

# Towards Automated Technologies in the Referencing Quality of Wikidata

Seyed Amir Hosseini Beghaeiraveri

Supervised by Alasdair J.G. Gray and co-supervised by Fiona J. McNeill

Heriot-Watt University

United Kingdom

sh200@hw.ac.uk

## ABSTRACT

Wikidata is a general-purpose knowledge graph with the content being crowd-sourced through an open wiki, along with bot accounts. The Wikidata data model enables assigning references to every single statement. Currently, there are more than 1 billion statements in Wikidata, of which about 70% have got references. Due to the rapid growth of Wikidata, the quality of Wikidata references is not well covered in the literature. To cover the gap, we suggest using automated tools to verify and improve the quality of Wikidata references. For verifying reference quality, we develop a comprehensive referencing assessment framework based on Data Quality dimensions and criteria. Then, we implement the framework as automated reusable scripts. To improve reference quality, we use Relation Extraction methods to establish a reference-suggesting framework for Wikidata. During the research, we managed to develop a subsetting approach to create a comparison platform and handle the big size of Wikidata. We also investigated reference statistics in 6 Wikidata topical subsets. The results of the latter investigation indicate the need for a wider assessment framework, which we aim to address in this dissertation.

## CCS CONCEPTS

• **Information systems** → **Resource Description Framework (RDF)**; • **Software and its engineering** → **Software libraries and repositories**; • **Social and professional topics** → **Quality assurance**.

## KEYWORDS

semantic web, data quality, reference quality, Wikidata, relation extraction and linking, subsetting, topical subset

### ACM Reference Format:

Seyed Amir Hosseini Beghaeiraveri, Supervised by Alasdair J.G. Gray and co-supervised by Fiona J. McNeill. 2022. Towards Automated Technologies in the Referencing Quality of Wikidata. In *Companion Proceedings of the Web Conference 2022 (WWW '22 Companion)*, April 25–29, 2022, Virtual Event, Lyon, France. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3487553.3524192>

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](https://permissions.acm.org).

*WWW '22 Companion*, April 25–29, 2022, Virtual Event, Lyon, France

© 2022 Association for Computing Machinery.

ACM ISBN 978-1-4503-9130-6/22/04...\$15.00

<https://doi.org/10.1145/3487553.3524192>

## 1 INTRODUCTION

Knowledge Graphs (KG) are large datasets representing real-world information in the form of data graphs [23]. In a KG, real-world entities are presented as nodes of a graph, and relationships between entities as edges. Research on knowledge graphs and their application is increasing both in academia [27] and industry [38]. Knowledge graphs play an essential role in web search [46], life sciences [36], and explainable AI [31].

Semantic Web knowledge graphs are a class of KGs that are constructed using Semantic Web standards such as RDF [16], RDFS [14], OWL [3], and SPARQL [22] developed by the World Wide Web Consortium (W3C). Almost all open general-purpose KGs, such as Wikidata<sup>1</sup> [48], DBPedia<sup>2</sup> [2], and YAGO [18], were built on these standards. These KGs compose the backbone of the Linked Data [9], the collection of interrelated datasets on the Web using W3C standards [12]. Between major general-purpose KGs, Wikidata is becoming more important in industry and research [1]. Wikidata's volume has increased ten times in the last five years, reaching more than 100 GB compressed. Wikidata collects its data with the participation of volunteer users. Wikidata can also be populated through *bots*. Bots are automatic tools that grab information from various related sources such as External datasets, web pages, CSV tables, and scientific papers, then add them to the knowledge graph. Bots usually add or edit a large amount of data in the knowledge base at once. Improper use of bots can lead to data quality problems in Wikidata [13].

The quality of data [39, 49] has been extensively studied in the context of Linked Data and knowledge graphs [10, 17, 19, 20, 24, 25, 35, 43, 51]. Data quality is defined as the “fitness for use” and subsequently depends on the use case and the user perspective [28]. Data quality in Linked Data and KGs is measured using different criteria, dimensions, and categories [51]. One of the main categories of data quality is trustworthiness [51]. The trustworthiness of data has a direct relationship with provenance and references [19]. In large open-source knowledge graphs, where the data comes from underlying sources, the provenance of the entities and their relationships is vital [40]. Considering the incremental use of knowledge graphs in AI, documenting the provenance of data has benefits such as increasing trust in the data, facilitating error detection, and evidence-based decision making [32].

One way to document the provenance and increase the trust in data is the use of *references* at different levels [19]. References specify the source of data in the knowledge base. Wikidata is the only general-purpose knowledge base that allows and recommends

<sup>1</sup>[https://www.wikidata.org/wiki/Wikidata:Main\\_Page](https://www.wikidata.org/wiki/Wikidata:Main_Page) - accessed 08 December 2021

<sup>2</sup><https://wiki.dbpedia.org/> - accessed 08 December 2021

users to provide references in the statement level [19]. However, the quality of the references in Wikidata has been rarely investigated. To the best of our knowledge, Piscopo et al.'s study [40], and its recently developed version by Amaral et al. [1], are the only studies in this regard. Due to the daily increase in the size of Wikidata, the increasing use of Wikidata in science and technology, the essential role of references in shaping Wikidata<sup>3</sup> and its data quality, and the shortcomings of the two current studies, a gap is being felt in studying Wikidata reference quality.

To fill the research gap, we aim to use automated technologies to assess and improve the quality of references in Wikidata. Our approach has two parts. In the first part, we develop a comprehensive reference quality assessment framework based on Data Quality dimensions and criteria. Then we implement the framework in the form of re-usable programs. In the second part, we use Relation Extraction methods to suggest references to Wikidata. During this research, we have developed a subsetting approach to creating a comparison platform and handling the big size of Wikidata. We also started investigating references' quality by performing statistical analysis on 6 Wikidata topical subsets. The results of the statistical research indicate the need for a comprehensive quality assessment framework for references.

## 2 PROBLEM

To the best of our knowledge Wikidata is the only KG with the capability of specifying references for each individual statement. During the literature review phase (the first half of 2020), the only available study on the quality of Wikidata references was Piscopo et al.'s paper [40]. The paper had studied the Wikidata dump 2016. In October 2016, Wikidata was ten times smaller than 2020. The paper had investigated the Wikidata English external sources only. Although some of these limitations have been improved by Amaral et al.'s work in 2021 [1], their study still covers only two quality dimensions. To fill this gap, we chose the main direction of the research to *evaluate* and *improve* the quality of Wikidata references, using efficient and reusable automated technologies. The initial questions of our research are as follows:

- RQ1** How can the quality of references be quantified considering different aspects of data quality? We hypothesize that a comprehensive quality assessment framework based on the data quality criteria can score the quality of references. Implementing such a framework will automatize the quality assessment process, and data owners can do evaluation tasks regularly.
- RQ2** How much has the reference quality of Wikidata changed since 2016, considering the rapid increase in data volume and diversity of sources? We hypothesize that the reference quality has increased with time.
- RQ3** What percentage of Wikidata references have been added by bots? How can we compare the quality of bot added references to human added references? Considering the increasing use of bots in recent years, does it affect the quality

of Wikidata references? The hypothesis is that the use of bots increases reference quality.

- RQ4** How effective and accurate is it to use Information Extraction and Linking (REL) methods to provide new references for Wikidata? We hypothesize that using REL methods can provide references to Wikidata in different topics with higher quality compared to the current reference adder bots.

## 3 STATE OF THE ART

The problem and the research questions of this dissertation are a combination of three main fields: Data Quality Criteria, Wikidata Reference Quality, and Relation Extraction. Data quality and Relation Extraction are widely covered in the literature, while Wikidata Reference Quality is rarely covered.

### 3.1 Data Quality

There are lots of studies on the quality of data. Data quality is generally considered a multidimensional concept. One of the earliest is Wang and Strong' multidimensional study [49] which categories data quality into four main categories, each consisting of one or more dimensions: *Intrinsic* (dimensions that are independent of the user's context), *Contextual* (dependent on the task at hand and the context of the data consumer), *Representational* (dimensions that describes in which extent data represented to the data consumers), and *Accessibility* (the form in which the data is available and to how it can be accessed by data consumers). In the context of web-based information, Bizer et al. [11] proposed a quality assessment framework that allows filtering good quality information on the web. The framework benefits from a wide range of metrics expressed in the form of graph patterns.

Data quality is also heavily investigated in Linked Data. Zaveri et al. [51] surveyed 21 different approaches up to 2012. They identified a core set of 23 different data quality dimensions categorized in 6 different categories. Their study is the most comprehensive aggregation of data quality dimensions in the context of Linked Data. More recently, Färber et al. [19] evaluated five large KGs (Freebase, Wikidata, YAGO, Cyc, and DBpedia) over 34 data quality metrics. They extended the criteria of Wang [49] into 11 dimensions and 34 metrics. In each metric, they carefully analyzed the KGs and assigned them a score between 0 and 1. They also developed an algorithm in which users can admit a weight to each metric and compute the quality score based on their needs. Debattista et al. [17] examined about 3.7 billion triples taken from 37 Linked Data datasets via 27 metrics based on the Zaveri et al. survey. They also provided a Principal Component Analysis (PCA) over their evaluation results to find the minimum number of metrics that can inform users about the quality of Linked Data datasets.

### 3.2 Trustworthiness and Reference Quality

In the literature, the ability to mention the provenance of data and its quality is considered as *trustworthiness* [1, 17, 19]. Färber et al. [19] defined trustworthiness as a combination of three concepts mentioned in [49]: *Believability* (the extent to which data are accepted or regarded as true, real, and credible), *Objectivity* (the extent to which data are unbiased, and impartial), and *Reputation* (the extent to which data are trusted or highly regarded

<sup>3</sup>According to the Wikidata project, "Wikidata is not a database that stores facts about the world, but a secondary knowledge base that collects and links to references to such knowledge" [47].

in terms of their source or content). Färber et al. [19] considered the criteria *trustworthiness on statement level* and Debattista et al. [17] considered metric *traceability of the data*, both checking only the existence of reference usage in datasets. However, the two studies did not provide an analytical review of how references are used.

### 3.3 Wikidata Reference Quality

Among data quality dimensions, accuracy and trustworthiness have been less investigated in Wikidata [41]. Wikidata recommends the provided references should be relevant and authoritative [47]. Piscopo et al. [40] examined the authoritativeness and the relevance of Wikidata English external sources. They first evaluated a small set of sample references (<300 statements) through microtask crowdsourcing. The samples evaluation results were then given to a machine-learning algorithm to measure the relevance and authoritativeness of all grabbed data. The machine-learning algorithm showed that about 70% of Wikidata external sources are relevant and 80% of them are authoritative. The experiment recently has been developed and reproduced by Amaral et al. [1] on Wikidata snapshot of 16 April 2021, considering both English and non-English external sources. However, the recent study is still limited to relevance and authoritativeness. In [41], Piscopo et al. compared Wikipedia and Wikidata in terms of external references. They showed Wikidata has a more diverse pool of sources (in terms of origin country) and benefits from external datasets (such as library catalogs) more than Wikipedia. Curotto and Hogan [15] published a proposal that aims to index English Wikipedia references as a source for Wikidata statements.

### 3.4 Relation Extraction and Linking (REL)

Relation Extraction and Linking (REL) is the task of extracting and representing relations from unstructured text and linking their elements to the properties and entities of a KG [34]. Different REL tools usually use different approaches based on their goals. Usually, they all start with an Entity Extraction (the task of finding entities of real-world in unstructured texts) phase. The process then continues with relationship parsing and linking the relations to the reference KG. Finally, REL tools demonstrate the extracted data in RDF format. REL has a broad application in KG population [50], specially in domains like Medicine [44], Sport [42], and Terrorism [26]. Another important application of REL is Question Answering (Q&A) [52]. Martinez-Rodriguez et al. [34] provided a comprehensive overview of using IE and REL in the Semantic Web.

Theoretically, REL methods can provide references for structured facts through searching the fact's similar relations in unstructured texts. FRED [21] is a REL approach to extract binary and n-ary relations from raw texts. FRED uses Stanford CoreNLP [33] for identifying entities in the text. FRED creates its IRIs instead of linking extracted relations to an existed KG. FRED is also not limited to a specific domain. Another potential method for detecting references is DeepDive [37]. DeepDive is an efficient statistical learning approach to build knowledge graphs from extensive collections of web pages. The ability to render web pages makes DeepDive a potential tool for providing external URI sources.

## 4 METHODOLOGY

To address the research questions, we split the research into two sub-projects: (i) assessing the quality of existing Wikidata references and (ii) designing a Reference Suggestion Framework (RSF) via REL methods.

To address RQ1, RQ2 and RQ3, we consider a comprehensive reference quality assessment framework based on data quality dimensions. The existing studies only considered the relevance and authoritativeness of references. Our framework covers a wide range of data quality dimensions. We provide formal definitions for metrics specifically in the context of referencing. We also define the metrics as objectively as possible, easy to compute, and reusable for other RDF-based KGs. Having a comprehensive reference assessment framework enables us to evaluate current Wikidata references and third-party reference suggestion bots.

To address RQ4, we design an automated framework that suggests qualified sources for new or existing non-referenced Wikidata facts. We call non-referenced claims Target Statements. Most of the Wikidata items (Q-IDs) have a corresponding Wikipedia article, including several references in different types such as web pages, scholarly articles, or books. These Wikipedia references are potential sources for Target Statements of that particular Wikidata item. However, Wikipedia references are often unstructured texts. Our solution is to parse Wikipedia references via REL methods. Then, we search the extracted structured data to find triples equal to the Target Statements. If such a triple is found, the unstructured text will be suggested as a source for the Target Statement. Compared to the existing reference adder bots, our approach is topic-independent. Reference adder bots are simple single-purpose scripts that add references only in the claim-creation time. They can populate only one specific topic, e.g., proteins, politicians, etc. The variety of sources they can parse is also limited to CSV files or external structured datasets.

### 4.1 Referencing Quality Assessment Framework

Starting with the literature review, we noticed the lack of a referencing quality assessment framework for the graph-based datasets. We targeted Wikidata because it is the only open general purpose KG with the capability of adding references at the statement level. We also needed a comparison platform to evaluate the quality of Wikidata references. Initially, we intended to base the comparison on the separation of human-added versus bot-added references. Tracing bots/human activities requires querying the edit history of Wikidata. The edit history of Wikidata is a huge dataset (more than 3TB compressed) and requires more resources than Wikidata itself. We then proposed the concept of Topical Subsets [7, 8] to build coherent subsets of Wikidata and overcome the high volume issues. We performed a statistical analysis on references in 6 Wikidata WikiProject subsets [7]. The statistical analysis helped us define new reference-specific metrics in the amount-of-data and relevancy dimensions. We have now defined 40 reference-specific quality metrics in 21 dimensions, covering all major data quality categories in the Zaveri et al. survey. Currently, we are implementing the metrics in Python scripts. The entire framework will be a Python module called Referencing Quality Scoring System (RQSS). The next step is

to evaluate the framework on Wikidata topical and random subsets. In summary, we will have the following tasks:

*RQSS Implementation (Python).* The first goal is to finish the implementation of the reference quality assessment metrics in Python. Along with the framework, RQSS includes other layers to extract required collections from datasets and present the output to users. The implementation task has already begun [4]. At the moment of writing, we implemented four metrics.

*RQSS Evaluation via Subsets and Unit Tests.* We provide a comparison platform using Wikidata subsets and evaluate, debug, and improve RQSS using the subsets. We have extended the functionality of the wikidata subsetting tool (WDumper [5]) to extract several small random subsets of Wikidata. Combined with small topical subsets of [7] (e.g., Ships 2021), we can efficiently evaluate the RQSS metrics and compare the results on both meaningful and random datasets. The functionality of methods is also being tested via python unit tests.

*Analyzing RQSS Results.* After ensuring RQSS's proper functionality, we will apply it to some up-to-date topical and random subsets. We will then compare the scores of the subsets. We discuss the reasons for high or low scores by considering different hypotheses and pieces of evidence.

*Analyzing Bot-Added references via RQSS (optional).* We try to build an additional comparison platform by separating human-added and bot-added references in each subset. For this task, we consider two solutions: (i) using the Wikidated 1.0 [45] dataset and (ii) querying the Wikidata edit history dumps<sup>4</sup> directly. The separation results can address RQ2. If the task succeeds, we run RQSS on the separated references to analyze the difference between human-added and bot-added references. The final results can address RQ3.

## 4.2 Reference Suggestion Framework (RSF)

For this sub-project, we reviewed the state-of-the-art Relation Extraction methods to find the best approach in terms of reference suggestion. We have nominated two REL approaches: FRED [33] and DeepDive [37]. First, we will evaluate the candidates' efficiency on a gold standard set. We will then develop a framework that provides appropriate sources from Wikipedia references for Wikidata claims. We will examine the performance of the framework on-demand and on-batches. The results determine the possibility of having an online version of the framework. There will be considerations for those Target Statements that do not have a corresponding Wikipedia article or the corresponding article has no references. A similar approach to DeFacto [32] can be considered, i.e., using search engines and keywords to provide the potential sources. We will evaluate RSF using our RQSS metrics. We will compare the results to existing Wikidata references and also compare RSF to traditional reference adder bots.

## 5 RESULTS

This dissertation has had two contributions to the Semantic Web community. The first contribution is the evaluation of WDumper as an effective tool in topical subsetting [8]. We extracted four use case

topical subsets via WDumper. We provided the performance details of the test runnings. We also assessed the outputs of WDumper by performing different SPARQL queries on the input and output. We performed all experiments on Wikidata 2016 and 2020 dumps. Using two different dumps allowed us to measure the change of data volume from 2016 to 2020 in different topics. We described WDumper advantages and limits in topical subsetting: WDumper accurately extracts the desired data (including Wikidata qualifiers and references). However, WDumper is not flexible in defining subsets.

Our second contribution is a statistical analysis of references in six Wikidata WikiProjects [7]. We extracted six subsets of Wikidata corresponding to Astronomy, Gene Wiki, Taxonomy, Law, Music, and Ships WikiProjects [6]. We investigated the statistics of reference nodes, the ratio of referenced statements, the usage of reference-specific properties, and the distribution of triples per node. We also introduced the concept of shared references as an important factor in the quality of references. The statistical study showed the need for a comprehensive quality assessment framework.

The project in hand is the reference quality assessment framework. We have completed the formal definitions of 40 reference quality metrics and implemented four of the metrics. We tested the four implemented metrics on on WikiProjects Ships 2016, Ships 2021 [6], and two random Wikidata subsets with two different sizes.

## 6 CONCLUSIONS

This dissertation delves into the automated, reusable, and comprehensive technologies for evaluating and improving the quality of Wikidata references. The increased use of Wikidata in industry and research, besides the importance of trust in data, motivates us to investigate this problem. We considered four main research questions and two research projects to address them. The main goal of this work is to establish a comprehensive automated quality assessment framework to evaluate Wikidata references. Having an automated reference quality assessment system helps data owners to evaluate the quality of references regularly and more efficiently. Data scientists can also compare datasets quantitatively. The main limitation regarding automation is the subjectivity of some data quality dimensions, e.g., Relevancy. Involving human opinions in the calculation of quality metrics makes automation difficult. We also propose to use Relation Extraction and Linking methods to suggest higher quality references in Wikidata. As future work, we consider generalizing our quality assessment framework to all RDF knowledge graphs. Although our metric definitions are general, the implementation of the metrics is currently specific to the Wikidata data model.

## ACKNOWLEDGMENTS

I acknowledge the School of Mathematical and Computer Science at Heriot-Watt University, which dedicates the James Watt Ph.D. Scholarship to this dissertation. I also appreciate the efforts of my supervisors, Dr. Alasdair Gray and Dr. Fiona McNeill. I also acknowledge the fruitful discussions with the participants of BioHackathon Europe subsetting projects in 2020 and 2021 [29, 30].

<sup>4</sup><https://dumps.wikimedia.org/backup-index.html> - accessed 17 December 2021

## REFERENCES

- [1] Gabriel Amaral, Alessandro Piscopo, Lucie-Aimée Kaffee, Odinaldo Rodrigues, and Elena Simperl. 2021. Assessing the quality of sources in Wikidata across languages: a hybrid approach. *arXiv preprint arXiv:2109.09405* (2021).
- [2] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. DBpedia: A Nucleus for a Web of Open Data. In *The Semantic Web (Lecture Notes in Computer Science)*. Springer, Berlin, Heidelberg, 722–735. [https://doi.org/10.1007/978-3-540-76298-0\\_52](https://doi.org/10.1007/978-3-540-76298-0_52)
- [3] Jie Bao et al. 2012. OWL 2 Web Ontology Language Document Overview (Second Edition). <https://www.w3.org/TR/owl2-overview/> Publisher: World Wide Web Consortium - accessed 08 December 2021.
- [4] Seyed Amir Hosseini Beghaeiraveri. 2021. *RQSS-Framework*. <https://github.com/seyedahbr/RQSSFramework>
- [5] Seyed Amir Hosseini Beghaeiraveri. 2021. *WDumper*. <https://github.com/seyedahbr/wdumper>
- [6] Seyed Amir Hosseini Beghaeiraveri. 2021. *Wikidata Subsets of 6 Wikiproject (Gene Wiki, Taxonomy, Astronomy, Music, Law, Ships)*. <https://doi.org/10.5281/zenodo.5117928>
- [7] Seyed Amir Hosseini Beghaeiraveri, Alasdair Gray, and Fiona McNeill. 2021. Reference Statistics in Wikidata Topical Subsets. In *Proceedings of the 2nd Wikidata Workshop (Wikidata 2021) (CEUR Workshop Proceedings, Vol. 2982)*. CEUR, Virtual Conference, October. [https://researchportal.hw.ac.uk/files/53252708/Reference\\_Statistics\\_in\\_Wikidata\\_Topical\\_Subsets\\_corrected\\_version.pdf](https://researchportal.hw.ac.uk/files/53252708/Reference_Statistics_in_Wikidata_Topical_Subsets_corrected_version.pdf) ISSN: 1613-0073.
- [8] Seyed Amir Hosseini Beghaeiraveri, Alasdair J. G. Gray, and Fiona J. McNeill. 2021. Experiences of Using WDumper to Create Topical Subsets from Wikidata. In *CEUR Workshop Proceedings, Vol. 2873*. CEUR-WS, 13. <https://researchportal.hw.ac.uk/files/45184682/paper13.pdf> ISSN: 1613-0073.
- [9] Tim Berners-Lee. 2006. Linked Data - Design Issues. <https://www.w3.org/DesignIssues/LinkedData> accessed 08 December 2021.
- [10] Christian Bizer. 2007. *Quality-driven information filtering in the context of web-based information systems*. PhD Thesis. Freie Universität Berlin.
- [11] Christian Bizer and Richard Cyganiak. 2009. Quality-driven information filtering using the WIQA policy framework. *Journal of Web Semantics* 7, 1 (Jan. 2009), 1–10. <https://doi.org/10.1016/j.websem.2008.02.005>
- [12] Christian Bizer, Tom Heath, and Tim Berners-Lee. 2011. Linked Data: The Story so Far. <https://doi.org/10.4018/978-1-60960-593-3.ch008>
- [13] Wikidata Bots. 2021. Wikidata:Bots - Wikidata. <https://www.wikidata.org/wiki/Wikidata:Bots> accessed 08 December 2021.
- [14] Dan Brickley and R.V. Guha. 2014. RDF Schema 1.1. <https://www.w3.org/TR/rdf-schema/> Publisher: World Wide Web Consortium - accessed 08 December 2021.
- [15] Paolo Curotto and Aidan Hogan. 2020. Suggesting Citations for Wikidata Claims based on Wikipedia's External References.. In *Wikidata@ISWC*.
- [16] Richard Cyganiak, David Wood, and Markus Lanthaler. 2014. RDF 1.1 concepts and abstract syntax. (2014). <https://www.w3.org/TR/rdf11-concepts/> Publisher: World Wide Web Consortium - accessed 08 December 2021.
- [17] Jeremy DeBattista, Christoph Lange, Sören Auer, and Dominic Cortis. 2018. Evaluating the quality of the LOD cloud: An empirical investigation. *Semantic Web* 9, 6 (2018), 859–901. Publisher: IOS Press.
- [18] MS Fabian, Kasneci Gjergji, WEIKUM Gerhard, et al. 2007. Yago: A core of semantic knowledge unifying wordnet and wikipedia. In *16th International World Wide Web Conference, WWW*. 697–706.
- [19] Michael Färber, Frederic Bartscherer, Carsten Menne, and Achim Rettinger. 2017. Linked data quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO. *Semantic Web* 9, 1 (Nov. 2017), 77–129. <https://doi.org/10.3233/SW-170275>
- [20] Christian Fürber and Martin Hepp. 2011. SWIQA - A Semantic Web Information Quality Assessment Framework. *ECIS 2011 Proceedings* (Oct. 2011). <https://aisel.aisnet.org/ecis2011/76>
- [21] Aldo Gangemi, Valentina Presutti, Diego Reforgiato Recupero, Andrea Giovanni Nuzzolese, Francesco Draicchio, and Misael Mongiovi. 2017. Semantic web machine reading with FRED. *Semantic Web* 8, 6 (2017), 873–893.
- [22] Steve Harris and Andy Seaborne. 2013. SPARQL 1.1 Query Language. <https://www.w3.org/TR/sparql11-query/> Publisher: World Wide Web Consortium - accessed 08 December 2021.
- [23] Aidan Hogan, Eva Blomqvist, Michael Cochez, et al. 2021. Knowledge Graphs. *Synthesis Lectures on Data, Semantics, and Knowledge* 12, 2 (Nov. 2021), 1–257. <https://doi.org/10.2200/S01125ED1V01Y202109DSK022> Publisher: Morgan & Claypool Publishers.
- [24] Aidan Hogan, Andreas Harth, Alexandre Passant, Stefan Decker, and Axel Polleres. 2010. Weaving the pedantic web. In *LDOW*.
- [25] Aidan Hogan, Jürgen Umbrich, Andreas Harth, Richard Cyganiak, Axel Polleres, and Stefan Decker. 2012. An empirical survey of Linked Data conformance. *Journal of Web Semantics* 14 (July 2012), 14–44. <https://doi.org/10.1016/j.websem.2012.02.001>
- [26] Uraivan Inyaem, Phayung Meesad, Choochart Haruechaiyasak, and Dat Tran. 2010. Construction of fuzzy ontology-based terrorism event extraction. In *2010 Third International Conference on Knowledge Discovery and Data Mining*. IEEE, 391–394.
- [27] Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S. Yu. 2021. A Survey on Knowledge Graphs: Representation, Acquisition, and Applications. *IEEE Transactions on Neural Networks and Learning Systems* (2021), 1–21. <https://doi.org/10.1109/TNNLS.2021.3070843>
- [28] Joseph M. Juran. 1962. *Quality control handbook*. McGraw Hill. Issue: 658.562 Q-1q.
- [29] Jose Emilio Labra-Gayo et al. 2021. *Project 21 - Biohackathon 2021 - KG subsets*. <https://github.com/kg-subsetting/biohackathon2021> original-date: 2021-11-08T13:27:08Z.
- [30] Jose Emilio Labra-Gayo, Ammar Ammar, Dan Brickley, et al. 2021. Knowledge graphs and wikidata subsetting. (March 2021). <https://doi.org/10.37044/osf.io/wu9et> Publisher: BioHackrXiv.
- [31] Freddy Lecue. 2020. On the role of knowledge graphs in explainable AI. *Semantic Web* 11, 1 (Jan. 2020), 41–51. <https://doi.org/10.3233/SW-190374> Publisher: IOS Press.
- [32] Jens Lehmann, Daniel Gerber, Mohamed Morsey, and Axel-Cyrille Ngonga Ngomo. 2012. DeFacto - Deep Fact Validation. In *The Semantic Web - ISWC 2012 (Lecture Notes in Computer Science)*. Springer, Berlin, Heidelberg, 312–327. [https://doi.org/10.1007/978-3-642-35176-1\\_20](https://doi.org/10.1007/978-3-642-35176-1_20)
- [33] Christopher D Manning, Mihai Surdeanu, John Bauer, et al. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*. 55–60.
- [34] Jose L. Martinez-Rodriguez, Aidan Hogan, and Ivan Lopez-Arevalo. 2020. Information extraction meets the semantic web: a survey. *Semantic Web Preprint* (2020), 1–81.
- [35] Pablo N. Mendes, Hannes Mühleisen, and Christian Bizer. 2012. Sieve: linked data quality assessment and fusion. In *Proceedings of the 2012 Joint EDBT/ICDT Workshops*. 116–123.
- [36] David N. Nicholson and Casey S. Greene. 2020. Constructing knowledge graphs and their biomedical applications. *Computational and Structural Biotechnology Journal* 18 (Jan. 2020), 1414–1428. <https://doi.org/10.1016/j.csbj.2020.05.017>
- [37] Feng Niu, Che Zhang, Christopher Ré, and Jude W. Shavlik. 2012. DeepDive: Web-scale Knowledge-base Construction using Statistical Learning and Inference. *VLDS* 12 (2012), 25–28. Publisher: Istanbul.
- [38] Natasha Noy, Yuqing Gao, Anshu Jain, Anant Narayanan, Alan Patterson, and Jamie Taylor. 2019. Industry-scale knowledge graphs: lessons and challenges. *Commun. ACM* 62, 8 (2019), 36–43. Publisher: ACM New York, NY, USA.
- [39] Leo L. Pipino, Yang W. Lee, and Richard Y. Wang. 2002. Data quality assessment. *Commun. ACM* 45, 4 (April 2002), 211–218. <https://doi.org/10.1145/505248.506010>
- [40] Alessandro Piscopo, Lucie-Aimée Kaffee, Chris Phethean, and Elena Simperl. 2017. Provenance information in a collaborative knowledge graph: an evaluation of Wikidata external references. In *International semantic web conference*. Springer, 542–558.
- [41] Alessandro Piscopo and Elena Simperl. 2019. What we talk about when we talk about wikidata quality: a literature survey. In *Proceedings of the 15th International Symposium on Open Collaboration*. ACM, Skövde Sweden, 1–11. <https://doi.org/10.1145/3306446.3340822>
- [42] Ana B Rios-Alvarado, Ivan Lopez-Arevalo, and Victor J Sosa-Sosa. 2013. Learning concept hierarchies from textual resources for ontologies construction. *Expert Systems with Applications* 40, 15 (2013), 5907–5915.
- [43] Anisa Rula, Matteo Palmonari, and Andrea Maurino. 2012. Capturing the age of linked open data: Towards a dataset-independent framework. In *2012 IEEE Sixth International Conference on Semantic Computing*. IEEE, 218–225.
- [44] David Sánchez and Antonio Moreno. 2007. Learning medical ontologies from the web. In *AIME Workshop on Knowledge Management for Health Care Procedures*. Springer, 32–45.
- [45] Lukas Schmelzeisen, Corina Dima, and Steffen Staab. 2021. Wikidated 1.0: An Evolving Knowledge Graph Dataset of Wikidata's Revision History. (2021).
- [46] Amit Singhal. 2012. Introducing the Knowledge Graph: things, not strings. <https://blog.google/products/search/introducing-knowledge-graph-things-not-strings/> Library Catalog: [www.blog.google](http://www.blog.google).
- [47] Wikidata Verifiability. 2021. Wikidata:Verifiability - Wikidata. <https://www.wikidata.org/wiki/Wikidata:Verifiability> accessed 08 December 2021.
- [48] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM* 57, 10 (Sept. 2014), 78–85. <https://doi.org/10.1145/2629489>
- [49] Richard Y. Wang and Diane M. Strong. 1996. Beyond accuracy: What data quality means to data consumers. *Journal of management information systems* 12, 4 (1996), 5–33. ISBN: 0742-1222 Publisher: Taylor & Francis.
- [50] René Witte, Ninus Khamis, and Juergen Rilling. 2010. Flexible Ontology Population from Text: The OwlExporter.. In *LREC*, Vol. 2010. 3845–3850.
- [51] Amrapali Zaveri, Anisa Rula, Andrea Maurino, Ricardo Pietrobon, Jens Lehmann, and Soeren Auer. 2016. Quality assessment for Linked Data: A Survey. *Semantic web* 7, 1 (2016), 63–93. Publisher: IOS Press.
- [52] Lei Zou, Ruizhe Huang, Haixun Wang, et al. 2014. Natural language question answering over RDF: a graph data driven approach. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*. 313–324.