. section 2);
3. A 3-phase rigorous extraction process based on a simple and extensible framework (simplify the task for users and can be extended to other languages and dialects after some tweaking, cf. section 4);
4. A set of statistical measures produced along with the RDF dataset to further clarify the results (cf. section 5).

The remainder of the paper is organized as follows: Section 2 examines some of the related works. Section 3 discusses the different frameworks, Application Programming Interface (API), and technologies used by Marefa. The extraction framework and its constituent modules (including algorithms) are detailed in Section 4. Thereafter, Section 5 elaborates on the experiment conducted and gives some analysis of the obtained results. Finally, a conclusion along with future work ideas are included in Section 6.

2. RELATED WORKS

Over the previous decades, there were few similar works to Marefa that aim to improve the presence of Arabic datasets on the web of data. Although they share the same goal, Marefa is the first tool that targets Arabic catalogs to help publishers embrace the Web of data. In this section, the authors examined some of these works and their contributions to the Arabic linked data.

2.1. Arabic DBpedia

DBpedia represents one of the major projects initiated aiming to build the first version of the web of data. It uses a rigorous extraction framework to create RDF triples by extracting the structured data contained in the infoboxes of the Wikipedia articles of over 100 languages (Ismayilov, Kontokostas, Auer, Lehmann, & Hellmann, 2017; Lehmann et al., 2015). The extraction framework is built around 24 distinct extractors, each responsible for extracting a particular bit of information, grouped under four general categories: raw, mapping-based, feature, and statistical extraction (Lehmann et al., 2015). DBpedia in its 2016-10 release contained 13 billion RDF triples describing 6.6 million entities. DBpedia works toward becoming the central interlinking point in the web of data. By examining the state of the LOD project (cf. Figure 1), it became evident that this goal was practically achieved due to the significant number of inter/outer links DBpedia shares with other datasets (Ismayilov et al., 2017). Also, the DBpedia internationalization effort such as Greek (Kontokostas et al., 2012) and Arabic (Al-Feel, 2013, 2015) supported this claim.
This title is available in InfoSci-Computer Systems and
Software Engineering eJournal Collection, InfoSci-
Networking, Mobile Applications, and Web Technologies
Disciplines Computer Science, Security, and Information
Technology, InfoSci-Journal Disciplines Engineering, Natural,
and Physical Science, InfoSci-Select. Recommend this
product to your librarian:

www.igi-global.com/e-resources/library-
recommendation/?id=163

Related Content

Mediating RDF/S Queries to Relational and XML Sources
I. Koffina, G. Serfiotis, V. Christophides and V. Tannen (2006). International Journal
on Semantic Web and Information Systems (pp. 68-91).
www.igi-global.com/article/mediating-rdf-queries-relational-
xml/2827?camid=4v1a

Taxonomies of Problems and Generic Skills
Technologies: Models and Ontologies (pp. 107-123).
www.igi-global.com/chapter/taxonomies-problems-generic-
skills/44928?camid=4v1a

Towards Practical ABox Abduction in Large Description Logic Ontologies
Jianfeng Du, Guilin Qi, Yi-Dong Shen and Jeff Z. Pan (2012). International Journal on
Semantic Web and Information Systems (pp. 1-33).
www.igi-global.com/article/towards-practical-abox-abduction-
large/70741?camid=4v1a
Identifying Polarized Wikipedia Articles
Advancing Information Management through Semantic Web Concepts and Ontologies (pp. 272-284).

www.igi-global.com/chapter/identifying-polarized-wikipedia-articles/71860?camid=4v1a