WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

Bachelor Thesis

How scientific are Wikidata's external references?

Mag. Marco Marsoner

Date of Birth: 27.04.1993
Student ID: h01207119

Subject Area: Information Business

Studienkennzahl: 033/561

Supervisor: Univ.-Prof. Dr. Axel Polleres

Date of Submission: 4th of December 2022

Department of Information Systems and Operations, Vienna University of Economics and Business, Welthandelsplatz 1, 1020 Vienna, Austria

DEPARTMENT FÜR INFORMATIONSVERARBEITUNG UND PROZESSMANAGEMENT DEPARTMENT OF INFORMATION SYSTEMS AND OPERATIONS

Contents

List of Figures

Abstract

Wikidata is a rapidly growing user edited open knowledge-graph. It provides easy access to structured data all over the globe. Since Wikidata allows contradictory information, references are very important to support statements and track the source of an information. We examined the scientific value of references that point towards information outside of Wikimedia projects. After conducting a literature review on Wikidata reference analysis, we extracted a sample of Wikidata reference properties and references via the WDQS SPARQL endpoint. We classified the sampled references based on identifiers derived from related work and Wikidata initiatives. Our results show that external references are rather scientific, reference properties are not consistently used throughout Wikidata and that Wikidata items have no links, relink towards other items, or might need renumbering.

"*Wikidata is a free, collaborative, multilingual, secondary database, collecting structured data to provide support for Wikipedia, Wikimedia Commons, the other wikis of the Wikimedia movement, and to anyone in the world.*"
[Wikimedia, 2019]

1 Introduction & Motivation

It is one of the main projects from the Wikimedia foundation. Wikidata was launched 2012 [Wikimedia, 2022h] and celebrates its 10th birthday this year in October. It is one of the most important sources of structured data in the form of a free and open knowledge database on the internet [Wikimedia, 2019]. Wikidata acts as "*central storage for of its Wikimedia sister projects including Wikipedia, Wikivoyage, Wiktionary, Wikisource, and others.*" [Wikimedia, 2019] Like other Wikimedia foundation projects, the content is created collaboratively, available to anyone and based on the software MediaWiki, which was initially developed for Wikipedia [Wikimedia, 2022b]. Consequently, Wikidata provides a huge potential to access (huge amounts of) data easily all around the world and not just to the other Wikimedia projects (e.g. providing information to the info boxes of Wikipedia [Wikimedia, 2022i]). To simplify integration, Wikidata entities are linked to entries in several digital libraries.

Wikidata is often described as a knowledge graph, with its information being accessible encoded in the RDF (Resource descriprtion format) format, which encodes Wikidata statements about items as subject-predicate-object triples.. Wikidata is a secondary database and does not itself entail 'primary' knowledge. Therefore, the Wikidata Knowledge graph, just like pages in Wikipedia, relies on references to support the entailed claims and statements. References should point to the source which is the origin for the provided statement. Statements are supported by and linked to at least one source according to the internal Wikidata guidelines [Wikimedia, 2022a].

Wikidata expands quickly starting 2013 to 42.3 million items by the end of 2017b [Wikimedia, 2022d] to about 100 million items in 2022 [Wikimedia, 2022c]. Due to this expansion, more emphasis is put on ensuring the quality of its data. Wikidata aims to cover a wide range of topics through user collaborations. Since the content is primarily created and edited by users, Wikidata references should be relevant (supporting the claim), authoritative and accessible according to their policies [Wikimedia, 2022e]. Furthermore, the referenced sources should provide context and supportive arguments for statements. However, the evaluation of references itself is the responsibility of the

Wikidata user community. Many researchers and practitioners investigated different features and chracteristics of Wikidata. Most of the work around Wikidata focuses on the quality of the sources, as this is a major issue for Wikidata. Adequate, relevant and trustworthy references are becoming more and more important nowadays, and they could improve the reputation of Wikimedia projects in general. Missing sources or inappropriate references can affect the reliability and prevent the reuse of data.

Along these lines, the research interest of this thesis lies in the analysis of the quality and structure of the external references within Wikidata, especially concerning their scientific character and background. The aim is to address this topic via an algorithmic approach that can be used to automatically export and analyse Wikidata references. Our approach will be building on findings gained in various publications regarding this topic. Albeit, there have been numerous studies on Wikidata and its references, no study has solely examined the scientific background of sources in more detail with a special focus on the used properties and identifiers. The evaluation of Wikidata sources is broad. Compared to previous studies, the approach proposed in this thesis is able to identify and assess the scientific references and citations and takes input from Wikidata's Source Meta Data [Wikidata, 2022e] project into consideration.

This thesis is structured as follows: Firstly, there is a short chapter on the structure of Wikidata titled 'Background'. The research questions are described and the methods used to answer them are presented in Section 3. Additionally, the theoretical approach and practical implementation of the methodology are explained. Secondly, insights into the current state of research on the topic are provided. In doing so, relevant and current literature and existing research are taken into account: The paper presents existing research on algorithm-based (reference) analysis of Wikidata entries. Thirdly, we try to develop an own approach for the export of external references extracting a sample of reference nodes using the Wikidata Query Service (WDQS, `query.wikidata.org`) endpoint and examine the scientific character of the extracted references based on well-defined characteristics/requirements. Next, we conduct an analysis of all used reference properties and examine the sampled reference nodes in more detail. Finally, we point out emerged issues and findings during our analysis which could be the input for further research. At the end, the findings of the thesis are summarised in a conclusion.

2 Background

As already mentioned, Wikidata follows the resource description framework (RDF) format. This means information is encoded via statements which are subject-predicate-object triples. These triples are formed via items, properties and values in Wikidata.

Items represent either real or surreal things, people or concepts and properties postulate relationships. Both are uniquely identified by Uniform Resource Identifiers (URIs). The labels of item-URIs start with the letter 'Q' followed by a number and property-URIs begin with a 'P' followed by a number. Values are most commonly items but can also have other values for instance numbers.

A claim is the combination of a property with at least one value and a qualifier that provides information on an item. If this property-value pair is enriched by additional information (e.g. references, ranks), the result forms a statement. A claim without a qualifier is called a 'snak', which basically represents a basic triple consisting of an item and a property-value pair.

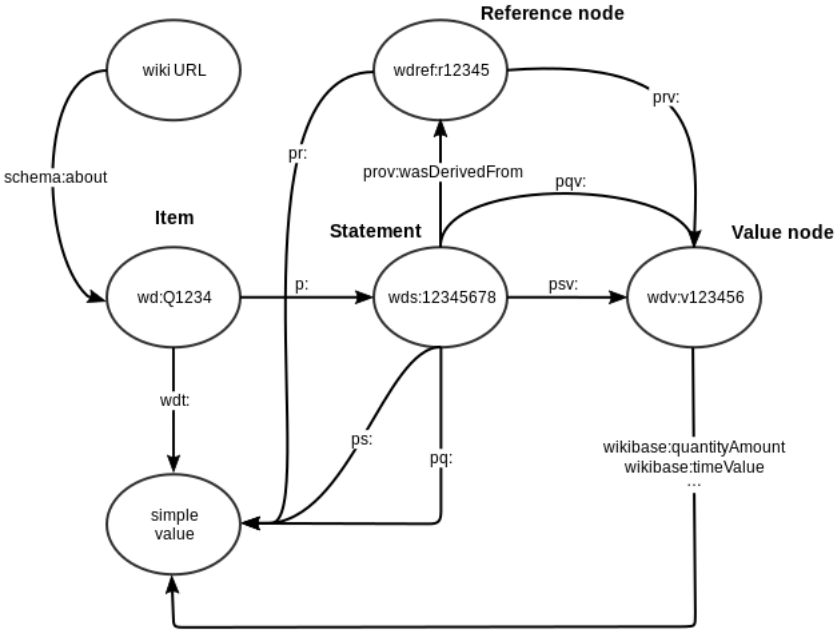

Figure 1: RDF mapping vector [Schönitzer,Michael F., 2017]

References are appended to statements as additional information. The reference node is connected to a value. This allows for many different references for one statement. Wikidata allows contradicting statements in case of

controversial, uncertain or debatable information. Therefore, the reference has to support the whole statement. This structure can be illustrated in many ways. However, Figure 1 shows the original structure which is provided in the documentation of Wikidata [Wikidata, 2022a]. Figure 2 shows an example based on Gustav Klimt inspired by a graphic found in [Amaral et al., 2021].

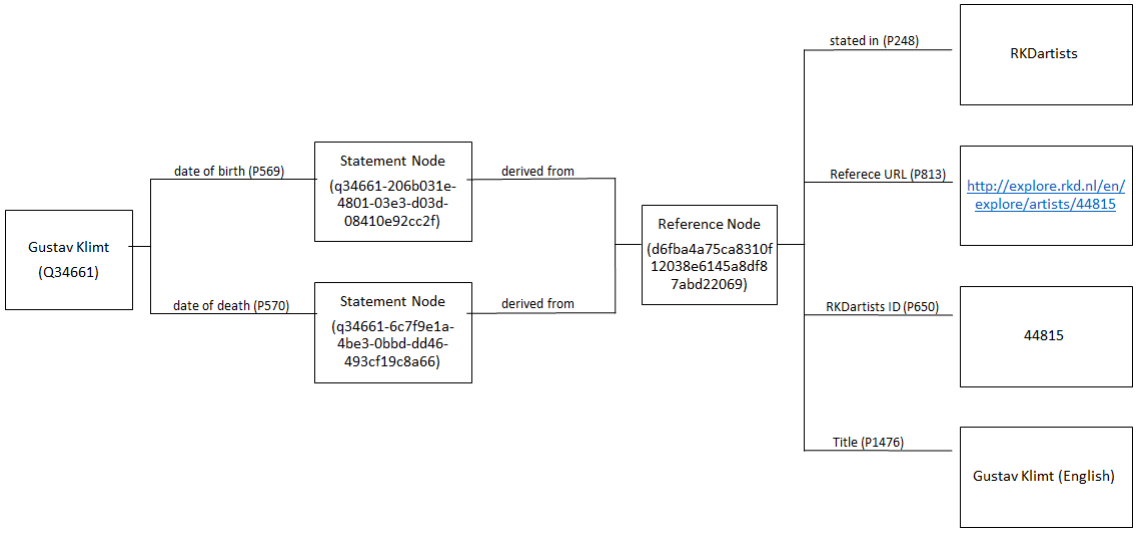

Figure 2: Wikidata item connected with information via statements (relationships) and supported by references. Inspired by [Amaral et al., 2021]

The graphics illustrate both the human-readable representation of the underlying statements about Gustav Klimt (Q34661[1]) as well as the abstract structure of the Wikidata data model. The reference nodes entail a variety of different properties. However, reference nodes themselves can be identified by the 'derived from'-connection to the respective statement. Reference nodes have their own identifier, which are hashes. These hashes are the same if a reference node entails the identical properties and values [Wikidata, 2022a]. Furthermore, a reference node can be connected to multiple statement nodes [Wikidata, 2022a]. Both findings are illustrated by the provided example in Figure 2.

There are no binding guidelines on how and when to use those reference properties. Nevertheless, there is an initiative to unify and clarify the reference usage across Wikimedia projects (not only Wikidata). This project is called 'Wikicite' [MediaWiki, 2022a]. The goal of Wikicite is "*to develop open citations and linked bibliographic data to serve free knowledge. WikiCite is a series of conferences and workshops in support of that*

[1]https://w.wiki/6A5Y

goal"[MediaWiki, 2022a]. One project of this initiative is the 'WikiProject Source Metadata' (see [Wikidata, 2022e]). The aim of the project is to: "

- *to act as a hub for work in Wikidata involving citation data and bibliographic data as part of the broader WikiCite initiative.*
- *to define a set of properties that can be used by citations, infoboxes, and Wikisource. [..]*
- *to create a large open bibliographic database within Wikidata.* [Wikidata, 2022e]

This project aims at building clear structures and boundaries within the Wikimedia citations. In order to achieve these goals they have several ongoing activities. One of the main project sub-streams focuses on scholarly articles in the Wikidata knowledge graph called 'Bibliographic metadata for scholarly articles in Wikidata' [Wikidata, 2022f]. The project stream came up with a template that includes a list of properties with regard to scientific literature [Wikidata, 2022d]. This list entails identifiers for scholarly articles, scholarly journals, proceedings, proceedings series, supplements, theses, books, authors, publishers, founders and others [Wikidata, 2022d]. This list serves as main identifier list because it entailed all previously mentioned identifiers and many more. Furthermore, this list was set up by a project team including Wikimedia employees [Wikidata, 2022f] with the aim to identify and enter scholarly information on Wikidata. The number of properties will be explained in more detail later on.

3 Our Approach

The research questions aim to fill the research gap of analysing the scientific value of external Wikidata references with an algorithmic based approach including data extracted via SPARQL queries over the WDQS endpoint. Based on this objective and the overall research question of this thesis "How scientific are Wikidata's external references" from the title, we derived the following four sub-questions:

1. What is the current state of research on algorithmic based extraction and analysis of Wikidata (external) references?
2. How are reference properties used within the Wikidata knowledge graph?
3. How can references be characterised as 'scientific' within Wikidata?

4. How can we assess the quality of such scientific references within Wikidata?

The outlined research questions will be answered using two research methods: Firstly a systematic literature review to examine the current state of research on the topic (Question 1). Secondly, an algorithm will be developed to query and extract references via the WDQS endpoint and analyse them afterwards (Questions 2, 3 and 4). We relied on a sample to look at the research questions. The sampling methodology will be described along the description of the algorithm and extraction. After the extraction, we did some manual checks to validate the extracted data. This procedure will also be explained within the algorithm section. Furthermore, the exported data set was parsed and classified using Python.

4 Related Work

We conducted a literature review based on the approach put forth by [Webster and Watson, 2002]. According to this, we followed the following steps to identify relevant literature in the field of information systems:"

1. *The major contributions are likely to be in leading journals. It makes sense, therefore to start with them. Look through journal databases to accelerate the identification of relevant articles[..].*

2. *Go backward by reviewing the citations for the articles identified in step 1 to determine prior articles you should consider.*

3. *Go forward [..] to identify articles citing the key articles identified in the previous steps. Determine which of these articles should be included in the review.*" [Webster and Watson, 2002]

4.1 Implementation

Step 1 includes looking through 'leading journals'. We implemented this search by querying key words in three sources or databases: ABI/Inform Global | T&I (ProQuest), INFODATA Informationswissenschaft (IDAT) and Google Scholar. Those databases cover a wide variety of journals including relevant literature for the outlined subject. We selected those resources based on the two factors availability and relevance. Therefore, we identified accessible data sources and databases within our resources. Next, we searched for free available data sources to widen our selection list with potentially

relevant sources. Thirdly, we applied a crucial selection criterion to this list of data sources. This criterion was research relevance for one of the following topics: information systems, information technology or research conduct. Consequently, we concluded that the three selected sources would be the best fit to cover all relevant literature to answer the research questions.

The above mentioned sources were searched based on structured search queries. The aim was to maximize the relevance of the output. On the one hand, every search included the keyword 'Wikidata'. It is the main theme of the paper and the most important factor to identify relevant papers. On the other hand, our search consisted of a word which could be potentially used to describe or name references. The following words were used as synonyms: 'reference', 'citation', 'quotation', 'quote', 'source' and 'remark'. The search of these key words was not limited to the the title. Therefore, an occurrence in the abstract or main text was considered relevant as well. The queries brought up more than ten thousand results.

There were still some potentially irrelevant papers, which were filtered using four criteria (language, length, publication date and relevance). Only papers in English were included in the next search queries to exclude papers that cannot be read by the authors and are highly likely not in leading journals. In addition, publications that did not exceed five pages were not included in the outcome of the queries. Those papers were considered as too short (e.g. seminar or conference papers that are not peer-reviewed or cannot outline their approach in just five pages) and classified as irrelevant. On the one hand, there will be a backward search of the references of the examined papers (which would include papers which are not five pages or longer) and, on the other hand, Wikidata was established in 2012 [Wikimedia, 2022h]. Consequently, we limited the outcome of our literature search by publication date and decided on 2019 based on the publication counter of Scholia [Scholia, 2022]. The remaining list of articles, consisting of about 300 papers, was then roughly assessed based on the topic outlined in the abstract and/or match of the keywords and ranked by their expected adequacy to the research questions. Furthermore, redundancy was avoided so no article or paper was included more than once in the final list of papers which consisted of eight publications.

After reading the abstract of the identified articles, more possibly relevant papers were identified and added to the literature list. However, the list was then heavily limited to papers that included a clear description of the applied methodology (e.g. the used algorithm) and focused on Wikidata references (instead of other Wikimedia projects e.g. Wikipedia). A classification methodology for the references was not an essential critierion. Hence, the ultimate outcome of the literature review with similar task and

approaches were four papers. Those were identified as highly relevant for the examination of the outlined research questions with which the outcome of the own approach will be compared.

In addition to the four Wikidata related papers, we included a fifth non-Wikidata but Wikipedia related publication which was used to identify classification criteria for scientific references. The paper was chosen after we identified a gap regarding this issue in the literature on Wikidata. The paper was chosen after applying a simple variation of the already described approach.

We queried the aforementioned databases with three keywords: 'Wikipedia', 'identifiers' in conjunction with a synonym for 'reference' (see above). The outcome was filtered in accordance with the same criteria as for Wikidata (English, more than five pages, publication not before 2019, relevance). The top five listed papers of all databases were examined and judged based on their relevance and desired input in quantity (number of classification criteria) and quality (description and justification of the classification criteria). Furthermore, we had a look at the amount of work that referenced the respective article. The ultimately chosen paper was listed in the top spot in Google Scholar for the word 'citation'.

4.2 Identified relevant previous publications

As outlined above, we classified four papers as relevant previous work for the chosen research field. These papers provide useful starting points, insights and guidance for the development of the approach. The insights taken into account concerned the structure of the data source (Wikidata), classifiers for scientific references and as comparision basis (e.g. diverging results). Accordingly, we provide a brief summary of the concerned papers starting with the four Wikidata publications listed chronologically and followed by an article on measuring scientific citations within Wikipedia, including their data source, code and identifiers to classify a source as scientific:

1. [Piscopo et al., 2017b]: The authors compare the use of external references between Wikipedia and Wikidata. For this purpose, they extracted all external references from Wikidata via the Wikidata SPARQL on the 16th of April 2017 endpoint with the help of a team member the Wikidata development team at Wikimedia Deutschland, which is a national organization to promote the interests of Wikimedia in Germany (see also `https://en.wikipedia.org/wiki/List_of_Wikimedia_chapters`. Otherwise, they would have had to deal with the restrictions of the WDQS endpoint (e.g. 2 minute time-out barrier, limit of requests

within a given time). The Wikipedia articles were downloaded and the respective HTML code was parsed with Python and references wered identified based on the <ref></ref> tags. The resulting CSV files (Wikidata and Wikipedia references) were compared and plotted with Python packages. The code is publicly available (see `https://github.com/pvougiou/Wikidata-Referencing`). The paper does not tackle the topic regarding the scientific character of external references within Wikidata. Hence, there are no identifiers or advised properties which can be used for the classification analysis. However, the paper gives a good overview of the data structure of Wikidata articles which will also be described in more detail later on. The authors point out that while Wikipedia advises secondary sources because it does not want to entail original research, Wikidata as secondary database desires primary sources (see [MediaWiki, 2022b] for Wikipedia and [Wikimedia, 2022e] for Wikidata citation guidelines) .

2. [Piscopo et al., 2017a]: This paper by the same main author as the previous mentioned work (Alessandro Piscopo) analyzes the relevance and authoritativeness ([Wikimedia, 2022e]) of Wikidata references which are the only requirements for sources. Relevance means that the reference "*must provide evidence for the claim it is linked to*"[Piscopo et al., 2017a]. Authoritativeness refers to sources that are "*deemed trustworthy, up-to-date, and free of bias for supporting a particular statement on Wikidata*"[Wikimedia, 2022e]. Additionally, the authors trained a machine learning model (Random Forest, Naive Bayes and Support Vector Machine) to predict the two properties. The paper provides a good description and introduction to Wikidata with a focus on references and their data structure. The authors used a Wikidata dump from the 1st of Ocotber 2016 as data source for their examination. The data set included the editing history and entailed all English Wikidata entries. Using a sample from this dump, a crowd-sourced evaluation was performed to use the judgements regarding relevance and authoritativeness of the participants as basis to train a machine learning model. The crowdsourcing tasks included that the participants had to judge the relevance and authoritativeness of the sampled sources. The tasks were completed by the participants using a questionnaire. The aim of this model was to predict the reference quality across Wikidata based on the factors that the crowdsourcing workers did identify:

   - "*URL reference uses*
   - *Domain reference uses*

- *Source HTTP code*
- *Statement property*
- *Statement item*
- *Statement object*
- *Subject parent class*
- *Property parent class*
- *Object parent class*" [Piscopo et al., 2017a]

The scope of the paper was limited to English references and external URL references (Wikidata Property 'P852'). In order to identify English references, the top-level domains of the extracted URLs were filtered according to a list of endings. The paper does not directly tackle the topic of the scientific character of the exported references. However, the classification system to judge the authoritativeness of sources does include 'academic and scientific organisation' because the Wikidata guidelines state that publications of those institutions are authoritative [MediaWiki, 2022b]. Despite this classification criterion, the paper does not entail further information on how those organisations are identified. Furthermore, museums and libraries are excluded from this classification with no explanation. Table 7 of the paper illustrates the publisher type for all external references (URLs) on Wikidata. According to the authors, 12.4% of all references were published by academic & research institutions, 11.2% by other academic organisations and 0.2% by academic publishers [Piscopo et al., 2017a]. The referenced Github repository (`https://github.com/Aliossandro/WD_references_analysis`) entails the code and data of this approach. However, the classification of sources was performed by the crowdsourcing workers. Consequently, this literature piece does not provide clear identifiers or advised properties which can be used for the classification analysis. The taken approach which focuses on some sequences within the domain (e.g. '.ac') was incorporated into the methodology. But, the authors used this approach to filter their sample in this paper.

3. [Nielsen et al., 2017]: The authors developed a tool named 'Scholia'. The Python source code for the project is publicly available at `https://github.com/WDscholia/scholia` and the web service can be reached via the following URL: `https://scholia.toolforge.org/`). The article explains how the tool works. Scholia accesses the WDQS endpoint via SPARQL queries. The purpose of the tool is to "*create on-the-fly*

scholarly profiles for researchers, organizations, journals, publishers, individual scholarly works, and for research topics" [Nielsen et al., 2017]. Besides the functionality, the basic structure of Wikidata and the contained references and author information is described. The properties used for scientific publications are also explained. In addition to author-specific reference properties, the focus of the evaluations is on the identification of scientific literature in Wikidata overall using the 'instance of' property (wdt:P31) in combination with the value 'scientific article' (wd:Q13442814). Some descriptive statistics on bibliographic content (amount of scientific articles, authors etc.) and the WDQS queries that were used to get the data. An informative table on the scientific content within Wikidata is Figure 3, according to which about 2.3 million scientific articles (with the combination 'instance of' 'scientific article') existed in Wikidata in 2017. However, no other identifiers were considered in this evaluation, although other possibly helpful reference properties (DOI, PMID, PMCID, arXiv, ORCID, Google Scholar, VIAF, Crossref, funder ID, ZooBank and Twitter) are listed. Therefore, one might assume that there are many more scientific articles which do not entail this property-value pair (P31-Q13442814). Furthermore, this query is not aimed at reference nodes but on statements in general, meaning that there are about 2.3 million scholarly articles as Wikidata objects. The project is also connected to the Wikicite initiative [Taraborelli, D et al., 2016]. One of Wikicite's purpose was to clearly indicate scientific articles with the above mentioned property and value [Wikidata, 2022f].

Count	Description
2,380,009	Scientific articles
93,518	Scientific articles linked to one or more author items
5,562	Scientific articles linked to one or more author items and no author name string (indicating that the author linking may be complete)
3,379,786	Citations, i.e., number of uses of the P2860 property
16,327	Distinct authors (author items) having written a scientific article
13,332	Distinct authors having written a scientific article with author gender indicated

Figure 3: "*Statistics on bibliographic information in Wikidata on 2 August 2017*" [Nielsen et al., 2017]

4. [Amaral et al., 2021]: Based on the previous paper from [Piscopo et al., 2017a],

the authors enhanced the analysis of Wikidata references by eliminating the filter criteria: The paper did not limit its (data) scope to English references represented by an external URL. Nevertheless, the aim was to analyse the relevance and authoritativeness of the references. But the authors added a third examined dimension: ease of access. Ease of access could have also been considered as part of verifiability because only an accessible source satisfies this criterion[MediaWiki, 2022b]. So, this paper goes beyond the scope of the previous work in sample size, criteria and methods (auxiliary methods to extract URLs from statements and deep learning modules). However, the basic approach of the authors was very similar: a crowd-sourcing exercise, descriptive statistics on the references and train a machine learning based on the outcome of the crowd-sourcing exercise and predict the three observed dimensions of Wikidata references. The data set used was a Wikidata Dump from the 16th of April 2020. The authors extracted a random 20% sample out of this dump for their use. They identified six languages within this sample and extracted 385 random reference nodes per language. These 2310 reference nodes where shown to the crowd workers. They focused on two Wikidata properties 'stated in' (P248) and 'reference URL' (P854), which will also be a focus of our work. Furthermore, the paper identifies that

"*most stated in sources are related to scientific publications, with over half of all pointing towards PubMed Central and Europe PubMed Central, which are archives of life sciences journal literature, and Crossref, which deals with information on scientific publications. After that, there are biology bases, such as NCBI Gene and UniProt.*" [Amaral et al., 2021]

Therefore, we were able to adopt some properties which can be used to classify scientific sources (P698 for PubMed ID, P932 for PubMed Central ID, P2322 for article ID which corresponds to an item at Crossref [Wikidata, 2022f], P685 for NCBI taxonomy ID) that go beyond DOI and domain pattern. The Python scripts which were used for the preprocessing of the data and training of the model are provided on Github (`https://github.com/gabrielmaia7/wikidata-reference-analysis`). The desired data of the dump was inserted into new SQL tables. To sum up, this paper gives a more detailed insight into external references of Wikidata and also their scientific character. However, the information on the scientific background of references is limited to one paragraph regarding the identifiers and two tables regarding the distribution of domain names (e.g. `www.ebi.ac.uk`) and website suffixes (e.g. in this

case '.ac').

5. [Singh et al., 2021]: The authors provide a thorough data set on citations including identifiers extracted from the English Wikipedia. Furthermore, the paper states that some references which point towards scientific articles and publications did not include the corresponding DOI. Therefore, identifiers are a reliable indication for scientific publications. However, this principle does not work vice versa meaning that if a reference does lack an identifier it is not a scientific citation. Consequently, the authors recommend an approach that goes beyond the used identifiers of scientific databases like Crossref and most famously Altmetric. Crossref mainly supports DOI and URLs which are mentioned by publishers or within forums. Their webpage does not explicitly mention identifiers [Martyn Rittmann, 2020]. Altmetrics publicly lists the supported identifiers: DOIs, PubMed IDs, ISBNs, Handles, arXiv IDs, ADS IDs, SSRN IDs, RePEC IDs, URNs and ClinicalTrials.gov records [Altmetric, 2021]. The authors propose a citation classification including a citation identifier look-up to find additional identifiers like DOIs. By applying this approach, the paper maps DOIs to existing references where they were missing and provides the resulting data set. The approach identifies theses DOIs by querying the Crossref API for the title of references and link matches. Our aim is to apply a similar approach including all given identifiers and identifying new ones.

With regard to the first research question, it can be stated that the literature on algorithmic-based analysis of Wikidata (external) references is already advanced. With several other notable papers (see Subsection 4.3) that dealt with this topic in the past, algorithmic-based analysis of Wikidata is definitely advanced. Most of the projects provide a Github folder to build on their existing work and extend the developed approaches. The approaches also differ regarding data basis (WDQS, dump) and coding (e.g. different Python packages and the use of Postgres).

4.3 Excluded previous publications

Based on the aforementioned criteria (English, more than five pages and publication date), we assessed the papers. However, we disregarded some papers based on the relevance and not based on those criteria. These are the following:

- [Beghaeiraveri et al., 2021]: This short article performed a reference quality analysis on Wikidata Topical Subsets. The results are presented in various statistics and compared across two different Wikidata

dumps. There is no clear definition of 'reference quality' and the authors propose the implementation of a reference scoring system. This paper was excluded due to the lack of a detailed approach to classify sources apart from exporting references and present some descriptive statistics.

- [Hosseini Beghaeiraveri, 2022]: The authors elaborate on their previous work [Beghaeiraveri et al., 2021] and develop a reference quality assessment framework to improve the quality of Wikidata references. Quality means that the reference is accessible and verifies the statement that is connected to. The framework can incorporated via a Python module (Referencing Quality Scoring System, see `https://github.com/seyedahbr/RQSSFramework`). Furthermore, a reference suggestion framework is introduced to propose references for Wikidata claims. Unfortunately, scientific character was not one of the metrics of the examination. Neither Believability, Objectivity nor Reuptation include identifiers for scientific papers. Despite the interesting exportation of data, this paper was excluded due to the lack of in depth analysis of references and/or scientific character of sources.

- [Haller et al., 2022]: In this paper, the authors investigate the linkage of Wikdata to other data sources. Furthermore, an analysis of links to external datasets and ontologies is conducted. However, the paper does not include external Wikidata references or identifiers regarding their scientific character which is why this paper was excluded from the Related Work section.

- [Lewoniewski et al., 2017]: This paper concerns the analysis of Wikipedia references across languages. It also includes some identifiers for scientific sources. However, the paper was older, shorter and especially less detailed than the ultimately chosen paper.

5 Data - WDQS SPARQL endpoint and sampling

We accessed the Wikidata Query Service SPARQL endpoint, which provides the same results as `query.wikidata.org`), with Python scripts. This was done to repeat multiple queries in a short amount of time and get data that is up to date. In order to run the queries, we accessed the endpoint via the following link `https://query.wikidata.org/sparql` using the SPARQLWrapper library. Despite the recommendation in the outlined user policy by

Wikimedia [Wikimedia, 2022f], we did not use the request library. Nevertheless, we added "*an informative User-Agent string with contact information*" [Wikimedia, 2022f] with the use of the Python requests package to avoid getting banned.

Furthermore, we used the time library to implement the necessary pauses in between the API requests. The WDQS user manual for the SPARQL endpoint does not list an amount of requests that every client is allowed in a specific time frame. The two provided limits are 60 seconds of processing time every 60 seconds per client which is identified via IP address and user agent information in the header and a maximum of 30 error queries within a minute [Wikimedia, 2022g]. In addition, every query that needs more than 60 seconds will time out. When developing our queries and the algorithm, we had to keep these limitations in mind. It is hard to estimate the processing time for random items via the API. Hence, we ran into some issues with our first trials. Therefore, we can advise a more conservative approach when setting the 60 seconds timeout pauses.

The maximum timeout limit of 60 seconds was not an issue for any queries, not even regarding the counting of reference properties. This shows the performance and potential of the WDQS SPARQL endpoint, given prior knowledge of the Wikidata structure of course. Even for the generally most often used properties (see [Wikidata, 2020]: this list is not limited to properties used within reference nodes) within Wikidata, the query was extremely fast (e.g. 88,232,004 P248 'stated' in counts).

The result of query was parsed individually. The return format of the request was set to JSON. The desired values were extracted and saved into a pandas data frame. The basis for all extractions was the following function:

```
sparql = "https://query.wikidata.org/sparql"
user_agent = "User-Agent:                    @        )"

def get_results(endpoint_url, query):
    sparql = SPARQLWrapper(endpoint_url, agent=user_agent)
    sparql.setQuery(query)
    sparql.setReturnFormat(JSON)
    return sparql.query().convert()
```

Figure 4: Basic function for the WDQS endpoint export

We started with the same query as shown in Figure 3. It is a simple and quick query to count the number of scholarly articles that are linked with the proposed value pair of P31 and Q13442814. Hence, our first request including the query[2] looked like this:

Since there are no binding unified approaches towards reference properties. Our second query aimed at counting the use of all reference properties.

[2]https://w.wiki/6A4g

```
query = """
    SELECT
    (count(*) AS ?CNT)
    WHERE {
      ?S wdt:P31 wd:Q13442814 .
    }
    """
results = get_results(sparql, query)

print(["Count of scientific articles",results['results']['bindings'][0]['CNT']['value']])
```

```
['Count of scientific articles', '38423441']
```

Figure 5: Query to count all scholarly articles within Wikidata with the corresponding property-value pair. Notice that this query does not entail scholarly articles within reference nodes but those articles in statements in general (wdt:P31 instead of pr:P31; see Figure 1 for a better understanding)

Consequently, we slightly adapted the previous query and made a loop. Despite knowing that there is no reference property with the value P1 (see [Wikidata, 2022c] and `https://www.wikidata.org/wiki/Property:P1`) as of 10th of October 2022, our query started with this item and ran up until P12000. We chose this property number because the chronologically sorted list of all properties, ended with 11,079 as of 10th of October 2022 [Wikidata, 2022c]. Furthermore, the property dashboard which is part of one of the Wikidata services shows properties up to 11,999 as of 10th of October 2022 [Wikidata, 2022b]. Assuming that there could be some deleted properties and a gap of a couple of properties and the partly incorrect listing, the limit was set to P12000 to have a high number of tolerance. Therefore, we ended up with the following request including a loop ending at the count of 12,000 and the query for the count:

```
while i <= 12000:
    query = """
    SELECT
    (count(*) AS ?CNT)
    WHERE {
      ?R pr:P"""+str(i)+""" ?reference .
    }
    """
```

Figure 6: Query to count the records of all properties within reference nodes

Based on all aforementioned information about the data model, we decided to extract all reference nodes including all properties. The main property that identifies external references is P854 (reference URL, `https:`

//www.wikidata.org/wiki/Property:P854). However, this is not always the case as the previously mentioned list [Wikidata, 2022d] shows in the category 'Scholarly Article' subcategory 'Data'(e.g. P953 'full work available at URL'). Hence, we did not limit the queries of reference nodes to a specific property. Additionally, this approach is supported by the fact that a reference node entails various attributes including the identifiers. These identifiers are important to classify an external reference as scientific. For instance, if there was a reference to an article which is neither part of Wikidata nor Wikipedia that entails a scientific contribution with a DOI or some other external scientific ID. This could be considered an external scientific reference since the origin of the reference is external and can be classified as scientific within the respective reference node. Furthermore, some additional data on the references allows us to analyse the references in more detail (e.g. property usage, reference date anomalies).

As already pointed out, we took a random sample of 200,000 (2 times 100,000) items and looked up their respective queries. The random numbers for the items were generated using the Python 'random' library. The lower limit was set to 1 and the upper limit was set to 99,999,997 because the last item with a valid number was Q999999996 (concerning Hitachi Citizen Sports Park https://www.wikidata.org/wiki/Q99999996) on the 8th of September 2022 at 21:12 CET. This is also pretty close to the 99,880,487 claimed items within Wikidata as of 13th of October. We seperated the 100 thousand items into 10 chunks of 10,000 thousand items to get some additional timeouts. The algorithm was developed to pause after every 50 queries or in case of an error for 60 seconds. We did not measure the consumed time. However, the code was executed a second time to increase the number of reference nodes and increase the sample size.

Despite implementing error codes and fallback solutions, we double-checked our queries via random sampling. We did either run the query again manually via the Query Service GUI or directly accessed the selected items and their references in Wikidata. This procedure was conducted for every described query.

We calculated our sample size based on the following criteria. We aimed for a confidence level of 95% (z-value: 1.96) and a confidence interval/margin of error of about 0.25% (moe). Furthermore, we assumed that the expected share of scientific external references will be somewhere around 25% (p) based on the findings from [Piscopo et al., 2017a]. This paper identified that 23.8% of external references came from a source/author that could be classified as scientific. Hence, we calculated based on a population (n) of roughly 99.88M items within Wikidata the following sample size (k):

```
rnumbers = set()
while len(rnumbers) < 1000000:
    rnumbers.add(random.randint(1,99999997)) # highest number 08.09.2022 21:12 Hitachi Citizen Sports Park (Q99999996)

rnumbers = list(rnumbers)

chunks = [rnumbers[x:x+10000] for x in range(0, len(rnumbers), 100000)]
```

```
for i in chunks:
    for l in i:
        query = """SELECT DISTINCT
        ?R
        ?RefP
        ?reference
        WHERE {
        wd:Q"""+str(l)+""" ?P ?X .
        ?X prov:wasDerivedFrom ?R .
        ?R ?RefP ?reference .
         }
        """
```

Figure 7: Generate a random sample of 100,000 items and divide them into 10 chunks of 10,000 items and the SPARQL-query to extract the reference nodes for a random sample of 100,000 times

$$k = \frac{\frac{z^2 * p * (1-p)}{e^2}}{1 + \frac{z^2 * p * (1-p)}{e^2} * N} = \frac{\frac{1.96^2 * 0.25 * (1-0.25)}{0.0025^2}}{1 + \frac{1.96^2 * 0.25 * (1-0.25)}{0.0025^2 * 99.88M}} \sim 115,235 \quad (1)$$

Since this value is above 100,000, we increased our sample size and ran the algorithm a second time.

6 Reference (Property) Analysis

The analysis was mainly conducted with Python scripts and the pandas package. We filtered and sorted the exports in order to get insightful outcome. Visualizations (e.g. bar charts) have been created using the matplotlib library. In order to do some manual checks and comparisons we exported some parts of our analysis as CSV files. Based on the above described queries and algorithm our main findings are listed in the following subchapters.

6.1 General Observations - Reference properties

The first query was regarding the count of scholarly articles. The number of Wikidata items that are scholarly articles increased since the query from [Scholia, 2022] by quite a margin. There are 38,423,441 scholarly articles part of Wikidata as of 15th of October 2:41 a.m. This is an increasement by more than 30 million articles in the last five years. However, the number of reference nodes that entail this property value is 1. There is exactly one

reference node that identifies a scholarly article in this way. This result can be achieved by replacing 'wdt:' with 'pr:' in the shown query in Figure 5. Since Wikidata items are internal references, they are not part of this analysis.

We tried to get data via database reports provided by Wikidata linked on their homepage `https://www.wikidata.org/wiki/Wikidata:Database_reports` which are supposed to run daily. However, the provided link of the respective Database report 'Wikidata Datamodel Reference' (`https://grafana.wikimedia.org/dashboard/db/wikidata-datamodel-references?orgId=1`) returned an error message ('Dashboard not found') for our requests on the 28th of September at 3:12 p.m. and 15th of October 1:03 p.m. Hence, we were unable to compare this data with another source.

Our second query[3] aimed at an overview of the used properties within reference nodes. We counted more than 5,267 different reference properties with a total of 335,960,448 records. The top 10 reference properties (see Figure 8) represent 89.2% of the population.

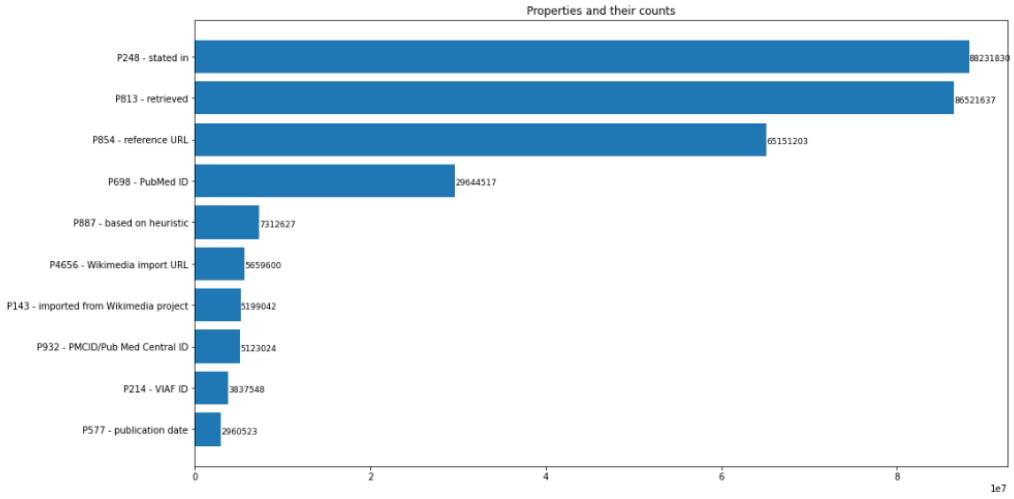

Figure 8: Top 10 Wikidata reference properties

The two top most used reference properties (P248 and P813) are used more than 80M times. Those properties can be universally used and reference to either external or internal reference values. There are 65,151,203 reference URLs in Wikidata which point towards external references. PubMed ID with 29,644,517 and PubMed Central ID with 5,123,024 show that there are many references linked with those identifiers. One potential reason that those IDs are more apparent in Wikidata is that there is a bot linking the IDs to Wikidata items [Wikimedia, 2021]. Hence, it can be analysed that

[3]https://w.wiki/6A5F

there is clear tendency towards three data properties (stated in, retrieved and reference URL). The most popular identifiers for scientific references are PubMedID and PMCID, which indicates that in the fields of biomedical and life sciences Wikidata entries link to more scientific sources than in other areas.

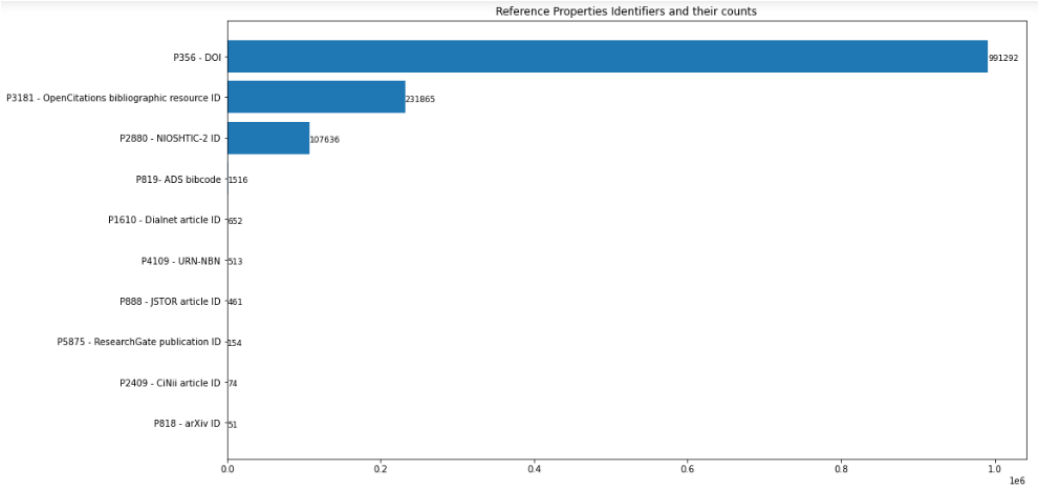

Figure 9: Top 10 Wikidata reference properties for scientific identifiers (without PubMed ID and PMC ID)

6.2 Scientific character of external references

Our data set consisted of the reference node hashes, reference properties and reference values from 200,000 Wikidata items. E.g. for the Wikidata item Q83886083 'DEEP2-GRS 12013317', the first line of our output looked like this: ['72f480a6a79284a882fa0931c570b9f702f, P248, `http://www.wikidata.org/entity/Q68959020`']. The concerned reference value that is connected via 'stated in' (P248) to the reference node is an article published in 'The Astrophysical Journal Supplement Series' [Matthews et al., 2013]. The article is referenced as internal Wikidata item without a DOI in the reference node. However, the Wikidata item entry entails the respective DOI number. Hence, it can be classified as scientific reference. But it is a reference to a Wikidata entry and therefore not an external reference.

The 204,884 unique reference nodes of our data set entailed 862 unique reference properties and 235,278 unique reference values. In total, we extracted 926,668 records/reference values for 619,273 reference nodes. This would mean that every item in our sample has on average only about 1

unique reference node. This little number is not implausible for four main reasons:

1. The Wikidata item has no reference node. Our sample entailed a type of railway (https://www.wikidata.org/wiki/Q16777248 or a calendar date (https://www.wikidata.org/wiki/Q69206026). Both links have been accessed on the 13th of October at 4:38 p.m.. Especially the later item has typically no reference. An analysis whether those items should be Wikidata entries is not within the scope of this work.

2. Our random sample might have selected a Wikidata item which redirects to another Wikidata entry. This is a common procedure within Wikimedia projects to avoid duplicates and multiple entries for the same item. E.g. the link https://www.wikidata.org/wiki/Q67108909 (accessed on the 13th of October at 4:40 p.m.) redirects one to another Wikidata item with the number Q59371514.

3. Furthermore, a closer look at our previous example item Q83886083 provides us the most obvious answer: This Wikidata item has ten references within its entry. However, there are only two unique references associated with this item. Hence, there can be a high number of items with little unique references.

4. Additionally, we did not limit the factor uniqueness on the query. This means for instance, that reference nodes are unique to an item might be also a unique reference node for another Wikidata item. Consequently, the aforementioned Astrology article (Q68959020) has been used 229,468 times as reference value [4].

As already described in the previous section, we used to the proposed identifiers to classify an external reference as scientific. We extracted all external references based on the list of data attributes provided in the Bibliography template [Wikidata, 2022d]. Figure 10 provides an overview of the property counts within our sample. The outcome is very similar to Figure 8 that entailed the counts of all Wikidata reference properties. Hence, our sample can be seen as representative in this regard. We identified 112,441 unique external reference nodes and 167,799 reference values out of which 109,735 are reference URLs.

The property list entailed the most common (potential) external reference property 'P854 - reference URL'. However, we also parsed the list for all

[4] https://w.wiki/6A5i

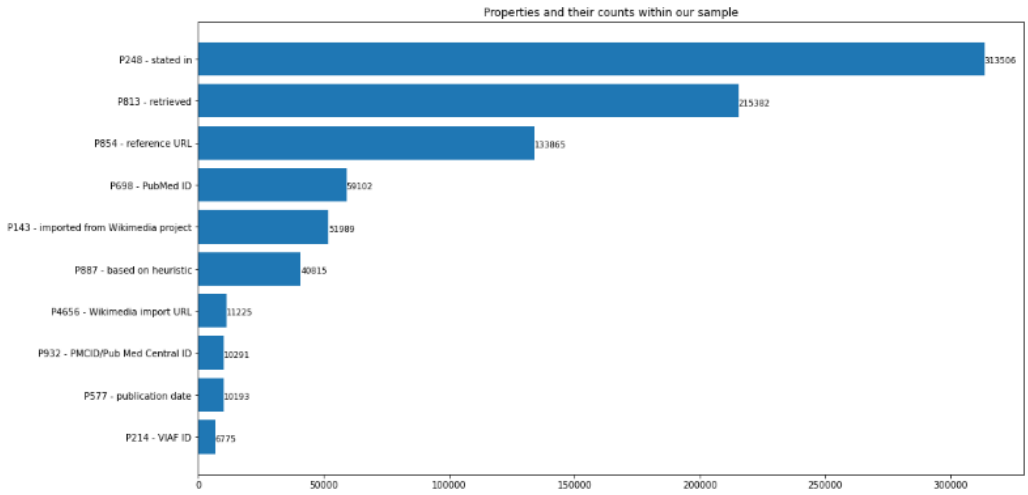

Figure 10: Top 10 Wikidata reference properties within our sample

links that are directing to a webpage by filtering the reference values for 'https://', 'http://' or just 'www.'. In a second step we excluded all links that pointed towards Wikipedia or Wikidata pages. Our data frame did not include any reference nodes with the property 'P31'. Furthermore, we examined the values for references that connected via 'stated in' (P248) and 'retrieved' (P813) properties to the reference nodes. Retrieved statements did only include timestamps and 'stated in' values did only reference towards Wikidata items. The outcome was considered our final sample of external reference nodes which was further analysed. The top 10 properties can be seen in Figure 11 below.

We came to the conclusion that most of our reference values are URLs. Due to the high amount of reference URLs within external reference nodes (approximately 97.6% of all), we applied a two-fold approach regarding URLs. We identified scientific external references via two methods: Firstly, based on the identifiers provided in the list developed by the Wikidata Meta Source project [Wikidata, 2022d]. Those identifiers cover also the identifiers used in all described publications within in this paper. Secondly, with regard to filtering based on the URL, we filtered for most common academic research databases that entail their name in the URL and do not just redirect towards other sites like Google Scholar. We used the following filter criteria: '.ac.', 'crossref' (`https://www.crossref.org/`, 'eric.ed' (`https://eric.ed.gov/`), 'ieee.org' (`https://ieeexplore.ieee.org/Xplore/home.jsp`) and 'doi.org'.

We identified 68,123 or 60.6% (or 33.3% of all reference nodes) scientific

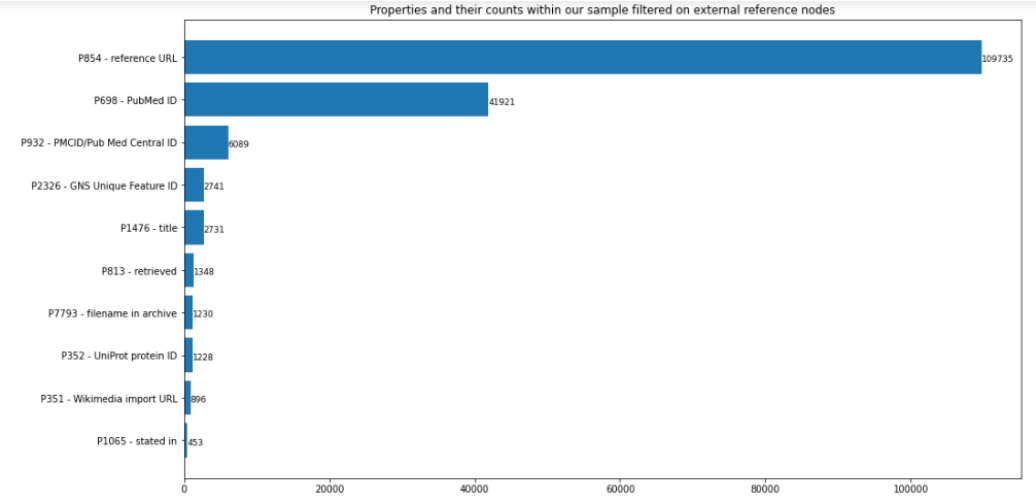

Figure 11: Top 10 Wikidata reference properties within our sample filtered on external reference nodes

external reference nodes. Figure 12 illustrates this statistic.

We identified most of the external references based on the URL classification method applied to the domain name (2nd method) and not based on the 1st method which uses the listed properties. Figure 13 provides an overview of the occurrences of the scientific property identifiers within our external reference nodes. The figure shows that the two most common scientific property identifiers are also the most common identifiers within our sample. Furthermore, it shows that around 48 thousand reference nodes were able to be classified as scientific based on identifiers. However, as already stated the majority of the 68,123 scientific external reference nodes were classified based on the URL classification method. There were no occurrences of 'eric.ed' or 'ieee.org'. 54,339 URLs entailed the '.ac.' suffix, 13,218 'crossref' and 142 'doi.org'. This is a clear difference to the amount of identified DOI properties (330) which can be explained that the identifier can also be in a reference node without a DOI-specific URL.

7 Discussion

Our findings on the proportion of external scientific references differ from the data in previous papers. It would therefore make sense to apply the two-part approach, which uses URL endings and the identifier list of the template from WikiProject Source MetaData, to a Wikidata dump. Furthermore, the DOI example shows that properties and URLs are not always used in the

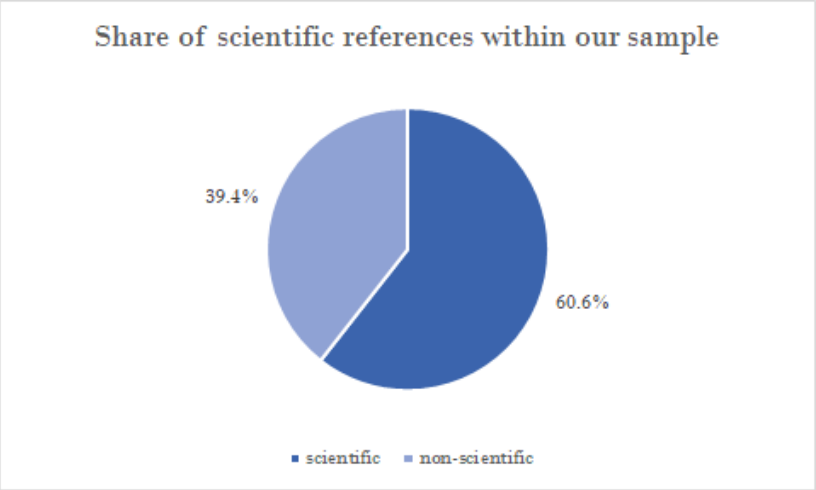

Figure 12: Share of scientific external reference nodes

same way. Despite the size of the sample, an overall detailed examination of the reference nodes based on a Wikidata dump makes sense. Hence, we propose a thorough examination of these topics with the whole population and not just a sample.

Furthermore, this dump can be used to analyse the amount of Wikidata items listed as 'instance of' 'scholarly articles' and their records within reference nodes. An increasing number of internal Wikidata references that represent scientific papers could make Wikidata a scholarly database and act as shown by [Scholia, 2022] as bibliographic tool. However, this tendency of drafting scholarly articles, journals and books as Wikidata items will in the long term decrease the existence of external references. It is unclear whether this is an aim of Wikimedia or Wikidata. Nevertheless, a closer look at the existence and referencing towards Wikidata and Wikimedia items could shed some light on this topic.

As already mentioned in our analysis, we found that Wikidata items which have an entry that only redirects towards another Wikidata item. "*Redirects are recorded but currently have no additional semantics implemented.*" [Wikidata, 2022a]. Therefore, it would be interesting to examine the number of redirects within Wikidata. We could have identified those by checking whether the returned JSON was empty. Unfortunately, we underestimated the number of redirects. When we did our manual checks, we realised that this could be a more apparent issue than expected. We manually checked 188 items and 39 items (approximately 21.7%) did not have any reference nodes. Six out of those 39 items were redirects, 15 items with no reference nodes and eight numbers that did not contain a Wikidata item. These records indicate

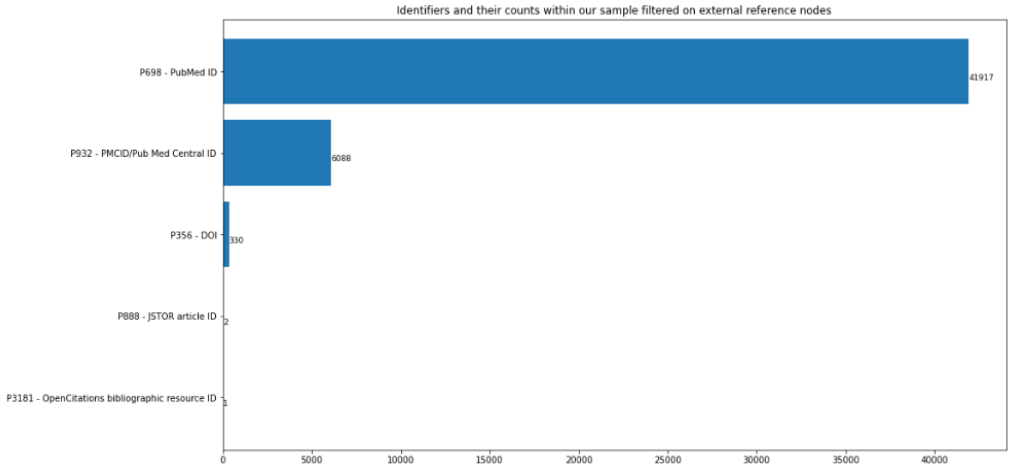

Figure 13: Identifiers and their counts within our sample filtered on external reference nodes

that the real number of Wikidata items might be below the proclaimed 99.88 million items. So, we advice further research on redirects, empty numbers and unreferenced Wikidata entries.

8 Conclusion

We conducted a thorough literature review to understand the current state of research on algorithmic based analysis of Wikidata's external references. Five papers (four regarding Wikidata and one regarding Wikipedia), which were identified based on multiple criteria, are summarized and presented in more detail. The papers include: a bibliographic tool that uses Wikidata items [Nielsen et al., 2017], a comparison between Wikipedia and Wikidata external references [Piscopo et al., 2017b], a sampling approach to analyse the relevance and authoritativeness of Wikidata's external references W[Piscopo et al., 2017a], an advanced approach which adds the accessibility as third factor to the aforementioned approach [Amaral et al., 2021] and a try to measure the scientific character of Wikipedia references [Singh et al., 2021]. Those articles build the most relevant contributions in this research field and show the development of Wikidata analysis. We ran various queries on the WDQS SPARQL endpoint to show the distribution of properties within reference nodes. The three most often used reference properties are: P248 - 'stated in' (88,231,830 records), P813 - 'retrieved' (86,521,637) and P854 'reference URL' (65,151,203). PubMed ID with 29,644,517 and PubMed Central ID with 5,123,024 counts are the two most prominent properties that can

be used to classify references as scientific. Hence, external references form a significant part of Wikidata sources and also scientific identifiers are common within Wikidata. Based on the overall number of Wikidata items of about 99.88 million entries, we took a sample of reference nodes. We extracted the distinct reference nodes of 200,000 random Wikidata items. We classified 60.6% (68,123) of external reference nodes as scientific. Every third (33.3%) distinct reference node can be classified as external and scientific. This shows a higher number of scientific references than previous approaches by [Piscopo et al., 2017a] and [Amaral et al., 2021]. We showed a twofold approach to classify external references and took the latest recommendations of the Wikicite initiative into account. Future work should apply this method to a Wikidata dump and identify further gaps between properties and reference URLs (like DOI property and doi URLs). Furthermore, there is significant increase of scholarly articles which are represented as Wikidata items (08/2017: 2.38 million compared to 10/2022: 38.42 million). Hence, there is a high number of reference nodes that point towards Wikidata items which represent scientific articles. Therefore, we propose further research into these topics.

9 Abbreviations - Acronyms

- ADS = Astrophysics Data System
- API = Application Programming Interface
- CSV = Comma-separated Values
- DOI = Digital Object Identifier
- GUI = Graphical User Interface
- HTML = Hypertext Markup Language
- HTTP = Hypertext Transfer Protocol
- IP = Internet Protocol
- ISBN = International Standard Book Number
- JSON = JavaScript Object Notation
- NCBI = National Center for Biotechnology Information
- ORCID = Open Researcher and Contributor ID
- PMID = PubMed Identifier
- PMCID = PubMed Central Identifier
- RePEC = Research Papers in Economics
- RDF = Resource Description Format
- SQL = Structured Query Language
- SSRN = Social Science Research Network
- URI = Uniform Resource Identifiers
- URL = Uniform Resource Locator
- URN = Uniform Resource Name
- VIAF = Virtual International Authority File
- WDQS = Wikidata Query Service

References

[Altmetric, 2021] Altmetric (2021). Scholarly identifiers supported by altmetric. https://help.altmetric.com/support/solutions/articles/6000240585-scholarly-identifiers-supported-by-altmetric. Last checked on Oct 25, 2022.

[Amaral et al., 2021] Amaral, G., Piscopo, A., Kaffee, L.-A., Rodrigues, O., and Simperl, E. (2021). Assessing the quality of sources in wikidata across languages: a hybrid approach. *Journal of Data and Information Quality (JDIQ)*, 13(4):1–35.

[Beghaeiraveri et al., 2021] Beghaeiraveri, S. A. H., Gray, A. J., and McNeill, F. (2021). Reference statistics in wikidata topical subsets. In *Wikidata@ ISWC*.

[Haller et al., 2022] Haller, A., Polleres, A., Dobriy, D., Ferranti, N., and Rodríguez Méndez, S. J. (2022). An analysis of links in wikidata. In *European Semantic Web Conference*, pages 21–38. Springer.

[Hosseini Beghaeiraveri, 2022] Hosseini Beghaeiraveri, S. A. (2022). Towards automated technologies in the referencing quality of wikidata. In *Companion Proceedings of the Web Conference 2022*, pages 324–328.

[Lewoniewski et al., 2017] Lewoniewski, W., Węcel, K., and Abramowicz, W. (2017). Analysis of references across wikipedia languages. In *International Conference on Information and Software Technologies*, pages 561–573. Springer.

[Martyn Rittmann, 2020] Martyn Rittmann (2020). Crossref - event data. https://www.crossref.org/services/event-data/. Last checked on Oct 25, 2022.

[Matthews et al., 2013] Matthews, D. J., Newman, J. A., Coil, A. L., Cooper, M. C., and Gwyn, S. D. (2013). Extended photometry for the deep2 galaxy redshift survey: A testbed for photometric redshift experiments. *The Astrophysical Journal Supplement Series*, 204(2):21.

[MediaWiki, 2022a] MediaWiki (2022a). Wikicite. https://meta.wikimedia.org/wiki/WikiCite. Last checked on Oct 25, 2022.

[MediaWiki, 2022b] MediaWiki (2022b). Wikipedia verifiability. https://en.wikipedia.org/wiki/Wikipedia:Verifiability. Last checked on Oct 25, 2022.

[Nielsen et al., 2017] Nielsen, F. Å., Mietchen, D., and Willighagen, E. (2017). Scholia, scientometrics and wikidata. In *European Semantic Web Conference*, pages 237–259. Springer.

[Piscopo et al., 2017a] Piscopo, A., Kaffee, L.-A., Phethean, C., and Simperl, E. (2017a). Provenance information in a collaborative knowledge graph: an evaluation of wikidata external references. In *International semantic web conference*, pages 542–558. Springer.

[Piscopo et al., 2017b] Piscopo, A., Vougiouklis, P., Kaffee, L.-A., Phethean, C., Hare, J., and Simperl, E. (2017b). What do wikidata and wikipedia have in common? an analysis of their use of external references. In *Proceedings of the 13th International Symposium on Open Collaboration*, pages 1–10.

[Scholia, 2022] Scholia (2022). Wikidata publications q2013. https://scholia.toolforge.org/topic/Q2013. Last checked on Oct 25, 2022.

[Schönitzer,Michael F., 2017] Schönitzer,Michael F. (2017). Wikibase rdf mapping diagram. https://www.mediawiki.org/wiki/Wikibase/Indexing/RDF_Dump_Format#/media/File:Rdf_mapping-vector.svg. Last checked on Oct 25, 2022.

[Singh et al., 2021] Singh, H., West, R., and Colavizza, G. (2021). Wikipedia citations: A comprehensive data set of citations with identifiers extracted from english wikipedia. *Quantitative Science Studies*, 2(1):1–19.

[Taraborelli, D et al., 2016] Taraborelli, D et al. (2016). Wikicite 2016 report. https://upload.wikimedia.org/wikipedia/commons/2/2b/WikiCite_2016_report.pdf. Last checked on Nov 25, 2016.

[Webster and Watson, 2002] Webster, J. and Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS quarterly*, pages xiii–xxiii.

[Wikidata, 2020] Wikidata (2020). Wikidata:database reports/list of properties/1-1000. https://www.wikidata.org/wiki/Wikidata:Database_reports/List_of_properties/1-1000. Last checked on Oct 25, 2022.

[Wikidata, 2022a] Wikidata (2022a). Wikibase rdf format. https://www.mediawiki.org/wiki/Wikibase/Indexing/RDF_Dump_Format. Last checked on Oct 25, 2022.

[Wikidata, 2022b] Wikidata (2022b). Wikidata, category:property dashboard. https://www.wikidata.org/w/index.php?title=Category:Property_dashboard&from=1100. Last checked on Oct 25, 2022.

[Wikidata, 2022c] Wikidata (2022c). Wikidata:database reports/list of properties/all. https://www.wikidata.org/wiki/Wikidata:Database_reports/List_of_properties/all. Last checked on Oct 25, 2022.

[Wikidata, 2022d] Wikidata (2022d). Wikidata:template:bibliographic_properties. https://www.wikidata.org/wiki/Template:Bibliographic_properties. Last checked on Oct 25, 2022.

[Wikidata, 2022e] Wikidata (2022e). Wikidata:wikiproject source metadata. https://www.wikidata.org/wiki/Wikidata:WikiProject_Source_MetaData. Last checked on Oct 25, 2022.

[Wikidata, 2022f] Wikidata (2022f). Wikidata:wikiproject source metadata/bibliographic metadata for scholarly articles in wikidata. https://www.wikidata.org/wiki/Wikidata:WikiProject_Source_MetaData/Bibliographic_metadata_for_scholarly_articles_in_Wikidata. Last checked on Oct 25, 2022.

[Wikimedia, 2019] Wikimedia (2019). Wikidata homepage. https://www.wikidata.org/wiki/Wikidata:Main_Page. Last checked on Oct 25, 2022.

[Wikimedia, 2021] Wikimedia (2021). Wikidata,user:citationgraph bot. https://www.wikidata.org/wiki/User:Citationgraph_bot. Last checked on Oct 25, 2022.

[Wikimedia, 2022a] Wikimedia (2022a). Help:statements. https://www.wikidata.org/wiki/Help:Statements. Last checked on Dec 03, 2022.

[Wikimedia, 2022b] Wikimedia (2022b). Mediawiki history. https://www.mediawiki.org/wiki/MediaWiki_history. Last checked on Dec 03, 2022.

[Wikimedia, 2022c] Wikimedia (2022c). Wikidata statistics. https://www.wikidata.org/wiki/Special:Statistics. Last checked on Dec 03, 2022.

[Wikimedia, 2022d] Wikimedia (2022d). Wikidata stats. https://wikidata-todo.toolforge.org/stats.php. Last checked on Dec 03, 2022.

[Wikimedia, 2022e] Wikimedia (2022e). Wikidata verifiability. https://www.wikidata.org/wiki/Wikidata:Verifiability. Last checked on Oct 25, 2022.

[Wikimedia, 2022f] Wikimedia (2022f). Wikimedia:user-agent policy. https://meta.wikimedia.org/wiki/User-Agent_policy#Python. Last checked on Oct 25, 2022.

[Wikimedia, 2022g] Wikimedia (2022g). Wikimedia:wikidata query service/user manual. https://www.mediawiki.org/wiki/Wikidata_Query_Service/User_Manual#SPARQL_endpoint. Last checked on Oct 25, 2022.

[Wikimedia, 2022h] Wikimedia (2022h). Wikipedia article on wikidata. https://en.wikipedia.org/wiki/Wikidata. Last checked on Oct 25, 2022.

[Wikimedia, 2022i] Wikimedia (2022i). Wikipedia article relationship wikipedia - wikidatainfobxes. https://en.wikipedia.org/wiki/Wikipedia:Wikidata#Infoboxes. Last checked on Oct 25, 2022.