Are Linked Datasets fit for Open-domain Question Answering? A Quality Assessment

Harsh Thakkar
Enterprise Information Systems
University of Bonn
Bonn, Germany
hthakkar@uni-bonn.de

Kemele M. Endris
Enterprise Information Systems
University of Bonn
Bonn, Germany
endris@iai.uni-bonn.de

Jose M. Gimenez-Garcia
Laboratoire Hubert Curien
Jean-Monnet University
St. Étienne, France
jose.gimenez.garcia@univ-st-etienne.fr

Jeremy Debattista
Enterprise Information Systems
University of Bonn / Fraunhofer IAIS
Bonn, Germany
debattis@cs.uni-bonn.de

Christoph Lange
Enterprise Information Systems
University of Bonn / Fraunhofer IAIS
Bonn, Germany
langec@cs.uni-bonn.de

Sören Auer
Enterprise Information Systems
University of Bonn / Fraunhofer IAIS
Bonn, Germany
auer@cs.uni-bonn.de

ABSTRACT

The current decade is a witness to an enormous explosion of data being published on the Web as Linked Data to maximise its reusability. Answering questions that users speak or write in natural language is an increasingly popular application scenario for Web Data, especially when the domain of the questions is not limited to a domain where dedicated curated datasets exist, like in medicine. The increasing use of Web Data in this and other settings has highlighted the importance of assessing its quality. While quite some work has been done with regard to assessing the quality of Linked Data, only few efforts have been dedicated to quality assessment of linked data from the question answering (QA) perspective. From the linked data quality metrics that have so far been well documented in the literature, we have identified those that are most relevant for QA. We apply these quality metrics, implemented in the Luzzu framework, to subsets of two datasets of crucial importance to open domain QA – DBpedia and Wikidata – and thus present the first assessment of the quality of these datasets for QA. From these datasets, we assess slices covering the specific domains of restaurants, politicians, films and soccer players. The results of our experiments suggest that for most of these domains, the quality of Wikidata with regard to the majority of relevant metrics is higher than that of DBpedia.

Keywords

Linked Open Data, Data Quality, Quality assessment, Question Answering, DBpedia, Wikidata

MINES MINES, FRANCE

DOI: 10.1145/1235

1. INTRODUCTION

Recent advancements in the fields of Web of Data and Data Science have led to increasing amounts of structured data published according to standards such as RDF(a), Linked Data, Schema.org[1], and to a wide range of tools to produce, manage and *consume* such data. Answering questions spoken or written in natural language (cf. [37] for an overview) is an increasingly popular setting in which apps for end users consume Web Data to satisfy sophisticated information needs. Users expect answers to be correct, or at least relevant, even when they have not phrased their question precisely or when their speech has been recorded with a low fidelity. Thus, to support, for example, query disambiguation, answer retrieval, results ranking, Web Data consumed in such settings therefore has to meet a certain level of quality. Quality can generally be defined as "fitness for use", but there are a lot of concrete factors that influence a dataset's fitness for use in question answering[2] settings and in specific application domains. Recently, a number of research activities have been concerned with automating the assessment of linked data quality by benchmarks such as Luzzu [7], RDFUnit [25], and others. In this paper, we compile a list of metrics that are applicable for assessing the quality of Linked Open Data (LOD) for question answering. We support this list by determining the quality of various subsets of Wikidata[3] and DBPedia[4], of which both are widely used in linked data based question answering, such as SINA [36], BioASQ [38], QALD tasks[39], and others.

1.1 User Scenario

[1]The amount not only of structured, but also of semi-structured and unstructured data available online is also steadily increasing; however, for the purpose of our work we assume that such data has first been translated to the RDF data model using standard tools, e.g. from the Linked Data Stack [2].

[2]In this section, we do not abbreviate "question answering" as "QA" to avoid confusion with "quality assessment".

[3]http://wikidata.org

[4]http://dbpedia.org

To put the reader into the context of our work, we present a scenario and discuss it from two different perspectives: the data consumer's (i.e. application developer's) perspective and the user's perspective.

Data consumer's perspective: Data Caterer GmbH, a data science company, has been providing search and discovery services in various domains, including films and soccer. To cover these domains, they have access to a wide range of cross-domain datasets. To serve users' needs, they specialise on mobile question answering applications on top of a generic, open domain core engine. Data Caterer would now like to enter the market of restaurant recommendations and is facing fierce competition from other providers such as Zomato[5]. To provide the best answers to the users' questions, it is crucial to use the right dataset, i.e. one that is rich in domain-specific information and correct. This task poses two challenges:

1. **How to perform an initial data quality assessment?** – Different datasets from different sources, which may have been created using different techniques, usually vary in quality w.r.t. level of detail, redundancy and correctness. Assessing the quality of the datasets in an initialisation step, before actually answering questions, supports the system in delivering faster output, which, moreover, is precise.

2. **Which subsets of a given dataset are specifically suitable for answering questions related to given topic?** – Cross-domain datasets, such as datasets obtained from Wikipedia, cover many domains, but how well do they cover the specific domain of the question, i.e., here, the restaurant domain? We need to assess the quality not of the dataset as a whole, but of certain subsets (slices) of it.

User's perspective: Walter enjoys savouring a variety of cuisines. He came to Berlin for Christmas holidays and wants to find the best Fränkische Bratwurst in town. He uses the Data Caterer mobile app, types in a question and gets the precise location of the Frankenstüberl restaurant as a quick response. His information needs are satisfied. The app also offers him to rate the answer, i.e. whether he is satisfied or not. Data Caterer uses this rating to score the subset of the source dataset from which Walter's query was answered. Based on previously computed quality results, Data Caterer will be able to serve subsequent users in an even better way, as it will always use the (sub)set that provides the best data in the domain of a question.

1.2 Objective

Our objective is to identify precisely defined and practically computable metrics for the quality (and thus fitness for use) of linked data in question answering settings. This requires answering the following research questions:

Obj 1 – What dimensions of data quality are relevant to open domain question answering systems?

Obj 2 – By what data quality metrics can the quality of data for open domain question answering systems be assessed precisely?

Obj 3 – How well do popular cross-domain datasets serve the needs of question answering applications in certain specific domains?

Obj 4 – How can data quality from the perspective of open domain question answering be practically assessed using available tools and techniques?

The remainder of this paper is structured as follows: In section 2, we present the state of the art for quality assessment of linked datasets and question answering in general.

2. STATE OF THE ART

In this section, we summarise the state of the art of data-driven question answering and data quality assessment, and highlight the gaps that our research aims to address.

The goal of a question answering system is to satisfy the user's information need (represented as a natural language query or keywords) by returning the best answer (a ranked list of documents, URIs, specific excerpts from text, etc.). *Closed domain* question answering systems use context/application specific datasets, ontologies and thesauri (for instance, MeSH[6]). The coverage or scope of information in the dataset is a specific application domain, e.g., medical, financial, geographical. Such datasets are typically collected, curated and validated by domain experts such as, in the medical domain, doctors, clinicians, or physicians, to ensure correctness of the results.

In contrast, *open domain* question answering relies on a wide spectrum of datasets, ontologies and thesaurus to answer any question regardless of its category. Using multiple data sources poses greater challenges to the evaluation and the performance of the question answering system, mainly because of differences in the structure (structured, semi-structured or unstructured) and quality of the data. Furthermore, [39] and [37] report a number of challenges, related to data quality issues, affecting the overall system performance of QA systems and advocate the need to formally consider these factors for question answering systems.

A number of studies have identified, defined and classified data quality dimensions and metrics [3, 4, 29, 43]. Zaveri et al. present an exhaustive state-of-the-art survey on quality dimensions and metrics relevant for assessing *linked* data [44]. However, **only a handful of these dimensions and metrics are relevant for answering questions** (tackling Obj 1, Obj 2 – cf. Section 1.2) over linked datasets on the Web, as we will explain in Section 3.

The question answering community has invested substantial effort into evaluating the performance of their systems, targeting both open and closed domain question answering. Regular challenges include the NTCIR Question Answering Challenge (QAC) [14], TREC [41], CLEF [21], BioASQ [38], and QALD series [27]. The overall performance of a questions answering system is typically evaluated by measures such as precision, recall, F-score, and also, in certain tasks, by the global F-score w.r.t. the total number of questions, as in QALD-5 [40]. However, **these metrics for evaluating question answering systems do not provide any information on the quality of data being used to answer the questions** (tackling Obj 3, Obj 4 – cf. Section 1.2).

The amount of Data that is published on the Web as

[5]https://www.zomato.com/

[6]Medial Subject Headings: http://www.ncbi.nlm.nih.gov/mesh

Linked Open Data (LOD) for maximum reusability is growing rapidly[7]. Hence, the need for developing automated tools for LOD quality assessment has been addressed by several works in the present decade [6, 12, 28, 18]. We narrow the focus of these general-purpose tools down to the large-scale automated assessment of quality metrics relevant to *question answering*.

3. QUALITY DIMENSIONS AND METRICS RELEVANT TO OPEN DOMAIN QUESTION ANSWERING

3.1 General Terminology

Before presenting those quality dimensions and metrics that we have identified as relevant to open domain question answering, we introduce the basic terminology used in the remainder of this section.

- A quality **dimension** is a characteristic of a dataset. The importance of a dimension depends on the context in which a dataset is intended to be used. For instance, availability of dataset is crucial for a question answering system that depends on a SPARQL endpoint for retrieving answers but not for a question answering system that consumes offline data indexed locally.
- Bizer and Cyganiak define a **metric** as a "procedure for measuring a[n ...] quality dimension" [4]. Metrics "rely on quality indicators and calculate an assessment score from these indicators using a scoring function" [4]. A metric can have an exact value, or the value can be estimated when sufficient data for an exact computation is not available, or approximated when exact computation would take too much time.

3.2 Overview

We take inspiration from the comprehensive study conducted by Zaveri et al. [44], who reviewed 30 different works on linked data quality and, from these, identified 18 different dimensions and 69 different metrics. We reviewed all of these dimensions and metrics and identified those that are most relevant to open domain question answering systems by analysing how, and to what end, such systems use data. We identified the following relevant dimensions: *Availability*, *Completeness*, *Timeliness*, *Interlinking*, *Data diversity*, *Semantic accuracy*, *Consistency*, *Trust and Provenance*, *Conciseness*, *Coverage*, *Licensing*. In our current work, we focused on those dimensions that had already been implemented for the Luzzu linked data quality assessment framework (cf. Section 4.3): Availability, Interlinking, Data diversity, Consistency, and Trust and Provenance. For any of these dimensions, the following subsections discuss its relevance to open domain question answering and summarises the relevant metrics.

3.3 Availability Dimension

Zaveri et al. define availability of a dataset as *"the extent to which data (or some portion of it) is present, obtainable and ready for use"* [44]. Flemming refers to availability as the proper functioning of all access methods [12]. We adopt a hybrid definition of both of these:

[7]LOD Cloud: http://lod-cloud.net/

Definition 1 (Availability)**.** *Availability is the degree of readiness of a dataset, ontology or thesaurus for consumption.*

Availability can be established, e.g., by providing the dataset as an RDF dump, exposing it via a SPARQL endpoint, or via some other API, package or library – i.e. in whatever form a question answering system may prefer to consume. Some of these forms may be provided by third parties. We use the following metrics to measure availability:

Metric A.1 (Dereferenceability)**.** *This metric measures the number of valid 303 or hash URIs among all resource URIs in a dataset, following the LOD best practices [35]. As typical datasets for question answering are large, an approximate metric using statistical sampling technique [8] was used.*

Metric A.2 (Dereferenceability of Forward Links)**.** *This metric assesses the extent to which a resource includes all triples from the dataset that have the resource's URI as the subject [23]. To do so, it computes the ratio between the number of triples that are "forward-links" a.k.a. "out-links" (i.e. that have the resource's URI as their subject) and the total number of triples in the RDF graph served as a description of each resource.*

Metric A.3 (No Misreported Content Types)**.** *This metrics checks whether RDF data is served with the right content type, e.g., whether RDF/XML content is served as application/rdf+xml (cf. [44]). This is done by checking the content type returned when dereferencing a URI and checking whether the content can be processed by a parser for the expected format.*

Metric A.4 (RDF Availability)**.** *This metric checks whether a syntactically valid RDF representation of a dataset can be downloaded (cf. [44]).*

Metric A.5 (Endpoint Availability)**.** *This metric checks whether a SPARQL endpoint for a dataset is available, i.e. queryable.*

Relevance: In the context of open domain question answering systems, the relevance of the availability dimension is system dependent. The dimension is crucial, for example, for systems that leverage a remote SPARQL endpoint. The output of such systems relies on the availability of the data in a specific form to be able to query the dataset and return the result. If the endpoint is not responsive or responds erroneously, then the system returns no answer and thus becomes unable to serve the user's request. However, systems that maintain offline copies of datasets and locally computed indexes are less prone to the availability of the original datasets.

More generally, the availability dimension is crucial for any system that relies on a remote third party package, tool or library that gives access to a specific data source.

3.4 Interlinking Dimension

Interlinking refers to connections between two terms with different names (URIs) that have the same meaning. For instance, "heart attack" and "myocardial infarction" are terms from different subdomains of medicine but have the same meaning. Such terms may occur in the same dataset, or

across different datasets. New datasets are typically interlinked with "reference" datasets to aid customers in finding further relevant information. Large, widely known datasets such as DBpedia, but also high-quality curated domain-specific datasets typically serve as such references. In LOD, interlinking is typically implemented as RDF triples that connect a subject and an object by the *owl:sameAs* property. The following definition is paraphrased from [44].

Definition 2 (Interlinking)**.** *Interlinking refers to the degree to which entities representing the same concept are associated to each other, within one or more datasets.*

We consider two metrics to measure the degree of interlinking in a dataset.

Metric I.6 (Interlink Detection)**.** *Guéret et al. reuse five network measures to assess a dataset's degree of interlinking [18]:*

The same-as *measure detects open owl:sameAs chains. owl:sameAs resources should be symmetric, in a sense that if* `X owl:sameAs Y` *then* `Y owl:sameAs X`*. The latter is called a closed* `owl:sameAs` *chain. The* descriptive richness *measure assesses the amount of new properties added to a resources through an owl:sameAs relation. The* degree *measure checks the number of in-links and out-links of a resource. A high degree measure means that agents can find information easily by means of traversal. The* centrality *measure checks the dependency of a resource in a dataset. Finally, the* clustering coefficient *measure aims at identifying how well resources are connected, by measuring the density of the resource neighbourhood. A network has a high clustering cohesion when a node has a large number of neighbouring nodes, all of which are connected to each other. This means that links may end up being meaningless [18]. We have described an approach to approximating the* clustering coefficient *measure in [8].*

Metric I.7 (External Link Data Providers)**.** *The* external links to data providers *metric assesses the extent of connectivity between the dataset under assessment and external sources. In [8], we have implemented a approximate version of this metric.*

Metric I.8 (Dereferenceable Backlinks)**.** *This metric measures the extent to which a resource includes all triples from the dataset that have the resource's URI as the object. This allows browsers and crawlers to traverse links in either direction [23]. The metric is defined as the ratio between the number of triples that are "back-links" a.k.a. "in-links" (i.e. that have the resource's URI as their object) and the total number of triples in the RDF graph served as a description of each resource.*

Relevance: Interlinking is relevant for open domain answering systems, since it is concerned with data integration. Consider the following scenario. Let us assume a user (who is a football fan) searching for a list of all German soccer players. Here, the instances of the terms *German* and *Soccer* should be interlinked with the instances *Deutschland* and *Football* respectively in order to return the correct and accurate list of all the German football players. Similarly, different datasets may use different ontologies as their vocabularies, which also use different URIs for terms that mean the same. Thus, it becomes crucially important for datasets *and* their vocabularies to be well interlinked to aid the process of disambiguation.

3.5 Data diversity Dimension

IRIs as the identifiers of entities in datasets are not necessarily human-friendly, or, if they are, they are often words in the dataset author's favourite language. However, human-comprehensible information in arbitrary formats and languages can be *attached* to entities via appropriate annotation properties.

Definition 3. *Data diversity refers to the availability of data in formats that are accessible to a wide range of human end users (numbers, images, audio, videos, etc.), and in different international languages.*

We use the following metrics to measure the data diversity of a dataset:

Metric D.9 (Human Readable Labelling)**.** *This metric measures the percentage of entities that have a human-readable label (rdfs:label). Although different datasets might different label properties (cf. [10]), we assume, for now, that the most commonly used standard label property (i.e.* `rdfs:label`*) is employed.*

Metric D.10 (Multiple Language Usage)**.** *This metric checks whether literal objects are available in different languages. We check all literals having a language tag and ignore those without a language tag. The value of this metric is the average number of different languages used per resource, throughout the dataset.*

Relevance: Data diversity is an important dimension for question answering systems because of the diverse needs of their human end users – who might, e.g., speak different languages. Moreover, rich natural language labels aid natural language approaches that disambiguate input words to uncover the semantics intended by the user and finally map them to the right entities in the underlying datasets. Generally, any question answering system has to deal with lexical and structural differences between the words used in questions and resources in the available datasets and terms in the ontologies used as vocabularies by these datasets. Labels of classes and properties play a great role in the process of matching question words with the dataset's terminology. We also refer to data diversity in terms of variety of the data formats available to the question answering system, other than text, for instance images, audio (e.g., for systems that support speech recognition), or videos. In addition, the final answer may need to be presented to the end users in natural language or as a visual graph – once more a situation in which human-readable labels play an important role.

3.6 Consistency Dimension

A knowledge base is consistent if it contains no contradicting information [22]. One can define the semantics of RDF data by defining classes and properties using RDFS and OWL. Reasoning can then use the semantics of the defined classes and properties to interpret the data and infer new knowledge – if the axioms that define the classes and properties and the triples in the datasets are consistent, or, in more formal terms, if the overall knowledge base has a model, i.e., an interpretation that satisfies all axioms. We adopt the definition of consistency given by [44] as follows:

Definition 4. *Consistency means that a knowledge base is free of (logical/formal) contradictions with respect to particular knowledge representation and inference mechanisms.*

Consistency metrics includes: entities as a member of disjoint classes, misplaced classes/properties, misuse of *owl:DatatypeProperty* and *owl:ObjectProperty*, ontology hijacking, and incorrect domain and range usage. We consider the following two metrics relevant for question answering systems:

Metric C.11 (Ontology Hijacking)**.** *Ontology hijacking refers to a dataset redefining classes or properties reused from an external vocabulary (which uses a different URI namespace).*

Metric C.12 (Misused OWL Datatype/Object Property)**.** *This metric is based on the metric defined by [23]. It detects properties that are defined as a owl:DatatypeProperty but is used like an object property, i.e. with resources in the object position, and vice versa. The metric computes a ratio of misused properties. Undefined properties are ignored by this metric.*

Relevance: Redefinition of external classes or properties by third parties could lead to an incorrect interpretation of the data by ontology-based question answering systems. When, for instance, as reported in [22], a term is defined as the domain of *rdf:type*, this results in every entity described on the Web being inferred as a member of that term. On the other hand, question answering systems highly rely on datatype properties, i.e., properties with literal-valued objects, to interpret user queries given in natural language and return answers expressed in natural language. Using an object property in place of data type properties leads a question answering system to consider IRIs as literal values, or to necessitate extra checks for every property, which might produce unexpected results.

3.7 Trust & Provenance Dimension

Trust has been widely studied in computer science research, on different fields, such as artificial intelligence [34], the World Wide Web [16], or multi-agent systems [31]. On the Semantic Web, it is seen as concept similar to confidence in data. Zaveri et al. [44] proposes the following definition for trustworthiness:

Definition 5 (Trustworthiness)**.** *The degree to which the information is accepted to be correct, true, real and credible.*

Several proposals for computing trust on the Semantic Web have been introduced in recent years, based on ranking values [17, 32], provenance information [20], or a combination of both [15, 9]. This information is usually attached to the original data and then used to derive a trust value, which roughly represents the belief or disbelief, whether it is on statements [9, 20] or entities [15, 17, 32]. However, current works do not address how to compute initial ranking or trust values; they are provided by users or just taken for granted. Hence, without an initial preprocessing to assign initial values, trust metrics are simply inapplicable in the current environment of Linked Data.

Provenance, on the other hand, is used to annotate data with metadata about its process of creation (when was it created, who did it, using what previous data, etc.). There are currently several proposals for annotating metadata: using named graphs [5, 13], using a framework to deal with multiple annotations on each triple [45], making use of the PROV-O ontology [26], attaching a provenance graph to the dataset [19], or devising a model to represent statements about statements [30, 11].

Implementing a provenance metric that takes into account such an heterogeneous ecosystem of provenance solutions is not straightforward. Here we present three simple metrics to evaluate provenance.

Metric P.13 (Basic Provenance)**.** *This metric, defined in [7], checks whether a dataset has at least a triple with the* `dc:creator` *or the* `dc:publisher` *property, to describe a dataset.*

Metric P.14 (Extended Provenance)**.** *This metric, defined in [7], checks if each entity in a dataset has the required provenance information such that an agent can identify the entity's origin.*

Metric P.15 (Provenance Richness)**.** *The Provenance Richness metric provides a measure of the information that a dataset has about itself, using the proportion of metadata statements in the dataset.*

Relevance: In the context of a QA system that extracts the information from Linked Data, trust and provenance are key factors to providing answers a user can have confidence in. In an environment where everyone can publish data, any given query can find disparate or even contradicting answers in different datasets. While trust provides a (general or personalised) measure of the trustworthiness a user can have in the data, provenance allows to relate data with their author and process of creation.

4. EVALUATION OF SELECTED METRICS

4.1 Datasets: DBpedia and Wikidata

For our pilot study we chose the most widely used cross-domain linked open datasets DBpedia and Wikidata.[8]

DBpedia is a community endeavour to extract structured information from Wikipedia into RDF for further consumption. For working with the dataset, one can either load it into a local triple store, or access it via a SPARQL endpoint[9].

Wikidata has lately emerged as one of the most famous user-curated source for structured information inside Wikipedia. Wikidata serves as the pivotal data management platform for Wikipedia and the majority of its sister projects [42]. Wikidata can be queried/accessed using a public endpoint[10] deployed by third parties. However, the uptime of this endpoint is debatable for reasons unknown to the authors, one of them possibly being heavy user (request) traffic.

There are the following key differences between these two datasets:

- DBpedia makes use of language dependent (i.e. human-readable) Wikipedia article identifiers for creating IRIs for concepts in every language (in Wikipedia language editions). DBpedia uses RDF and Named Graphs as its native data model.

[8]Freebase used to be another popular cross-domain dataset but support for it has expired, which is why we did not consider it; cf. https://www.freebase.com/.
[9]http://dbpedia.org/sparql
[10]https://query.wikidata.org/bigdata/namespace/wdq/sparql

Dimension	Metric
D1. Availability	A.1 Estimated Dereferenceability
	A.2 Estimated Dereferenceability of Forward Links
	A.3 No Misreported Content Types
	A.4 RDF Availability
	A.5 Endpoint Availability
D2. Interlinking	I.6 Estimated Interlink Detection
	I.7 Estimated External link Data Providers
	I.8 Estimated Dereference Backlinks
D3. Data diversity	D.9 Human Readable Labelling
	D.10 Multiple Language Usage
D4. Consistency	C.11 Ontology Hijacking
	C.12 Misused OWL Datatype Or Object Properties
D5. Trust and Provenance	P.13 Proportion of triples with *dc:creator* or *dc:publisher* properties
	P.14 Adoption of PROV-O Provenance
	P.15 Proportion of provenance statements

Table 1: Data quality assessment dimensions and metrics relevant to open domain question answering systems.

Slice statistics	Restaurants	Politicians	Films	Soccer players
Size (in MB)	5.7	528.2	1000	1700
# of triples	35,333	3,306,109	6,452,118	11,448,309
# of instances	725	36,221	90,063	104,562

Table 2: Statistics of each data slice obtained from DBpedia on various categories.

Slice statistics	Restaurants	Politicians	Films	Soccer players
Size (in MB)	5.4	2100	1500	1600
# of triples	33,748	11,990,654	8,873,545	9,530,527
# of instances	1,316	262,532	161,012	186,840

Table 3: Statistics of each data slice obtained from Wikidata on various categories.

- Wikidata, on the other hand, uses language-independent numeric identifiers and its own data model, which provides a furnished platform for representing provenance information.

Every subject on which Wikidata has structured data is called an *entity*, and every entity has a page. There are two types of entities: *terms*, which represent individuals and classes, and *properties*, which represent RDF properties. Every item page contains a *label*, a short *description*, a list of *aliases*, a list of *statements* and a list of *site links*. Wikidata statements are not just triples but can have additional quantifiers and references which support the claim. Each statement (typed as `rdfs:Statement`) has its own UUID identifier, which is used to connect to items, qualifiers and references.

We prepared four *slices* of DBpedia and Wikidata, respectively, to cover the categories *Restaurants*, *Politicians*, *Films* and *Soccer players*. The choice of categories for slices was based on internal voting of the authors from a list of popular topics.

We prepared the Wikidata slices by querying our local SPARQL endpoint which uses the RDF dumps[11] of November 30, 2015. For DBpedia we instead queried the publicly available SPARQL endpoint maintained by the DBpedia community.

The detailed statistics of the prepared slices are presented in Table 2 and Table 3 respectively. The slices prepared for our study, from both the datasets, are publicly available for research purposes and can be downloaded from https://goo.gl/Kn6Fom (DBpedia slices), https://goo.gl/5aTkLp (Wikidata slices).

The justification for evaluating quality metrics over slices of a dataset is: on the one hand, cross-domain datasets contain data on different domains, each of these datasets vary in quality of information (our **obj 3**). We would therefore like to compare the quality of datasets by looking at domain-specific subsets than the overall datasets. On the other hand, knowing the "fitness to use", as in where to find high quality data on a specific topic, helps federated question answering systems to choose the right source at query time.

We elaborate in brief on our key findings on both datasets in sections 4.4 and 6 respectively.

4.2 Experimental setup

For slice preparation we initially relied on querying the publicly available SPARQL endpoints of DBpedia and Wikidata. However, in the case of Wikidata, due to a lack of reliability of the endpoint (in terms of uptime and constant request timeout errors), we preferred deploying our custom endpoint locally and run a local slicing tool [12].

For preparing the slices and assessing their quality, we used a machine with the following specification:

- Triple store: Virtuoso OpenSource 7.2.1
- OS: Ubuntu 14.04 LTS
- Storage: 1 TB disk space
- RAM: 32 GB DDR3
- CPU: 12 core Intel Xeon E5-2690 v2 (@3.0 GHz)

A step by step guide for the endpoint setup and preparing slices is publicly available at https://goo.gl/y5xryk. Furthermore, we are planning to provide readers access to our virtual machine, as an interactive playground to run and evaluate datasets, on request.

4.3 The Luzzu Quality Assessment Framework

We carried out our evaluation using the Luzzu [7] quality assessment framework developed previously by some of the authors. Luzzu provides an integrated platform that: (1) assesses Linked Data quality using a library of generic and user-provided domain specific quality metrics in a scalable manner; (2) provides queryable quality metadata on the assessed datasets; (3) assembles detailed quality reports on assessed datasets. A number of metrics defined in [44] were

[11] http://tools.wmflabs.org/wikidata-exports/rdf/index.php?content=dump_download.php&dump=20151130

[12] https://github.com/keme686/LDSlicer

implemented for the Luzzu framework[13].

4.4 Results

In this section we present our findings from assessing the data slices from DBpedia and Wikidata presented in Section 4.1 with regard to the 5 dimensions and 15 metrics identified as relevant to the question answering domain in Section 3. The complete evaluation results are available online as a publicly accessible spreadsheet at https://goo.gl/ignzzI. We now discuss our observations.

4.4.1 Observations from slices statistics

We first discuss observations from the slice statistics; please refer to figures 1a, 1c, and 1c respectively.

- *Restaurants:* The sizes of the slices obtained from both datasets are very similar. The DBpedia slice has $1,585$ more triples than the Wikidata slice. However, it is interesting to see that the number of *instances* of restaurants in Wikidata is almost double that of DBpedia.
- *Politicians:* The politicians slice sizes of both the datasets shows high variation. Wikidata outperforms the DBpedia slices in the total number of triples and number of unique instances per slice by, $8,684,545$ and $262,532$ respectively. This advocates for the information richness of Wikidata "politicians" category is far better than that of DBpedia. Thus, from the slice statistics, it will be advisable to fire a query related to politics or politicians on the Wikidata dataset rather than DBpedia.
- *Films:* The case of the films slices of both datasets is similar in terms of the number of triples and the number of unique instances per slice. Wikidata provides more information, with almost 2.5 million more triples than DBpedia and $70,949$ more unique instances as compared to DBpedia.
- *Soccer players:* However, for soccer players slices, DBpedia is richer in information than Wikidata, with almost 2 million triples and $82,278$ more unique instances per slice respectively.

From the above observations, we can primarily infer that, the information richness of DBpedia is comparably reasonable to Wikidata for category *Restaurants*, very poor for categories *Politicians* and *Films*, and far better in case of the *Soccer players* category. Thus, for the user scenario we presented in Section 1.1, it will be debatable to consider one dataset over the other, as DBpedia has $1,585$ more triples than Wikidata, however, the latter has 591 more unique instances as compared to DBpedia in the restuarant category. Therefore, considering the sensitiveness of the application, one could choose one over the other, or employ a hybrid search on both the datasets.

4.4.2 Observations from computed scores of data quality assessment metrics

Next, we discuss observations from assessing the quality of the slices, by quality dimension:

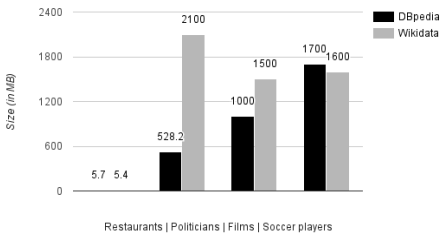
- ***Availability***: This dimension comprises of six metrics. The ***RDF Availability*** metrics gives a binary value: 0 meaning that an RDF dump is not accessible or 1 meaning that an RDF dump is accessible. This metric checks the usage of the *void:dataDump* property in the dataset's metadata and whether the linked URL is dereferenceable. For each slice we found 0 because the slices did not include VOID metadata.
  Similarly, ***Endpoint Availability*** checks the existence of the value of the *void:sparqlEndpoint* property, and its accessibility by firing an *ASK* query. For DBpedia, we manually checked this metric on the publicly available endpoint and found a VOID description [14], but it does not include the *void:dataDump* property. For Wikidata, we could not find such information in the public endpoint. The ***Dereferenceability*** A.1 and ***Dereferenceability Forward links*** A.2 metrics report almost constant scores for DBpedia slices, between $[0.012, 0.013]$ for A.1 and purely constant at 0.027 for A.2. Whereas for Wikidata slices, A.1 scores are reported to increase with the increase in the size of data slice; refer to figure 2. This implies that Wikidata slices are better in dereferentiability as compared to DBpedia slices. Also, there is no proportional relation between the size and the scores for either of the datasets. Hence, on a one-to-one comparison basis, the average scores of Wikidata slices for availability are comparably better than that of the DBpedia slices. However, Wikidata only minorly outperforms DBpedia and both the datasets return poor results in this dimension. One of the primary reason for such a low performance could be pin-pointed to the lack of regular maintainance (i.e. updation and cleaning) of the dataset's internal uri's/links and/or the lack of proper description in the subjects resource uri.
- ***Interlinking***: This dimension comprises of three metrics: ***Interlink Detection metric*** I.6, `Linking to External Providers metric` I.7 and `Dereferenceable Back Links metric` I.8 respectively. Comparing the scores of these metrics, for I.8 the assessment results show considerable improvement of Wikidata slices as compared to DBpedia (e.g. the politicians slice of DBpedia reports only 1.4% whilst the Wikidata politicians slice scores 9.8%). This implies that Wikidata is more traversible than compared to DBpedia, since resources in DBpedia report more broken/missing links which hinder its traversivity (via browser or crawlers). We were not be able to compute the scores for I.6 for Wikipedia and DBpedia. The reason behind this can be considered to be the lack of ontology information about Wikidata during the quality assessment process. Moreover, scores for the I.7 metric was not computed, further invetigation is needed for the same. Potential reason could be some configuration or parsing error in computation of the metric. The highest score for I.7 metric for Wikidata slices is reported for the politicians slice, i.e. 11, since the politican slice has the highest number of referred external links as compared to other slices.
- ***Consistency***: The scores of the metrics in this dimension, i.e. *Ontology Hijacking* C.11 and *Misused OWL Datatype Or Object Properties* C.12 are observed to be identical for all the slices. The reason is the same as mentioned above for interlinking: i.e. the

[13]Most had existed before commencing this work, in http://github.com/diachron/quality; those about trust we added for this evaluation; see https://github.com/jm-gimenez-garcia/ProvenanceRichnessLuzzu.

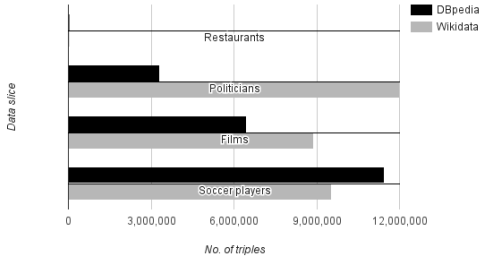
[14]http://dbpedia.org/void/page/Dataset

lack of ontological definitions in the slices. Since the slices are prepared from the whole DBpedia and Wikidata dataset using a specific SPARQL query patterns (e.g., `{?s a dbr:Restaurant. ?s ?p ?o}`), they do not have any schema information. Therefore these scores can be ignored for the quality comparison of both datasets. However, for completeness we propose to load the schema from the whole dataset and compute these scores on DBpedia and Wikidata as a whole.

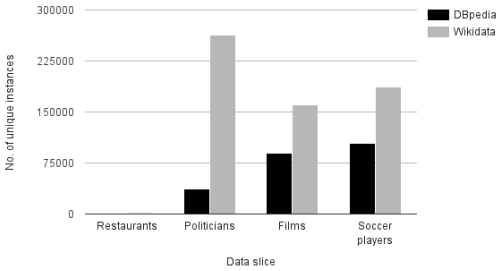
- ***Data diversity***: This dimension comprises of two metrics: *Human Readable Labelling* D.9 and *Multiple Language Usage* D.10. The one-to-one comparison of evaluation results for *Human Readable Labelling* on both datasets suggests that DBpedia remarkably outperforms Wikidata. The reason behind this is, for these metrics Luzzu checks the existence of `rdfs:label` for every instances types. This means, it also considers having no label for `rdf:Statement`s, their qualifiers and reference objects used for reification in Wikidata is a minus, which does not make sense to give labelling for reifications. This is one of the key differences between the data models of DBpedia and Wikidata, as mentioned in Section 4. However, Wikidata slices yield better scores for the *Multiple Language Usage* metric, as compared to DBpedia slices, on average. One of the reason better scores of Wikidata slices could be the language support (multilingual resource support) or popularity (in terms of relying on a particular dataset for updates). It can be observed from tables 4 and 5 that the average number of languages in Wikidata is greater or equal than that in DBpedia for all of the four slices studied here. However, the scores for D.10 of all the slices across the datasets the same. This could again be avoided for quality assessment judgement.
  For this dimension, we do not assess the diversity of data value types, such as images, audio, or videos. This is because we do not have implementation support in Luzzu at the moment for the same, but we are planning to implement them in the near future.
- ***Trust and Provenance***: We include three metrics for the Provenance dimension. *Basic Provenance* P.13 and *Extended Provenance* P.14 show that none of the datasets makes use of standard vocabularies for provenance. It is worth to be noted that, considering that the slicing method could provide a biased account of those metrics, we performed an evaluation against the *whole* DBpedia and Wikidata datasets with the same results. The metric *Provenance Richness* P.15 reports a negative result for DBpedia, while showing comparable figures for Wikidata. This result is expected because the Wikidata data model aims to provide statements about statements. Among the Wikidata slices, it can be observed that politicians have a greater provenance richness, while films and soccer players have lower values. It could be argued that it is more common to provide references for political activities, but the actual reason should be further studied.



(a) Data slice sizes.



(b) Number of triples per data slice.



(c) Number of unique instances per data slice.

Figure 1: A comparison of data slice statistics: DBpedia vs Wikidata.

5. RELATED WORK

A majority of the work in this area focuses on the broader domain of LOD quality assessment, with very little or no specific coverage of the question answering perspective. The organisers of community wide question answering challenges such as NTCIR Question Answering Challenge (QAC) [14], TREC [41], CLEF [21], BioASQ challenge [38], QALD series [27], etc., do discuss the extreme importance of having gold standard (manifold cross-validated) data at disposal, and also discuss about the challenges in curating such good quality datasets for the question answering domain. However, there is, so far, not much dedicated work for developing an autonomous framework for establishing and enforcing quality assurance from the questions answering perspective.

Preparing datasets of high quality (referred to as "fitness for use") requires administration from field experts. For instance, the preparation of CLEF eHealth track datasets demands a huge man-power effort from medical domain experts, such as clinicians, physicians, lab attendants (wet workers), etc. The judgements (relevance judgements) for a question set have to be cross verified by more than one subject expert with utmost care. This is very expensive in terms of time and finance required to curate a high quality dataset.

Crowdsourcing has also become a viable option in the

Dimension	Metric	Restaurants	Politicians	Films	Soccer Players
Availability	EstimatedDereferenceabilityMetric	0.013	0.013	0.012	0.012
	EstimatedDereferenceabilityForwardLinksMetric	0.027	0.027	0.027	0.027
	NoMisreportedContentTypesMetric	0	1	1	1
	RDFAvailabilityMetric	0	0	0	0
	EndPointAvailabilityMetric	0	0	0	0
Interlinking	EstimatedInterlinkDetectionMetric	–	–	–	–
	EstimatedLinkExternalDataProviders	–	–	–	–
	EstimatedDereferenceBackLinks	0.012	0.014	0.015	0.022
Consistency	OntologyHijacking	1	1	1	1
	MisusedOwlDatatypeOrObjectProperties	1	1	1	1
Data diversity	HumanReadableLabelling	0.953	0.985	0.997	1
	MultipleLanguageUsageMteric	1	2	3	3
Trust and Provenance	Basic Provenance	0	0	0	0
	Extended Provenance	0	0	0	0
	Provenance Richness	0	0	0	0

Table 4: Quality assessment evaluation of the four DBpedia slices. All figures in the table have been rounded to the third decimal point; precise figures can be retrieved from the spreadsheet mentioned in Section 4.4.

Dimension	Metric	Restaurants	Politicians	Films	Soccer Players
Availability	EstimatedDereferenceabilityMetric	0.051	0.063	0.049	0.063
	EstimatedDereferenceabilityForwardLinksMetric	0.093	0.053	0.051	0.65
	NoMisreportedContentTypesMetric	0	1	0	1
	RDFAvailabilityMetric	0	0	0	0
	EndPointAvailabilityMetric	0	0	0	0
Interlinking	EstimatedInterlinkDetectionMetric	NaN	NaN	NaN	NaN
	EstimatedLinkExternalDataProviders	5	11	9	8
	EstimatedDereferenceBackLinks	0.129	0.098	0.089	0.0826
Consistency	OntologyHijacking	1	1	1	1
	MisusedOwlDatatypeOrObjectProperties	1	1	1	1
Data diversity	HumanReadableLabelling	0.176	0.077	0.091	0.103
	MultipleLanguageUsageMteric	2	3	2	3
Trust and Provenance	Basic Provenance	0	0	0	0
	Extended Provenance	0	0	0	0
	Provenance Richness	0.055	0.0829	0.010	0.0252

Table 5: Quality assessment evaluation of the four Wikidata slices. All figures in the table have been rounded to the third decimal point; precise figure can be retrieved from the spreadsheet mentioned in Section 4.4.

present decade especially from the quality assessment perspective; consider dedicated research efforts such as DBpediaDQCrowd [1] or TripleCheckMate [24]. Also, general quality assessment tools and frameworks such as our own Luzzu [7], and others including one by Flemming [12], as well as Sieve [28], RDFUnit [25], LinkQA [18], and LiQuate [33] advocate the criticality of data quality assessment. However, for none of the quality assessment frameworks and tools mentioned earlier the specific requirements for benchmarking datasets from a question answering perspective have been discussed so far. Also, the quality requirements or guidelines to be followed while preparing a gold standard for question answering systems (keeping in mind the standard quality assessment tools and frameworks) has, so far, not materialised formally.

6. CONCLUSION AND FUTURE WORK

With question answering becoming increasingly reliant on open domain datasets, the quality of such datasets crucially influences the quality of answers to questions. So far, question answering systems have not yet been evaluated by the quality of the data they use, and while quality dimensions and metrics for linked data have been well documented, it has not been clear which of these are relevant for question answering scenarios. We have addressed these shortcomings of state-of-the-art research on question answering and data quality with a comprehensive review of linked data quality dimensions and metrics relevant to open domain question answering, and applied these metrics to subsets of the popular DBpedia and Wikidata datasets. The results of our experiments suggest that for most of these domains, the quality of Wikidata with regard to metrics relevant for open domain question answering is higher than that of DBpedia.

A major limitation of our work so far is that we have focused on those quality metrics that had been implemented for our Luzzu linked data quality assessment framework already, whereas further metrics would help to better cover the quality dimensions that we have identified as relevant to open domain question answering. Furthermore, we have so far only covered four subsets of DBpedia and Wikidata each, whereas these linked open datasets cover a lot more domains in which users would like their questions to be answered, and there exist further datasets that are widely used for open domain question answering.

In our near future work we plan to address these shortcomings, in particular ***i)*** to extend and automate the process of rigorously assessing data quality, carried out by Luzzu. We plan to implement the dimensions that our current evaluation has not yet covered, such as *Timeliness*, *Licensing*, and *Semantic accuracy*, and also those metrics whose values could not be calculated due to technical problems of our implementation, such as the estimated interlink detection metric for the slices of both DBpedia and Wikidata, and the approximate implementation of the external links to data providers metric for DBpedia slices. We will also ***ii)*** assess the fitness of further datasets in the LOD Cloud for open domain question answering by computing their quality metrics. Finally, we will ***iii)*** conform whether our dataset quality assessment correlates with the quality of question answering services over these datasets as perceived by the user, by having actual questions answered by data-driven question answering systems such as SINA [36].

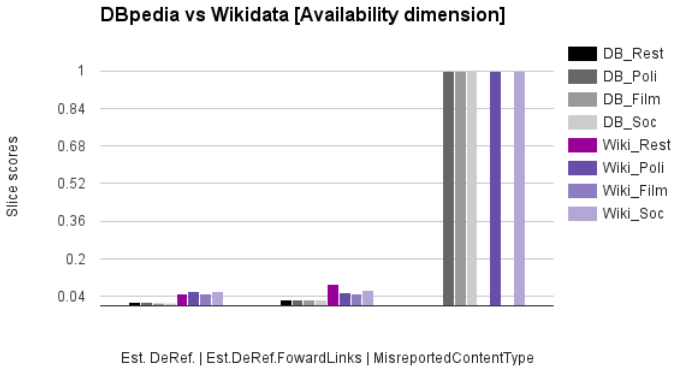


Figure 2: A comparison of the three varying availability metrics of DBpedia and Wikidata.

7. ACKNOWLEDGEMENT

This project is supported by funding received from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 642795 (WDAqua ITN) and from the Seventh Framework Program FP7 grant 601043 (http://diachron-fp7.eu).

References

[1] Maribel Acosta, Amrapali Zaveri, Elena Simperl, Dimitris Kontokostas, Sören Auer, and Jens Lehmann. Crowdsourcing linked data quality assessment. In *The Semantic Web–ISWC 2013*, pages 260–276. Springer, 2013.

[2] Sören Auer, Jens Lehmann, and Axel-Cyrille Ngonga Ngomo. Introduction to linked data and its lifecycle on the web. In *Reasoning Web*, pages 1–75, 2011.

[3] Carlo Batini, Cinzia Cappiello, Chiara Francalanci, and Andrea Maurino. Methodologies for data quality assessment and improvement. *ACM Computing Surveys (CSUR)*, 41(3):16, 2009.

[4] Christian Bizer and Richard Cyganiak. Quality-driven information filtering using the WIQA policy framework. *J. Web Sem.*, 7(1):1–10, 2009. . URL http://dx.doi.org/10.1016/j.websem.2008.02.005.

[5] Jeremy J. Carroll, Christian Bizer, Patrick J. Hayes, and Patrick Stickler. Named graphs, provenance and trust. In *Proceedings of the 14th international conference on World Wide Web, WWW 2005, Chiba, Japan, May 10-14, 2005*, pages 613–622, 2005. ISBN 1-59593-046-9. . URL http://doi.acm.org/10.1145/1060745.1060835.

[6] Jeremy Debattista, Santiago Londoño, Christoph Lange, and Sören Auer. LUZZU – A framework for linked data quality assessment. 2014. URL http://arxiv.org/abs/1412.3750.

[7] Jeremy Debattista, Sören Auer, and Christoph Lange. Luzzu – a framework for linked data quality analysis. 2015. URL http://eis.iai.uni-bonn.de/upload/paper/ICSC_2016_paper_17.pdf.

[8] Jeremy Debattista, Santiago Londoño, Christoph Lange, and Sören Auer. Quality assessment of linked datasets using probabilistic approximation. In *The Semantic Web. Latest Advances and New Domains*, pages 221–236. Springer International Publishing, 2015.

[9] Li Ding, Pranam Kolari, Tim Finin, Anupam Joshi, Yun Peng, and Yelena Yesha. On homeland security and the semantic web: A provenance and trust aware inference framework. In *AAAI Spring Symposium: AI Technologies for Homeland Security*, pages 157–160, 2005.

[10] Basil Ell, Denny Vrandecic, and Elena Paslaru Bontas Simperl. Labels in the web of data. In Lora Aroyo, Chris Welty, Harith Alani, Jamie Taylor, Abraham Bernstein, Lalana Kagal, Natasha Fridman Noy, and Eva Blomqvist, editors, *International Semantic Web Conference (1)*, volume 7031 of *Lecture Notes in Computer Science*, pages 162–176. Springer, 2011. ISBN 978-3-642-25072-9. URL http://dblp.uni-trier.de/db/conf/semweb/iswc2011-1.html#EllVS11.

[11] Fredo Erxleben, Michael Günther, Markus Krötzsch, Julian Mendez, and Denny Vrandecic. Introducing wikidata to the linked data web. In *The Semantic Web - ISWC 2014 - 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part I*, pages 50–65, 2014. ISBN 978-3-319-11963-2. . URL http://dx.doi.org/10.1007/978-3-319-11964-9_4.

[12] Annika Flemming. Quality Characteristics of Linked Data Publishing Datasources. http://sourceforge.net/apps/mediawiki/trdf/index.php?title=Quality_Criteria_for_Linked_Data_sources, 2010. [Online; accessed 13-February-2014].

[13] Giorgos Flouris, Irini Fundulaki, Panagiotis Pediaditis, Yannis Theoharis, and Vassilis Christophides. Coloring RDF triples to capture provenance. In *The Semantic Web - ISWC 2009, 8th International Semantic Web Conference, ISWC 2009, Chantilly, VA, USA, October 25-29, 2009. Proceedings*, pages 196–212, 2009. . URL http://dx.doi.org/10.1007/978-3-642-04930-9_13.

[14] Jun-ichi Fukumoto, Tsuneaki Kato, and Fumito Masui. Question answering challenge (qac-1): An evaluation of

question answering tasks at the ntcir workshop 3. In *New Directions in Question Answering*, pages 122–133, 2003.

[15] Yolanda Gil and Varun Ratnakar. Trusting information sources one citizen at a time. In *The Semantic Web - ISWC 2002, First International Semantic Web Conference, Sardinia, Italy, June 9-12, 2002, Proceedings*, pages 162–176, 2002. . URL http://dx.doi.org/10.1007/3-540-48005-6_14.

[16] Jennifer Golbeck. Trust on the world wide web: A survey. *Foundations and Trends in Web Science*, 1 (2):131–197, 2006. . URL http://dx.doi.org/10.1561/1800000006.

[17] Jennifer Golbeck, Bijan Parsia, and James Hendler. *Trust networks on the semantic web*. Springer, 2003. .

[18] Christophe Guéret, Paul Groth, Claus Stadler, and Jens Lehmann. Assessing linked data mappings using network measures. In *The Semantic Web: Research and Applications*, pages 87–102. Springer, 2012.

[19] Harry Halpin and James Cheney. Dynamic provenance for SPARQL updates. In *The Semantic Web - ISWC 2014 - 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part I*, pages 425–440, 2014. . URL http://dx.doi.org/10.1007/978-3-319-11964-9_27.

[20] Olaf Hartig. Trustworthiness of data on the web. In *Proceedings of the STI Berlin & CSW PhD Workshop*. Citeseer, 2008.

[21] Jesús Herrera, Anselmo Penas, and Felisa Verdejo. Question answering pilot task at clef 2004. In *Multilingual Information Access for Text, Speech and Images*, pages 581–590. Springer, 2005.

[22] Aidan Hogan, Andreas Harth, and Axel Polleres. Saor: Authoritative reasoning for the web. In *The Semantic Web*, pages 76–90. Springer, 2008.

[23] Aidan Hogan, Jürgen Umbrich, Andreas Harth, Richard Cyganiak, Axel Polleres, and Stefan Decker. An empirical survey of linked data conformance. *Web Semantics: Science, Services and Agents on the World Wide Web*, 14:14–44, 2012.

[24] Dimitris Kontokostas, Amrapali Zaveri, Sören Auer, and Jens Lehmann. Triplecheckmate: A tool for crowdsourcing the quality assessment of linked data. In *Knowledge Engineering and the Semantic Web*, pages 265–272. Springer, 2013.

[25] Dimitris Kontokostas, Patrick Westphal, Sören Auer, Sebastian Hellmann, Jens Lehmann, Roland Cornelissen, and Amrapali Zaveri. Test-driven evaluation of linked data quality. In *Proceedings of the 23rd international conference on World Wide Web*, pages 747–758. ACM, 2014.

[26] Timothy Lebo, Satya Sahoo, Deborah McGuinness, Khalid Belhajjame, James Cheney, David Corsar, Daniel Garijo, Stian Soiland-Reyes, Stephan Zednik, and Jun Zhao. Prov-o: The prov ontology. *W3C Recommendation*, 30, 2013.

[27] Vanessa Lopez, Christina Unger, Philipp Cimiano, and Enrico Motta. Evaluating question answering over linked data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 21:3–13, 2013.

[28] Pablo N Mendes, Hannes Mühleisen, and Christian Bizer. Sieve: linked data quality assessment and fusion. In *Proceedings of the 2012 Joint EDBT/ICDT Workshops*, pages 116–123. ACM, 2012.

[29] Felix Naumann. *Quality-driven query answering for integrated information systems*, volume 2261. Springer Science & Business Media, 2002.

[30] Vinh Nguyen, Olivier Bodenreider, and Amit P. Sheth. Don't like RDF reification?: making statements about statements using singleton property. In *23rd International World Wide Web Conference, WWW '14, Seoul, Republic of Korea, April 7-11, 2014*, pages 759–770, 2014. ISBN 978-1-4503-2744-2. . URL http://doi.acm.org/10.1145/2566486.2567973.

[31] Isaac Pinyol and Jordi Sabater-Mir. Computational trust and reputation models for open multi-agent systems: a review. *Artif. Intell. Rev.*, 40(1):1–25, 2013. . URL http://dx.doi.org/10.1007/s10462-011-9277-z.

[32] Matthew Richardson, Rakesh Agrawal, and Pedro Domingos. Trust management for the semantic web. In *ISWC – International Semantic Web Conference*, pages 351–368. Springer, 2003. .

[33] Edna Ruckhaus, Oriana Baldizán, and María-Esther Vidal. Analyzing linked data quality with liquate. In *On the Move to Meaningful Internet Systems: OTM 2013 Workshops*, pages 629–638. Springer, 2013.

[34] Jordi Sabater and Carles Sierra. Review on computational trust and reputation models. *Artif. Intell. Rev.*, 24(1):33–60, 2005. .

[35] Leo Sauermann and Richard Cyganiak. Cool URIs for the semantic web. W3C Interest Group Note, World Wide Web Consortium (W3C), December 2008. URL http://www.w3.org/TR/2008/NOTE-cooluris-20081203/.

[36] Saeedeh Shekarpour, Edgard Marx, Axel-Cyrille Ngonga Ngomo, and Sören Auer. Sina: Semantic interpretation of user queries for question answering on interlinked data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 30: 39–51, 2015.

[37] Saeedeh Shekarpour, Denis Lukovnikov, Ashwini Jaya Kumar, Kemele Endris, Kuldeep Singh, Harsh Thakkar, and Christoph Lange. Question answering on linked data: Challenges and future directions. 2016. URL http://arxiv.org/pdf/1601.03541v1.pdf.

[38] George Tsatsaronis, Michael Schroeder, Georgios Paliouras, Yannis Almirantis, Ion Androutsopoulos, Eric Gaussier, Patrick Gallinari, Thierry Artieres, Michael R Alvers, Matthias Zschunke, et al. Bioasq: A challenge on large-scale biomedical semantic indexing and question answering. In *AAAI Fall Symposium: Information Retrieval and Knowledge Discovery in Biomedical Text*, 2012.

[39] Christina Unger, André Freitas, and Philipp Cimiano. An introduction to question answering over linked data. In *Reasoning Web. Reasoning on the Web in the Big Data Era*, pages 100–140. Springer, 2014.

[40] Christina Unger, Corina Forascu, Vanessa Lopez, Axel-Cyrille Ngomo Ngonga, Elena Cabrio, Phillipp Cimiano, and Sebastian Walter. Question answering over linked data. CLEF2015 Working Notes, 2015. URL http://ceur-ws.org/Vol-1391/173-CR.pdf.

[41] Ellen M Voorhees. The trec question answering track. *Natural Language Engineering*, 7(04):361–378, 2001.

[42] Denny Vrandečić and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014.

[43] Richard Y Wang and Diane M Strong. Beyond accuracy: What data quality means to data consumers. *Journal of management information systems*, pages 5–33, 1996.

[44] Amrapali Zaveri, Anisa Rula, Andrea Maurino, Ricardo Pietrobon, Jens Lehmann, and Sören Auer. Quality assessment for linked data. *Semantic Web Journal*, preprint(preprint), 2015. URL http://www.semantic-web-journal.net/content/quality-assessment-linked-data-survey.

[45] Antoine Zimmermann, Nuno Lopes, Axel Polleres, and Umberto Straccia. A general framework for representing, reasoning and querying with annotated semantic web data. *J. Web Sem.*, 11:72–95, 2012. . URL http://dx.doi.org/10.1016/j.websem.2011.08.006.