ICIST2019
Kopaonik, Serbia
Mar 10-13, 2019

9th International Conference on
Information Society and Technology

Quality Issues of Open Big Data Ecosystems

Toward Solution Development

Guma Lakshen: *School of Electrical Engineering*
Valentina Janev, Sanja Vraneš: *Mihajlo Pupin Institute University of Belgrade*

Overview

Motivation: Study the **Quality** of Open Data and the **Benefits** for Industry

Approach:

- Surveys
- Selection of Data Quality Dimensions
- Testing with Arabic Open Data

Results:

- Survey on tools / methodologies
- Design of Quality Assessment Service as part of ALDDA

Main Contributions

Motivation: Study the Quality of Open Data and the Benefits for Industry

Big data is a **Structured / Semi-structured / Unstructured** Data Sets

With massive data volumes that cannot be easily captured, stored, manipulated, analyzed managed and presented by traditional Hardware, Software, and Database management technologies.

Data quality dimension
Used to describe a feature of data that can be measured or assessed against defined standards in order to determine the quality of data.

Additional factors include:

- ❑ Usability
- ❑ Flexibility
- ❑ Confidentiality
- ❑ Value Timing issues of the data.

Motivation: Linked Data Challenges

Challenges in Industry …….

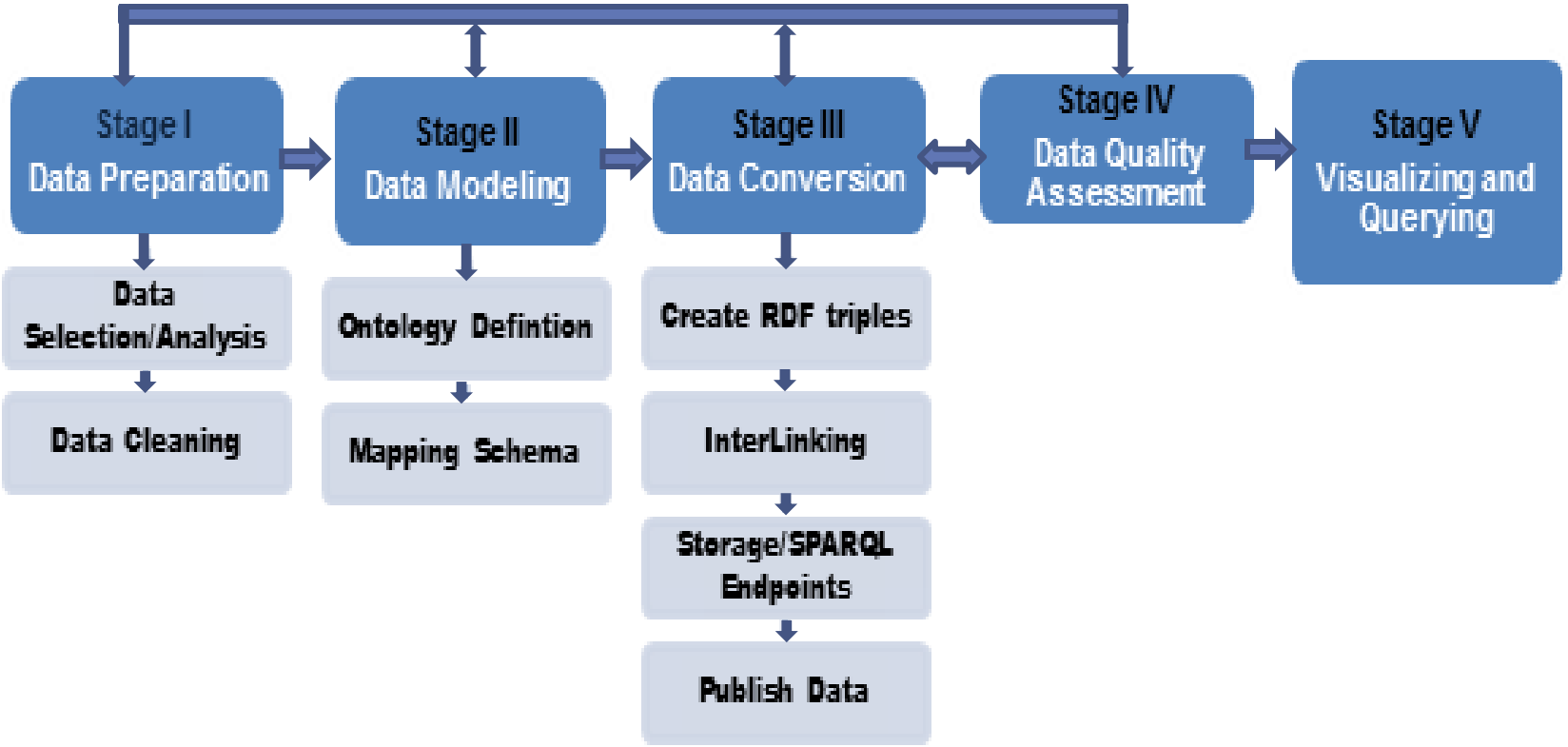
- Heterogeneity and incompleteness
- Diversity of data sources
- Huge data volume
- Short data timeline
- Non-existing and approved data
- quality standards
- Lack of structure
- Error-handling
- Privacy
- Timeliness
- Provenance
- Visualization

Additional Problems of data quality with Arabic datasets include…….

- Lack of validation routines
- Data valid, but not correct
- Mismatched syntax, formats, and structures
- Unexpected changes in source system
- Spider-web of interfaces
- Lack of referential integrity checks
- Poor system design
- Data conversion errors

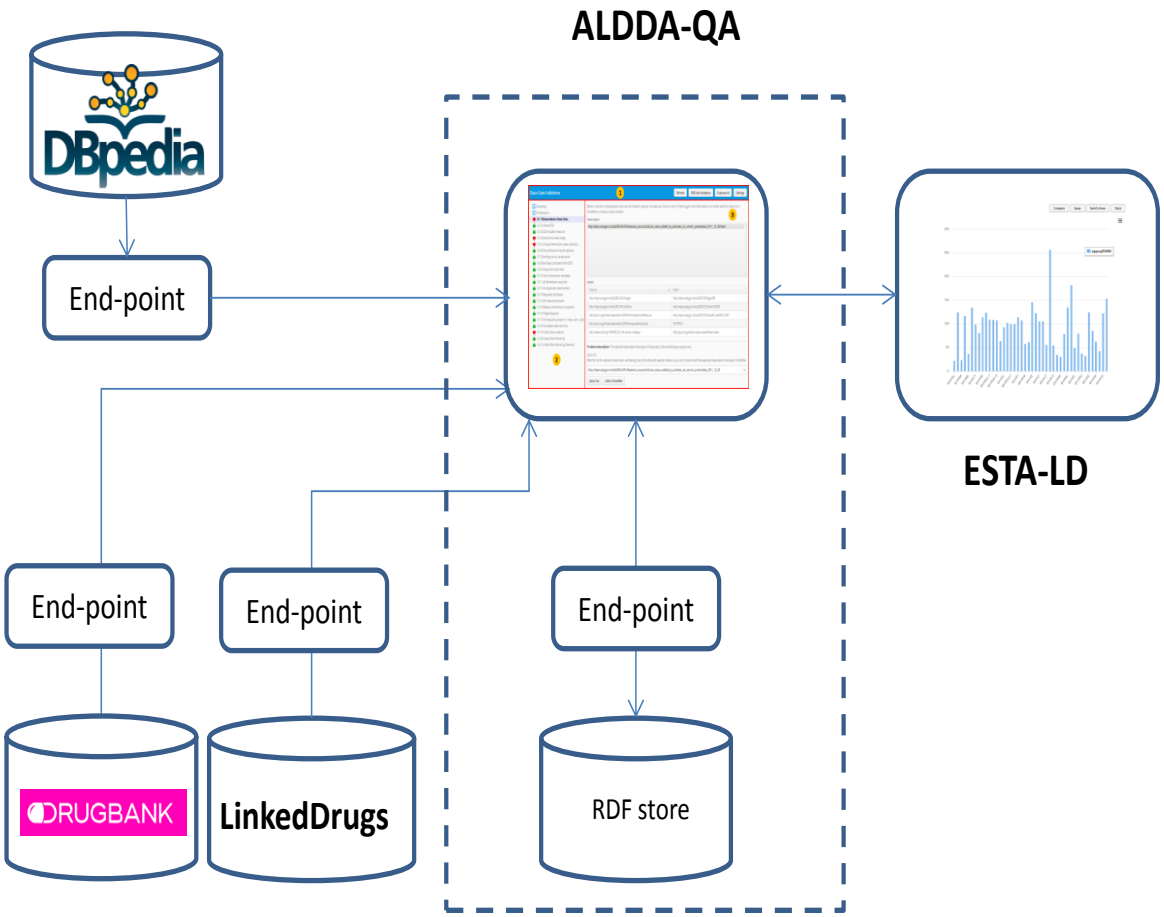
See WIMS 2018 paper
Challenges in Quality Assessment of Arabic DBpedia

Design of Quality Assessment Service



PIQA-LD (Pharmaceutical Data Quality Assessment-Linked Data) Framework

ALDDA– Quality Assessment



Results: Comparison between Linked Data Methodologies

Authors / Organization	Title / Steps	
W3C Government Linked Data Working Group (2014) [13]	*Best Practices for Publishing Linked Data*:	
	(1) Prepare stakeholders, (2) Select a dataset, (3) Model the data, (4) Specify an appropriate license, (5) Good URIs for linked data, (6) Use standard vocabularies,	*Initialization*
	(7) Convert data, (8) Provide machine access to data,	*Innovation*
	(9) Announce new data sets, (10) Recognize the social contract	*Validation &Maintenance*
Hyland et al. (2011) [8]	*A Cookbook for Publishing Linked Government Data on the Web*:	
	(1) Identify, (2) Model, (3) Name, (4) Describe,	*Initialization*
	(5) Convert, (6) Publish,	*Innovation*
	(7) Maintain	*Validation &Maintenance*
Hausenblas et al. (2016) [7]	*Linked Data Life Cycles*:	
	(1) Data awareness, (2) Modeling,	*Initialization*
	(3) Publishing, (4) Discovery, (5) Integration,	*Innovation*
	(6) Use-cases	*Validation &Maintenance*
Villazón-Terrazas et al. (2011) [14]	*Guidelines for Publishing Government Linked Data*:	
	(1) Specify, (2) Model,	*Initialization*
	(3) Generate, (4) Publish,	*Innovation*
	(5) Exploit	*Validation &Maintenance*
Auer, et all. (2012) [1]	*Managing the Life-Cycle of Linked Data with the LOD2 Stack*:	
	(1) Extraction,	*Initialization*
	(2) Storage, (3) Authoring, (4) Interlinking, (5) Classification,	*Innovation*
	(6) Quality, (7) Evolution/Repair, (8) Search/ Browsing/ Exploration	*Validation &Maintenance*
Jovanovik and Trajanov (2017) [9]	*Methodological guidelines for consolidating drug data*:	
	(1) Domain and Data Knowledge, (2) Data Modeling and Alignment,	*Initialization*
	(3) Transformation into 5-star Linked Data, (4) Publishing the Linked Data Dataset on the Web,	*Innovation*
	(5) Use-cases, Applications and Services	*Validation &Maintenance*

Selection of Data Quality Dimensions

- **Zaveri** et al. (Semantic Web Journal, 2012-2016) identified **18** quality dimensions and **69** metrics

A Data Quality Dimension *or characteristic is an aspect or* ***feature*** *of* ***information*** *and a way to* ***classify*** *information and* ***data quality needs****. 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.*

- ***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

Results of Analysis

Selected data quality dimensions used for assessing the quality of Arabic datasets

Dimension / Metrics Definition	Category	Sub-category
Accuracy (Intrinsic): *I* Is the degree of closeness between a value x and a value x′, considered as the correct representation of the reality that x aims to represent. If x is the number of the correct values, and x′ is the number of total values, then, Accuracy = x/ x′	Triple incorrectly extracted	• Object value is incorrectly/ incompletely extracted * • Special template not properly recognized • Wrong values in numerical data * *
	Data type problems	• 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): Data are consistent if it meets a set of constraints. If x is the number of consistent values, and x′ is the number of total values. Then, consistency= x/ x′	Representation of number values	• Inconsistency in representation of number values* *
Relevancy (Contextual): Is the data useful for the specified task? What kind of information is provided by a source? Does this information match the users' or system's requirements?	Irrelevant information extracted	• Extraction of attributes containing layout information * * • Redundant attribute values • Image related information* • Other irrelevant information

* Specific for Dbpedia, * * Specific for Arabic DBpedia

Results of Analysis

Known problems	Specific to Arabic DBpedia
1. 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. 2. 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. 3. Object values incompletely or incorrectly extracted. 4. Data type incorrectly extracted. 5. Some templates may be more abstract, thus cannot map to a specific class. 6. Some templates not used or missing inside the articles.	7. Presentation of characters as symbols via web browsers due to errors during the extraction process. 8. Wrong values in numerical data, due to the use of Hindu numerals in some Arabic sources. 9. 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". 10. 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 ." 11. Geo-names templates formatting problems when placed in the infobox. 12. Errors in <owl:sameAs> relations and problems in identifying the <owl:sameAs> relations due to heterogeneity in different data sources.

Results of Analysis

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

TOOL / ATTRIBUTE	License	Extensibility	Scalability	Last Update	Collaboration	Cleaning Support
RDFUnit Testing Suite https://github.com/AKSW/RDFUnit	Apache	SPARQL	✔	March 2018	✘	✘
Luzzu https://github.com/EIS-Bonn/Luzzu	-	JAVA, LQML	✔	July 2017	✘	✘
TripleCheckMate https://github.com/AKSW/TripleCheckMate	Apache	✘	Crowd Sourcing	March 2017	✔	✘
LOD Laundromat https://github.com/LOD-Laundromat	-	SPARQL	✔	May 2018	✔	✔
Sieve, http://sieve.wbsg.de	Apache	XML	✔	2014	✘	✔

ALDDA– Quality Assessment

Selection of Tools

- **Data Preparation / Modeling:** TopBraid Composer (TopQuadrant)
- **Data Conversion / Interlinking: TBD**
- **Data Quality Assessment:**
  - Vaadin, https://vaadin.com/framework, a Java framework for building web applications
  - Sesame, https://sourceforge.net/projects/sesame/, an open-source framework for querying and analyzing RDF data
  - Virtuoso, https://github.com/openlink/virtuoso-opensource
- **Visualization of statistics:** ESTA-LD (PUPIN)

Results

Issues identified

- The creation of the Arabic Chapter opened the door for development of new applications, however users from the Arabic countries are not aware yet!!! of the benefits and potentials of the Linked Data approach.
- The Arabic DBpedia dataset lacks continuous improvement, and it needs effective management in order to increase Arabic extracted triples.
- Solutions for fully automating the mapping process should be found that integrates quality assessment methods as well

Contributions

- Towards a Methodology for integrating the quality assessment in Linked Data Apps
- Integrating the Arabic datasets and design of Linked Data application for the pharma industry

Thank you for your attention