Challenges in Quality Assessment of Arabic DBpedia

Guma Abdulkhader Lakshen
School of Electrical Engineering,
University of Belgrade
Bul. Kralja Aleksandra 73
Belgrade, Serbia
jlackshen65@yahoo.com

Valentina Janev
Mihajlo Pupin Institute
University of Belgrade
Volgina 15
Belgrade, Serbia
Valentina.Janev@institutepupin.com

Sanja Vraneš
Mihajlo Pupin Institute
University of Belgrade
Volgina 15
Belgrade, Serbia
Sanja.Vranes@institutepupin.com

ABSTRACT

The development of Semantic Web technology has fueled the creation of a large amount of Linked Open Data. DBpedia is a good example of an open data repository extracted from the crowd sourced knowledge base, Wikipedia. Because of the way Wikipedia and DBpedia were created, the information available there is more vulnerable to grammatical errors, inconsistency, structures, and, data type problems. The introduction of the Arabic chapter of DBpedia created additional problems due to the nature of the language, when it comes to quality assessment issues. In this paper [1], we focus on identifying challenges in quality assessment of Arabic DBpedia, as well as analysis of existing tools and methodologies used for Linked Data quality assessments.

CCS CONCEPTS

• **Social and professional topics** → **Management of computing and information systems** → System management → Quality assurance
• **Computing methodologies** → **Artificial intelligence** → Knowledge representation and reasoning → Ontology engineering
Computing methodologies → Natural language processing

KEYWORDS

DBpedia, Quality assessment, Tools, Quality dimensions, Application

1 INTRODUCTION

Linked Open Data (LOD) is a growing movement for organizations to make their existing data available in a machine-readable format. The Linked Data approach, based on principles defined back in 2006 by Tim Berners-Lee [1], enables linking of datasets through references to common concepts. HTTP URIs (*Uniform Resource Identifiers*) serves to name the entities and concepts, as well as relations (links) to other related URIs. The Resource Description Framework (RDF) is the model used for representation of the information (entities and concepts), as well as a model that enables exchange and reuse of structured metadata. In the last decade, the Linked Data approach has been adopted by an increasing number of data providers leading to the creation of a global data space that contains many billions of assertions—the Linked Open Data cloud, http://lod-cloud.net/. The cloud has increased from 12 datasets in 2007 to 1,139 in January 2017. One of the central interlinking hubs of the Linked Open Data cloud is the DBpedia [2], a rich multi-lingual knowledge base that represents content from Wikipedia in Linked Data format. Because of its importance for the LOD, the DBpedia is a topic explored in many research studies.

This paper primarily refers to the challenge of quality of DBpedia, and in particular the Arabic Chapter of DBpedia [3]. Al-Feel [3] has defined a mapping methodology and constructed the Arabic Chapter by mapping templates and attributes from Wikipedia to the classes and properties in the DBpedia ontology. However, since the beginning, the author pointed to issues related to quality of data because, for instance, a wrong mapping can cause loss or errors in the knowledge base.

In this paper, we build upon the state-of-the art research on quality issues with DBpedia and, in Section III, define a complete list of problems with Arabic DBpedia. In Section IV, we analyze existing tools used for Linked Data quality assessments in order to propose a new service for quality assessment of Linked Data that will support developers from the Arabic world in quickly fixing errors in DBpedia dump before using it in a Linked Data application (e.g. in the pharmaceutical domain).

2 DBPEDIA

2.1 DBpedia Ontology

One of the objectives of the DBpedia is to structure Wikipedia data in a standard way and reduce redundancy. The DBpedia ontology (in RDF form) is a shallow, cross-domain ontology based on the most commonly used infoboxes within Wikipedia [2]. The latest edition of DBpedia (Dataset 2016-10) encompasses 760 classes, which form a subsumption hierarchy and are described by 1,105 object properties, 1,622 datatype properties,

[1] Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

WIMS '18, June 25–27, 2018, Novi Sad, Serbia

ACM ISBN 978-1-4503-5489-9/18/06…$15.00
https://doi.org/10.1145/3227609.3227675

132 specialized datatype properties, 414 owl:equivalentClass, 220 owl:equivalentProperty mappings to external vocabularies, and over 4.2 million instances [4, 5].

The mapping of multilingual Wikipedia infoboxes to DBpedia ontology, discussed in [6], does not fully fulfill the needs of the mapping of contents in the Arabic language. Therefore, this paper explores the possibilities to assess the quality of DBpedia when leveraged for building Linked Data applications on top of Arabic data resources.

2.2 DBpedia Quality Issues

The DBpedia knowledge base contains information on many different domains. The automatically extracted DBpedia data from Wikipedia, based on infoboxes, has obvious advantages in terms of automatization and ensures wide coverage, but it also poses some quality issues. In 2012, Mendes et al. [7] pointed out issues such as completeness, conciseness, and consistency. In 2014, Kontostas et al. [8] provided several automatic quality tests on Linked Open Data LOD datasets based on patterns modeling various error cases, and detected 63 million errors among 817 million triples. At the same time, Zaveri et al. [9], conducted a user-driven quality evaluation in which was stated that DBpedia has indeed quality problems (around 12% of the evaluated triples had issues), that can be summarized as follows: Incorrect/missing values, incorrect data types, and incorrect links. Based on the survey, the authors [10] developed a comprehensive methodological quality assessment framework based on 18 quality dimensions and 69 metrics. Based on the work of Zaveri et al. [10] and the ISO 25012 data quality model, Radulović et al. [11] developed a Linked Data Quality Model and tested the model with DBpedia with a special focus on accessibility quality characteristics.

Implementing algorithms for detection and correction of errors in DBpedia instance data [12], as well as for detection of incorrect mappings [13] can improve the quality of DBpedia. In [13], Rico et al. proposed a machine learning based approach to building a predictive model that can detect incorrect mappings. An alternative is to leverage the value of the crowd for identifying quality issues with DBpedia [14].

3.2 Towards Data Quality Dimensions

Zavari et all [9, 10] conducted a systematic survey in 2014 on literature related to Linked Data quality and identified a set of data quality dimensions that can be applied to assess the quality of linked data. A Data Quality Dimension or characteristic is an aspect or feature of information and a way to classify information and data quality needs [16]. Dimensions are used to define, measure, and manage the quality of the data and information. Each dimension of data quality consists of a set of attributes. Each attribute characterizes a specific data quality requirement and can be measured by different methods [17, 18, 19]. Zaveri et al. grouped the identified dimensions according to the classification introduced in [20]:

- *Accessibility*: Availability, licensing, interlinking, security, and performance
- *Intrinsic*: Syntactic validity, semantic accuracy, consistency, conciseness, and completeness
- *Contextual*: Relevancy, trustworthiness, understandability, and timeliness
- *Representational*: Representational conciseness, interoperability, interpretability, and versatility

Wang and Strong [24] and other authors conceive 'the data quality as fitness for use.' Hence, assessing the quality of data usually requires a large number of quality measures to be computed rather than one single measure. Based on previous results reported in literature and the end-user requirements (see Section 4.2), the authors selected three data quality dimensions for assessing the quality of Arabic DBpedia (see Table 1).

Table 1: A selected list of data quality dimensions (*Specific for DBpedia, ** Specific to Arabic DBpedia)

Category	Sub-category
Accuracy (Intrinsic)	Triple incorrectly extracted • Special template not properly recognized* • Wrong values in numerical data**
	Data type incorrectly extracted
	Implicit relationship between attributes • One/ Several fact encoded in one/several attributes* • Attribute value computed from another attribute value**
Consistency (Intrinsic)	• Inconsistency in representation of number values**
Relevancy (Contextual)	Irrelevant information extracted • Extraction of attributes containing layout information** • Redundant attribute values • Image related information* • Other irrelevant information

3 ANALYSIS OF ARABIC CHAPTER OF DBPEDIA

3.1 Motivation

DBpedia is heavily interlinked with other datasets and plays a central role in the Linked Open Data cloud. The authors' specific motivation for testing the quality of the Arabic Chapter of the DBpedia refers to an application in the drug domain titled "Arabic Linked Drug Data Application (ALDDA)." Indeed, the pharmaceutical industry was among the first to express an interest in validating the Linked Data approach for publishing and consolidating drug data. Unfortunately, the existing linked drug datasets does not include Arabic datasets to a large extent. (see for instance LinkedDrugs [21]). Hence, as a part of ALDDA, the end-user needs a quality assessment (QA) component.

Based on the initial analysis of the literature, the authors defined the metrics for accuracy, consistency, and relevance of the data, as most relevant dimensions for their application, see an extract in Table 1. The ALDDA-QA is a Java web application based on Vaadin, https://vaadin.com/framework, and Sesame, https://sourceforge.net/projects/sesame/, while ESTA-LD tool [22] is used for visualization of the statistics (see Figure 1).

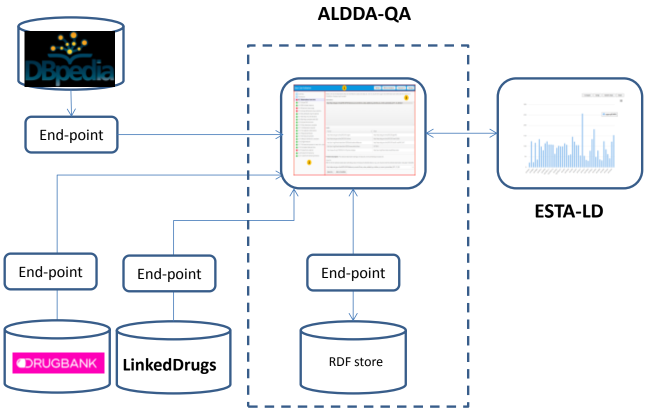

Figure 1: Quality Assessment Framework - Simplified illustration.

3.2 About Arabic language

Arabic is the Classical language of the 6th century and its modern descendants. The literary language, called Modern Standard Arabic, is the only official form of Arabic that is used in most written documents, as well as in formal spoken occasions, such as lectures and news broadcasts [15]. Almost 422 million speakers in 25 countries speak standard the Arabic language, making it one of the six most-spoken languages in the world. It is one of six official languages of the United Nations.

3.3 Identified Problems in the Arabic Chapter

When it comes to quality assessment of the DBpedia Arabic Chapter, there are problems specific to the Arabic language that result in:

1. Presentation of characters as symbols via web browsers due to errors during the extraction process.
2. Wrong values in numerical data, due to the use of Hindu numerals in some Arabic sources.
3. Occurrence of different names for the same attribute, for instance, the birth date attribute appears in various infoboxes by different names: one time as "(*eng*. birth date) تاريخ الميلاد" another time as "(*eng*. delivery date)تاريخ الولادة", third time as "(*eng*. birth) الميلاد".
4. Inconsistency of names between the infobox and its template; for instance, there is a template called "(*eng*. city) مدينة" while the infobox name is called "(*eng*. city information) مدينة معلومات ."
5. Geo-names templates formatting problems when placed in the infobox.
6. Errors in <owl:sameAs> relations and problems in identifying the <owl:sameAs> relations due to heterogeneity in different data sources.

However, some of the problems present in other DBpedia chapters are also identified in the Arabic Chapter. Specifically, the authors would like to point to:

7. Wrong Wikipedia Infobox information; for example, the height of minaret of the grand mosque in Mecca (the most valuable mosque for all Muslims) is given as 1.89 m, where the correct height is 89 m.
8. Mapping problems from Wikipedia, such as unavailability of infoboxes for many Arabic articles; for example, "Man-made river in Libya النهر الصناعي" which is considered as the biggest water pipeline project in the world, or not containing all the desired information.
9. Object values incompletely or incorrectly extracted.
10. Data type incorrectly extracted.
11. Some templates may be more abstract, thus cannot map to a specific class.
12. Some templates not used or missing inside the articles.

4 ALYSIS OF EXISTING QUALITY ASSESSMENT TOOLS

There were several attempts in the past to design and implement a generic tool for Linked Data quality assessment [7, 8, 23]. One of the first open-source frameworks for flexibly expressing quality assessment methods, as well as fusion methods, was Sieve [7], http://sieve.wbsg.de, released in 2012. As part of the Linked Data Integration Framework (LDIF, http://sieve.wbsg.de/), Sieve aims at supporting users to consume data from the LOD cloud. Taking into consideration that DBpedia is a core element in the LOD cloud, in 2014 the RDFUnit Testing Suite [8], https://github.com/AKSW/RDFUnit, was published as a tool that enabled users to run automatically-generated (based on a schema) and manually-generated test cases against an endpoint, e.g. DBpedia SPARQL endpoint. Realizing that there are a large variety of dimensions and measures of data quality, Luzzu, https://github.com/EIS-Bonn/Luzzu, was developed at the same time to allow knowledgeable engineers without Java expertise to create quality metrics in a declarative manner [23]. LOD Laundromat, http://lodlaundromat.org, was designed with the goal of helping crawling the LOD cloud, converting all its contents in a standards-compliant way (gzipped N-Triples), as well as removing all data stains, such as syntax errors, duplicates, and blank nodes. TripleCheckMate, https://github.com/AKSW/TripleCheckMate, is a tool for crowdsourcing the assessment of Linked Open Data. It was developed for evaluating the correctness of DBpedia. TripleCheckMate provides an easy-to-use user interface with multiple resource assignment methods and a ready-to-use error classification scheme. The quality assessment methods implemented in these tools can be grouped into automatic, semi-automatic, manual, or crowd-sourced approaches. Initial results of analysis and comparison of the selected tools is provided in Table 2.

Table 2: Comparison of open-source quality assessment tools according to several attributes

Tool	Extensibility	Last Update	Colla-boration	Cleaning Support
RDFUnit	SPARQL	03/2018	✕	✕
Luzzu	JAVA, LQML	07/2017	✕	✕
TripleCheckMate	✕	03/017	✔	✕
Laundromat	SPARQL	05/2018	✔	✔
Sieve	XML	2014	✕	✔

To the best of the authors' knowledge, these tools have not been tested with the Arabic DBpedia. ALDDA-QA might adopt some of the functionalities of the analyzed tools, however will concentrate on the needs of users from Arabic countries.

5 CONCLUSION AND FUTURE WORK

In summary, the authors have presented in this paper the problems regarding the Arabic DBpedia and proposed a solution for design of a quality assessment tool for Arabic linked datasets. The quality assessment method is driven by the three dimensions that have been identified as relevant to the Arabic DBpedia, or Linked Data in general. A comparison matrix of the existing quality assessment tools that the authors have elaborated on show different attributes/functionalities of the available tools. The tests conducted in the research showed that the Arabic DBpedia dataset lacks continuous improvement, and it needs effective management in order to increase Arabic extracted triples.

The future work will include implement of a stable and open-source version of the ALDDA-QA quality assessment framework that will allow the end-user to fully explore and, if possible, to repair the errors observed in the Arabic DBpedia. Thus, the end-user will benefit from the interlinking of private with public data and enrichment of local data with information from the Web.

ACKNOWLEDGMENTS

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

REFERENCES

[1] T. Berners-Lee. 2006. Design issues: Linked data. Retrieved August 10, 2017, from http://www.w3.org/DesignIssues/LinkedData.html

[2] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and I. Zachary. 2007. DBpedia: A Nucleus for a Web of Open Data. In *Proceeding of the 6th International Semantic Web Conference* (The Semantic Web). Lecture Notes in Computer Science, Vol. 4825. Springer-Verlag, London, 722-735. DOI: http://dx.doi.org/10.1007/978-3-540-76298-0_52

[3] H. Al-Feel. 2015. The roadmap for the Arabic chapter of DBpedia. In *Proceeding of the Mathematical and Computational Methods in Electrical Engineering*. Sleima, Malta:WSEAS Press, 115-125.

[4] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, and C. Bizer. 2015. DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web – Interoperability, Usability, Applicability,* Vol. 6, No. 2 (2015), 167-195. DOI: http://dx.doi.org/10.3233/SW-140134

[5] DBpedia - Towards a Public Data Infrastructure for a Large, Multilingual, Semantic Knowledge Graph. http://wiki.dbpedia.org/datasets/dbpedia-version-2016-10 last visited 28.09.2017

[6] C. Bratsas, L. Ioannidis, D. Kontokostas, S. Auer, C. Bizer, S. Hellmann, and I. Antoniou. 2011. DBpedia internationalization-a graphical tool for I18n infobox-to-ontology mappings. In *Proceeding of the International Semantic Web Conference* (ISWC2011 Demo), Bonn, Germany.

[7] P. N. Mendes, H. Mühleisen, and C. Bizer. 2012. Sieve: Linked Data Quality Assessment and Fusion. In *Proceedings of the 2012 Joint EDBT/ICDT Workshops*. ACM, New York, NY, USA, 2012, pp. 116–123. DOI: http://dx.doi.org/10.1145/2320765.2320803

[8] D. Kontokostas, P. Westphal, S. Auer, S. Hellmann, J. Lehmann, and R. Cornelissen. 2014. Test driven Evaluation of Linked Data Quality. In *Proceeding of the 23rd International Conference on World Wide Web*. New York, NY, USA, 2014, pp. 747–758. DOI: http://dx.doi.org/10.1145/2566486.2568002

[9] A. Zaveri, D. Kontokostas, M. A. Sherif, L. Bühmann, M. Morsey, S. Auer, and J. Lehmann. 2013. User-driven Quality Evaluation of DBpedia. In *Proceedings of the 9th International Conference on Semantic Systems*. New York, NY, USA, 2013, pp. 97–104.

[10] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, and S. Auer. 2016. Quality assessment for linked data: A survey. *Semantic Web – Interoperability, Usability, Applicability*, Vol. 7, No. 1 (2016), 63-93. DOI: http://dx.doi.org/10.3233/SW-150175

[11] F. Radulovic, N. Mihindukulasooriya, R. García-Castro, and A. Gómez-Pérez. 2018. A Comprehensive Quality Model for Linked Data. *Semantic Web – Interoperability, Usability, Applicability*, Vol. 9, No. 1 (2018), (2018), 3-24, Special issue on Quality Management of Semantic Web Assets (Data, Services and Systems). DOI: https://doi.org/10.3233/SW-170267

[12] W. Dominik and H. Paulheim. 2014. Detecting incorrect numerical data in DBpedia. In *Proceeding of the Extended Semantic Web Conference* (ESWC 2012): The Semantic Web: Research and Applications. Lecture Notes in Computer Science, Vol. 8465. Springer, Berlin, Heidelberg, 504-518. DOI: https://doi.org/10.1007/978-3-319-07443-6_34

[13] M. Rico et al. 2018. Predicting Incorrect Mappings: A Data-Driven Approach Applied to DBpedia. In. *Proceedings of the SAC 2018*: *Symposium on Applied Computing*, Pau, France,. April 9–13, 2018 (SAC 2018)

[14] M. Acosta, A. Zaveri, E. Simperl, D. Kontokostas, F. Flöck, and J. Lehmann. 2018. Detecting Linked Data Quality Issues via Crowdsourcing: A DBpedia Study. *Semantic Web – Interoperability, Usability, Applicability*, Vol. 9, No. 3 (2018), 303-335. DOI: https://doi.org/10.3233/SW-160239

[15] M. C. Bateson. 2003. *Arabic Language Handbook*. Georgetown University Press, ISBN 0-87840-386-8.

[16] A.F. Karr, A.P. Sanil, and D.L. Banks. 2006. Data quality: A statistical perspective. *Stat. Methodol.* 2006;3:137–173. DOI: https://doi.org/10.1016/j.stamet.2005.08.005.

[17] T.C. Redman. 2005. Measuring Data Accuracy A Framework and Review. In: Wang R.Y., Pierce E.M., Madnick S.E. (editors) *Information Quality*. M.E. Sharpe, Inc.; Armonk, NY, USA: 2005. pp. 21–36.

[18] C. Batini, C. Cappiello , C. Francalanci, and A. Maurino.2009. Methodologies for data quality assessment and improvement. *ACM Comput.Surv*. 41 (2009), 1–52. DOI: https://doi.org/10.1145/1541880.1541883

[19] L. Pipino, R.Y. Wang, D. Kopcso, and W. Rybolt, 2005. Developing Measurement Scales for Data-quality Dimensions. In: Wang R.Y., Pierce E.M., Madnick S.E., (editors) *Information Quality*. M.E. Sharpe, Inc.; Armonk, NY, USA: 2005. pp. 37–51.

[20] Y. R. Wang and D. M. Strong. Beyond accuracy: what data quality means to data consumers. Journal of Management Information Systems, 12(4):5–33, 1996.

[21] M. Jovanovik and D. Trajanov. 2017. Consolidating drug data on a global scale using linked data. *Journal of Biomedical Semantics*, 8(3). DOI: https://doi.org/10.1186/s13326-016-0111-z

[22] V. Mijović, V. Janev, D. Paunović, and S. Vraneš. 2016. Exploratory Spatio-Temporal Analysis of Linked Statistical Data. *Journal of Web Semantics, Web Semantics: Science, Services and Agents on the World Wide Web* 41C (2016), 1-8. DOI: https://doi.org/10.1016/j.websem.2016.10.002.

[23] J. Debattista, S. Auer and C. Lange. 2016. Luzzu -- A Framework for Linked Data Quality Assessment. In *Proceeding of the 2016 IEEE Tenth International Conference on Semantic Computing (ICSC)*, Laguna Hills, CA, 2016. IEEE, 124-131. DOI: https://doi.org/10.1109/ICSC.2016.48