A systematic literature review on Wikidata

Marçal Mora-Cantallops, Salvador Sánchez-Alonso and Elena García-Barriocanal

{marcal.mora, salvador.sanchez, elena.garciab}@uah.es

Universidad de Alcalá de Henares, Alcalá de Henares, Spain

Abstract:

Purpose-

To review the current status of research on Wikidata and, in particular, of articles that either describe applications of Wikidata or provide empirical evidence, in order to uncover the topics of interest, the fields that are benefiting from its applications and which researchers and institutions are leading the work.

Design/methodology/approach-

A systematic literature review is conducted to identify and review how Wikidata is being dealt with in academic research articles and the applications that are proposed. A rigorous and systematic process is implemented, aiming not only to summarize existing studies and research on the topic but also to include an element of analytical criticism and a perspective on gaps and future research.

Findings-

Despite Wikidata's potential and the notable rise in research activity, the field is still in the early stages of study. Most research is published in conferences, highlighting such immaturity, and provides little empirical evidence of real use cases. Only a few disciplines currently benefit from Wikidata's applications and do so with a significant gap between research and practice. Studies are dominated by European researchers, mirroring Wikidata's content distribution and limiting its Worldwide applications.

Originality/value-

Our results collect and summarize existing Wikidata research articles published in the major international journals and conferences, delivering a meticulous summary of all the available empirical research on the topic which is representative of the state of the art at this time, complemented by a discussion of identified gaps and future work.

Keywords: Wikidata, survey, literature review, empirical studies, applications

1. Introduction

Wikidata is an open, collaborative project started on October 30, 2012 by Wikimedia Deutschland, hosted and supported by the Wikimedia Foundation (Abián et al., 2018), continuously increasing its popularity since its creation (Vrandečić, 2013; Vrandečić and Krötzsch, 2014) and whose main goals are two: (1) to be the central storage for the structured data of all its Wikimedia sister projects (such as Wikipedia itself), avoiding duplicate and contradicting information, but also facilitating multi-language capabilities and management, and (2), to provide data to other third-party projects and initiatives, and allowing complex queries on the existing base of knowledge. Wikidata does not only store facts, but also the corresponding reference sources, allowing data validation and the creation of timelines (e.g. a country's population is a variable that can be referenced to the census and changes across time). Labels, aliases, and descriptions of entities in Wikidata are provided in more than 350 languages. The basic structure of Wikidata consists of *items* (that have a label, a description and any number of aliases, known as terms), *properties* and *values*, linked in *statements* that closely resemble an RDF triple. However, the model of Wikidata statements is slightly more complex, as they can be enriched with *qualifiers* (providing additional context for the claim) and *references* (which support the claim). As of October 2018, Wikidata has more than 60.000 registered authors (contributors with 10 or more edits) who, together with anonymous users and automatic bots, have contributed to more than 53.5 million data items[1]. Furthermore, the number of authors and articles has been steadily increasing since its conception[2].

Erxleben et al. (2014, p. 51) point out that "the relevance of Wikidata for researchers in semantic technologies, linked open data, and Web science […] hardly needs to be argued for". In 2014, however, they found that Wikidata had been hardly used in the semantic web community, even though the relative success of projects such as DBpedia (Bizer et al., 2009) and Freebase (Bollacker et al., 2008) hinted at the potential of Wikidata. The situation notably changed in the last few years, as Freebase was shutdown in 2015 and integrated into Wikidata (Pellissier Tanon et al., 2016), while in 2017 Wikidata was already found to be the most suitable source of information for person data (twice as many instances as DBpedia) or detailed information about countries, among others (Ringler and Paulheim, 2017). The question is... has Wikidata become relevant to researchers and practitioners too?

The purpose of this study is, thus, to review the current status of research on Wikidata and, in particular, we concentrate on articles that either describe applications of Wikidata or study the project empirically, to uncover the topics of interest, to assess its related research activity and to identify what researchers and institutions are leading the work, as detailed in Section 2. Our methodology is described in Section 3 and our results are presented in Section 4. In Section 5, all four research questions are discussed. Conclusions are finally presented in Section 6.

2. Research questions

The research questions addressed by this study are:

RQ1. How much research activity has there been since the introduction of Wikidata?

RQ2. What are the main topics covered by empirical studies on Wikidata?

RQ3. What Wikidata applications are proposed in the literature?

RQ4. Who is leading Wikidata research?

With respect to RQ1, we identified how many relevant papers were published per year as well as the journal or conference that published them. To answer RQ2, we considered the scope of the study (whether it is based on empirical evidence or proposes an application) and the topics or disciplines involved. In particular, applications will be reviewed in more detail in RQ3, as one of the main goals of Wikidata is to support third-party projects and initiatives and, thus, it becomes relevant to survey the current range of existing applications. Finally, with respect to RQ4, we considered individual researchers and their affiliations.

3. Methodology

In order to identify and review how Wikidata is addressed in academic research articles and what applications are proposed, a systematic literature review was conducted. We used a rigorous and systematic process, aiming not only to summarize existing studies and research on the topic but also to include an element of analytical criticism and a perspective on gaps and future research (Okoli, 2015). Systematic reviews employ carefully defined protocols to determine which studies are to be included, as well as to analyze their contribution in as unbiased a form as possible (Kitchenham, 2004; Webster and Watson, 2002). This study has been undertaken as a systematic literature review based on the original guidelines as proposed by Kitchenham (2004) combined with the guidelines in Software Engineering by Budgen and Brereton (2006).

3.1. Search strategy

Four online academic research databases (ACM Digital Library, IEEE Xplore, Springer Link and Science Direct) were scanned for relevant articles, complemented with a search in ISI Web of Science and Google Scholar to add any articles that had not been found in the previous four databases. ACM and IEEE were considered relevant due to their focus on information systems and computer science, while Springer Link and Science Direct give access to a number of important journals from the same fields. ISI Web of Science and Google Scholar are comprehensive citation search engines, and they were used to increase the overall reliability of the search results and to ensure that articles from other scholarly fields were also included. All searches were narrowed down to empirical studies (studies where Wikidata is the clear and main object of analysis) published in peer-reviewed full conference papers and journal articles that included the term “wikidata” in either their titles, abstracts or keywords. A preliminary total of 93 articles were retrieved (Table 1), which, after removal of duplicates, were reduced down to 82.

Source	# of papers
ACM Digital Library	35
IEEE Xplore	12
Springer Link	16
Science Direct	6
ISI Web of Science	21
Google Scholar	3

Table 1. Total number of papers identified from each database.

3.2. Inclusion and exclusion criteria

All items were recorded and manually checked to determine their relevance; a number of further criteria were specified to select the appropriate studies for inclusion in the review. To be included, papers had to (a) either describe a practical application or include empirical evidence directly related to Wikidata, (b) be published in journals or conference proceedings and include an abstract and future work and (c) be in English. Thus, exclusion criteria were (a) papers that used a Wikidata dump only for testing or papers that reported on technical details (e.g. implementation or migration), (b) papers not peer-reviewed and (c) papers in other languages. Additionally, when one study superseded an older one (extending or replacing previous work by the same authors or group), only the newest was kept. For example, Steiner (2014a) is extended (and therefore superseded) by Steiner (2014b).

Three additional studies were discarded due to their descriptive nature (they lacked research hypotheses). All three are relevant for Wikidata as a topic, with two of them being foundational (Vrandečić, 2013; Vrandečić and Krötzsch, 2014) and the third one describing the migration of data from Freebase (which was discontinued in 2015) to Wikidata (Pellissier Tanon et al., 2016). As these papers neither describe a particular application or carry out an empirical study on Wikidata, they are excluded of the final selection.

As a result of this selection, a total of 57 articles meeting the inclusion criteria were selected for the review (Table 2).

Phase	Criteria	Papers left
Search results	English only. Wikidata in title, abstract or keywords. No duplicates.	82
Removal of non peer-reviewed articles	Journal and conference papers only.	78
Focus exclusion upon full text	Main focus on Wikidata. Originality of the study (newer studies take priority).	60
Exclusion of descriptive articles.	Only empirical and/or application articles.	57

Table 2. Selection steps for the included studies.

3.4. Data collection and analysis

Each of the final 57 articles was read and then tabulated to show:

- Source and year of publication (addressing RQ1).
- Main topic areas and summary of the study for both applications and empirical studies (addressing RQ2 and RQ3).
- Authors, affiliations and their countries (addressing RQ4).

Table 3 shows the abridged results of the systematic review, complemented with a concept map depicting the main themes of research in Figure 2.

ID	Reference	Type	Topic	Aim
S1	(Abián et al., 2018)	Empirical	Knowledge organization	Comparison between DBpedia and Wikidata in regard to data quality dimensions.
S2	(Balaraman et al., 2018)	Application	Data quality	An approach to evaluate the completeness of entities in Wikidata.
S3	(Benedetti et al., 2018)	Application	NLP IR	CSA (Context Semantic Analysis) for inter-document similarity computation.
S4	(Bergamin and Bacchi, 2018)	Application	Knowledge integration	Restructuring bibliographic records to include them in the Wikidata model.
S5	(Brasileiro et al., 2016)	Empirical	Knowledge organization	Assess the taxonomic hierarchies in Wikidata.
S6	(Burgstaller-Muehlbacher et al., 2016)	Application	Medical	Gene Wiki initiative based on Wikidata as a semantic framework.
S7	(Chekol and Stuckenschmidt, 2018)	Application	Information retrieval	Proposal of a model for querying and maintaining temporal knowledge graphs.
S8	(Chisholm et al., 2017)	Application	NLG	Neural model for creating Wikipedia biographic summary sentences from Wikidata. Code available at: https:// github.com/andychisholm/mimo.
S9	(Cuong and Müller-Birn, 2016)	Empirical	Communities User roles and collaboration patterns	Study dynamic participation patterns across multiple user roles.
S10	(English, 2018)	Application	NLP	Using Wikidata to construct large annotated corpus. Available at: github.com/Building-Large-Annotated-Corpora.
S11	(Erxleben et al., 2014)	Application	Knowledge integration	RDF exports to connect Wikidata and the Linked Data Web.
S12	(Ferrada et al., 2018)	Application	Knowledge integration	Querying Wikimedia Images through Wikidata information and combining visual queries with semantic facts.
S13	(Geiß and Gertz, 2016)	Application	Information retrieval	Disambiguation model to identify person names in texts.
S14	(Geiß et al., 2018)	Application	Information retrieval	Tool to assign Wikidata entities to Location, Person or Organization, the most common classes of named entities.
S15	(Geiß et al., 2015)	Application	Information retrieval	Creation of a person-centric network from the information contained in Wikipedia and Wikidata. Available at https://dbs.ifi.uni-heidelberg.de/resources/data/#wikipediasocial.
S16	(Hall et al., 2018)	Application	Data quality	Automated editors (bots) detection.
S17	(Heindorf et al., 2016)	Application	Data quality	Machine learning-based approach to detect vandalism in Wikidata.
S18	(Hempelmann et al., 2016)	Application	Information retrieval	Fuzzy set membership for individuals assigned in multiple classes.
S19	(Hernández et al., 2016)	Empirical	Queries	Comparison of the efficiency of various database engines querying the Wikidata.
S20	(Hollink et al., 2018)	Empirical	Gender	Explored gender differences in the various Wikipedia language editions through Wikidata content.
S21	(Ingvaldsen and Gulla, 2015)	Application	NLP Recommendation Systems	News stream aggregating system based on Wikidata's semantic representation.
S22	(Kaffee et al., 2018)	Application	NLP NLG	Generate summaries for Wikipedia articles in under-served languages, given structured data as an input.
S23	(Kaffee et al., 2017)	Empirical	Community Language	State of languages in Wikidata and comparison to the real-World distribution.
S24	(Kaffee and Simperl, 2018)	Empirical	Community Language	Language distribution among Wikidata’s editors and relationship to Wikidata’s content.
S25	(Klein et al., 2016)	Application	Gender	Introduction of the “Wikidata Human Gender Indicators” (WHGI) to monitor biographical gender disparities.
S26	(Leva and Chemello, 2018)	Empirical	Community Collaboration	A case study of a GLAM-Wiki collaboration between institutions.

S27	(Lim et al., 2017)	Application	NLP Topic Modeling	A spatial and temporal variant of LDA to better detect more specific topics associated with specific days and locations.
S28	(Müller-Birn et al., 2015)	Empirical	Community User roles	Cluster analysis of participants' content editing activities.
S29	(Murase et al., 2019)	Application	NLP	Creating feature vectors by using inference results on an external knowledge base (Wikidata).
S30	(Nielsen et al., 2017)	Application	Information retrieval	Scholia, a tool to handle scientific bibliographic information through Wikidata. Available at https://github.com/fnielsen/scholia.
S31	(Nielsen, 2018)	Application	Knowledge integration	Linking the ImageNet WordNet synsets and Wikidata.
S32	(Olivieri et al., 2017)	Application	Data validation	An approach to evaluate the trustworthiness of online information modeled as RDF Triples.
S33	(Pellissier Tanon et al., 2018)	Application	NLP Question answering	Platypus, a natural language question answering system on Wikidata. Available at: https://askplatyp.us.
S34	(Pellissier Tanon and Kaffee, 2018)	Empirical	Knowledge organization	Analysis of the stability in Wikidata's schema.
S35	(Pfundner et al., 2015)	Application	Medical	Implementation of an automated system for keeping drug-drug interaction information in Wikipedia up to date using Wikidata.
S36	(Piscopo, Kaffee, et al., 2017)	Empirical	External references	Exploration of the relevance and authority of Wikidata's external references.
S37	(Piscopo, Phethean, et al., 2017)	Empirical	Community User roles Collaboration	Framework to evaluate the evolution of the ontology, in order to cluster editing activities and identify user roles in time windows.
S38	(Piscopo and Simperl, 2018)	Empirical	Community User roles Ontology	Investigate how the relationship between different type of users influenced the outcome quality in Wikidata.
S39	(Piscopo, Vougiouklis, et al., 2017)	Empirical	External references	Relationship between Wikipedia and Wikidata in terms of their external references.
S40	(Prasojo et al., 2016)	Application	Data quality	A completeness tool for Wikidata (COOL-WD). Available at http://cool-wd.inf.unibz.it/
S41	(Putman et al., 2016)	Application	Medical	Microbial specific data model, based on Wikidata, to represent microbial genomes.
S42	(Putman et al., 2017)	Application	Medical	WikiGenomes (wikigenomes.org), a web application based on Wikidata that facilitates the consumption and curation of genomic data by the entire scientific community.
S43	(Ringler and Paulheim, 2017)	Empirical	Knowledge organization	Quantification of differences, overlapping and complementary parts of public knowledge graphs (DBPedia, YAGO, Wikidata).
S44	(Sáez and Hogan, 2018)	Application	NLP IR	Automatic generation of Wikipedia's info-boxes from Wikidata.
S45	(Samuel, 2018)	Application	Data quality Vandalism	Obtaining the translation path of properties and visualizing them. Available at https://github.com/johnsamuelwrites/wdprop/tree/v10.
S46	(Sarabadani et al., 2017)	Application	Data quality Vandalism	Automated vandalism detection tools, with novel approach on feature engineering.
S47	(Sen et al., 2017)	Application	Information retrieval	Creation of thematic maps of information. Available at http://cartograph.info.
S48	(Spitz et al., 2016)	Application	Information retrieval	Disambiguation model to identify toponyms in texts.
S49	(Steiner, 2014b)	Application	Data quality	Monitoring of editing activities by bots and users in realtime.
S50	(Stinson et al., 2018)	Empirical	GLAM-Wiki Community	Explore GLAM-Wiki tactics, opportunities and collaboration between institutions.
S51	(Ta and Anutariya, 2015)	Application	NLP	Model for aligning and enriching both Wikipedia and Wikidata through info-boxes and properties.
S52	(Thakkar et al., 2016)	Empirical	Knowledge organization	Quality assessment of linked data from the question answering domain's perspective.

S53	(Turki et al., 2017)	Application	NLP Multilingual dictionary	Wikidata as a multi-lingual and multi-dialectal dictionary for Arabic dialects.
S54	(Vagliano et al., 2017)	Application	NLP Recommendation Systems	Recommendation approach using the semantic annotation of user reviews combined with Wikidata to extract useful and non-trivial information about the items to recommend.
S55	(Vougiouklis et al., 2018)	Application	NLP NLG	Neural networks for Natural Language Generation on top of Semantic Web triples.
S56	(Yang et al., 2018)	Application	NLP	Relation Linking System for Wikidata (RLSW) is proposed to link the relations in KGs to plain texts.
S57	(Zangerle et al., 2016)	Empirical	Data quality	Analysis of the Wikidata's property suggesting system.

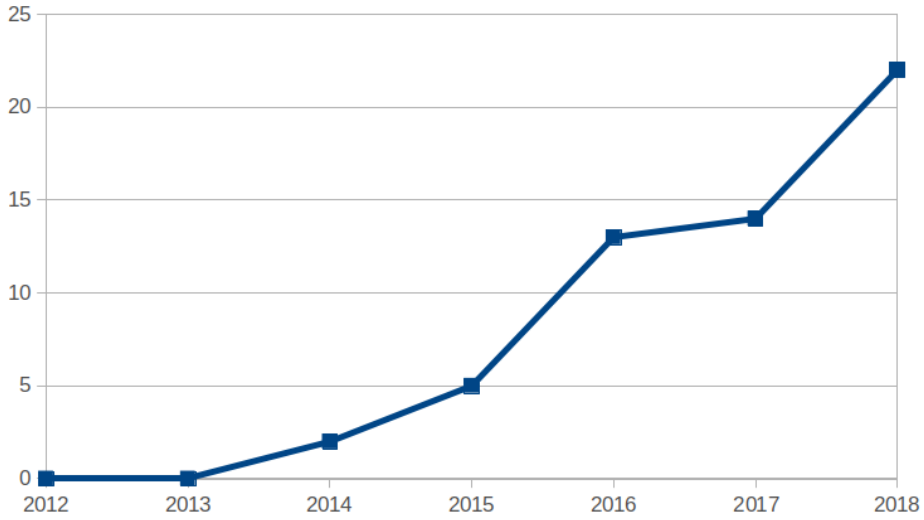
Table 3. Summary of the papers included in the systematic review.

4. Results

4.1. How much research activity has there been since the introduction of Wikidata?

Overall, 57 relevant studies were identified. The majority of them were published in conference proceedings (47, a 82.5%), while only 10 (17.5%) were found in journals. 18 papers analysed empirical evidence (31.6%) but most of the activity around Wikidata is related to applications (68.4%).

Since its introduction in 2012, the number of papers has been growing steadily; from two identified papers in 2014 to 22 in 2018 (studies available published before the 1st of December 2018; there is one 2019 study that is not included in Figure 1), illustrating the growing interest on the topic by the scientific community.



[Figure 1]

4.2. What are the main topics covered by the empirical studies on Wikidata?

The main topics identified are, in descending order of relevance: users and their editing practices, knowledge organization, external references, language and a few miscellaneous topics. We will study each in detail in the following subsections.

4.2.1. Users and editing practices

Understanding how users participate and collaborate to build a structured knowledge base has been the most found topic in our set of studies. Piscopo, Phethean, et al. (2017) investigated how the relationship between different types of users (bots or human editors, registered or anonymous) influenced the outcome quality in Wikidata, including the effects of tenure and interest diversity among the registered users. The conclusion of their study was that, to create high quality items, "the interaction between human and algorithmic users is necessary". Tenure and heterogeneous groups were also found to be positive influences, while, on the other hand, anonymous users were classified as detrimental for quality. In a more recent work by the same group (Piscopo and Simperl, 2018), a framework to evaluate the evolution of the ontology (in breath and depth) was proposed in order to cluster editing activities and identify user roles in monthly windows. In their exploration, they found the Wikidata ontology to be "large and messy, with numerous underpopulated classes and uneven depth". Two roles were identified, contributors and leaders, with the second category related positively to the depth of the ontology; no relation was found concerning the breadth of the ontology, however.

Wikidata's editors and contributing users are assisted by a property suggesting system; Zangerle et al. (2016) argued that this recommendation mechanism has the potential to improve data consistency and quality, so they evaluated the usefulness of such suggestions and compared them with other state-of-the-art recommendation approaches, finding that "the current recommendation algorithm works well in regard to recall and precision" and that incorporating contextual information into the computation of property recommendations could further improve its performance significantly.

User roles have also been of interest for Müller-Birn et al. (2015), who performed a cluster analysis of participants' content editing activities and compared them to the typical roles found in peer-production systems and collaborative ontology projects. After finding six editing patterns, they suggested that the majority of users were very specialized in their contributions, while only a minority - the most active group - participated all over the project. They also concluded that the Wikidata project finds itself between the approach of a "classic" peer-production system and a collaborative ontology. A later work by Cuong and Müller-Birn (2016) applied sequence analysis methods to study the dynamic participation patterns across the previously characterized roles in Wikidata and observed, among other findings, a relationship between users' joining time and the turbulence in their editing behavior, with more veteran editors changing roles or patterns more often than the ones that joined later.

Community practices, tactics and strategies were explored through the lens of the relationships between GLAMs (Galleries, Libraries, Archives and Museums) and the Wikimedia communities (such as the Wikidata one) by Stinson et al. (2018), as " Wikidata's breadth creates a sweeping landscape of opportunities for a large and flexible linked data hub, for describing collections or the knowledge contained within them", and identified multiple opportunities for collaboration between heritage institutions and Wikidata's communities, such as the one Leva and Chemello (2018) outline in their description of the partnership between the Fondazione BEIC (Biblioteca Europea di Informazione e Cultura) and Wikimedia Italia.

4.2.2. Knowledge organization

According to Brasileiro et al. (2016), "the quality of taxonomic structures is key to properly capturing knowledge in Wikidata" and, after assessing the taxonomic hierarchies in Wikidata, they identified a significant number of issues, such as problematic classification and taxonomic statements, related to an inadequate use of instantiation and subclassing in certain Wikidata hierarchies. For them, support to contributors would be beneficial in order to improve the quality of the Wikidata content. Pellissier Tanon and Kaffee (2018) also recognized the importance of the stability in Wikidata's schema to foster its data usage and analyzed the changes in labels of properties in six different languages, finding it stable and easily reusable.

For Ringler and Paulheim (2017), although "DBpedia, YAGO, or Wikidata, are often considered similar in nature and coverage, there are, in fact, quite a few differences". In their work, they quantified those differences and identified the overlapping and complementary parts of these three public knowledge graphs, finding them "hardly interchangeable", each with advantages and issues that depended on the desired application or domain. Abián et al. (2018) also compared Wikidata with DBpedia in regard to "the most relevant data quality dimensions" and highlighted how Wikidata has "an open centralised nature" and its multilingual capacity, while DBpedia is "more popular in the Semantic Web and the Linked Open Data communities". On the other hand, Thakkar et al. (2016) ran a quality assessment of linked data in DBpedia and Wikidata from the perspective of question answering, a popular application scenario for knowledge databases, and found "the quality of Wikidata with regard to the majority of relevant metrics [...] higher than that of DBpedia".

4.2.3. External references

Data quality is also influenced by the relevance and authoritativeness of its external references or sources. Piscopo, Kaffee, et al. (2017) analyzed this particular aspect of Wikidata quality and found external references to be "mostly relevant and authoritative", and explored models to predict non-relevant or non-authoritative references that could be useful for future applications. Piscopo, Vougiouklis et al. (2017) also investigated the relationship between two closely related Wikimedia projects, Wikipedia and Wikidata, from the external references perspective, finding little reuse of references across Wikidata and Wikipedia and less Anglo-American sources in the former (although their references "often point to the same domain" which might be a sign of actual diversification of knowledge across languages).

4.2.4. Languages

If multilinguality is an important topic for knowledge bases in general, it is of particular interest for Wikidata, as one of its aims is to serve the multilingual requirements of the Wikimedia projects. Kaffee et al. (2017) explored the state of languages in Wikidata and compared them to the real-World distribution, finding a large gap, as Wikidata's knowledge "is mostly available in a few languages, while most languages have close to no coverage [...] similar to Wikipedia". This line of work was complemented in Kaffee and Simperl (2018), investigating the language distribution across Wikidata's

editors and relating it to Wikidata's content and users' community, finding a relationship between language content and community language, but also editors that extend their activities to languages unknown to them as well.

4.2.5. Miscellaneous

One of Wikidata's goals is to be able to provide support for complex queries. Hernández et al. (2016) experimentally compared the efficiency of various database engines for the purpose of querying the Wikidata knowledge graph and revealed a number of strengths and weaknesses for each of the tested engines.

On the other hand, Hollink et al. (2018) explored gender differences in Wikipedia (through Wikidata content) with respect to the coverage of the Members of the European Parliament in multiple languages and found a very small gender-difference, with women covered by slightly more editions, which could be due to a sample limitation (with only one very specific profession) or it could arguably reflect how Wikidata's information aggregation process fosters diversity.

No other topics have been found, reflecting how most of the existing studies are focused in only a few fields of research.

4.3. What Wikidata applications are proposed in the existing literature?

4.3.1. Natural Language Processing and Generation

Knowledge bases are employed in several domains and applications; some of the most popular venues are the improvement of Natural Language Processing (NLP), language generation (NLG) and Information Retrieval (IR). Benedetti et al. (2018) proposed a knowledge-based technique, called CSA (Context Semantic Analysis), that could take advantage of Wikidata's knowledge for inter-document similarity computation and also showed how it could be applied to IR tasks. The NLP community also relies heavily in data for their research and only a few large-scale expertly annotated corpora are available, due to cost in time and money. English (2018) reviewed his experience using Wikidata to construct large annotated corpus under distant supervision, an application that could be beneficial for a greater community.

One common technique for text analysis is topic modeling; Lim et al. (2017) proposed a modification of the Latent Dirichlet Allocation (LDA) algorithm that incorporated contextual information (spatial and temporal) extracted and verified using Wikidata's knowledge base in order to improve topic detection. Validating their approach on a Twitter dataset, the preliminary results showed how the modified algorithms were "able to detect highly relevant and detailed topics associated with specific days and locations". Recommendation systems can also benefit from knowledge databases; for example, Wikidata is used by Ingvaldsen and Gulla (2015) to create a news stream aggregating system that automatically recognizes and disambiguates geo spatial and meaning bearing entities in news texts. In another work by Vagliano et al. (2017), semantic annotation of user reviews was combined with knowledge extracted from Wikidata in the movie, book and music domains, providing more diverse and novel recommendations than traditional techniques.

A neural model for mapping between structured and unstructured data, aiming to derive Wikipedia biographic summary sentences from the information contained in Wikidata was implemented by Chisholm et al. (2017). Kaffee et al. (2018) introduced a system that extends Wikipedia's *ArticlePlaceholders* (similar to stubs, but dynamically updated to accommodate Wikidata's information) with multilingual summaries automatically generated from Wikidata triples for under-served languages on Wikipedia, obtaining promising results as members of the targeted language communities ranked their texts as "close to the expected quality standards of Wikipedia", allowing most of their content to be potentially reused by the editors. Vougiouklis et al. (2018) extended this idea with an approach that does not require manually defined templates and applied neural networks to generate textual summaries from Wikidata triples. A prototype for generating info-boxes (which provide a summary of the most important meta-data relating to a particular entity described by a Wikipedia article) from Wikidata was created by Sáez and Hogan (2018), a method that could be applied to generate info-boxes for the supported languages automatically, without any need for manual input or templates. Info-boxes were also the object of study of Ta and Anutariya (2015), who propose a model for aligning and enriching both Wikipedia and Wikidata through the info-boxes of the former and the properties of the latter. And following the multilingual goal, Turki et al. (2017) proposed to convert Wikidata into a multi-lingual and multi-dialectal dictionary for Arabic dialects, that could not only be completed, verified, adjusted and used by users, but it could also contain semantic links, claiming that Wikidata could implement the direct functions and indirect functions of a dictionary "more effectively than any other similar project".

In order to leverage the knowledge in Wikidata to help machines understand plain texts (benefiting many NLP applications) Yang et al. (2018) described a system to combine Wikidata's information with plain texts in order to establish whether a word sequence should be linked to a relationship. Systems that can answer natural language questions also rely on knowledge bases; Platypus is a question answering system implemented by Pellissier Tanon et al. (2018) and explicitly based on Wikidata information that supports multiple natural languages thanks to this. Murase et al. (2019) proposed complementing language feature vectors with associative knowledge through inference on the Wikidata knowledge graph, in order to improve the response of human-like dialog systems when information is not explicitly mentioned by the human user.

4.3.2. Data quality and validation

Tracking data quality is key to the success of any knowledge base, and Wikidata is not an exception. In their article, Balaraman et al. (2018) presented an approach towards measuring completeness(defined as the degree to which all known information about an item is stated) in Wikidata. As total information about an item is arguably infinite, the authors compare item completeness to other similar items from the same domain (relative completeness) obtaining an "objective criteria for assessing quality" that could be used for resource allocation and project management. Prasojo et

al. (2016) also discuss how to manage and consume meta-information about completeness for Wikidata, in their case using a completeness tool called COOL-WD.

Wikidata, among others, has taken advantage of automation to build its database at a rate and scale unachievable by human contributors; still, understanding bot (and human) behavior in the community is an important topic that depends on accurate bot recognition. Steiner (2014b) introduced an application and a related API to monitor all edit activity on the 287 Wikipedias and Wikidata in realtime. Hall et al. (2018) developed a machine classifier to detect bots according to their editing patterns to support community patrolling activities and avoid potentially damaging behavior.

This possibility of having Wikidata edited by anyone (including bots), results in frequent vandalization, exposing all information systems to the risk of spreading spurious facts or making decisions on incorrect information. Heindorf et al. (2016) developed a machine-learning based approach to automatically detect vandalism in Wikidata, achieving good results, as did Sarabadani et al. (2017) when they extended previous works with additional focus in feature engineering. Facilitating vandalism detection in translations of labels and property descriptions of Wikidata is also the aim of Samuel (2018), who developed a tool for understanding and visualizing such translation patterns.

Validation of information is another process that benefits from knowledge bases, specially in a larger scale. Olivieri et al. (2017) proposed an approach to evaluate the trustworthiness of online information, modeling such information as RDF triples, matching its properties to a specific ontology (WordNet, in their case) and to Wikidata, obtaining feature vectors that can be used in a machine-learning pipeline to predict the veracity of a predicate.

4.3.3. Information retrieval

The emergence of information extraction or retrieval from knowledge graphs has aided their growth; data models, novel approaches for obtaining information from data and visualization techniques are relevant to be able to infer relevant knowledge from their databases. Chekol and Stuckenschmidt (2018), for example, proposed a probabilistic temporal model for knowledge graphs to allow the recording of extraction dates and to support time travel queries.

According to Geiß and Gertz (2016), "there is an increasing need for approaches in information retrieval [...] to uniquely identify real-world persons based on mentions in text documents". The authors proposed a disambiguation model to conduct such identification with high precision using the Wikipedia Social Network, a network build by the same group of authors (Geiß et al., 2015) from a combination of *interwiki* links, Wikidata’s and Wikipedia's categories. The same approach was followed with toponyms, instead of person names, in Geiß et al. (2016). Geiß et al. (2018) also introduced a tool for named entity recognition in Wikidata, to provide a classification of its entities into the most predominantly used classes (Location, Person and Organization). And although not exclusive to Wikidata, Hempelmann et al. (2016) relied on this knowledge base to describe a method to automatically assign degrees of fuzzy set membership to individuals that have been asserted to several classes, although their experiments did not produce satisfactory results.

On another subject, Nielsen et al. (2017) developed Scholia, a tool to handle scientific bibliographic information through Wikidata, creating automatic profiles for researchers, organizations, journals, publishers, papers and topics, showing the potential both as a repository and as a tool to obtain scientometric statistics.

Cartograph, a visualization system developed by Sen et al. (2017), is another interesting effort related to information retrieval, as it aims to harness the knowledge encoded in Wikipedia (through Wikidata) to create thematic maps of almost any data, visualizing non-spatial data using geographic approaches.

4.3.4. Knowledge integration

In the early stages of Wikidata, Erxleben et al. (2014) noticed how Wikidata had "hardly been used in the semantic web community" in spite of Wikidata's potential. They suggested that the reason was the unavailability of data in RDF, so they discussed and developed RDF encodings for Wikidata, as well as implemented a tool for creating file exports.

In more recent years, Bergamin and Bacchi (2018) proposed and illustrated the possibility of restructuring the UNIMARC bibliographic records (a bibliographic ontology, after all) to convert them to the Wikidata data model. This would allow, among other benefits, the exploitation of the technical solutions and services implemented in Wikidata. Another paper by Nielsen (2018) describes efforts to link the ImageNet WordNet synsets and Wikidata to leverage both KGs for solving machine learning problems. Nielsen includes a promising application that would use Wikidata in an image classification setting.

Wikidata's information can also be linked to multimedia content such as the one available in Wikimedia Commons, as proposed in IMGpedia; Ferrada et al. (2018) presented a web interface to browse and explore the resulting dataset both in a user-friendly manner and allowing visuo-semantic queries that combine facts from Wikidata with visual similarity from IMGpedia.

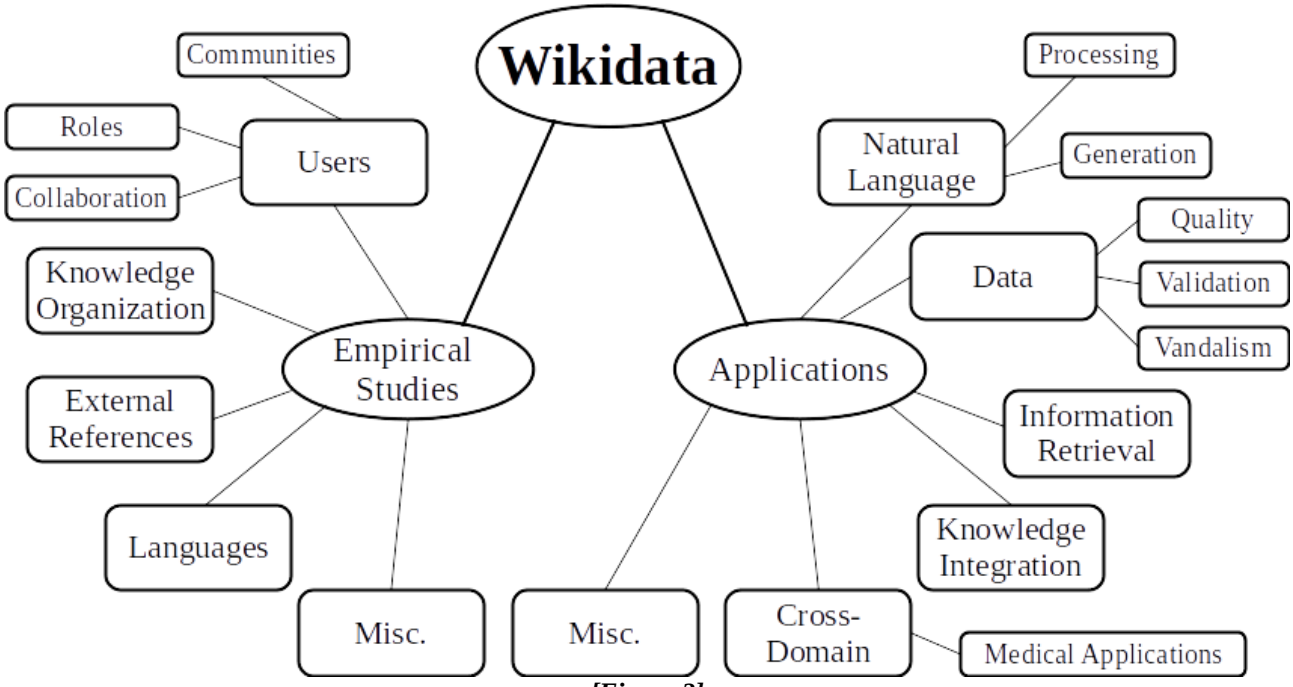
4.3.5. Medical applications

Wikipedia is an important source of medical information for both patients and medical professionals (Heilman et al., 2011), so improving it's quality and completeness could have a positive impact on global health. Four articles have been found that are directly linked to medicine (or *bioinformatics*) in our set. Pfundner et al. (2015) created a prototypical implementation of an automated system for keeping drug-drug interactions up to date in Wikipedia through their work on Wikidata, showing it as a viable option in the long-term. Burgstaller-Muehlbacher et al. (2016) created a fully open and extensible data resource for human and mouse molecular biology and biochemistry data using Wikidata as its semantic framework. Putman et al. (2016) developed a microbial specific data model, based on Wikidata, to represent microbial genomes and uploaded all the resulting content in Wikidata, as they consider it to be "a tremendous potential platform for managing the process of collaboratively understanding microbial genomics" as these are use cases where a lot of data from many fragmented sources must be collected and aggregated to unleash its potential and to obtain the full picture. In the same line of work, Putman et al. (2017) described a web application, called WikiGenomes and based

on Wikidata, to empower researcher's curation efforts of genomic data that integrates the resulting knowledge into Wikidata, enabling it to be accessed by anyone.

4.3.6. Miscellaneous

A single study has been found outside the scope of the five main applications, highlighting how there is still a lot more ground to be covered. In particular, Klein et al. (2016) took advantage of Wikidata's content to derive the “Wikidata Human Gender Indicators” (WHGI), a biographical dataset to monitor gender disparities across time, space, culture, occupation and language, "an approach not possible before Wikidata".



[Figure 2]

4.4. Who is leading Wikidata research?

Geographically speaking, studies on Wikidata are dominated by European researchers, who have been involved in 42 (73.7%) of the studies, followed by North American scientists (12 studies, 21.1%), with a total of 13 studies by the rest of the World combined. German researchers, in particular, have been involved in 17 studies; the other countries with more than five published works are the US (12), the UK (10), Italy (7) and France (6) .

Two institutions have been particularly prolific: 10 papers have been authored by researchers affiliated to the University of Southampton in the UK and 5 to the Université of Lyon in France. It is worth noting that both institutions have collaborated together in two publications (Kaffee et al., 2018; Vougiouklis et al., 2018). As could be expected, among the six researchers with more than three publications four are affiliated to the University of Southampton: Simperl (8), Kaffee (7), Piscopo (5) and Voigiouklis (4). The other two are Geiβ and Gertz, from the Heidelberg University, who co-authored four papers in our set.

A few non-academic institutions have also participated; Google and Wikimedia stand out due to their tight relationship to Wikidata. Google Inc. and Google Germany contribute with three publications, while Wikimedia researchers from various branches (Foundation, Research, Deutschland and Italia) worked in a total of five.

5. Discussion

The current review has shown how research on Wikidata is on the rise; we focused in surveying the existing empirical studies and the proposed applications for this rather recent knowledge database. A relatively large number of papers has been published in the recent years, specially in 2018, although the vast majority of them have been restricted to conferences and only a few are available in high impact journals. This is not only an indication of the novelty and technicality of the topic, but also a sign of Wikidata's research immaturity. This is further amplified by the fact that most of the papers are descriptions, proposals or implementations of applications, models or tools that take advantage of Wikidata's structure or knowledge graph, demonstrating how present efforts are mostly restricted to finding uses for Wikidata instead of conceptualizing its *raison d'être* or going further and deeper in some of its potential fields of application, which might bring new approaches and contribute to a real breakthrough in Wikidata's research, use and purpose.

In regard to empirical studies, the most relevant topics for the scientific community so far have been Wikidata's users (and their related edition practices, which impact in data quality), knowledge organization (or quality of the ontology itself), its links to external sources or references and Wikidata's multilingual affordances. As for applications, most of the works are dedicated to natural language (either in processing or generation), data quality and information retrieval. Such applications, however, are mainly reflexive; they are mostly limited to Wikidata itself (improving its data, expanding its capabilities or integrating more knowledge) and are rarely linked to disciplines outside information systems. Only a few timid efforts from the medical field have been captured by our review (as ontologies have particular relevance among the healthcare community), but this is only the tip of a much larger knowledge iceberg. Many other fields, such as education or the media, could greatly benefit from Wikidata.

Therefore, empirical work has been, so far, strongly focused in understanding the patterns, tendencies and characteristics of the editor community, although this work is limited by the lack of consideration of the social networks they form. The editorial process itself, which allows for participation of any user regardless of his or her familiarity with semantic technologies, should be investigated more thoroughly, as it might have a direct impact in the data quality but it should also influence the design of Wikidata and its interfaces. In the same manner as editor work could be aided by further research on user interface, Wikidata's usage could be enhanced by multimedia content; inclusion of visual and audio descriptors, image-based reasoning or expanding queries to include these categories could add countless applications to its knowledge base. Following the data quality line, task automation capabilities, with special focus in multilingual support, should be further developed, both to balance Wikidata's content and to predict and correct

undesired effects such as non-relevant references or vandalism. As Wikipedia grows larger and larger, ensuring that its content is trustworthy (or, at least, not deliberately false) will also become a larger problem; developing detection methods to help editors in their struggle to improve Wikidata content is crucial for its future success. Besides vandalism, this includes detection of incomplete or unbalanced data, language or demographic gaps, etc.

It is undeniable, however, that to have a sustained impact, it is vital that the Wikidata community carries the work done so far further in the structured world of Wikidata and integrates as much knowledge as possible from a variety of other sectors, which is another key for the long-term future success of Wikidata and its derived applications. Thus, there is still much work to do for it to become a viable source of integrated knowledge for users and practitioners. As of today, most of the existing research and, in particular, applications, are centered around another growing field: Natural Language Processing and Generation. However, this is not the only application that could greatly benefit from a large-scale integrated knowledge base; information extraction and retrieval, fact checking, content enrichment, recommendation systems, alert systems and others could as well. Wikidata also provides a framework that allows for collaborative work, aiding collective efforts to make sense of large amounts of data (such as the microbial genetic data in one of the included articles). In spite of this, only a few (and rather limited in scope) projects have been found that connect institutions with Wikidata, either sharing knowledge or creating value from the relationship. Wikidata is already showing its capabilities, but such applications need to be harnessed in order to unleash the full potential of Wikidata. The number of case studies is low and limited to test cases; there is still a long way for unlocking Wikidata for practitioners and institutions at a large scale. However, as our survey highlights, the included applications are mostly preliminary work, published in conferences and with many acknowledged limitations and future work to be conducted. Ultimately, there is still a notable gap between the rise of Wikidata's research and its translation to broader industrial or commercial applications.

Surprisingly, we have not found any works on one of the near-future potential problems of Wikidata. As Wikidata grows continuously larger, queries will likely become more resource demanding and using the whole network of information will soon become unfeasible. What techniques should be applied to filter the KG without losing its potential or information? What novel approaches should be introduced to deal with millions of items and a much larger number of links? As network analysis scales exponentially, and unless processing power follows the same progression, Wikidata's applications need to be ready.

Last, regarding who is leading Wikidata research, we found a large majority of European researchers, led by German researchers, which might be linked to the German origin of Wikidata (conceived by Wikimedia Deutschland), although more research from US researchers was expected (as Wikimedia is based in California). Two non-German institutions, however, have produced the most papers on the subject, University of Southampton in the UK and Université of Lyon in France. More worrying is the lack of research in the rest of the World, a distribution which mirrors the language

distribution of Wikidata items, with a disproportionate amount of English, Dutch, French, German, Italian, Spanish and even Swedish articles in comparison to their real-world speakers (Kaffee et al., 2017).

6. Conclusion

Despite the optimism around Wikidata's potential and the rise in research activity, the current systematic review shows how the field is still in the early stages. In summary:

- Most Wikidata research is published in conferences, demonstrating the immaturity of the field.
- Current works are biased towards Information Science in both empirical studies and related applications, with limited penetration in other disciplines.
- Empirical studies target editing behavior and profiles and data quality, but few provide empirical evidence of real use cases.
- Among the applications described in the literature, most are devoted to NLP. Other disciplines are under-explored.
- As of today, Wikidata's research is dominated by European researchers, mirroring Wikidata's content distribution. Worldwide research is needed to truly unleash Wikidata's potential.
- There is a significant gap between researchers and practitioners, emphasized by the lack of multidisciplinary applications. A few examples of works combining medicine with information systems have been found, but other knowledge-based fields such as education (e.g. e-learning) or media (e.g. news generation or fact checking) could easily benefit from Wikidata in the light of its potential.
- As people coordinate when editing Wikidata, forming communities (by language, interest, etc.), the relationship between their social network and the produced content should be taken into account in future research because it is relevant to understand editor behavior, user roles and influence on the knowledge (social) graph.

The current review has a number of limitations. As with most reviews, it is limited by the search terms used and the journals included in the manual search process, which target a specific set of journals and conference proceedings. This is consistent with the best practices (Kitchenham, 2004) of other researchers looking at research trends, but it also implies that we might have missed some relevant studies, in particular if they are published in journals or conferences outside our scope. Thus, our results must be qualified as applying only to Wikidata research articles published in the major international journals and conferences, providing a snapshot of empirical research on Wikidata which is representative of the state of the art at this time. Additionally, it could be appropriate to refine our classification of topics, either proposing a higher level classification or splitting some categories into sub-categories as required. In spite of this, the number of studies in some topics is still arguably too low as to give them a categorical entity.

References

Abián, D., Guerra, F., Martínez-Romanos, J. and Trillo-Lado, R. (2018), "Wikidata and DBpedia: A Comparative Study", *Semantic Keyword-Based Search on Structured Data Sources*, Springer, Cham, pp. 142–154.

Balaraman, V., Razniewski, S. and Nutt, W. (2018), "Recoin: Relative Completeness in Wikidata", *Companion Proceedings of the The WWW '18*, Geneva, Switzerland, pp. 1787–1792.

Benedetti, F., Beneventano, D., Bergamaschi, S. and Simonini, G. (2018), "Computing inter-document similarity with Context Semantic Analysis", *Information Systems*.

Bergamin, G. and Bacchi, C. (2018), "New ways of creating and sharing bibliographic information: an experiment of using the Wikibase Data Model for UNIMARC data", *JLIS.IT*, Vol. 9 No. 3, pp. 35–74.

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R. and Hellmann, S. (2009), "DBpedia - A crystallization point for the Web of Data", *Journal of Web Semantics*, Vol. 7 No. 3, pp. 154–165.

Bollacker, K., Evans, C., Paritosh, P., Sturge, T. and Taylor, J. (2008), "Freebase: A collaboratively created graph database for structuring human knowledge", *Proceedings of the SIGMOD '08*, pp. 1247–1250.

Brasileiro, F., Almeida, J.P.A., Carvalho, V.A. and Guizzardi, G. (2016), "Applying a Multi-Level Modeling Theory to Assess Taxonomic Hierarchies in Wikidata", *Proceedings of the 25th Int. Conf. on WWW*, Geneva, Switzerland, pp. 975–980.

Budgen, D. and Brereton, P. (2006), "Performing systematic literature reviews in software engineering", *Proceeding of the ICSE '06*, p. 1051.

Burgstaller-Muehlbacher, S., Waagmeester, A., Mitraka, E., Turner, J., Putman, T., Leong, J., Naik, C., et al. (2016), "Wikidata as a semantic framework for the Gene Wiki initiative", *DATABASE-THE JOURNAL OF BIOLOGICAL DATABASES AND CURATION*.

Chekol, M.W. and Stuckenschmidt, H. (2018), "Towards Probabilistic Bitemporal Knowledge Graphs", *Companion Proceedings of the The WWW '18*, Geneva, Switzerland, pp. 1757–1762.

Chisholm, A., Radford, W. and Hachey, B. (2017), "Learning to generate one-sentence biographies from Wikidata", *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, Vol. 1, pp. 633–642.

Cuong, T.T. and Müller-Birn, C. (2016), "Applicability of Sequence Analysis Methods in Analyzing Peer-Production Systems: A Case Study in Wikidata", *Social Informatics*, Springer, Cham, pp. 142–156.

English, S.M. (2018), "An Extensible Schema for Building Large Weakly-Labeled Semantic Corpora", *Procedia Computer Science*, Vol. 128, pp. 65–71.

Erxleben, F., Günther, M., Krötzsch, M., Mendez, J. and Vrandečić, D. (2014), "Introducing Wikidata to the Linked Data Web", *The Semantic Web – ISWC 2014*, Springer, Cham, pp. 50–65.

Ferrada, S., Bravo, N., Bustos, B. and Hogan, A. (2018), "Querying Wikimedia Images Using Wikidata Facts", *Companion Proceedings of the The Web Conference 2018*, Geneva, Switzerland, pp. 1815–1821.

Geiß, J. and Gertz, M. (2016), "With a Little Help from My Neighbors: Person Name Linking Using the Wikipedia Social Network", *Proceedings of the 25th International Conference Companion on WWW*, Geneva, Switzerland, pp. 985–990.

Geiß, J., Spitz, A. and Gertz, M. (2015), "Beyond Friendships and Followers: The Wikipedia Social Network", *Proceedings of the ASONAM '15*, ACM, New York, NY, USA, pp. 472–479.

Geiß, J., Spitz, A. and Gertz, M. (2018), "NECKAr: A named entity classifier for wikidata", *Lecture Notes in Computer Science*, Vol. 10713 LNAI, pp. 115–129.

Hall, A., Terveen, L. and Halfaker, A. (2018), “Bot Detection in Wikidata Using Behavioral and Other Informal Cues”, *Proc. ACM Hum.-Comput. Interact.*, ACM, New York, NY, USA, Vol. 2 No. CSCW, p. 64:1--64:18.

Heilman, J.M., Kemmann, E., Bonert, M., Chatterjee, A., Ragar, B., Beards, G.M., Iberri, D.J., et al. (2011), “Wikipedia: a key tool for global public health promotion.”, *Journal of Medical Internet Research*, Vol. 13 No. 1, p. e14.

Heindorf, S., Potthast, M., Stein, B. and Engels, G. (2016), “Vandalism Detection in Wikidata”, *Proceedings of the CIKM ’16*, ACM, New York, NY, USA, pp. 327–336.

Hempelmann, C.F., Petrenko, M. and Matthews, G. (2016), “Automatic discovery of degrees of fuzzy set membership in ontologies”, *2016 Annual Conference of the NAFIPS*, pp. 1–5.

Hernández, D., Hogan, A., Riveros, C., Rojas, C. and Zerega, E. (2016), “Querying Wikidata: Comparing SPARQL, Relational and Graph Databases”, *The Semantic Web – ISWC 2016*, Springer, Cham, pp. 88–103.

Hollink, L., van Aggelen, A. and van Ossenbruggen, J. (2018), “Using the Web of Data to Study Gender Differences in Online Knowledge Sources: The Case of the European Parliament”, *Proceedings of the 10th ACM Conf. on Web Science*, ACM, New York, NY, USA, pp. 381–385.

Ingvaldsen, J.E. and Gulla, J.A. (2015), “Taming news streams with linked data”, *2015 IEEE 9th Int. Conf. on RCIS*, pp. 536–537.

Kaffee, L.-A., Elsahar, H., Vougiouklis, P., Gravier, C., Laforest, F., Hare, J. and Simperl, E. (2018), “Mind the (Language) Gap: Generation of Multilingual Wikipedia Summaries from Wikidata for ArticlePlaceholders”, *The Semantic Web*, Springer, Cham, pp. 319–334.

Kaffee, L.-A., Piscopo, A., Vougiouklis, P., Simperl, E., Carr, L. and Pintscher, L. (2017), “A Glimpse into Babel: An Analysis of Multilinguality in Wikidata”, *Proceedings of the OpenSym ’17*, ACM, New York, NY, USA, p. 14:1--14:5.

Kaffee, L.-A. and Simperl, E. (2018), “Analysis of Editors’ Languages in Wikidata”, *Proceedings of the OpenSym ’18*, ACM, New York, NY, USA, p. 21:1--21:5.

Kitchenham, B. (2004), *Procedure for Undertaking Systematic Reviews, Computer Science Department, Keele University and National ICT Australia Ltd, Joint Technical Report*.

Klein, M., Gupta, H., Rai, V., Konieczny, P. and Zhu, H. (2016), “Monitoring the Gender Gap with Wikidata Human Gender Indicators”, *Proceedings of the OpenSym ’16*, ACM, New York, NY, USA, p. 16:1--16:9.

Leva, F. and Chemello, M. (2018), “The effectiveness of a Wikimedian in permanent residence: the BEIC case study”, *JLIS.IT*, Vol. 9 No. 3, pp. 141–147.

Lim, K.H., Karunasekera, S., Harwood, A. and Falzon, L. (2017), “Spatial-based topic modelling using wikidata knowledge base”, *2017 IEEE Int. Conf. on Big Data*, pp. 4786–4788.

Müller-Birn, C., Karran, B., Lehmann, J. and Luczak-Rösch, M. (2015), “Peer-production System or Collaborative Ontology Engineering Effort: What is Wikidata?”, *Proceedings of the OpenSym ’15*, ACM, New York, NY, USA, p. 20:1--20:10.

Murase, Y., Koichiro, Y. and Nakamura, S. (2019), “Associative knowledge feature vector inferred on external knowledge base for dialog state tracking”, *Computer Speech & Language*, Vol. 54, pp. 1–16.

Nielsen, F.Å. (2018), “Linking ImageNet WordNet Synsets with Wikidata”, *Companion Proceedings of the The WWW ’18*, Geneva, Switzerland, pp. 1809–1814.

Nielsen, F.Å., Mietchen, D. and Willighagen, E. (2017), “Scholia, Scientometrics and Wikidata”, *The Semantic Web: ESWC 2017 Satellite Events*, Springer, Cham, pp. 237–259.

Okoli, C. (2015), “A Guide to Conducting a Standalone Systematic Literature Review”, *Communications of the Association for Information Systems*, Vol. 37.

Olivieri, A.C., Shabani, S., Sokhn, M. and Cudré-Mauroux, P. (2017), “Assessing data veracity through domain specific knowledge base inspection”, *2017 ICACSIS*, pp. 291–296.

Pellissier Tanon, T., de Assunção, M.D., Caron, E. and Suchanek, F.M. (2018), “Demoing Platypus – A Multilingual Question Answering Platform for Wikidata”, *The Semantic Web: ESWC 2018 Satellite Events*, Springer, Cham, pp. 111–116.

Pellissier Tanon, T. and Kaffee, L.-A. (2018), “Property Label Stability in Wikidata: Evolution and Convergence of Schemas in Collaborative Knowledge Bases”, *Companion Proceedings of the The WWW ’18*, Geneva, Switzerland, pp. 1801–1803.

Pellissier Tanon, T., Vrandečić, D., Schaffert, S., Steiner, T. and Pintscher, L. (2016), “From Freebase to Wikidata: The Great Migration”, *Proceedings of the WWW ’16*, Geneva, Switzerland, pp. 1419–1428.

Pfundner, A., Schnoeberg, T., Horn, J., Boyce, R.D. and Samwald, M. (2015), “Utilizing the Wikidata System to Improve the Quality of Medical Content in Wikipedia in Diverse Languages: A Pilot Study”, *JOURNAL OF MEDICAL INTERNET RESEARCH*, Vol. 17 No. 5.

Piscopo, A., Kaffee, L.-A., Phethean, C. and Simperl, E. (2017), “Provenance Information in a Collaborative Knowledge Graph: An Evaluation of Wikidata External References”, *The Semantic Web – ISWC 2017*, pp. 542–558.

Piscopo, A., Phethean, C. and Simperl, E. (2017), “What Makes a Good Collaborative Knowledge Graph: Group Composition and Quality in Wikidata”, *Social Informatics*, Springer, Cham, pp. 305–322.

Piscopo, A. and Simperl, E. (2018), “Who Models the World?: Collaborative Ontology Creation and User Roles in Wikidata”, *Proc. ACM Hum.-Comput. Interact.*, ACM, New York, NY, USA, Vol. 2 No. CSCW, p. 141:1--141:18.

Piscopo, A., Vougiouklis, P., Kaffee, L.-A., Phethean, C., Hare, J. and Simperl, E. (2017), “What Do Wikidata and Wikipedia Have in Common?: An Analysis of Their Use of External References”, *Proceedings of the OpenSym ’17*, ACM, New York, NY, USA, p. 1:1--1:10.

Prasojo, R.E., Darari, F., Razniewski, S. and Nutt, W. (2016), “Managing and Consuming Completeness Information for Wikidata Using COOL-WD”, *COLD@ISWC*.

Putman, T.E., Burgstaller-Muehlbacher, S., Waagmeester, A., Wu, C., Su, A.I. and Good, B.M. (2016), “Centralizing content and distributing labor: a community model for curating the very long tail of microbial genomes”, *DATABASE-THE JOURNAL OF BIOLOGICAL DATABASES AND CURATION*.

Putman, T.E., Lelong, S., Burgstaller-Muehlbacher, S., Waagmeester, A., Diesh, C., Dunn, N., Munoz-Torres, M., et al. (2017), “WikiGenomes: an open web application for community consumption and curation of gene annotation data in Wikidata”, *DATABASE-THE JOURNAL OF BIOLOGICAL DATABASES AND CURATION*.

Ringler, D. and Paulheim, H. (2017), “One Knowledge Graph to Rule Them All? Analyzing the Differences Between DBpedia, YAGO, Wikidata & co.”, *KI 2017*, Springer, Cham, pp. 366–372.

Sáez, T. and Hogan, A. (2018), “Automatically Generating Wikipedia Info-boxes from Wikidata”, *Companion Proceedings of the The WWW ’18*, Geneva, Switzerland, pp. 1823–1830.

Samuel, J. (2018), “Analyzing and Visualizing Translation Patterns of Wikidata Properties”, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Springer, Cham, pp. 128–134.

Sarabadani, A., Halfaker, A. and Taraborelli, D. (2017), “Building Automated Vandalism Detection Tools for Wikidata”, *Proceedings of the WWW ’17*, Geneva, Switzerland, pp. 1647–1654.

Sen, S., Swoap, A.B., Li, Q., Boatman, B., Dippenaar, I., Gold, R., Ngo, M., et al. (2017), “Cartograph: Unlocking Spatial Visualization Through Semantic Enhancement”, *Proceedings of the IUI ’17*, ACM, New York, NY, USA, pp. 179–190.

Spitz, A., Geiß, J. and Gertz, M. (2016), “So Far Away and Yet So Close: Augmenting Toponym Disambiguation and Similarity with Text-based Networks”, *Proceedings of the GeoRich ’16*, ACM, New York, NY, USA, p. 2:1--2:6.

Steiner, T. (2014a), “Bots vs. Wikipedians, Anons vs. Logged-ins”, *Proceedings of the WWW ’14*, ACM, New York, NY, USA, pp. 547–548.

Steiner, T. (2014b), “Bots vs. Wikipedians, Anons vs. Logged-Ins (Redux): A Global Study of Edit Activity on Wikipedia and Wikidata”, *Proceedings of the OpenSym ’14*, ACM, New York, NY, USA, p. 25:1--25:7.

Stinson, A.D., Fauconnier, S. and Wyatt, L. (2018), “Stepping Beyond Libraries: The Changing Orientation in Global GLAM-Wiki”, *JLIS.IT*, Vol. 9 No. 3, pp. 16–34.

Ta, T.H. and Anutariya, C. (2015), “A Model for Enriching Multilingual Wikipedias Using Infobox and Wikidata Property Alignment”, *Semantic Technology*, Springer, Cham, pp. 335–350.

Thakkar, H., Endris, K.M., Gimenez-Garcia, J.M., Debattista, J., Lange, C. and Auer, S. (2016), “Are Linked Datasets Fit for Open-domain Question Answering? A Quality Assessment”, *Proceedings of the WIMS ’16*, ACM, New York, NY, USA, p. 19:1--19:12.

Turki, H., Vrandečić, D., Hamdi, H. and Adel, I. (2017), “Using WikiData as a Multi-lingual Multi-dialectal Dictionary for Arabic Dialects”, *2017 IEEE/ACS 14th Int. Conf. on Computer Systems and Applications*, pp. 437–442.

Vagliano, I., Monti, D., Scherp, A. and Morisio, M. (2017), “Content Recommendation Through Semantic Annotation of User Reviews and Linked Data”, *Proceedings of the K-CAP 2017*, ACM, New York, NY, USA, p. 32:1--32:4.

Vougiouklis, P., Elsahar, H., Kaffee, L.-A., Gravier, C., Laforest, F., Hare, J. and Simperl, E. (2018), “Neural Wikipedian: Generating Textual Summaries from Knowledge Base Triples”, *Journal of Web Semantics*, Vol. 52–53, pp. 1–15.

Vrandečić, D. (2013), “The Rise of Wikidata”, *IEEE Intelligent Systems*, Vol. 28 No. 4, pp. 90–95.

Vrandečić, D. and Krötzsch, M. (2014), “Wikidata: A Free Collaborative Knowledgebase”, *Commun. ACM*, ACM, New York, NY, USA, Vol. 57 No. 10, pp. 78–85.

Webster, J. and Watson, R.T. (2002), “Analyzing the Past to Prepare for the Future: Writing a Literature Review”, *MIS Quarterly*, Vol. 26 No. 2, pp. xiii--xxiii.

Yang, X., Ren, S., Li, Y., Shen, K., Li, Z. and Wang, G. (2018), “Relation linking for wikidata using bag of distribution representation”, *Lecture Notes in Computer Science*, Vol. 10619 LNAI, pp. 652–661.

Zangerle, E., Gassler, W., Pichl, M., Steinhauser, S. and Specht, G. (2016), “An Empirical Evaluation of Property Recommender Systems for Wikidata and Collaborative Knowledge Bases”, *Proceedings of the OpenSym ’16*, ACM, New York, NY, USA, p. 18:1--18:8.

1 https://stats.wikimedia.org/wikispecial/EN/TablesWikipediaWIKIDATA.htm
2 https://stats.wikimedia.org/wikispecial/EN/ChartsWikipediaWIKIDATA.htm