

# RQSS: Referencing Quality Scoring System for Wikidata

Seyed Amir Hosseini Beghaeiraveri<sup>a,b</sup>, Alasdair Gray<sup>a</sup> and Fiona McNeill<sup>b</sup>

<sup>a</sup> *School of Mathematical and Computer Science, Heriot-Watt University, Edinburgh, Currie EH14 4AS, UK*  
*E-mails: sh200@hw.ac.uk, A.J.G.Gray@hw.ac.uk*

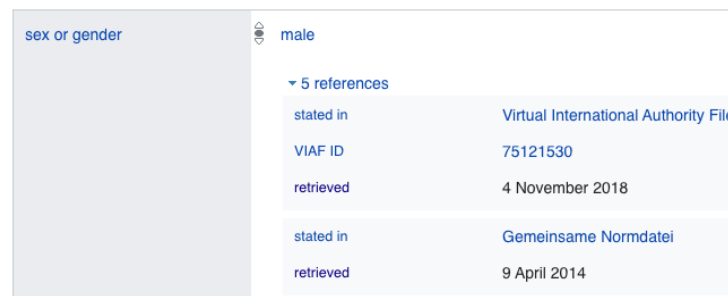
<sup>b</sup> *School of Informatics, The University of Edinburgh, Edinburgh, EH8 9AB, UK*  
*E-mails: seyed.hosseini@ed.ac.uk, f.j.mcneill@ed.ac.uk*

**Abstract.** Wikidata is a collaborative multi-purpose Knowledge Graph (KG) with the unique feature of adding provenance data to the statements of items as a reference. More than 73% of Wikidata statements have provenance metadata; however, few studies exist on the referencing quality in this KG, focusing only on the relevancy and trustworthiness of external sources. While there are existing frameworks to assess the quality of Linked Data, and in some aspects their metrics investigate provenance, there are none focused on reference quality. We define a comprehensive referencing quality assessment framework based on Linked Data quality dimensions, such as completeness and understandability. We implement the objective metrics of the assessment framework as the Referencing Quality Scoring System - RQSS. The system provides quantified scores by which the referencing quality can be analyzed and compared. RQSS scripts can also be reused to monitor the referencing quality regularly. Due to the scale of Wikidata, we have used well-defined subsets to evaluate the quality of references in Wikidata using RQSS. We evaluate RQSS over three topical subsets: Gene Wiki, Music, and Ships, corresponding to three Wikidata WikiProjects, along with four random subsets of various sizes. The evaluation shows that RQSS is practical and provides valuable information, which can be used by Wikidata contributors and project holders to identify the quality gaps. Based on RQSS, the average referencing quality in Wikidata subsets is 0.58 out of 1. Random subsets (representative of Wikidata) have higher overall scores than topical subsets by 0.05, with Gene Wiki having the highest scores amongst topical subsets. Regarding referencing quality dimensions, all subsets have high scores in accuracy, availability, security, and understandability, but have weaker scores in completeness, verifiability, objectivity, and versatility. Although RQSS is developed based on the Wikidata RDF model, its referencing quality assessment framework can be applied to KGs in general.

**Keywords:** Reference Quality, Data Quality, Wikidata, Knowledge Graphs, Subsetting, Topical Subsets, Random Subsets, Big Data, RQSS, Provenance, Linked Data, Quality Assessment Framework

## 1. Introduction

Approaching its tenth birthday, Wikidata [1] is now the paramount general-purpose user-contributed KG in research and industry [2]. By August 2022, Wikidata have had nearly 100 million data items and more than 1.7 billion statements [3]. Besides being collaborative and multilingual, Wikidata has the unique ability to assign one or more sources to each statement [4]. According to its introduction, Wikidata is a secondary database that collects statements along with their provenance [5]. Providing provenance in Wikidata is called referencing. In Wikidata, “references are used to point to specific sources that back up the data provided in a statement” [6]. Figure 1 shows a referencing in Wikidata, where Albert Einstein’s *sex or gender (P21) claim* has been referenced with two reference sets, one with three and another with two reference triples. More than 73% of Wikidata statements have at least one



Property	Value
sex or gender	male
5 references	
stated in	Virtual International Authority File
VIAF ID	75121530
retrieved	4 November 2018
stated in	Gemeinsame Normdatei
retrieved	9 April 2014

Fig. 1. An example of referencing in Wikidata for Albert Einstein's sex or gender statement.

reference.<sup>1</sup> Wikidata references can help AI tools detect errors and make decisions based on the supporting evidence [7]. Having references also makes Wikidata a believable and verifiable knowledge base for end users.

Linked Data Quality is a multi-dimensional concept [4, 8–15] including availability, completeness, etc., in which, providing the source of facts is considered part of *believability* and *verifiability* dimensions (see Section 2.3)[4, 13, 15]. Providing the provenance increases the trust in data [4, 13, 15]. Despite the high percentage of referencing in Wikidata and a large portion of metadata, e.g., referencing reification nodes, dedicated to references, few studies have delved into referencing quality in this knowledge base. The only reference-specific research on Wikidata was by Piscopo et al. in 2017 [16], and was extended in 2021 by Amaral et al. [2]. These studies evaluated two subjective data quality dimensions, relevancy and authoritativeness of Wikidata references which correspond to the relevancy and believability in our study. However, there are other aspects of quality we can define in the context of references, such as completeness, accuracy, and understandability. In this regard, the research question is how can the quality of references be quantified considering different aspects of data quality. To the best of our knowledge, there is no assessment framework for evaluating the referencing quality of Linked Data or Semantic Web KGs, including Wikidata. We aim to address this gap by defining and implementing a comprehensive framework for assessing referencing.

Although some KGs, e.g. DBpedia [17], support referencing on the resource (item) level, Wikidata is the only KG that supports referencing at the statement (facts and claims about items) level among open general-purpose KGs. Wikidata has an active user community contributing to and refining content and benefits from *bot* accounts; automatic tools designed to populate and maintain data in bulk. These features motivate us to investigate the quality of references in Wikidata. Based on Linked Data quality criteria and reference-specific requirements [13, 18], we formally define a referencing assessment framework with 40 metrics in 22 data quality dimensions classified in 6 data quality categories. Of these 40 metrics, 34 metrics are objective, i.e., can be measured without human expert opinions. Objective metrics can also be implemented as an automated routine enabling dataset holders to monitor data quality regularly, with no (or less) modification needed due to changing conditions and opinions, and with the most accuracy and certainty. Thus, we implement the objective metrics of the referencing assessment framework as an automatic tool called the *Referencing Quality Scoring System - RQSS*.

There is no KG comparable to Wikidata in terms of size and topic coverage. Due to the large volume of data, evaluating the entire Wikidata over 40 metrics requires expensive hardware and unexpected processing time. We use subsets of Wikidata to evaluate the assessment framework and implemented tools. Along with facilitating the processing of Wikidata's large volume, subsets provide a comparison platform to review differences in referencing quality scores in different thematic parts of Wikidata [19]. We use three topical subsets [18] and four random subsets of Wikidata in different sizes. Topical subsets allow us to analyze Wikidata referencing in multiple topics, while random subsets enable us to approximate the referencing quality of the entire Wikidata. Thus, by evaluating RQSS over Wikidata subsets, we provide a comprehensive statistical overview of the Wikidata referencing quality.

This study is the most comprehensive evaluation of Wikidata references in different dimensions and complements previous subjective research [2, 16]. Our contributions are (i) defining the first comprehensive referencing qual-

<sup>1</sup><https://wikidata-todo.toolforge.org/stats.php> - accessed 17 August 2022. The page has not produced sensible information recently.

ity assessment framework for Linked Data based on the Wikidata data model, (ii) developing RQSS which is the referencing quality scoring system to automatically monitor the referencing quality of Wikidata datasets, and (iii) providing statistical scores of Wikidata subsets referencing quality during the evaluation of RQSS. In Section 2, we review related work on data quality and state-of-the-art Wikidata reference quality assessments. Section 3 presents the referencing assessment framework, its dimensions, and metric definitions. Section 4 is an overview of the implemented metrics and the structure of RQSS. In Section 5 we provide the evaluation results of RQSS over Wikidata topical and random subsets. Section 6 presents the limitations we faced during the study and the countermeasures we deployed to overcome those. In Section 7 we discuss the main points of the study and a summary of lessons we learned during this research. Finally, in Section 8 we present our conclusion and discuss future work.

## 2. State of the Art

The research question and objectives require a complete survey on the Linked Data quality criteria and Wikidata referencing quality literature. Linked Data quality has been studied widely but referencing quality in Linked Data is rarely investigated.

### 2.1. Data Quality

Data quality is defined as “fitness for use” [20]. In the literature, the quality of data is considered a multidimensional concept. Wang and Strong [21] categorised data quality into four main categories, each consisting of one or more dimensions: *Intrinsic* (dimensions that are independent of the user’s context), *Contextual* (dependent on the task at hand and the context of the data consumer), *Representational* (dimensions that describe how understandable data is represented to the data consumers), and *Accessibility* (the form in which the data is available and how it can be accessed by data consumers). Bizer et al. [22] proposed a quality assessment framework to filter high-quality information on the web. They represented the framework metrics in the form of graph patterns.

There are lots of studies on the quality of Linked Data. Zaveri et al. [13] provided the most comprehensive aggregation of data quality dimensions by surveying 21 data quality papers up to 2012. From this core set, they identified 23 data quality dimensions categorized into 6 categories. Färber et al. [4] extended the criteria of Wang and Strong [21] into 11 dimensions and 34 metrics and then evaluated five KGs: Freebase [23], Wikidata, YAGO [24], Cyc [25], and DBpedia [17]. The score of each metric in their evaluation is between 0 to 1. With this scoring system, users can assign a weight to each metric based on their quality priorities. Debattista et al. [15] examined nearly 3.7 billion triples from 37 Linked Data datasets. They used 27 metrics based on the Zaveri et al. survey. They also provided a Principal Component Analysis (PCA) over their evaluation results to find the minimum number of metrics that can inform users about the quality of Linked Data datasets. None of these studies has done a comprehensive investigation of referencing quality metrics in Wikidata.

Wikidata quality has been investigated broadly. Piscopo et al. [26] surveyed 28 papers on Wikidata quality mostly published in 2017. They stated that trustworthiness needs to be investigated further in Wikidata. Shenoy et al. [27] proposed a framework to recognize low-quality statements in Wikidata. They created a historical dataset of removed Wikidata statements by finding the differences amongst 311 weekly dumps in sequence and applied the removing pattern to current statements to identify low-quality statements. Abian et al. [28] investigated the imbalances of Wikidata in gender, recency and geological data considering user needs. They used Wikipedia page view information to conclude user needs and applied them to Wikidata random items to find the gaps.

### 2.2. Trust and Referencing

The ability to provide the provenance of data is placed under the *trust* category [13]. In the literature, the trust category consists of different dimensions such as *believability*, *reputation*, *objectivity* and *verifiability*. Färber et al. [4] defined the *trustworthiness* dimension as a combination of the Wand and Strong three dimensions [21]: *believability* (the extent to which data are accepted or regarded as true, real, and credible), *objectivity* (the extent to which data are unbiased and impartial), and *reputation* (the extent to which data are trusted or highly regarded

Table 1

Linked data quality categories and dimensions as collected in [13]. Categories and dimensions in **bold** are applicable to references and are defined in this report.

Category	<b>Accessibility</b>	<b>Intrinsic</b>	<b>Trust</b>	<b>Dynamicity</b>	<b>Contextual</b>	<b>Representational</b>
Dimension	<b>Availability</b> <b>Licensing</b> <b>Security</b> <b>Interlinking</b> Performance	<b>Accuracy</b> <b>Consistency</b> <b>Conciseness</b>	<b>Reputation</b> <b>Believability</b> <b>Verifiability</b> <b>Objectivity</b>	<b>Currency</b> <b>Volatility</b> <b>Timeliness</b>	<b>Completeness</b> <b>Amount-of-data</b> <b>Relevancy</b>	<b>Representational-conciseness</b> <b>Representational-consistency</b> <b>Understandability</b> <b>Interpretability</b> <b>Versatility</b>

in terms of their source or content). Trustworthiness at the statement level was a metric in Färber et al. [4] and Debattista et al. [15]. However, both studies checked only the existence of reference usage in datasets and did not investigate how and in what manner references are being used.

### 2.3. Wikidata References Quality

The studies on Wikidata referencing quality are few and limited. In its quality rules, Wikidata recommends the provided references should be relevant (i.e., directly applicable and support the content or context of the associated fact) and authoritative (i.e., deemed trustworthy, up-to-date, and free of bias) [29]. Piscopo et al. [16] examined the authoritativeness and the relevance of Wikidata's English external sources. They first evaluated a small set of sample references (<300 statements) through microtask crowdsourcing. The results of this sampling were then given to a machine-learning algorithm that measured the relevance and authoritativeness of all English external sources. The final results showed that about 70% of Wikidata's external sources are relevant and 80% are authoritative. This approach has recently been reproduced and extended on Wikidata snapshot of 16 April 2021 [2]. The recent study considered both English and non-English external sources. However, it is still limited to relevance and authoritativeness. Piscopo et al. [30] showed that Wikidata has a more diverse pool of external references (in terms of origin country) than Wikipedia as well as benefits from external datasets (such as library catalogues). Curotto and Hogan [31] proposed an approach to index English Wikipedia references as a source for Wikidata statements. However, this proposal considers no plan to evaluate the quality of the indexed references.

## 3. Referencing Quality Assessment Framework

A robust evaluation of data quality requires rigorous and formally defined criteria. There are different dimensions to categorize data quality criteria based on measurement objectives. Although the definition of data quality criteria varies in various contexts, e.g. Linked Data and structured data, data quality dimensions are consistent. Considering references as metadata, data quality dimensions are applicable but appropriate reference-specific criteria should be defined for each dimension.

In this section, we select quality dimensions definable in the context of references and then define reference-specific quality metrics for each dimension. We base our dimension selection on the Zaveri et al. survey [13], which is, to the best of our knowledge, the most comprehensive collection of Linked Data quality metrics. At the beginning of each category and dimension, a brief survey of the Linked Data definition and metrics is provided. Then, the informal definition of the metrics is presented. The formal definitions, discussions, and additional considerations in computing the metrics can be found in Appendix A. Table 1 shows these dimensions with those that apply to references shown in bold.

### 3.1. Referencing Quality Metrics

In terms of computation, there are two types of metrics in this framework: objective and subjective. Subjective metrics cannot be computed without human opinion intervention. We highlight those metrics as (*Subjective*) in the text. All metrics are designed to return a number between 0 and 1 as the mean result, although in the majority of them, providing the distribution is helpful in analyzing the data.

## 1 **Category I. Accessibility** 1

2 This category includes dimensions that are related to access and retrieval of data. There are five dimensions in  
3 this category: availability, licensing, interlinking, security, and performance [13]. In the context of referencing, only  
4 performance is not applicable. 4

### 5 6 **DIMENSION 1. AVAILABILITY** 6

7 According to Zaveri et al., “Availability of a dataset is the extent to which information (or some portion of it) is  
8 present, obtainable, and ready for use” [13]. Several metrics are defined for availability in terms of Linked Data. It  
9 can be measured via the accessibility of the server and existence of SPARQL endpoints [4, 10], the existence of RDF  
10 dumps [4, 10], the uptime of URIs [4, 10], and proper dereferencing of URIs (in-links, back-links, or forward-links)  
11 [4, 10, 12, 15]. The suitability of data for consumers is also another (subjective) metric considered in literature  
12 [4, 10]. In the context of references, we define the following metric for the availability: 12

13  
14 *Metric 1. Availability of External URIs* The ratio of dissolvable external URIs to the total number of external URIs. 14

### 15 16 **DIMENSION 2. LICENSING** 16

17 “Licensing is defined as the granting of permission for a consumer to re-use a dataset under defined conditions”  
18 [13]. In datasets, the licensing criteria are the existence of human-readable [12, 15] or machine-readable license  
19 [4, 12, 15], permissions to use the dataset [32] (as cited in [13]), and indication of attribution [32] (as cited in [13]).  
20 In the context of references, we define the following metric for the licensing status of external URIs: 20

21  
22 *Metric 2. External URIs Domain Licensing* The ratio of human and/or machine-readable licensed external URIs to  
23 the total number of external URIs. 23

### 24 25 **DIMENSION 3. SECURITY** 25

26 “Security is the extent to which access to data can be restricted and hence protected against its illegal alteration  
27 and misuse” [13]. Security is not covered as much as other Accessibility dimensions. According to Zaveri et al.  
28 [13], Flemming’s study [32] is the only work that includes a definition for this dimension. While governmental or  
29 medical datasets often hold sensitive information accessed by numerous users, rendering them prime targets for  
30 potential attackers, Flemming’s tool lacks any metric to assess this aspect. Zaveri et al. (based on Wang and Strong  
31 [21]) mentioned secure access to data (e.g. via SSL or login credentials) and proprietary access to data as metrics  
32 of security. In the context of references, secure access to external IRIs is important. An unsecured external link  
33 decreases the trust in the provenance of data and causes security threats such as man-in-the-middle [33]. Therefore,  
34 the following metric can be considered for security in the context of references: 34

35  
36 *Metric 3. Security of External URIs* The ratio of external URIs that support TLS/SSL [34] connections to the total  
37 number of external URIs. 37

### 38 39 **DIMENSION 4. INTERLINKING** 39

40 In Linked Data, “interlinking refers to the degree to which entities that represent the same concept are linked to  
41 each other, be it within or between two or more linked data sources” [13]. This dimension is measured by data net-  
42 work parameters like interlinking degree, clustering coefficient, centrality, and sameAs chains [35]. Another metric  
43 is `owl:sameAs` links either to internal entities [4] or external URIs [4, 12, 15]. Färber et al. also considered the  
44 validity of external `owl:sameAs` links as a metric in this dimension [4]. Interlinking is one of the four fundamental  
45 principles of Linked Data [36]. We evaluate this dimension by a metric such as follows: 45

46  
47 *Metric 4. Interlinking of Reference Properties* The ratio of reference properties that are connected to another property  
48 in an external ontology to the total number of reference properties. 48

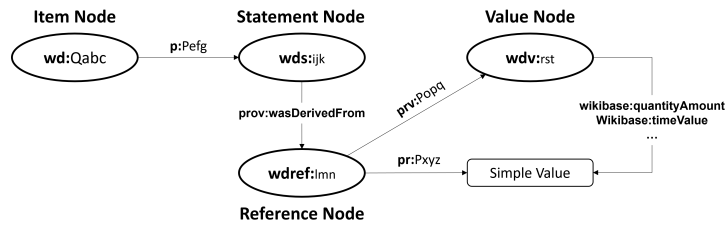


Fig. 2. The RDF model of Wikidata references, derived from [39].  $abc$  is an arbitrary Q-ID.  $efg$  is an arbitrary fact-specific P-ID.  $opq$  and  $xyz$  are arbitrary reference-specific P-IDs. In Wikidata, each fact has a corresponding *Statement Node* used to present the context of the fact. If the statement is referenced, for each reference there is a *Reference Node*. Reference Nodes can have *Simple Values* (literal and URI), or they can point to *Full Values*. A full value points to additional metadata about the value, such as ranges, precision, or timezone.

## DIMENSION 5. PERFORMANCE

In Linked Data, the performance of the dataset deals with the degree of responsiveness to a high number of requests. According to Zaveri et al., “performance refers to the efficiency of a system that binds to a large dataset, that is, the more performant a data source the more efficiently a system can process data” [13]. The measures of evaluating this dimension are the usage of hash-URIs instead of slash-URIs [32] (as cited in [13]), low latency [8, 15, 32], high throughput [15], and scalability of a data source [32] (as cited in [13]). This dimension is not meaningful in the context of references.

## Category II. Intrinsic

The intrinsic category contains dimensions that are independent of the user’s context. This category focuses on whether information correctly and compactly represents real-world data and whether the information is logically consistent in itself [13]. Dimensions that belong to this category are accuracy, consistency, and conciseness [13].

## DIMENSION 6. ACCURACY

According to Zaveri et al., “Accuracy is defined as the extent to which data is correct, that is, the degree to which it correctly represents the real world facts and is also free of syntax errors. Accuracy is classified into (i) syntactic accuracy, which refers to the degree to which data values are close to its corresponding definition domain, and (ii) semantic accuracy, which refers to the degree to which data values represent the correctness of the values to the actual real-world values” [13]. Accuracy is an important aspect of data quality as it is sometimes considered a synonym of quality in the literature [4]. Bizer and Cyganiak [22] suggest outlier detection methods (e.g. distance-based, deviations-based, and distribution-based methods [37]) as metrics of accuracy. Checking the use of proper data types for literals and assuring that literals are abiding by the data types is also used as a metric for accuracy [4, 10, 15]. By evaluating the quality of five open KGs, Färber et al. [4], based on Batini et al. [38], considered two syntactic metrics (syntactic validity of RDF documents and syntactic validity of literals) and one semantic metric (semantic validity of triples) for measuring the accuracy. We use these three metrics in the context of references.

*Metric 5. Syntactic Validity of Reference Triples* The ratio of statement nodes whose referencing metadata sub-graph matches the Wikidata data model, to the total number of statement nodes. Figure 2 shows the Wikidata referencing data model.

*Metric 6. Syntactic Validity of Reference Literals* The ratio of reference literal values that match the Wikidata specified literal rules to the total number of literals. Figure 3 shows an example of a regular expression specified to reference-specific property *title* (P1476).

*Metric 7. Semantic Validity of Reference Triples (Subjective)* The ratio of reference triples that based on their corresponding statement, exactly match a gold standard set of <statement,references> to the total number of reference triples.

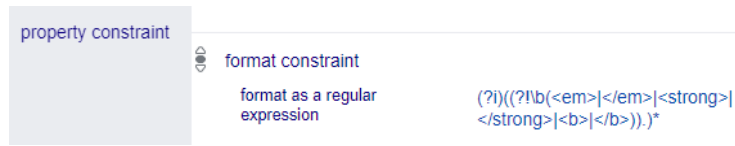


Fig. 3. One of the regular expressions of property *title* (P1476) in Wikidata.



Fig. 4. Qualifiers of the property scope value of property *stated in* (P248) constraints show that it can be used in references and/or qualifiers.

**DIMENSION 7. CONSISTENCY**

Combining the definition of multiple studies, Zaveri et al. stated that a knowledge base is consistent if it is “free of (logical/formal) contradictions with respect to particular knowledge representation and inference mechanisms.” [13]. Assessing this dimension depends on the knowledge inference methods (e.g., OWL or RDFS) used for inference in the knowledge base. The rate of entities that are members of disjoint classes [4, 10, 15], is one of the common criteria for this dimension. Other common metrics for checking consistency in Linked Data are usage of undefined classes [10, 15], ontology hijacking [10, 15], and OWL inconsistencies [10, 15], the extent of values compliance with the domain/range of data types [4, 15], and misuse of predicates [40]. In the context of references, consistency can be measured by three metrics: (i) use of consistent (reference-specific) predicates, (ii) compatibility of values with the domain and range of reference-specific properties, and (iii) compatibility of different references of an item/statement.

*Metric 8. Consistency of Reference Properties* The ratio of reference properties specified to be used in reference triples to the total number of reference properties. In Wikidata *property constraint* (P2302) carries another metadata about where the property should be used. This metadata is placed under the *property scope* (P5314) qualifier of the *property scope constraint* (Q53869507) values. Figure 4 shows the scope constraints of the property *stated in* (P248).

*Metric 9. Range Consistency of Reference Triples* The ratio of reference properties whose values are consistent with the specified ranges by Wikidata to the total number of reference properties. In Wikidata, ranges of a property can be fetched from the *class* (P2308) qualifier of the *property constraint* (P2302) statements that have the *value-type constraint* (Q21510865). Figure 5 shows the value allowed types for the property *stated in* (P248).

*Metric 10. Multiple References Consistency (Subjective)* The ratio of multiple-referenced statements whose references are consistent with each other to the total number of multiple-referenced statements.

**DIMENSION 8. CONCISENESS**

According to Zaveri et al., “conciseness refers to the redundancy of entities, be it at the schema or the data level. Conciseness is classified into (i) intensional conciseness (schema level) which refers to the case when the data does not contain redundant attributes and (ii) extensional conciseness (data level) which refers to the case when the data does not contain redundant objects” [13]. Redundancy in both schema and instance levels is covered in the Mendes et al. [9] framework. Debattista et al. [15] considered instance-level redundancy in their investigation of Linked Data. In the context of references, redundancy in the instance level is not considered a negative point in the quality of references (because different but equivalent references increase the trust in data). Note that Redundancy at the instance level is different from exact duplication. Exact duplication occurs when an entire triple is repeated in a dataset due to serialization errors. Such duplications are rare and can be ignored.

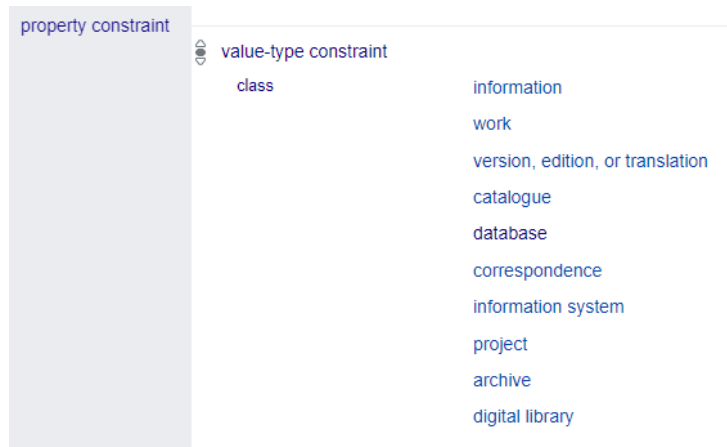


Fig. 5. Qualifiers of the value-type constraint value of property *stated in* (P248) constraints show the classes that can be used as values for this property.

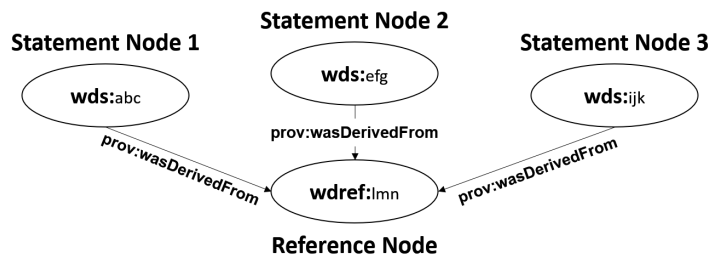


Fig. 6. Reference sharing in Wikidata data model. Statement nodes 1, 2, and 3 are all derived from the same source.

We consider redundancy in both schema and instance levels. The existence of different predicates for pointing to the same provenance information is the schema-based metric of conciseness. To illustrate the conciseness in references instance-level, we also provide a metric to measure reference sharing [18].

**Metric 11. Schema-level Conciseness of Reference Properties (Subjective)** The ratio of reference properties with another equivalent reference property to the total number of reference properties.

**Metric 12. Ratio of Reference Sharing** The ratio of reference nodes who are shared with more than one statement to the total number of reference nodes. Figure 6 shows reference sharing in the Wikidata data model.

### Category III. Trust

This category contains dimensions that illustrate the perceived trustworthiness of the dataset [13]. These dimensions are reputation, believability, verifiability, and objectivity [13]. In KGs, having references at different levels is a metric of trustworthiness [4]. When we aim to define trustworthiness in the context of references, we emphasize external sources presented as references.

#### DIMENSION 9. REPUTATION

Zaveri et al. defined reputation as “a judgment made by a user to determine the integrity of a data source” [13]. Reputation is the social aspect of trust in the Semantic Web [41], thus, the reputation criteria try to measure the opinions of users about datasets [42, 43]. Investigating the opinions of users can be done explicitly through questionnaires and decentralized voting such as Gil and Artz’s study [43]. On the other hand, implicit methods like relying on page ranks can be used as a metric for reputation [42, 43]. Golbeck and Hendler [41], proposed an



algorithm for computing the reputation of objects considering the incoming links to the object. We use the following metric to measure the referencing reputation of the dataset:

*Metric 13. External URIs Reputation* The average of the external URIs page ranks.

#### DIMENSION 10. BELIEVABILITY

Zaveri et al. define believability as “the degree to which the information is accepted to be correct, true, real and credible” [13]. Believability sometimes is considered as a synonym for *trustworthiness* [4, 15, 44]. Färber et al. considered trustworthiness as a collective dimension of believability, reputation, objectivity, and verifiability [4]. This dimension indicates the degree to which the user trusts the accuracy of data without evaluating it [8].

Believability considers the data consumer side in the trust category and is closely related to the reputation of the dataset [4]. Believability is a highly subjective dimension that needs to acquire the data users’ opinion [45, 46]. However, there are different objective metrics to measure believability, e.g., the use of trust ontologies in data [47] and clarifying the provenance of data [4, 15]. In the context of references, we define the metric for the believability dimension based on the fact that references are added more by humans or machines.

*Metric 14. Human-added References* The ratio of human-added reference triples to the total number of reference triples.

#### DIMENSION 11. VERIFIABILITY

Verifiability is defined as the “degree by which a data consumer can assess the correctness of a dataset” [13]. Verifiability indicates the possibility of verifying the correctness of the data [4]. A dataset is verifiable if there exists concrete means of assessing the correctness of data. Therefore, providing the provenance of facts [4, 15] and the use of digital signatures to sign RDF datasets [48] are suggested metrics for this dimension. Subjective methods like using unbiased trusted third-party evaluators are also suggested in the literature [8].

In the context of references, the document type of a reference is the subject of measurement. We score external sources (external or internal) based on their document type, and define the metric as follows:

*Metric 15. Verifiable Type of References* The average of type verifiability scores of the external sources. The pre-defined document types with grades from high to low are scholarly articles, well-known trusted knowledge bases, books and encyclopedic articles, and finally magazines and blog posts.

#### DIMENSION 12. OBJECTIVITY

Objectivity is defined as “the degree to which the interpretation and usage of data is unbiased, unprejudiced and impartial” [13]. As believability focuses on the subject side (data consumer), objectivity considers the object side (data provider) of the dataset [4]. Verifiability has a direct impact on objectivity [49]. Bizer [8] considered three subjective criteria to measure objectivity, including the neutrality of the publisher, confirmation of facts by various sources, and checking the bias of data. In the context of references, we define objectivity as the ratio of statements that have more than one provenance.

*Metric 16. Multiple References for Statements* The ratio of multiple-referenced statements (statements with more than one reference) to the total number of referenced statements.

### Category IV. Dynamicity

Dimensions of this category monitor the freshness and frequency of data updates [13]. These dimensions, according to Zaveri et al. [13] are currency, volatility, and timeliness. [4, 21] considered dynamicity as the timeliness dimension in the contextual category. Bizer [8] however, considered dynamicity as the timeliness dimension in the intrinsic category. More recently, Ferradji et al. [50] measured currency, volatility, and timeliness in Wikidata. Measuring the dimensions of this category is based on date/time values. There are different properties in the context of references to capture the date/time of a reference. In PROV-O [51] properties like `prov:generatedAtTime` and

prov:Time can be used. Wikidata uses *retrieved (P813)* for demonstrating the retrieval date of an external URI. In Wikidata, the edit history is also another way to capture reference modification dates.<sup>2</sup>

#### DIMENSION 13. CURRENCY

According to Zaveri et al., “currency measures how promptly the data is updated” [13]. This dimension is usually measured by computing the distance between the latest time data modified and the observation time [9]. Sometimes the release time of data is also included in the calculation [14]. Another way to measure this is to consider the time that it takes for a change made to a dataset for a known real-world event [13]. For example, the time that Wikidata takes to update a wrestler’s statement for his new Olympic medal is a currency measurement.

Using up-to-date references is very important in some cases, e.g., medical facts. In the context of references, currency can be measured via two metrics: the freshness of reference triples and the freshness of external URIs.

*Metric 17. Freshness of Reference Triples* The average time elapsed since the last update of reference triples, relative to their total existence duration.

*Metric 18. Freshness of External URIs* The average time elapsed since the last update of external URIs, relative to their total existence duration.

#### DIMENSION 14. VOLATILITY

According to Zaveri et al., “volatility refers to the frequency with which data varies in time” [13]. While currency focuses on the updates of data, volatility reports the frequency of change in data. Volatility can give the user an expectation of the near update. Volatility besides the currency can be a metric for the validity of data [13]. The `changeFreq` attribute of Semantic Sitemap [52] is a suggested metric for volatility [32] (as cited in [13]). Based on the `changeFreq` attribute of the external URIs, we define a metric for the volatility of external URIs.

*Metric 19. Volatility of External URIs* The average of the frequency-of-update scores, based on the `<changeFreq>` attribute in external URIs.

#### DIMENSION 15. TIMELINESS

“Timeliness measures how up-to-date data is, relative to a specific task” [13]. This dimension is a combination of currency and volatility and specifies data as up-to-date as it should be. Since the definition of timeliness is related to the task at hand, we define the metric *timeliness of external URIs* as the difference between volatility and currency.

*Metric 20. Timeliness of External URIs* The fraction of the external URI freshness score to their volatility.

### Category V. Contextual

The contextual category includes dimensions that mostly depend on the context of the task at hand [13]. There is more variability in the literature as to which dimensions belong to this category. Färber et al. [4] considered timeliness and trustworthiness with relevancy in this category. According to Zaveri et al. [13], *correctness*, *amount of data*, and *relevancy* belong to the contextual category. We follow the Zaveri et al. categorization.

#### DIMENSION 16. COMPLETENESS

Completeness indicates the extent to which the dataset covers real-world structures and instances. It is an extensive dimension that contains several sub-categories in some sources, e.g., Furber et al. [11] and Mendes et al. [9] that considered completeness in the schema and data instances. Zaveri et al [13] provided a comprehensive definition, according to which, “completeness refers to the degree to which all required information is present in a particular dataset. In terms of Linked Data, completeness comprises the following aspects: (a) Schema completeness, the degree to which the classes and properties of an ontology are represented, thus can be called "ontology completeness",

<sup>2</sup>A SPARQL query service for Wikidata history has been explained in [https://www.wikidata.org/wiki/Wikidata:History\\_Query\\_Service](https://www.wikidata.org/wiki/Wikidata:History_Query_Service) - last edited 11 May 2023

(b) Property completeness, measure of the missing values for a specific property, (c) Population completeness is the percentage of all real-world objects of a particular type that are represented in the datasets and (d) Interlinking completeness has to be considered especially in Linked Data and refers to the degree to which instances in the dataset are interlinked” [13]. Zaveri et al. definition reflects the criteria used to measure completeness in Linked Data. These criteria are schema completeness, property completeness, population (data instances) completeness, and interlinking completeness. In the context of references, we provide metrics for schema, property, and population completeness.

#### *Metric 21. Class/Property Schema Completeness of References*

- Class Schema Completeness of References: The ratio of classes in the dataset with defined reference-specific properties at the schema level to the total number of classes.
- Property Schema Completeness of References: The ratio of properties in the dataset with defined reference-specific properties at the schema level to the total number of properties.

#### *Metric 22. Schema-based Property Completeness of References*

The average completeness ratio of reference properties in the dataset relative to their schema-defined reference properties for each property. The completeness ratio of a given reference property represents the proportion of statements with its corresponding schema-defined property to the total number of referenced statements with that given specific reference property.

*Metric 23. Property Completeness of References* The average completeness ratio of reference properties in the dataset relative to their corresponding fact classes at the instance level. The ratio indicates the proportion of referenced facts with a specific reference property to the total number of facts with the corresponding property at the instance level.

*Metric 24. Population Completeness of References (Subjective)* The ratio of referenced statements in the dataset where the statements come from a selected set of facts to the total number of statements with the same facts properties.

### DIMENSION 17. AMOUNT-OF-DATA

According to Zaveri et al., “Amount-of-data refers to the quantity and volume of data that is appropriate for a particular task” [13]. In the context of linked data, this dimension represents the coverage of the dataset for a specific task. It includes statistics on the number of entities, the number of properties, and the number of triples [13]. In the context of references, this dimension can include quantitative statistics of references. Beghaeiraveri et al. [18] provided a statistical review of 6 Wikidata subsets that are relevant to this dimension. They investigated the number of reference nodes, the total number of reference triples, the distribution of triples per reference node, the usage frequency of reference-specific properties, and the percentage of shared references. For all of these concepts, we formally define a quantitative metric in the Amount-of-data dimension. In these metrics, having quantitative statistics and the distribution of scores helps users estimate the coverage of references.

*Metric 25. Ratio of Reference Nodes per Statement* The ratio of distinct reference nodes to the total number of statements in the dataset, indicates the richness of reference metadata in capturing diverse sources for facts.

*Metric 26. Ratio of Reference Triples per Statement* The ratio of distinct reference triples to the total number of statements in the dataset, provides an overview of the referencing depth and richness in capturing multiple details for each fact.

*Metric 27. Ratio of Reference Triples per Reference Node* The complement of the ratio of distinct reference nodes to the total number of reference triples in the dataset, representing the average number of triples associated with each reference node and indicating the level of detail in referencing.

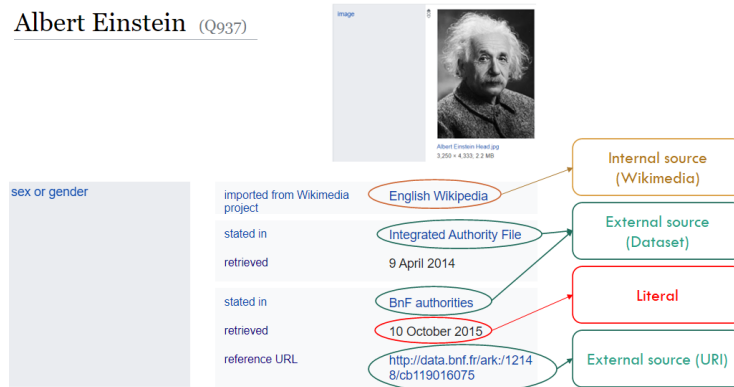


Fig. 7. Different types of reference values in Wikidata for *Albert Einstein (Q937)*

**Metric 28. Ratio of Reference Literals per Reference Triple** The ratio of distinct reference literals to the total number of reference triples in the dataset. Note that the Wikidata data model has three types of reference values: external sources, internal sources, and literals (Figure 7).

#### DIMENSION 18. RELEVANCY

According to Zaveri et al., “Relevancy refers to the provision of information which is in accordance with the task at hand and important to the users’ query” [13]. In Linked Data, relevancy metrics are checking the existence of meta-information attributes and the extent of using relevant external links and/or relevant `owl:sameAs` predicates [8]. Farber et al. [4] measured the relevancy of facts in KGs by looking at whether there is a ranking system on facts in the KG. Relevancy is one of the main conditions of Wikidata references [29]. According to Wikidata guidelines, references “should point to specific sources that back up the data provided in a statement” [29]. Few efforts are measuring the relevance of references in Wikidata. Judging of the relevance of a reference is highly subjective [4]. Due to the subjective nature of the concept, Piscopo et al. [16] proposed an approach to evaluate the relevance of Wikidata English external sources through microtask crowdsourcing followed up with a machine-learning algorithm. Recently, they extended the approach by supporting different languages, increasing the sample size, using a more recent Wikidata dump, and enhancing the machine-learning algorithm [2]. Their machine-learning-trained model is useful for measuring our relevancy metrics. We provide two metrics for the relevance of the references: one considers all reference triples and the other considers shared references.

**Metric 29. Relevance of Reference Triples (Subjective)** The ratio of reference triples deemed relevant to their associated facts to the entire reference triples.

**Metric 30. Relevance of Shared References (Subjective)** The complement of the ratio of shared reference triples that are deemed irrelevant to their corresponding fact to the total fact-reference triples.

#### Category VI. Representational

Representational dimensions indicate the proper presentation and ease of understanding of data to the user. According to Zaveri et al. [13], in Linked Data these dimensions are *representational-conciseness*, *representational-consistency*, *understandability*, *interpretability*, and *versatility*. Farber et al. [4] considered two dimensions *ease of understanding* (equivalent to understandability) and *interoperability* (composite of interpretability, representational consistency, concise representation). We follow the Zaveri et al. categorization.

#### DIMENSION 19. REPRESENTATIONAL-CONCISENESS

According to Zaveri et al., in the context of Linked Data, “representational-conciseness refers to the representation of the data which is compact and well-formatted on the one hand and clear and complete on the other hand” [13]. Literature measures this by keeping URIs short and free of SPARQL parameters [12, 15] and also avoiding the use

of RDF reification, containers, and collections [4, 12, 15]. As references are statements about statements, reification is inevitable [4]. However, short URIs in external sources can help machines process references.

*Metric 31. External Sources URL Length* The average of the length scores of the external sources URLs. Higher scores are given to shorter URLs.

#### DIMENSION 20. REPRESENTATIONAL-CONSISTENCY

Consistency in representation refers to “the degree to which the format and structure of the information conform to previously returned information as well as data from other sources” [13]. Representational consistency metrics assess the degree of using existing terms in the context [4] and established terms that already are used in the dataset [15]. In the context of referencing, despite there being no standard vocabulary, there are well-known general ontologies, e.g., Dublin Core Metadata [53] and the W3C PROV-O [51]. In addition, some ontologies use their specific properties for references, e.g., Genealogy.<sup>3</sup> Wikidata reference properties are in the form of P-IDs. Property labels also are specific; Wikidata does not use other well-known vocabularies. Since this dimension indicates the importance of using a steady and consistent manner (vocabularies and properties) to represent data [13], we define a metric based on the diversity of properties used in reference triples.

*Metric 32. Diversity of Reference Properties* The complement of the ratio of distinct reference properties to the total number of reference triples. For accurate insight, the diversity is measured based on the number and variety of reference properties used across all reference triples.

#### DIMENSION 21. UNDERSTANDABILITY

Understandability deals with the readability and accessibility of data for humans. According to Zaveri et al., “understandability refers to the ease with which data can be comprehended, without ambiguity, and used by a human information consumer” [13]. Metrics for evaluating understandability in Linked Data look for the percentage of entities, classes and properties with human-readable metadata, e.g., using `rdfs:label` and/or `rdfs:comment` [4, 15], the existence of example SPARQL queries for the dataset [32], the existence of a regular expression that expresses the URIs of the dataset [4, 15], the existence of a vocabulary list for the dataset [15], and using mailing lists and message boards [32]. In the context of references, we assess human readability by checking how many reference predicates have labels or comments and to which extent the external sources are handy, i.e., easy to access.

*Metric 33. Human-readable labelling of Reference Properties* The ratio of reference properties in the dataset that have associated human-readable labels to the total number of distinct reference properties.

*Metric 34. Human-readable Commenting of Reference Properties* The ratio of reference properties in the dataset that have associated human-readable descriptions to the total number of distinct reference properties.

*Metric 35. Handy External Sources* The average of the external source references reachability scores, with higher scores given to sources that are easy to reach for human users.

#### DIMENSION 22. INTERPRETABILITY

According to Zaveri et al., “Interpretability refers to technical aspects of the data, that is, whether the information is represented using an appropriate notation and whether it conforms to the technical ability of the consumer” [13]. Interpretable data increases the reusability and facilitates the integration with other datasets [13]. This dimension also considers technical aspects of data representation [4] and is a way to measure how exploring data is easy for machines. The interpretability criteria in Linked Data are using well-defined and unique identifiers across the dataset [8, 15], and avoiding the usage of RDF blank nodes [4, 12, 15]. In the context of references, we define a metric based on avoiding blank node usage in references.

---

<sup>3</sup><http://gov.genealogy.net/ontology.owl> - accessed 15 April 2024

Table 2

The classification of referencing quality assessment metrics based on the target of evaluation. Metrics in *italic* are subjective.

Target	Metrics
RDF structure (properties, triples, nodes)	Interlinking of Reference Properties, Syntactic Validity of Reference Triples, Syntactic Validity of Reference Literals, <i>Semantic Validity of Reference Triples</i> , Consistency of Reference Properties, Range Consistency of Reference Triples, <i>Schema-level Consciencess of Reference Properties</i> , Ratio of Reference Sharing, Multiple References for Statements, Property Completeness of References, <i>Population Completeness of References</i> , Ratio of Reference Nodes per Statement, Ratio of Reference Triples per Statement, Ratio of Reference Triples per Reference Node, Ratio of Reference Literals per Reference Triple, Diversity of Reference Properties, Human-readable labelling of Reference Properties, Human-readable Commenting of Reference Properties, Usage of Blank Nodes in References, Multilingual labelling of Reference Properties, Multilingual Commenting of Reference Properties
Metadata (schemas, historical metadata, sources metadata)	External URIs Domain Licensing, External URIs Reputation, Human-added References, Freshness of Reference Triples, Freshness of External URIs, Volatility of External URIs, Timeliness of External URIs, Class/Property Schema Completeness of References, Schema-based Property Completeness of References
Source content	Availability of External URIs, Security External URIs, <i>Multiple References Consistency</i> , Verifiable Type of References, <i>Relevance of Reference Triples</i> , <i>Relevance of Shared References</i> , External Sources URL Length, Handy External Sources, Multilingual Sources, Multilingual Referenced Statements

*Metric 36. Usage of Blank Nodes in References* The complement of the ratio of blank nodes in the union set of all reference nodes, reference properties, and objects in the dataset to the total number of elements in that union set.

#### DIMENSION 23. VERSATILITY

According to Zaveri et al., “Versatility refers to the availability of the data in an internationalized way, the availability of alternative representations of data and the provision of alternative access methods for a dataset.” In Linked Data, versatility has metrics such as providing different serialization for data [4, 15] and multilingualism [4, 15, 54]. In the context of references, multilingualism helps various language speakers verify the facts. Furthermore, non-English cultures and language facts require sources in their language.

*Metric 37. Multilingual labelling of Reference Properties* The ratio of reference properties in the dataset that have associated labels in languages other than English to the total number of distinct reference properties.

*Metric 38. Multilingual Commenting of Reference Properties* The ratio of reference properties in the dataset that have associated descriptions in languages other than English to the total number of distinct reference properties.

*Metric 39. Multilingual Sources* The ratio of non-English sources, including both internal and external references, to the total number of non-literal sources in the dataset.

*Metric 40. Multilingual Referenced Statements* The ratio of facts in the dataset that has at least one non-English source reference to the total number of facts.

### 3.2. Alternative Metric Categorizations

As Section 3.1 represents the metrics in Zaveri et al. categorizations (Table 1), the metrics can be classified in alternative categorizations based on their novelty in the context of references and the part of the referencing they focus on. Table 2 shows the classification of all defined metrics based on the metric targets, i.e., the part of referencing on which the quality review is conducted. Table 3 separates our referencing quality metrics into three categories -in terms of the coexistence with traditional Linked Data quality criteria. Note that the novel metrics are still packed in traditional Linked Data dimensions and categories. For example, the Human-added References metric is a new metric which has not already been in Link Data quality criteria; however, as it investigates the believability of a reference to the users, it fits in the Believability dimension.

Table 3

The categorization of referencing quality assessment metrics based on their relation with traditional Linked Data criteria. Metrics in *italic* are subjective.

Relationship	Metrics
Direct use of Linked Data quality criteria (with minor adjustments)	Availability of External URIs (Availability), External URIs Domain Licensing (Licensing), Security External URIs (Security), Syntactic Validity of Reference Literals (Accuracy), <i>Semantic Validity of Reference Triples</i> (Accuracy), Range Consistency of Reference Triples (Consistency), External URIs Reputation (Reputation), Freshness of Reference Triples (Currency), Freshness of External URIs (Currency), Volatility of External URIs (Volatility), Timeliness of External URIs (Timeliness), Class/Property Schema Completeness of References (Completeness), <i>Population Completeness of References</i> (Completeness), External Sources URL Length (Representational-conciseness), Human-readable labelling of Reference Properties (Understandability), Human-readable Commenting of Reference Properties (Understandability), Usage of Blank Nodes in References (Interpretability), Multilingual labelling of Reference Properties (Versatility), Multilingual Commenting of Reference Properties (Versatility)
Using the idea behind Linked Data quality criteria (major changes)	Interlinking of Reference Properties (Interlinking), Syntactic Validity of Reference Triples (Accuracy), Consistency of Reference Properties (Consistency), <i>Schema-level Consciences of Reference Properties</i> (Consciences), Schema-based Property Completeness of References (Completeness), Property Completeness of References (Completeness), <i>Relevance of Reference Triples</i> (Relevancy), <i>Relevance of Shared References</i> (Relevancy), Multilingual Sources (Versatility), Multilingual Referenced Statements (Versatility)
Novel metrics	<i>Multiple References Consistency</i> (Consistency), Ratio of Reference Sharing (Consciences), Human-added References (Believability), Verifiable Type of References (Verifiability), Multiple References for Statements (Objectivity), Ratio of Reference Nodes per Statement (Amount-of-data), Ratio of Reference Triples per Statement (Amount-of-data), Ratio of Reference Triples per Reference Node (Amount-of-data), Ratio of Reference Literals per Reference Triple (Amount-of-data), Diversity of Reference Properties (Representational-consistency), Handy External Sources (Understandability)

#### 4. Referencing Quality Scoring System (RQSS)

The Referencing Quality Scoring System (RQSS) is a data quality assessment methodology [13] that aims to measure the referencing quality of the Wikidata and other Wikibase-hosted datasets.<sup>4</sup> The main constituent of RQSS is the assessment framework defined in Section 3. As a system, RQSS has four components: *Extractor*, *Metadata Extractor*, *Framework Runner*, and *Presenter*. Figure 8 shows these components and (part of) data flow between them. In the following paragraphs, we explain the details of the system.

**Input** RQSS data pipeline starts with an RDF dataset based on the Wikidata data model. The input dataset can be the entire Wikidata or a subset of it.<sup>5</sup> In addition to the input dataset, RQSS needs other metadata: revision history metadata such as reference editors and the reference editing date-time, and schema information. These data come directly from the Wikidata knowledge base public SPARQL endpoint and its HTML pages.

**Extractor and Metadata Extractor** Extractor fetches the referencing-related sets required for calculating metrics from the input dataset. For example, to calculate the availability and security dimensions, the Extractor retrieves all external source URIs. As the Extractor retrieves the input dataset referencing data, the Metadata Extractor deals with external referencing data required for metrics, e.g., a summary of referencing metadata in Wikidata Entity-Schemas, which is required by completeness metrics such as Metric 21 and 22.

**Framework Runner** This module calculates the referencing quality metrics. For each dimension of the assessment framework, the Framework Runner takes the required data from the Extractor and Metadata Extractor and then calculates the score of the dimension's metrics. The user can apply different weights to each metric (the default weights are 1) depending upon the user's own perspective of the importance of each metric. The Framework Runner then returns the final weighted average of the scores. For some metrics, the Framework Runner also returns the disaggregated scores. For example, the score of the completeness metrics is the average completeness ratio of

<sup>4</sup><https://wikiba.se/> - accessed 15 April 2024

<sup>5</sup>Full Wikidata dumps can be downloaded from <https://dumps.wikimedia.org/wikidatawiki/entities/> - accessed 14 April 2024

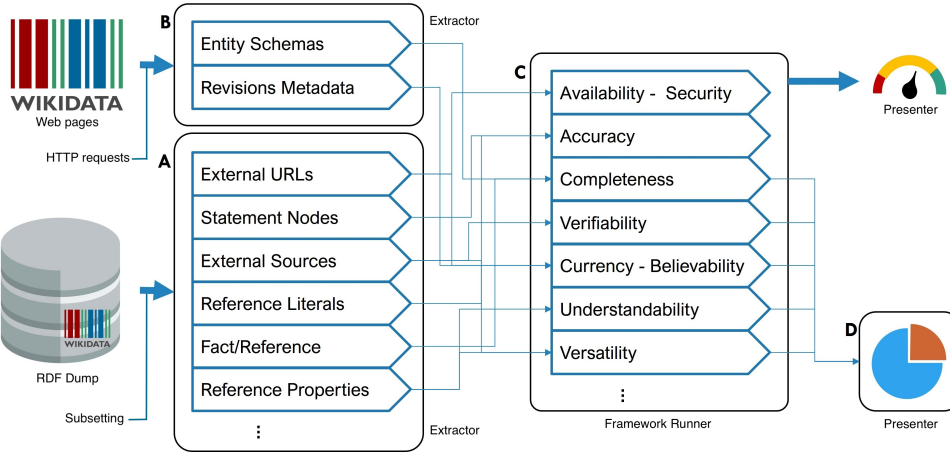


Fig. 8. Main components of RQSS and part of its data pipeline. *Extractor* (component A) fetches referencing data such as external URIs, statement nodes, etc. from the input dataset (which should be based on the Wikidata/Wikibase data model). The *Metadata Extractor* (component B) independently retrieves information such as EntitySchema (E-IDSs) summary and historical data from Wikidata. The extracted data is then given to the *Framework Runner* (component C), which calculates reference quality metrics in different dimensions and returns a referencing quality score of the input dataset as a weighted average between 0 and 1. In addition to the score, the Framework Runner also produces disaggregated scores (for some dimensions), which are then converted into visual charts by the *Presenter* (component D).

multiple reference properties. In that case, the Framework Runner returns the completeness ratio of each property besides the metric score.

*Presenter* To facilitate understanding of the data behaviour in large datasets, the Presenter draws different visual charts for those metrics that the Framework Runner returns disaggregated scores.

*RQSS Implementation* To automate the assessment of referencing quality in Wikidata and other Wikibase-hosted datasets, we implement the objective metrics of the RQSS assessment framework in a reusable environment. An automatic implementation facilitates monitoring the referencing quality regularly and helps users to judge the quality quantitatively. We implement RQSS in Python. Python is well-designed for Big Data science research and easy to write and debug. The code repository of the implementation is available on GitHub [55]. In the current version v1.0.2, all main components of Figure 8 are implemented. The input dataset (entire Wikidata or a subset) must be available through a SPARQL endpoint. The Extractor fetches the data by performing multiple SPARQL queries on the endpoint. Each metric is implemented as an independent class. The Metadata Extractor is embedded inside the metric classes and performs HTTP requests from different Wikidata web pages to fetch the required metadata. Extraction, as well as metrics, can be performed independently and simultaneously.

## 5. RQSS Evaluation Over Wikidata Subsets

Due to the limitations of our available resources, we cannot apply RQSS to the whole of Wikidata, which currently has more than 100 GB of data containing 1.2 billion statements representing 100 million items. RQSS is used to compute the scores and present the graphical charts of three topical and four random Wikidata subsets. Through subsetting, we establish a comparison platform and gain valuable insight into the referencing quality in different topics and also Wikidata as a whole.



Table 4

Initial statistics of the Wikidata subsets: The number of items, statement nodes, reference nodes, and referenced statements (statements with at least one reference).

Subset	Items	Statements	References	Referenced Statements
Gene Wiki	9,203,257	97,062,660	9,742,813	63,521,696 (65%)
Music	982,730	12,743,480	1,585,122	6,348,140 (50%)
Ships	128,815	1,116,976	61,996	301,290 (27%)
Random 100K #1	86,916	1,225,313	94,966	946,523 (77%)
Random 100K #2	86,865	1,226,097	94,982	940,552 (76%)
Random 500K	433,364	6,117,915	453,273	4,704,898 (77%)
Random 1M	864,665	12,231,380	894,093	9,392,549 (77%)

### 5.1. Subsetting Overview

We extract three topical subsets corresponding to three Wikidata WikiProjects: Gene Wiki [56], Music, and Ships [18].<sup>6</sup> These projects are active in curating references and have various sizes, covering a wide range of scientific and cultural fields of activities in Wikidata for investigating references. Besides topical subsets, we extract four random subsets in varying sizes as a random sampling of Wikidata without considering a specific topic. All subsets are extracted from the Wikidata full JSON dump of 3 January 2022 using the evaluated subsetting tool WDumper [57, 58].<sup>7</sup> Our subsetting approach is item-based, i.e., selecting the desired items (Q-IDs) and extracting all statements of those items [19]. For topical subsetting, we use the approach of [18]. For random subsetting, we tweaked the WDumper code to extract items from the dump by Q-IDs [58]. We then deployed a Python script to generate random Q-IDs and created two specification files with one hundred thousand Q-IDs, one with five hundred thousand Q-IDs, and one with one million Q-IDs.<sup>8</sup> Wdumper is configured to retrieve all referencing and provenance metadata for the selected items. To optimize the subset size, we ignore metadata irrelevant to referencing, such as item labels, item descriptions, and item qualifiers. All subsets are indexed and queried locally via Blazegraph 2.1.6. The specification files of topical and random subsets can be found in the GitHub repository of the paper [59]. The RDF files for each of the subsets can be found in [60].

Table 4 shows for each subset the number of items, statements, references, and statements that have at least one reference. We note that the referencing rate in random subsets is generally higher than in the topical subsets. We also observe that items are missing from each of the random subsets, i.e. none of the random subsets contains the expected number of items, but this rate is consistent across the four subsets. Wikidata item identifiers start with Q, followed by an incremental number. At the end of December 2021, the maximum Q-ID in Wikidata was 110,272,953. The random generator script is set to generate the given number of random Q-IDs (100K, 500K, or one million) between Q1 and Q110272953.<sup>9</sup> However, after the extraction, we recognized that the number of extracted items in the random subsets is 15% less than expected. We hypothesise that about 15% of Wikidata Q-IDs are not resolvable anymore.

#### 5.1.1. Random Subsets Topic Coverage

Table 5 shows the intersection between the random subsets, i.e., the number of overlapping items. Considering the sum-up size of each pair of subsets, the amount of overlap is negligible. However, the uniformity of referencing and missing item rates in the four random subsets with different sizes reveals the need for a deeper look at the main classes of instances inside the subsets. We call this process finding *topic coverage*; identifying classes with a higher number of item instances, similar to Wikidata [3, §(What is in Wikidata)].<sup>10</sup> To achieve this, we query all classes

<sup>6</sup>Gene Wiki WikiProject: [https://www.wikidata.org/wiki/Wikidata:WikiProject\\_Gene\\_Wiki](https://www.wikidata.org/wiki/Wikidata:WikiProject_Gene_Wiki), Music WikiProject: [https://www.wikidata.org/wiki/Wikidata:WikiProject\\_Music](https://www.wikidata.org/wiki/Wikidata:WikiProject_Music), and Ships WikiProject: [https://www.wikidata.org/wiki/Wikidata:WikiProject\\_Ships](https://www.wikidata.org/wiki/Wikidata:WikiProject_Ships) - accessed 14 April 2024

<sup>7</sup>The Wikidata full JSON dump of 3 January 2022 can be downloaded from <https://academictorrents.com/details/229cf2331ad43d4706efd435f6d78f40a3c438> - accessed 14 April 2024

<sup>8</sup>The script can be found in [https://github.com/seyedahbr/wdumper/blob/12f0ddf/extensions/create\\_random\\_spec.py](https://github.com/seyedahbr/wdumper/blob/12f0ddf/extensions/create_random_spec.py) - accessed 14 April 2024

<sup>9</sup>The script can be found in [https://github.com/seyedahbr/RQSS\\_Evaluation/blob/5178f83/scripts/create\\_random\\_spec.py](https://github.com/seyedahbr/RQSS_Evaluation/blob/5178f83/scripts/create_random_spec.py) - accessed 15 April 2024

<sup>10</sup>Note that the pie chart belongs to December 2019 when Wikidata had about 71 million items.

Table 5  
The number of overlapping items in random subsets.

	Random 100K #2	Random 500K	Random 1M
Random 100K #1	62	372	779
Random 100K #2		399	802
Random 500K			3,861

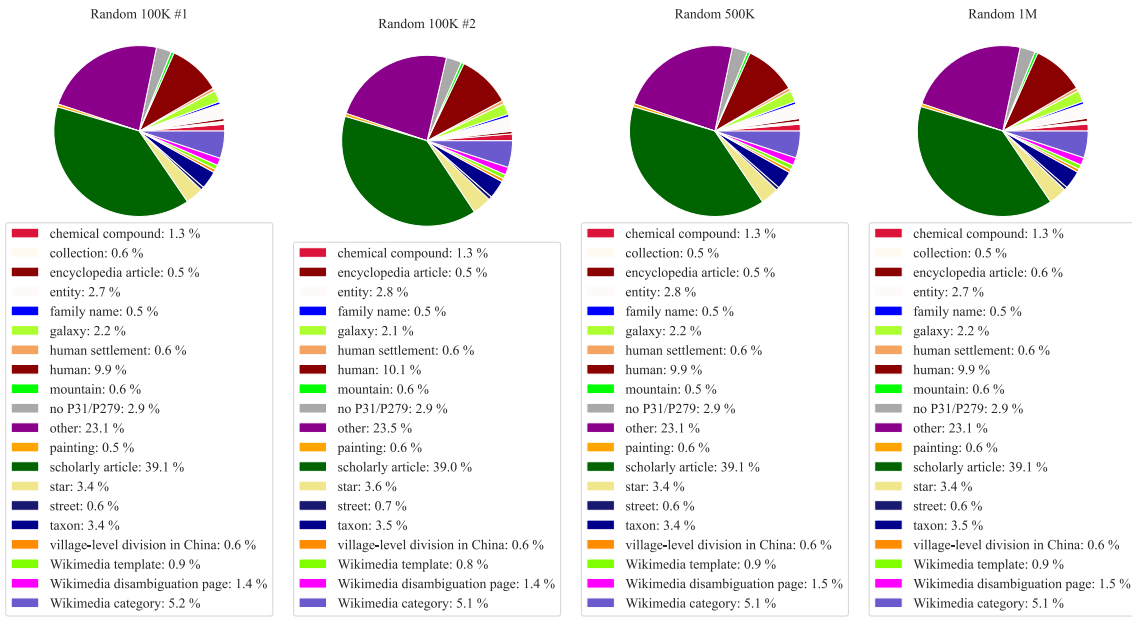


Fig. 9. Topic coverage of the four random subsets. Note that the colours are consistent across the four charts.

of items in the subset and then sort the classes based on the number of items that belong to that class. In the end, to guarantee that classes are disjoint, we remove the duplicated items in low-listed classes, i.e., if an item instance appeared in a top-listed class, it will not be counted in the low-listed class.<sup>11</sup>

Figure 9 shows the topic coverage of the four random subsets. All four subsets have a similar topic coverage. In all subsets, the majority belongs to the *scholarly article* (*Q13442814*) class. The next most frequent classes are *galaxy* (*Q318*) and *star* (*Q523*) (subclass of *astronomical object* (*Q6999*)). The order of frequency in all random subsets follows the same pattern of Wikidata topic coverage in [3, §(What is in Wikidata)]. This topic coverage shows that our random sampling is uniform, and the extracted random subsets are a good approximation of the entire Wikidata.<sup>12</sup>

## 5.2. Comprehensive Metric-by-Metric Analysis of Referencing Quality

In this section, we analyse the quality scores obtained by running RQSS over topical and random subsets in detail metric by metric. We also evaluate the correctness of RQSS by matching the obtained results with the previous knowledge from Wikidata. During this evaluation, we will discuss valuable information from the data composition in Wikidata.

<sup>11</sup>The script can be found in [https://github.com/seyedahbr/RQSS\\_Evaluation/blob/5178f83/scripts/topic\\_coverage.py](https://github.com/seyedahbr/RQSS_Evaluation/blob/5178f83/scripts/topic_coverage.py) - accessed 15 April 2024

<sup>12</sup>The lists of the distinct items in each random subset can be found in [https://github.com/seyedahbr/RQSS\\_Evaluation/tree/5178f8379ddde6b1a9c09ff69905ade1149b58b5/data/TopicCoverageLists/DistinctItems](https://github.com/seyedahbr/RQSS_Evaluation/tree/5178f8379ddde6b1a9c09ff69905ade1149b58b5/data/TopicCoverageLists/DistinctItems) - accessed 15 April 2024

Table 6

RQSS results of availability of external URIs

(Availability), external URIs domain licensing (Licensing), and security of external URIs (Security).

Subset	External URIs	URI Domains	Score (Metric 1)	Score (Metric 2)	Score (Metric 3)
Gene Wiki	2,559,493	10,138	0.9754	0.0635	0.9664
Music	215,161	21,593	0.8754	0.0480	0.8068
Ships	20,737	924	0.9647	0.0541	0.9294
Random 100K #1	48,618	2,057	0.9755	0.0700	0.9648
Random 100K #2	48,279	2,110	0.9739	0.0611	0.9641
Random 500K	240,183	5,952	0.9750	0.0633	0.9597
Random 1M	478,035	9,342	0.9760	0.0597	0.9589

Table 7

RQSS results for interlinking of reference properties.

Subset	Reference Properties	Score (Metric 4)
Gene Wiki	855	0.1274
Music	1,194	0.1122
Ships	97	0.2886
Random 100K #1	586	0.0972
Random 100K #2	607	0.0889
Random 500K	969	0.0804
Random 1M	1,159	0.0733

### 5.2.1. Availability: Availability of External URIs, Licensing: External URIs Domain Licensing, and Security: Security of External URIs

Table 6 shows the details of the availability, licensing and security of external URIs in each subset (Metrics 1, 2, and 3). To check the availability of external URIs, RQSS forces a 10-second request and 60-second response time-out. For security, RQSS sets HTTP requests to verify TLS certificates. To check whether a license exists for URI domains, RQSS probes the HTML home page of the domain to find any trace of licensing terms.<sup>13</sup>

Availability and security scores are high while licensing is low. Random subsets get better scores than topical subsets in general. The results of random subsets are similar due to their similar topic coverage. Between topical subsets, Gene Wiki has the highest, and Music has the lowest scores.

### 5.2.2. Interlinking: Interlinking of Reference Properties

Table 7 shows the RQSS results for interlinking of reference properties (Metric 4). To check the interlinking, RQSS seeks the number of values for *equivalent property* (*P1628*) statement of each reference property from Wikidata as of 19 August 2022. While scores for all subsets are low, topical subsets have relatively better scores. Ship's score is notably higher than all subsets. As a project with more human than bot edits, Ships project contributors have been provided more equivalents for their project reference properties. A simple query on WDQS shows that from 6968 reference-specific properties, only 158 (0.02%) have an equivalent property.<sup>14</sup> Figure 10 shows the distribution of equivalents in reference properties between properties with one or more equivalent values. Although there are reference properties with 11 equivalent values (e.g. *main subject* (*P921*)), the average is 2 to 3.

### 5.2.3. Accuracy

*Syntactic Validity of Reference Triples* RQSS deploys the PyShEx evaluator tool [61] to verify the reification of all referenced statements, reference nodes and reference values. We use a ShEx schema<sup>15</sup> that starts from the statement

<sup>13</sup>See the "licensing\_keywords" list in <https://github.com/seyedahbr/RQSSFramework/blob/94f960c/RQSSFramework/Licensing/LicenseExistenceChecking.py> - accessed 15 April 2024

<sup>14</sup>The query can be found in [https://github.com/seyedahbr/RQSS\\_Evaluation/blob/v1.0.1/queries/interlinking-on-WDQS.sparql](https://github.com/seyedahbr/RQSS_Evaluation/blob/v1.0.1/queries/interlinking-on-WDQS.sparql) - the number of items has been queried on 26 November 2023

<sup>15</sup><https://github.com/seyedahbr/RQSSFramework/blob/main/RQSSFramework/ShExes.py>

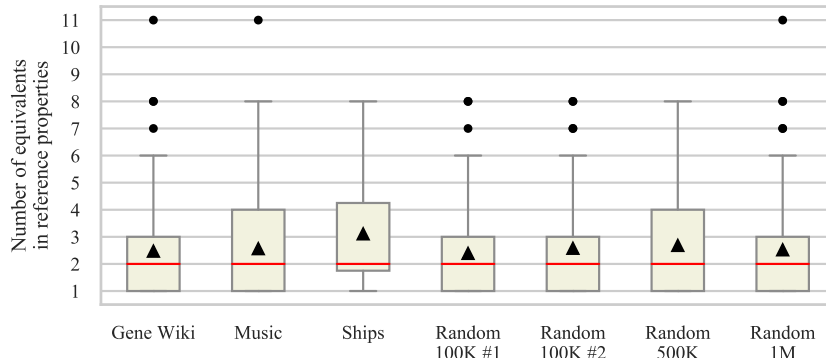


Fig. 10. The distribution of reference properties equivalents (between those with  $\geq 1$  equivalents). Red lines are medians, triangles are means, and circles are outliers.

Table 8  
RQSS results for reference triple syntax accuracy.

Subset	Statement Nodes	Failures	Score (Metric 5)
Gene Wiki	97,062,660	124783	0.9987
Music	12,743,480	2,798	0.9997
Ships	1,116,976	51	0.9999
Random 100K #1	1,225,313	580	0.9995
Random 100K #2	1,226,097	624	0.9994
Random 500K	6,117,915	2,482	0.9995
Random 1M	12,231,380	4,945	0.9995

node and verifies links, value types, and prefixes. The schema is general, i.e., not specific to any P-ID or Q-ID. Table 8 shows the number of statement nodes (as the starting points of the evaluation), the number of evaluation failures, and the final scores. The scores are high. According to the runtime prompts, the majority of the failures are caused by blank statement nodes that we think are created during RDF serialization.

*Syntactic Validity of Reference Literals* After extracting all (reference property, literal) pairs, we matched the literals with the regular expressions obtained from the *format as a regular expression (P1793)* qualifiers of each property given from Wikidata on 7 June 2022. Table 9 shows the total number of reference properties (with literal values), the total number of literal values, the total number of regular expressions in all properties, the total number of failures in regular expression matching, and the final score of each subset. The ‘Invalid’ column shows the number of invalid regular expressions. In the ‘Regexes’ column, the numbers inside the parentheses show how many regular expressions each property has on average. Unlike the random subsets, the average is less than one in topical subsets. However, there are reference properties with more than 20 regular expressions. Some properties do not have regular expressions at all. The ‘No Regex’ column shows the total number of literals affected by these properties. ‘Invalid’ regular expressions and ‘No Regex’ literals are ignored in calculating the scores. For the rest, the results show complete accuracy. The number of no regex literals has a high variation in different subsets. The reason for this variance is the use of the *retrieved (P813)* property in references, which is one of the most widely used reference properties in Wikidata that does not have any *format as a regular expression (P1793)* qualifier.

Figure 11 shows the top three reference properties in terms of having literal values in each subset. External ID properties have the majority in all subsets except Ships. In Ships and the two 100K random subsets, *retrieved (P813)* has a high share resulting in a large number of literals with no regex. In Music, *subject named as (P1810)* has the same role. The distribution of literals in random subsets is very similar. If we consider random subsets as an approximation of the entire Wikidata, about 50% of literals in Wikidata belong to *PubMed ID (P698)* values.

Table 9  
RQSS results for reference literal syntax accuracy.

Subset	Reference Properties	Literals	Regexes	Invalid	Failures	Score (Metric 6)	No Regex
Gene Wiki	705	4,608,209	684 (0.97)	5	0	1.0	70,751 (2%)
Music	1,036	704,514	1,049 (1.01)	15	0	1.0	95,533 (13%)
Ships	69	2,128	63 (0.91)	1	0	1.0	968 (45%)
Random 100K #1	543	51,004	590 (1.08)	6	0	1.0	5,334 (10%)
Random 100K #2	569	50,449	589 (1.03)	8	0	1.0	5,212 (10%)
Random 500K	902	243,147	939 (1.04)	10	0	1.0	15,472 (6%)
Random 1M	1,082	479,231	1,132 (1.04)	16	0	1.0	27,085 (5%)

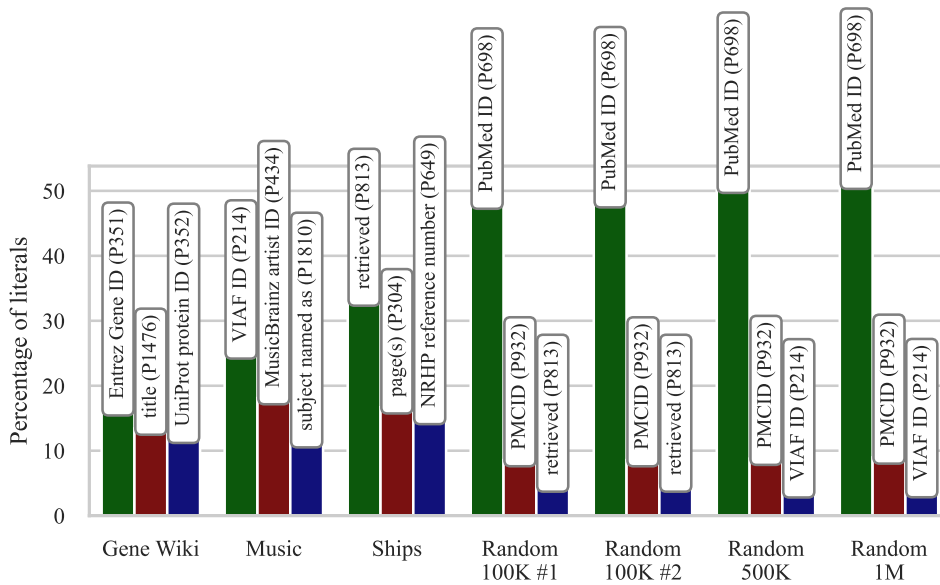


Fig. 11. The top three reference properties with the highest percentage of literals in each subset.

#### 5.2.4. Consistency

**Consistency of Reference Properties** Table 10 shows the RQSS results for reference specificity of reference properties (Metric 8). We check the reference-specificity of properties that are used in references using *property scope* (*P5314*) qualifiers from Wikidata on 7 June 2022. Having no such qualifier is considered non-reference-specific as well. The lowest score comes to Gene Wiki where more than a quarter of reference properties are not reference-specific. We believe the improper use of bots is the cause of this low score in Gene Wiki. In Ships, where there is less bot activity, the freshness of references is relatively low (See Section 5.2.10). Therefore, the low score may be due to the lack of regular data curation. In random subsets, the score is about 0.87. From the total of 84,944,052 distinct referenced statements in all subsets, 15,840,379 (19%) are referenced with the non-reference-specific properties, in which *PubMed ID* (*P698*) (11%) and *UniProt protein ID* (*P352*) (5%) have the majority. Both properties do not have *property scope* (*P5314*) qualifier.

**Range Consistency of Reference Triples** We extract all ⟨reference property, reference value⟩ pairs from the subsets and the ranges (*value-type constraint* (*Q21510865*)) of each property from Wikidata as of 18 June 2022. Table 11 shows the results of matching the class of values with the specified ranges. The second column is the number of reference properties that have ranges specified. The third column shows total reference object values. The fourth column shows the total number of range classes in all properties. Column five is the number of values where their type does not match with the specified range. Column six shows the metric score. The last column is the total number

Table 10  
RQSS results for consistency of reference properties.

Subset	Reference Properties	Score (Metric 8)
Gene Wiki	855	0.7298
Music	1,194	0.8072
Ships	97	0.7319
Random 100K #1	586	0.8788
Random 100K #2	607	0.8896
Random 500K	969	0.8627
Random 1M	1,159	0.8714

Table 11  
RQSS results for range consistency of reference triples.

Subset	Reference Properties	Reference Values	Ranges	Failures	Score (Metric 9)	No Ranges
Gene Wiki	122	14,528,575	462	8,150,998	0.4389	1,571,716 (11%)
Music	134	1,475,080	689	1,170,486	0.2064	45,740 (3%)
Ships	20	55,083	140	38,181	0.3068	678 (1%)
Random 100K #1	33	96,581	154	63,066	0.3470	2,670 (3%)
Random 100K #2	28	97,352	140	63,034	0.3525	3,263 (3%)
Random 500K	53	464,968	241	302,038	0.3504	13,032 (3%)
Random 1M	63	917,746	306	595,109	0.3515	25150 (3%)

of reference values whose properties have no ranges specified; We ignore these values in scoring. Results show a low consistency. The best scores belong to Gene Wiki, where bot accounts have high activity [18]. However, Gene Wiki also has the highest ratio of no range specified amongst all subsets. Music and Ships, on the other hand, have the lowest scores. This difference between the two groups of topical subsets shows another positive impact of bots: automated tools comply with the properties range more than humans. Random subsets have a 0.35 score on average. The reference Comparing the second column of Table 11 with the same column of Table 10 shows properties that have specified ranges are very limited in all subsets. However, having more properties with a specified range and choosing references in the specified range can indicate the participants' level of expertise (whether human or bot) in referencing.

#### 5.2.5. Conciseness: Ratio of Reference Sharing

Similar to [18], we count all incoming connections to each reference node to see if the reference node is used as a reference for more than one statement. Table 12 shows the ratio of reference sharing for each subset. As a factor of conciseness, reference sharing is a positive point. The ratio for random subsets is higher than for topical subsets. We believe it is related to scholarly articles as the majority of random subsets (as well as Wikidata). There are many reference nodes with the value of an article shared between all related items. Amongst topical subsets, Gene Wiki has the highest score; another evidence of bot activities in this subset. Column 'Maximum' in the table shows the highest number of incoming edges to a reference node. Column 'Mean' shows the average number of incoming nodes. While the average number of incoming nodes is 14, there are reference nodes shared between thousands of statements.

#### 5.2.6. Reputation: External URIs

We use Pydnsbl to check whether URI domains are among the public black-listed domains on the web.<sup>16</sup> Table 13 shows the number of URIs, URI domains, the score of the metric (considering the ratio of black-listed domains), and the number of URIs affected by the black-listed domains. The scores are high meaning there are few blacklisted URIs in the external sources; 13 affected URIs between 3,610,506 URIs, e.g., `jatim.litbang.pertanian.go.id`.

<sup>16</sup><https://pypi.org/project/pydnsbl/0.5.4/> - accessed 15 April 2024

Table 12  
RQSS results for reference sharing.

Subset	Reference Nodes	Maximum	Mean	Score (Metric 12)
Gene Wiki	9,742,813	1,281,307	13	0.4924
Music	1,585,122	1,378,301	12	0.2982
Ships	61,996	96,591	16	0.2710
Random 100K #1	94,966	41,667	14	0.7021
Random 100K #2	94,982	43,171	14	0.6969
Random 500K	453,273	206,837	15	0.6998
Random 1M	894,093	418,196	15	0.7031

Table 13  
RQSS results for the reputation of external URIs (Pydnsbl).

Subset	URIs	URI Domains	Score (Metric 13)	Affected URIs
Gene Wiki	2,559,493	10,138	0.9998	3
Music	215,161	21,593	0.9996	7
Ships	20,737	924	1.0	0
Random 100K #1	48,618	2,057	1.0	0
Random 100K #2	48,279	2,110	1.0	0
Random 500K	240,183	5,952	0.9996	3
Random 1M	478,035	9,342	1.0	0



Fig. 12. ‘View History’ tab of *Albert Einstein (Q937)* on 20 September 2022. The second record shows an addition of a reference to a claim.

### 5.2.7. Believability: Human-added References

In the absence of an effective solution to retrieve the revision history of Wikidata, RQSS reads the HTML history pages of items on the Wikidata website front end. Figure 12 shows the ‘View History’ tab of *Albert Einstein (Q937)* on 20 September 2022. In these HTML pages, there is a record for each edit in which the date-time of the edit, the editor’s account and a brief description of the edit are available. In terms of references, the metadata provided on these pages is limited. One can only check the addition, deletion, or change of a reference for a specific statement property. There is no data on what reference value has been changed. Also, there is no distinction between different statements with the same property. With these limitations in mind, RQSS retrieve all ⟨item, referenced statement property⟩ pairs from the subsets. Then, RQSS investigates the *last* editor user account that added/edited a reference for that specific property of that item using an XPath query [62]. Note that there is an upper date limit set to 3 January 2022 (the release date of the subsetted Wikidata dump). We consider an added/edited reference human-added if there is no sub-string `bot` in the editor’s account username.

Table 14 shows the number and the percentage of referenced items, the number of referenced facts (distinct properties used) of the referenced items, the score of the metric, and the number of fact properties in which there

Table 14

RQSS results for human-added references. Computing Gene Wiki scores timed out after three unsuccessful attempts and more than 90 days of processing.

Subset	Referenced Items	Referenced Facts (Distinct Properties)	Score (Metric 14)	No Historical Metadata
Gene Wiki	8,022,583 (87%)	49,552,129		
Music	862,053 (88%)	6,030,622	0.5028	1,868,355 (31%)
Ships	68,495 (53%)	286,307	0.7888	102,658 (36%)
Random 100K #1	70,458 (81%)	526,658	0.4316	440,174 (83%)
Random 100K #2	70,754 (81%)	526,028	0.4313	439,646 (83%)
Random 500K	351,923 (81%)	2,627,460	0.4294	2,193,210 (83%)
Random 1M	702,033 (81%)	5,243,722	0.4312	4,379,482 (83%)

is no historical metadata for them. While the initial (item, referenced statement property) pairs have been extracted quickly, the results of Gene Wiki were not available after three unsuccessful attempts and more than 90 days of processing due to the huge number of external HTTP requests and HTML rendering required. The scores vary between random and topical subsets. Due to the presence of active bots in the Gene Wiki WikiProject, such as Pathwaybot<sup>17</sup> and ProteinBoxBot<sup>18</sup>, we hypothesize that there are more bot-added references than human-added references in the Gene Wiki subset. For the same reason, i.e. the lack of active bots in the corresponding WikiProject, Ships have the highest human-added reference ratio. The ratio for random subsets is 0.43 on average, which is less than both topical subsets. It also justifies the higher rate of reference sharing in random subsets versus Music and Ships. The percentage of referenced facts with no historical metadata is also high in all random subsets. Note that if we consider curating a large amount of data in one action as the main feature of bots, some human user accounts (without bot prefixes or suffixes) may also show the same behaviour. Identifying those accounts requires pattern recognition over the Wikidata revision history which is not the scope of this paper.

#### 5.2.8. Verifiability: Type of References

We retrieve all IRI-based reference node values from the subsets. For Q-ID values, we get the type of value from Wikidata on 21 August 2022. For external URI values, we only check if the URI belongs to our well-known datasets list obtained through the authors' experience.<sup>19</sup> Table 15 shows the disaggregated statistics of source types and the verifiability scores. However, in both subsets, the main weakness is the high number of external URIs that are not well-known datasets (and get zero scores); this is the strong point in Gene Wiki and random subsets. The 'Unclassified' column shows the number and percentage of external sources for which RQSS can not classify their type. Note that many external links can be blog posts, encyclopedic articles, or even scholarly articles, but investigating the content of the external links is subjective. Identifying the verifiability category of a URL necessitates content rendering and the application of a concept recognition algorithm employing a trained model based on human opinions. However, this falls outside the scope of the current research. Music and Ships contains a large number of such external sources, which explains the reason for their low score.

#### 5.2.9. Objectivity: Multiple References for Statements

RQSS counts the number of reference nodes connected to each statement node via `prov:wasDerivedFrom` links (Figure 2). Table 16 shows the scores of objectivity based on the statements with multiple references. Although multiple referencing is low in all subsets, random subsets have lower scores. Less than one per cent of referenced statements have more than one reference in random subsets. The higher rate of multiple referencing can be related to more human contributions versus bot contributions, as found in the Music and Ships subsets. Figure 13 shows the distribution of references in statements having two or more references. Gene Wiki has the best average, and most

<sup>17</sup><https://www.wikidata.org/wiki/User:Pathwaybot> - accessed 15 April 2024

<sup>18</sup><https://www.wikidata.org/wiki/User:ProteinBoxBot> - accessed 15 April 2024

<sup>19</sup>The list of datasets can be found in <https://github.com/seyedahbr/RQSSFramework/blob/018c535/RQSSFramework/utills/lists.py> - accessed 15 April 2024



Table 15  
RQSS results for the type of sources.

Subset	URI Sources	Scholarly Article	Well-Known Dataset	Book, Encyclopedia, or Encyclopedic Article	Magazine, Blog, or Blog Post	Unclassified	Score (Metric 15)
Gene Wiki	2,899,958	206,449 (7%)	1,618,047 (56%)	473	51	1,074,938 (37%)	0.4897
Music	768,682	32	24,190 (3%)	1570	207	742,683 (96%)	0.0247
Ships	59,209	1	333	18	1	58,856 (99%)	0.0043
Random 100K #1	58,944	2,383 (4%)	36,405 (61%)	37	8	20,111 (34%)	0.5039
Random 100K #2	59,069	2,418 (4%)	36,041 (61%)	55	4	20,551 (34%)	0.4990
Random 500K	278,710	7,476 (3%)	179,340 (64%)	106	23	91,765 (33%)	0.5096
Random 1M	550,455	14,233 (3%)	358,289 (65%)	215	40	177,678 (32%)	0.5142

Table 16  
RQSS results for having multiple references for statements.

Subset	Referenced Statement	Multiple Referenced Statements	Score (Metric 16)
Gene Wiki	63,521,696	2,307,545	0.0363
Music	6,348,140	395,296	0.0622
Ships	301,290	16,068	0.0533
Random 100K #1	946,523	8,594	0.0090
Random 100K #2	940,552	8,567	0.0091
Random 500K	4,704,898	44,929	0.0095
Random 1M	9,392,549	90,684	0.0096

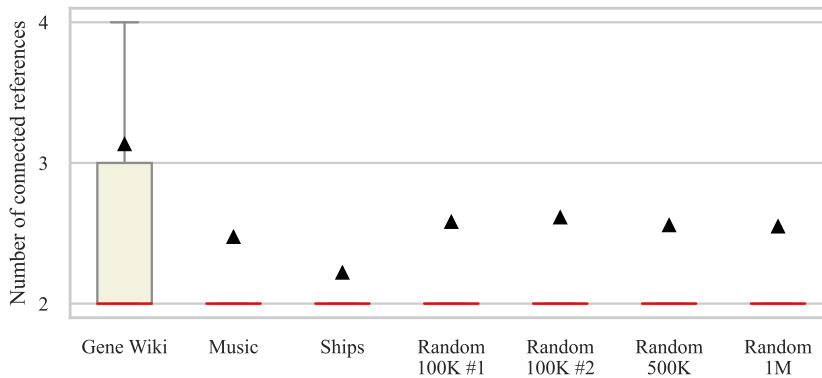


Fig. 13. The distribution of references connected to statements (between statements with  $\geq 2$  reference). Red lines are medians and triangles are means. Outliers are ignored due to readability.

of its multiple-referenced statements have between 2 and 4 references. Note that there are statements in Gene Wiki that have more than 100 references.

5.2.10. Currency

*Freshness of Reference Triples* As mentioned in Section 5.2.7, we do not have access to the historical metadata of a single triple. Instead, RQSS requests the “view history” HTML page of each item, then renders its content using XPath queries seeking all reference creation and modification times. Figure 12 shows an example; the Revision history of *Albert Einstein (Q937)* in which the creation of q reference for Albert Einstein’s *Golden ID (P7502)* statement can be seen. For each  $\langle \text{item, referenced fact} \rangle$  pairs, RQSS extracts the first creation time of each fact as its

Table 17

RQSS results for fact-reference freshness. Computing Gene Wiki scores timed out after three unsuccessful attempts and more than 90 days of processing.

Subset	Referenced Items	Referenced Facts (Distinct Properties)	Score (Metric 17)	No Historical Metadata
Gene Wiki	8,022,583 (87%)	49,552,129		
Music	862,053 (88%)	6,030,622	0.9245	1,947,806 (32%)
Ships	68,495 (53%)	286,307	0.9693	104,111 (36%)
Random 100K #1	70,458 (81%)	526,658	0.9459	442,960 (84%)
Random 100K #2	70,754 (81%)	526,028	0.9467	442,303 (84%)
Random 500K	351,923 (81%)	2,627,460	0.9450	2,207,080 (84%)
Random 1M	702,033 (81%)	5,243,722	0.9456	4,406,737 (84%)

Table 18

RQSS results for freshness of external URIs.

Subset	External URIs	Score (Metric 18)	No Last-Modified Header
Gene Wiki	2,559,493	0.0338	2,026,803 (79%)
Music	215,161	0.0758	196,460 (91%)
Ships	20,737	0.1239	19,687(95%)
Random 100K #1	48,618	0.1116	46,827 (96%)
Random 100K #2	48,279	0.0842	46,585 (96%)
Random 500K	240,183	0.1029	231,803 (96%)
Random 1M	478,035	0.1116	461,554 (96%)

*startTime*, and the latest reference creation or revision of the fact as the *modifTime*. The upper date limit is set to 3 January 2022. The results of fact-reference freshness are shown in Table 17. Similar to Section 5.2.7, the results of Gene Wiki were not available after three unsuccessful attempts and more than 90 days of processing due to the huge number of external HTTP requests and HTML rendering required. The percentage of missing historical data is similar to Section 5.2.7 (Table 14). The freshness scores, which include only found referenced facts, are high, and there is not much difference between different subsets.

*Freshness of External URIs* To calculate the freshness of external URIs, RQSS checks the `Last-Modified` header of the HTTP response of each URI. The *startTime* is set for 29 October 2012 (the Wikidata launch date) for all URIs. Table 18 shows the result of external URIs freshness. There is a very high percentage of external URIs without `Last-Modified` header, consequently the scores are very low. There is no relation between the found `Last-Modified` header percentage and the score. Gene Wiki has the lowest score despite lots of its external URIs having `Last-Modified` header.

### 5.2.11. Volatility and Timeliness

To compute Metric 19, RQSS uses the Ultimate Sitemap Parser Python package.<sup>20</sup> The package is utilized to perform a detailed analysis of XML sitemap files within a website's root domain. For a given external source URL domain, Ultimate Sitemap Parser navigates through the sitemap structure of the URL's domain, extracting essential metadata such as the `<changefreq>`. However, downloading, decompressing, and searching XML sitemaps is time-consuming. A complete analysis of the sitemap structure can take two to ten minutes. Considering thousands of distinct domains in even the smallest subset, we were not able to compute volatility results in a reasonable amount of time. As Metric 20 is the distance between freshness and volatility, timeliness results are also not computed.

### 5.2.12. Completeness

*Class/Property Schema Completeness of References* RQSS deploys PyShEx schema loader to parse Wikidata Entity Schema ShEx-C [63] raw texts and create a summary of the schema-level referenced classes, referenced fact

<sup>20</sup><https://pypi.org/project/ultimate-sitemap-parser/> - accessed 15 April 2024

Table 19  
RQSS results for class and property schema completeness in referencing.

Subset	Classes	Fact Properties	Score (Metric 21)	
			$m_{classSchemaCom}$	$m_{propertySchemaCom}$
Gene Wiki	17,184	4,206	0.0004	0.0147
Music	1,381	3,506	0.0014	0.0088
Ships	1,133	701	0.0008	0.0370
Random 100K #1	3,484	4,141	0.0025	0.0132
Random 100K #2	3,498	4,191	0.0022	0.0121
Random 500K	8,299	5,917	0.0010	0.0096
Random 1M	11,908	6,630	0.0007	0.0088

properties, and the used reference properties on 9 July 2022. On the date, there were 319 Entity-Schemas of which 13 had reference schema information. In total 16 classes and 63 properties had reference schemas. Table 19 shows the results of schema-level class/property completeness in the context of references. The scores for both ratios are low due to the low number of Entity-Schemas and schema-level referenced classes/properties. Although the Entity-Schema concept is new in Wikidata, the scores show the weakness of schema-level referencing information in this KG.

*Schema-based Property Completeness of References* Using the Entity-Schema summaries (Section 5.2.12) RQSS extracts all ⟨statement, reference property⟩ pairs from subsets and checks each pair over E-ID summaries. To provide an example, consider that at the schema level (Wikidata EntitySchemas) it has been mentioned that the *CAS Registry Number (P231)* properties should be referenced with at least one *InChIKey (P235)* reference property. Now, consider the instance level, there are ten of the P231 facts. If four of these ten P231 are referenced by property P235, then the completeness ratio of reference property P235 w.r.t. its references schema property P231 is 0.4. The metric score then will be the average of all completeness ratios of schema property-schema reference property pairs. There is a total of 193 ⟨fact property, reference property⟩ pairs in the schema level. Table 20 shows the details of comparing schema-level referencing metadata with the instance-level. The second column is the total number of ⟨statement, reference property⟩ pairs. The third column shows the number of statements without reference. The ‘Score’ column shows results with and without considering non-referenced statements in the instance level into account. A 0.60 score means the average completeness ratio of the 193 schema-level ⟨fact property, reference property⟩ (*comRefPropS* values in Metric 22) pairs is 60%. The scores of Gene Wiki are considerably higher than all subsets. Part of that is due to the activity of its community in defining Entity-Schemas and their attention to referencing. The Majority of the current Entity-Schemas belong to Gene Wiki classes.<sup>21</sup> That does not necessarily mean the instance-level data are following schema-level. That might be due to writing Entity-Schemas based on the instance-level data in the project. Both are useful as they help users to understand what kind of references they should expect on the topic. While in the previous metrics, the scores of the random subsets are similar, here, the scores increase as the random subset size increases. That is because the number of averaging factors is constant, while their values grow with the increase of instance-level data. For all subsets, there are 193 averaging factor pairs. As the subset size increases, there are more adjustable instance-level ⟨statement, reference property⟩ pairs to the 193 schema-level pairs. Thus, the *comRefPropS* values increase and due to a fixed 193 pairs, the total score rises. Figure 14 shows the distribution of all 193 *comRefPropS* values. In all subsets, there are a variety of *comRefPropS* values between 0 and 1. The details of *comRefPropS* values can be found at [64].

*Property Completeness of References* RQSS extracts all ⟨fact property, reference property⟩ pairs from subsets and checks if a fact with fact property *X* referenced by a reference property *Y* in the instance level, how many of other fact property *X* are referenced using reference property *Y*. As an example, consider that at the instance level, there are ten *CAS Registry Number (P231)* facts. If four of these P231 facts are referenced by property P235, then the completeness ratio of reference property P231 w.r.t. its fact property P231 is 0.4. The metric score then will be

<sup>21</sup>[https://www.wikidata.org/wiki/Wikidata:Database\\_reports/EntitySchema\\_directory](https://www.wikidata.org/wiki/Wikidata:Database_reports/EntitySchema_directory) - accessed 15 April 2024

Table 20  
RQSS results for schema-based property completeness of references.

Subset	$\langle$ statement, reference property $\rangle$ pairs	Non-Referenced Facts	Score (Metric 22)	
			Without Non-Referenced Facts	With Non-Referenced Facts
Gene Wiki	180,955,497	33,540,964	0.6098	0.5354
Music	12,148,520	6,395,340	0.1203	0.0632
Ships	490,748	815,686	0.1177	0.0523
Random 100K #1	2,754,858	278,790	0.4331	0.3647
Random 100K #2	2,722,602	285,545	0.4252	0.3584
Random 500K	13,681,074	1,413,017	0.4946	0.4195
Random 1M	27,304,697	2,838,831	0.5369	0.4645

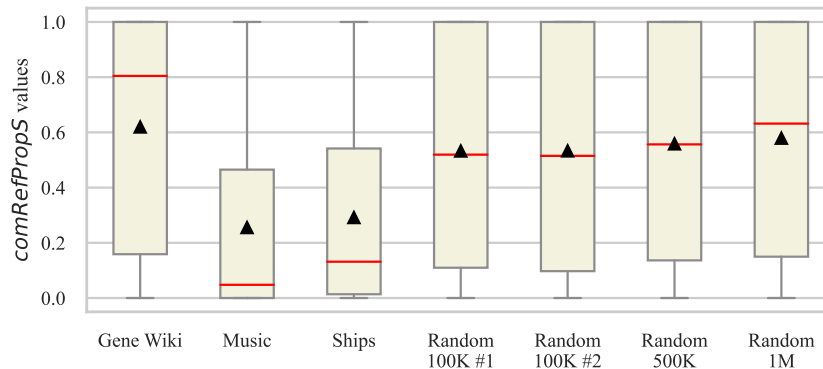


Fig. 14. The distribution of completeness ratios of the 193 schema-level  $\langle$ fact property, reference property $\rangle$  ( $comRefPropS$  values). Red lines are medians, and triangles are means.

the average of all instance-level property-reference property pairs completeness ratios. Table 21 shows the result of property completeness of references. The fourth column shows the number of  $\langle$ statement, reference property $\rangle$  pairs ( $comRefProp$  values in Metric 23), which are the averaging factors. Comparing the results with Section 5.2.12, Gene Wiki has no longer the highest but one of the lowest scores. Random subsets have better scores than topical subsets. The score falls with the increase in size due to the variable number of averaging factors because the averaging factors are not fixed and increase with the size of the subset. Unlike Metric 22, the entire Wikidata would probably get lower scores. It shows that the instance-level reference property completeness in Wikidata is weaker than schema-based reference property completeness. Figure 15 shows the distribution of averaging factors ( $comRefProp$  values). The distribution shows topical subset  $comRefProp$  values are less scattered. Detailed statistics of  $\langle$ fact property, reference property $\rangle$  pairs can be found on [64].

### 5.2.13. Amount-of-data

By extracting the number of statement nodes, reference nodes, reference triples and reference literals, RQSS computes the amount of data ratios. Besides that, RQSS retrieves the number of outgoing reference triples and outgoing literal values for each reference node. Figure 16 shows the scores of the four Amount-of-data metrics. Gene Wiki has the highest score in all metrics except for the Metric 25. Note that the definition of Metric 27 inverts the ratio and subtracts it from one to map the ratio into a number between 0 and 1. Figure 17 shows the distribution of triples and literals per reference node. The average of triples per reference node of Gene Wiki is 3.5, which is higher than other subsets as Metric 27 score shows. Random subsets have identically the same distribution over both ratios and their metric scores, as well as their distribution, are very close to Gene Wiki, showing that the Wikidata as a whole is in good condition concerning the amount of data.

Table 21  
RQSS results for property completeness of references.

Subset	⟨statement, reference property⟩ pairs	Non-Referenced Facts	⟨fact property, reference property⟩ pairs	Score (Metric 23)	
				Without Non-Referenced Facts	With Non-Referenced Facts
Gene Wiki	180,955,497	33,540,964	14,582	0.2942	0.1587
Music	12,148,520	6,395,340	15,823	0.2196	0.0975
Ships	490,748	815,686	1,637	0.3243	0.1673
Random 100K #1	2,754,858	278,790	8,227	0.4711	0.3318
Random 100K #2	2,722,602	285,545	8,264	0.4597	0.3214
Random 500K	13,681,074	1,413,017	14,037	0.3945	0.2429
Random 1M	27,304,697	2,838,831	17,324	0.3616	0.2128

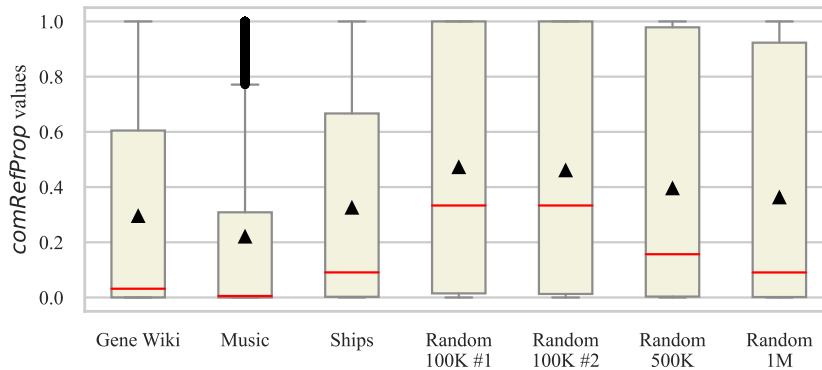


Fig. 15. The distribution of completeness ratios (fact property, reference property) (*comRefProp* values) at instance-level. Red lines are medians, and triangles are means. Circles on the Music bar are outliers.

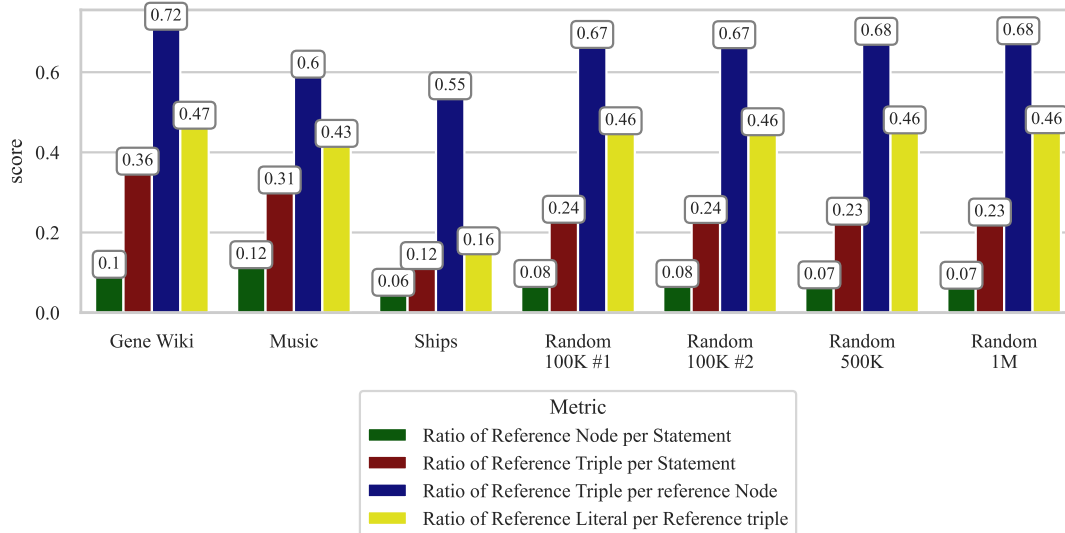


Fig. 16. RQSS results for metrics: Ratio of Reference Node per Statement (Metric 25), Ratio of Reference Triple per Statement (Metric 26), Ratio of Reference Triple per reference Node (Metric 27), and Ratio of Reference Literal per Reference triple (Metric 28).

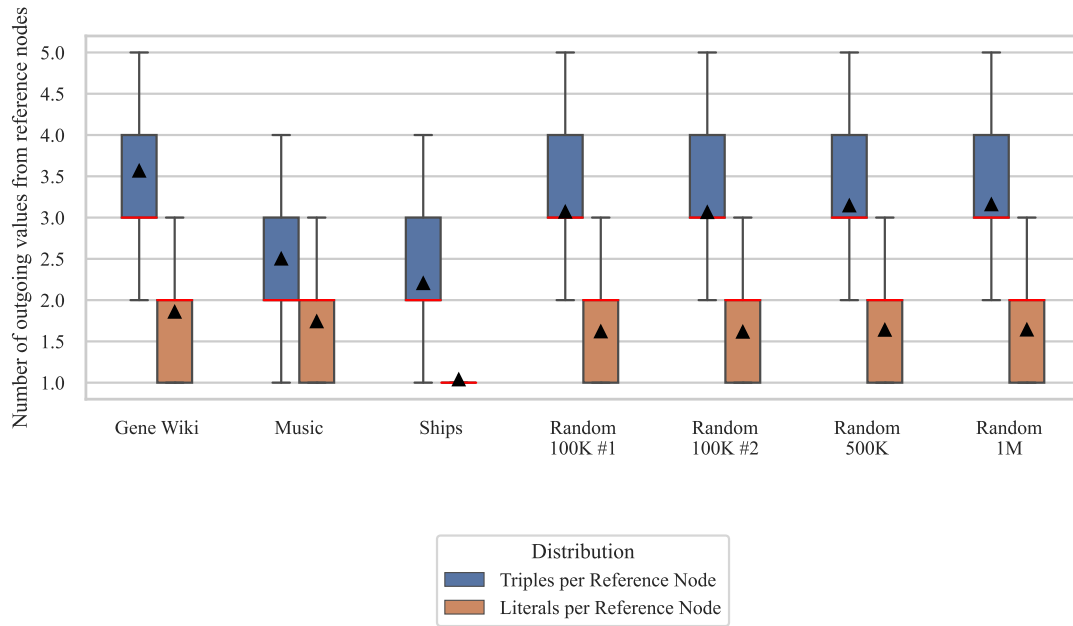


Fig. 17. The distribution of triples and literals per reference node. Red lines are medians and triangles are means. Outliers are ignored due to readability.

Table 22  
RQSS results for URI length of external sources.

Subset	External URIs	len ≤ 80	80 < len ≤ 2083	2083 < len ≤ 4096	len > 4096	Score (Metric 31)
Gene Wiki	2,559,493	1,212,860	1,346,633	0	0	0.8684
Music	215,161	164,166	50,995	0	0	0.9407
Ships	20,737	19,250	1,487	0	0	0.9820
Random 100K #1	48,618	21,721	26,897	0	0	0.8616
Random 100K #2	48,279	21,447	26,832	0	0	0.8610
Random 500K	240,183	107,025	133,158	0	0	0.8613
Random 1M	478,035	213,267	264,768	0	0	0.8615

#### 5.2.14. Representational-conciseness

RQSS decodes each external URI to percent encoding and counts the number of characters. Table 22 shows the details of External URI lengths in each subset and the scores. There are no URIs longer than 2083 in any of the subsets. Music and Ships score better than Gene Wiki and random subsets. The results show an inverse relation between referencing URI lengths and the activity of bots.

#### 5.2.15. Representational-consistency

Table 23 shows the results for reference property diversity. The scores of all subsets are higher than 0.9. Smaller random subsets have lower scores. In smaller random subsets, the property diversity of references is not far from larger subsets due to a broad type of statements (which is the nature of random selection), and the number of their triples is much less. Figure 18 shows the top five properties with the highest frequency of use in each subset. The frequency of property usage in topical subsets is similar to [18] and shows that sources in Music and Ships are more internal (Wikimedia-based projects). The distribution of frequency and type of properties in random subsets is similar. Apart from *Entrez Gene ID (P351)* and *UniProt protein ID (P352)* which are specific Gene Wiki reference properties, random subsets and Gene Wiki have similar frequency and type of used properties. Note that *PubMed*

Table 23  
RQSS results for the diversity of reference properties.

Subset	Reference Properties	Reference Triples	Score (Metric 32)
Gene Wiki	855	34,727,916	0.9999
Music	1,194	3,961,595	0.9996
Ships	97	136,518	0.9992
Random 100K #1	586	291,334	0.9979
Random 100K #2	607	290,854	0.9979
Random 500K	969	1,424,752	0.9993
Random 1M	1,159	2,822,601	0.9995

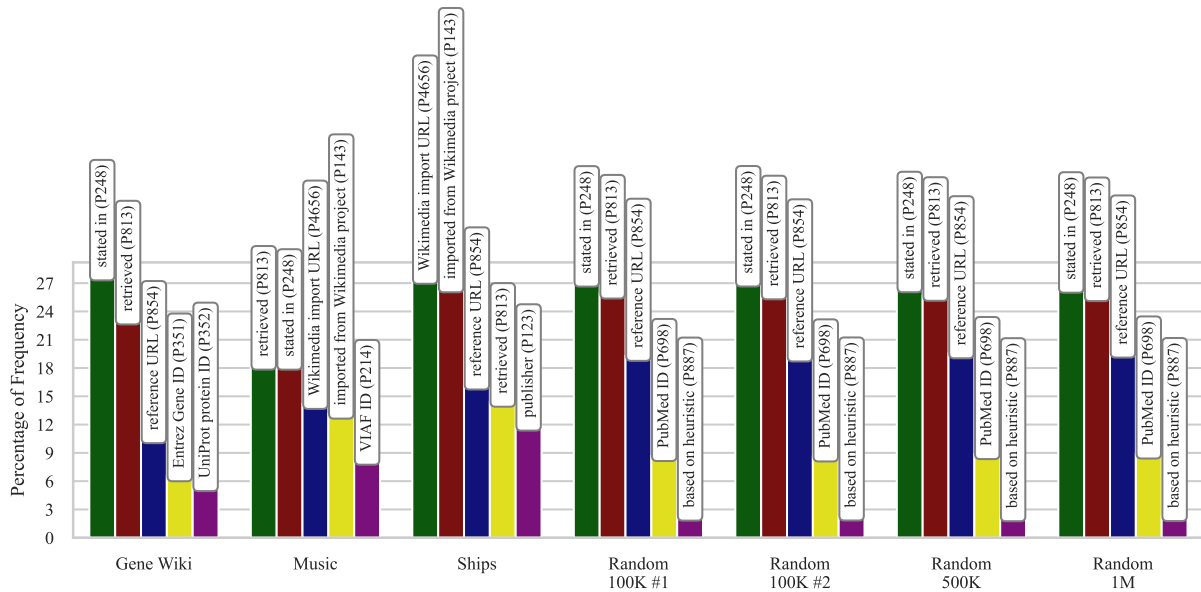


Fig. 18. Five properties with the highest frequency of use in each subset.

*ID (P698)*, which is one the most frequent literal accepting properties in the random subsets, is also the fourth most frequent property in general.

5.2.16. Understandability

*Human-readable labelling/Commenting of Reference Properties* RQSS queries the number of labels and comments of each reference property from Wikidata on 28 August 2022. Table 24 shows the result of human-readable labelling and commenting on reference properties. All reference properties in Gene Wiki and Ships have human-readable labels and comments. The results of other subsets are also high, and there are less than five properties with no tags and comments (e.g. *P2580*, *P6656*, and *P3043*). Figure 19 shows the distribution of the number of labels and comments in reference properties. The Ships subset has the best average and most uniform distribution. The average and the distribution of other subsets are similar.

*Handy External Sources* RQSS extracts all external sources (external URIs plus external sources that are Wikidata items) from the subsets. For external URIs, RQSS checks the existence of an anchor in the middle of the path part of the URI. For external sources that are Wikidata items, RQSS checks if the item is an instance of an online database (*Q7094076*) or if there is a value for its *full work available at URL (P953)*, *SPARQL endpoint (P5305)*, or *API endpoint (P6269)* properties on Wikidata on 21 August 2022. Table 25 shows the scores of handy external sources. The scores of all subsets are high, Music has the highest score, and topical subsets have better scores than random subsets. Two larger random subsets have better scores because they have lower offline sources but more URLs (with

Table 24  
RQSS results for human-readable labelling and commenting of reference properties.

Subset	Reference Properties	Labelling Score (Metric 33)	commenting Score (Metric 34)
Gene Wiki	855	1.0	1.0
Music	1,194	0.9983	0.9966
Ships	97	1.0	1.0
Random 100K #1	586	0.9965	0.9948
Random 100K #2	607	0.9967	0.9950
Random 500K	969	0.9979	0.9958
Random 1M	1,159	0.9974	0.9956

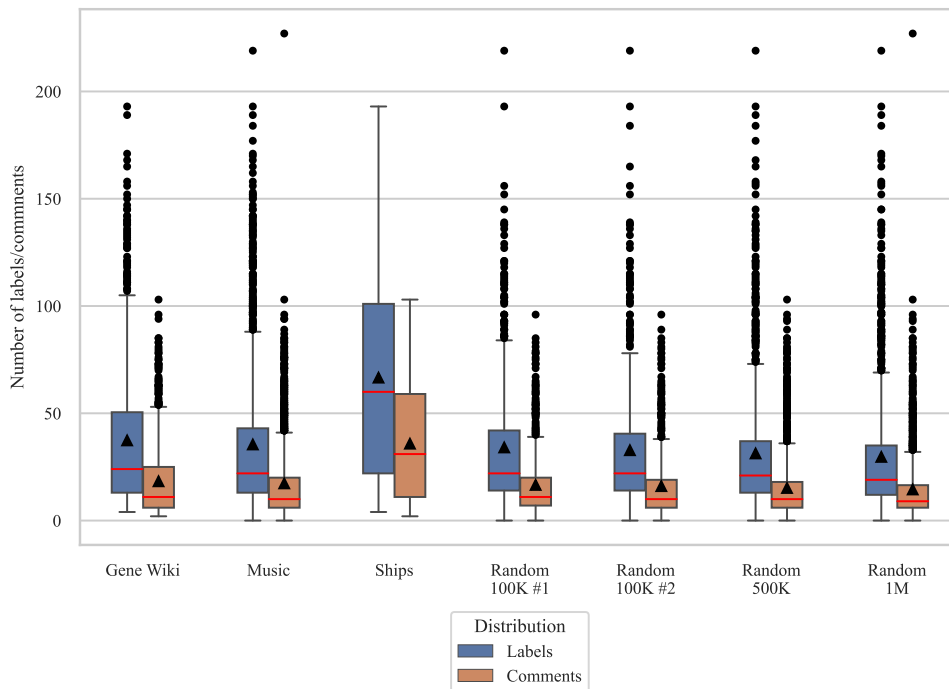


Fig. 19. The distribution of the number of labels and comments in reference properties. Red lines are medians, triangles are means, and circles are outliers.

no anchors). Figure 20 shows the share of each handy external source type in the final score. As Figure 20 shows, Music is the only subset with more than 10% of external URLs with anchors (in other subsets, this type has less than 1% of the share). The most frequent type in all subsets is the URLs with no anchors.

#### 5.2.17. Interpretability: Usage of Blank Nodes in References

RQSS checks the number of blank nodes amongst reference nodes and reference value nodes (Figure 2). Table 26 shows the number of nodes in each reification part, the number of blank nodes, and the scores. The results show quite a low number of blank nodes only in reference values. Note that the 'Value Nodes' column is the distinct counting of reference values. That is different from the 'Reference Values' column in Table 11 which was a property-value counting and was not a distinct counting. In topical subsets, the distinct reference value nodes are lower than the reference nodes, showing that some reference values are shared between reference nodes.

#### 5.2.18. Versatility

**Multilingual Labelling/Commenting of Reference Properties** RQSS queries the number of non-English labels and comments of each reference property from Wikidata on 28 August 2022. Table 27 shows the result of multilingual



Table 25  
RQSS results for handy external sources.

Subset	External Sources	Score (Metric 35)
Gene Wiki	2,788,210	0.7115
Music	268,081	0.7404
Ships	22,859	0.7295
Random 100K #1	57,127	0.7078
Random 100K #2	57,224	0.7032
Random 500K	260,408	0.7237
Random 1M	511,510	0.7266

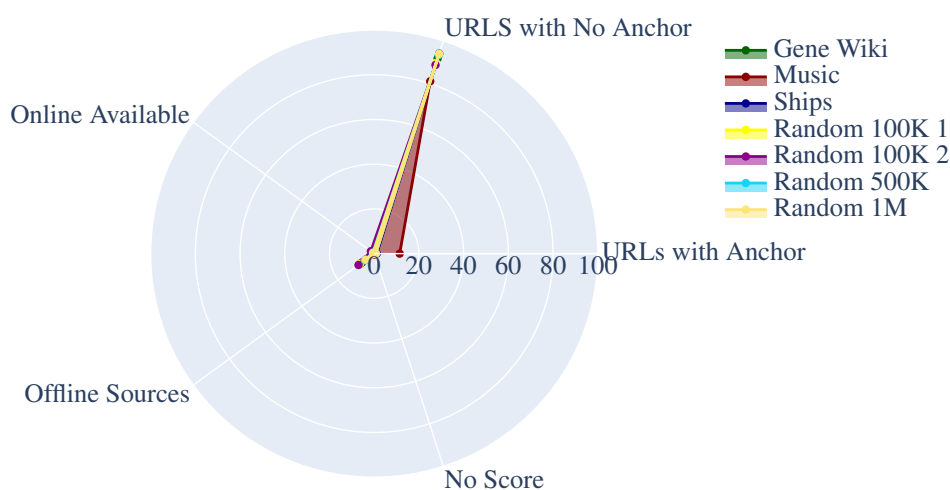


Fig. 20. The share (percent) of different handy external source types.

labelling and commenting on reference properties. Compared to Section 5.2.16, the scores of multilingual metadata are lower. However, high scores show that Wikidata is rich in non-English labelling/commenting. Figure 21 shows the distribution of the number of non-English labels and comments in reference properties, which is identical to Figure 19.

**Multilingual Sources** RQSS retrieves all internal and external sources from the subsets. For those sources that are Wikidata items, RQSS checks the *language of work or name (P407)* and then the *ISO 639-1 code (P218)* properties directly from Wikidata. For URL sources, RQSS checks the lang attribute of the html tag of the URL. Extracting the languages has been between 29 August to 16 September 2022. Table 28 shows the results of multilingualism in internal and external sources. Music has the highest score and the second lowest not-found languages. That can be due to having international data on music tracks, signers, albums etc. Random subsets have many not-found languages but better results than Ships and Gene Wiki. The multilingualism ratio decreases with the increase of subset size in random subsets. Despite having a high diversity of non-English languages, Gene Wiki has the lowest score as it widely uses well-known biomedical dataset IDs/sources in references, which are published in English.

Table 26

RQSS results for blank nodes in referencing reification.

Subset	Reference Nodes	Value Nodes	Blank Reference Nodes	Blank Value Nodes	Score (Metric 36)
Gene Wiki	9,742,813	7,239,594	0	6	0.9999
Music	1,585,122	1,449,236	0	13	0.9999
Ships	61,996	61,302	0	0	1.0
Random 100K #1	94,966	109,358	0	0	1.0
Random 100K #2	94,982	108,939	0	0	1.0
Random 500K	453,273	518,994	0	0	1.0
Random 1M	894,093	1,023,517	0	2	0.9999

Table 27

RQSS results for multilingual labelling and commenting of reference properties.

Subset	Reference Properties	Labelling Score (Metric 37)	commenting Score (Metric 38)
Gene Wiki	855	1.0	0.9988
Music	1,194	0.9983	0.9958
Ships	97	1.0	1.0
Random 100K #1	586	0.9965	0.9931
Random 100K #2	607	0.9967	0.9934
Random 500K	969	0.9979	0.9938
Random 1M	1,159	0.9974	0.9948

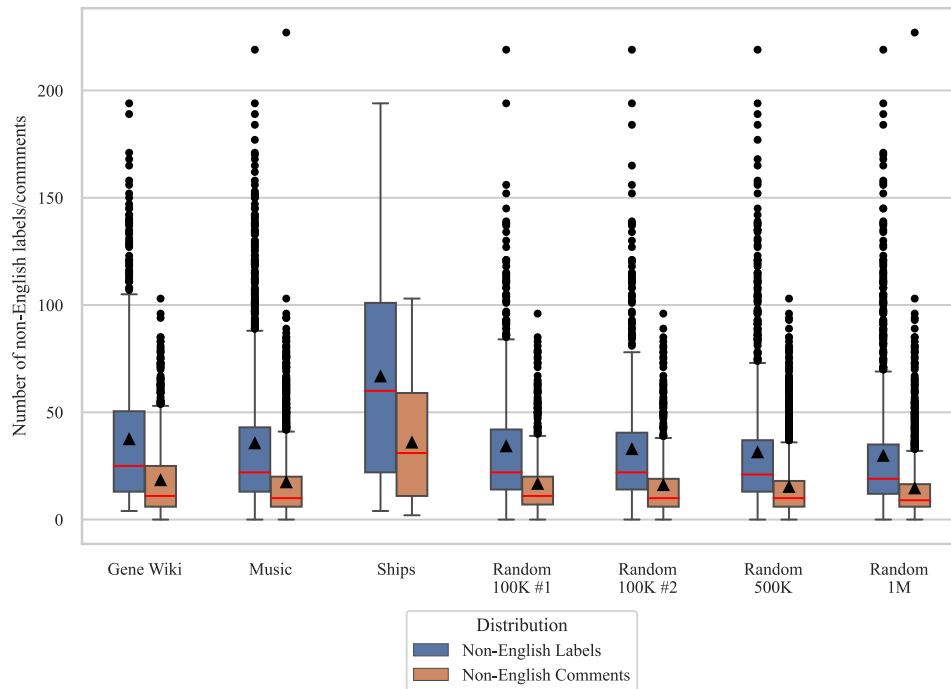


Fig. 21. The distribution of the number of non-English labels and comments in reference properties. Red lines are medians, triangles are means, and circles are outliers.

Table 28  
RQSS results for multilingual internal/external sources.

Subset	Sources	Non-English Languages	Score (Metric 39)	No Language Found
Gene Wiki	2,900,380	215	0.2017	1,674,149 (58%)
Music	769,290	316	0.4844	79,730 (10%)
Ships	59,242	77	0.2200	2,468 (4%)
Random 100K #1	59,270	143	0.2602	37,317 (63%)
Random 100K #2	59,396	137	0.2659	37,443 (63%)
Random 500K	279,454	208	0.2510	176,688 (63%)
Random 1M	551,439	239	0.2450	348,302 (63%)

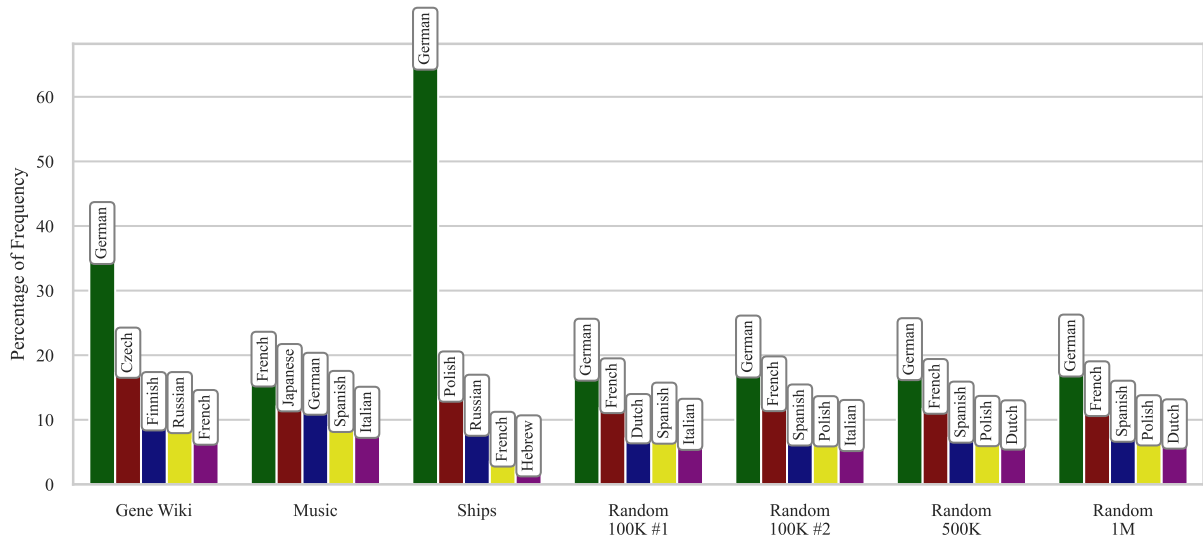


Fig. 22. Five most frequent non-English languages used in sources.

Figure 22 shows the five most frequent non-English languages used in sources. In Gene Wiki and Ships, German is dominant. In other subsets, non-English languages have a more uniform usage.

**Multilingual Referenced Statements** RQSS starts with extracting the (statement ID, reference value) pairs (IRI values, either internal or external), and matching the languages of sources using Section 5.2.18 data. Table 29 shows the number of referenced statements with internal or external sources and the ratio of multilingualism in each subset. The scores of Music and Ships are considerably higher than other subsets, especially the other topical subset Gene Wiki. The results show another impact of bot activities: bots added mostly English sources. In the random subsets and Gene Wiki, where bots are more active, despite having a good variety of non-English sources, a small fraction of statements use non-English references.

## 6. Challenges and Limitations

In addition to the statistical analytics and referencing scores, this comprehensive and in-depth study of Wikidata references brings several challenges, the solution of which requires novel techniques. The first and most important is querying the massive size of Wikidata. The public SPARQL endpoint is neither intended, nor suitable, for performing quality tests. Storing, processing and querying the 100 GB Wikidata dumps is beyond most computing resources available to researchers. Aiming to establish a local SPARQL endpoint on a full Wikidata dump, we were not able to deploy the Wikibase Docker containers due to the lack of root privileges (i.e. requisite administrative permissions

Table 29  
RQSS results for multilingual referenced statements.

Subset	Referenced Statements (Internal/External Sources)	Score (Metric 40)
Gene Wiki	63,234,184	0.0393
Music	5,937,119	0.3799
Ships	300,626	0.3142
Random 100K #1	940,887	0.0595
Random 100K #2	934,848	0.0613
Random 500K	4,677,314	0.0606
Random 1M	9,336,331	0.0602

for installing applications and running commands) and sufficient hardware resources, especially permanent storage space on our server.<sup>22</sup> - accessed 15 April 2024 We also could not find the proper guidance or tool for establishing a local Wikibase Docker image on an RDF N-Triples subset of Wikidata. At the time the experiments were done, the subsetting tools could create RDF outputs only, and the Wikibase software supported bulk data import only in JSON. Besides technical issues, many quality-driven queries with this amount of data require several hours (even days) of execution. Our approach to overcome the high volume is subsetting, but some subsets (such as the Gene Wiki) are still very large, consisting of 9 million triples and 12GB of data. Due to the interconnectivity (as the nature of a graph data model), shrinking subsets beyond a certain point will not conquer the problem. With the current triplestore technologies, it is necessary to use powerful hardware such as a high amount of RAM and SSD storage. The solution is to perform an initial evaluation of the entire Wikidata followed by periodical investigations only on newly added/edited data.

The size problem and technical limitations with Wikibase Docker (lack of root privileges and sufficient resources) meant that we had to query lots of metadata (e.g. languages of sources in Metric 39 or equivalence of reference properties in Metric 4) directly from the Wikidata public endpoint. It is not a good practice because there is a seven-month period between our data dump and the date of the experiment. The best practice would be to include all metadata in the subsets or index the 03 January 2022 full dump in a local triplestore and query it. The first solution is not possible with current subsetting tools. The second solution, however, requires expensive infrastructure.<sup>23</sup>

The lack of a permanent and easy access method to the Wikidata revision history impacted this study. Our approach utilised the HTML history web pages, which are inaccurate due to missing information. Wikimedia revision dump files are more than 3TB compressed, making it far harder than Wikidata dumps to process locally. Accessing the revision history is required for any quality study, and establishing permanent ways to access the historical metadata is the data provider's responsibility. In several metrics, we hypothesize the variation in scores is related to the amount of bot versus human activities, but distinguishing bots from humans requires pattern recognition of activities, which requires access to the detailed revisioning metadata. The same is true about freshness and date-time metadata.

In several metrics where accessing accurate data is impossible, we use proxies. For example, in Metric 13, we use the concept of black-listed domains as the reputation proxy. This approach has limitations: as the number of black-listed domains is low, the metric returns unrealistically high scores. A better solution would be to have a ranking system for Wikidata's external sources individually. A ranking algorithm can update the visits of external sources periodically and deliver better insight into the reputation of external sources.

The problem of subjective metrics is another matter of importance. One of these metrics is relevancy. The high relevance of references can increase the quality score of other objective metrics. In subsets such as Ships, many reference values are Wikidata ship instance items that are relevant to the statement they reference, but good referencing practice would be to link to external sources to verify the data [29]. For example, the claim for the power of a nu-

<sup>22</sup>The Wikibase Docker image can be found in <https://hub.docker.com/layers/wikibase/wikibase/1.35.4-wmde.2/images/sha256-9f665d6053138aa48f7b7af64f11b9e07f604dd78bab90cda0bdab7078956c18?context=explore>

<sup>23</sup>A Google Cloud computation engine with sufficient resources would cost more than \$571 per month. Estimated by Google Cloud Pricing Calculator: <https://cloud.google.com/products/calculator/#id=32eca290-7628-48af-9988-20508f4bc861> - accessed 27 November 2023

clear ship engine should refer to governmental documentation, encyclopedia articles, or military magazines, not an item within Wikidata. In such cases, we need an approach to distinguish non-relevant and non-sensible provenance values.

## 7. Lessons Learned

Despite the limitations discussed in Section 6, this research reveals important promising results. The findings of this study provide a resounding affirmative to the question: “can the quality of referencing in Wikidata be assessed effectively by relying on the Linked Data quality definitions and metrics”, by defining a framework consisting of 40 quality metrics across different data quality dimensions, coming both from Linked Data quality literature and novel definitions. The most important achievement of this research is that statistical analysis can identify data quality weaknesses in the context of referencing. The results revealed that while Wikidata exhibits high scores in areas like accuracy and security of references, there are opportunities for improvement in dimensions such as completeness, verifiability, objectivity, and multilingualism. For multilingualism, which is a flagship defining characteristic of Wikidata, our results indicate low performance. Our analysis critiques these scores and suggests the most efficient ways of improvement. Although having low scores in criteria such as the completeness of referencing is expected (and hard to improve due to the data volume and rapid growth of Wikidata), in other dimensions such as interlinking, the quality can be improved by treating a small amount of data, i.e., only reference properties. The quality scores also uncovered interrelationships between different quality dimensions. For example, we observed the human-added ratio has a strong indirect effect on verifiability (verifiable type of sources) and a direct effect on objectivity (multiple references per fact). Another relationship was that having multiple references for facts affects multilingualism positively. The comprehensive review gives us a good insight into the subjective versus quantitative criteria. Given the rapid advancements of Large Language Models (LLMs) and their capacity to access real-time data from the Web, an intriguing direction for future research is to explore the feasibility of integrating subjective criteria into LLMs. This approach could potentially alleviate the challenges associated with collecting human opinions in a high scale.

Another question that RQSS, as the main deliverable of this study, addresses is “to what extent is there a difference in the quality of references provided by humans and bots?”, where our initial hypothesis was that a strong bot activity would lead to higher overall referencing quality scores. The research found that this hypothesis is wrong. While bots perform well in tasks such as adding new provenance metadata and adhering to schemas, they lag in dimensions such as using referencing-specific properties consistently, maintaining freshness of references, representational conciseness, and providing multilingual sources. The human-added referencing ratio is lower in random subsets compared to topical subsets except Gene Wiki, where the highly bot-active exhibited similar patterns to random subsets in many metrics.

One of the primary lessons gleaned from this research is the importance of subsetting in assessing the quality of a KG. By examining both topical and random subsets in a unified comparison, our study illuminates the quality of referencing within specific Wikidata WikiProjects (such as Gene Wiki, Music, and Ships), which represent thematic aspects of the Wikidata knowledge base, alongside random subsets that reflect the entirety of the KG. This approach provides valuable insights into the referencing quality across different thematic areas and the whole Wikidata, and can be used in future quality assessments. Besides subsets, the framework can be deployed on other Wikidata projects such as Scholarly Articles, Astronomy, or Law, to allow maintainers and editors to identify weaknesses in the quality of references based on the scores. It can also be directly applied to other KGs hosted in Wikibase instances that follow the Wikidata model, e.g., the EU Knowledge Graph [65].

## 8. Conclusions

In this study, we investigated the referencing quality of a collaborative KG, Wikidata. We first defined a comprehensive framework for assessing referencing metadata based on previously defined Linked Data quality dimensions. We used the Wikidata data model to define formal referencing quality metrics. We implemented all objective

metrics as the Reference Quality Scoring System – RQSS – and then deployed RQSS over three topical and four random Wikidata subsets. We gathered valuable information on the referencing quality of Wikidata. RQSS scores show that Wikidata is rich in the accuracy, availability, security, and understandability of referencing, but relatively weak in completeness, defined schemas, verifiability, objectivity and multilingualism of referencing. In more detail, in the accessibility category, Wikidata subsets have an average of 0.95 for availability and 0.92 for security, but 0.06 for licensing and 0.12 for interlinking. In the intrinsic category, the average score is 0.99 for accuracy, 0.56 for consistency and 0.65 for conciseness. In the trust category, the average score of subsets for reputation is 0.99, for believability is 0.5, for verifiability is 0.35, but for objectivity is 0.02. In the currency category, the average is 0.94 for the freshness of facts-reference pairs but 0.09 for the freshness of external URIs. In the contextual category, the average of schema completeness is less than 0.01, however, for schema-based property completeness the average is 0.39 and for instance-based property completeness the average is 0.35, and for amount-of-data, the average is 0.34. In the representational category, the average of subsets scores is 0.88 for representational-conciseness, 0.99 for representational-consistency, 0.85 for understandability, 0.99 for interoperability, and 0.59 for versatility. RQSS reveals the interrelation between different referencing quality dimensions and highlights efficient ways to address the weaknesses in referencing quality in Wikidata, especially in reference properties.

The results show several metrics return a score very close to 0 or 1 in all subsets. These metrics can be divided into three categories:

1. Metrics that return high scores in Wikidata random and topical subsets, but might behave differently in other non-Wikidata Wikibase-derived datasets. Syntactic Validity of Reference Triples, Usage of Blank Nodes in References, and Labelling-Commenting metrics (both English and multilingual) belong to this category. In current Wikidata dumps, due to active maintenance, negative scores in such metrics are rare. However, these metrics are essential for the framework when the end users try to assess a non-Wikidata but a Wikibase-derived dataset or aim to find those rare inconsistencies.
2. Metrics that return low scores in Wikidata because the measuring target is very recent. Schema-based metrics in the Completeness dimension belong to this category. The concept of EntitySchemas in Wikidata is recent compared with the KG lifetime. Again, the presence of these metrics is required to be able to monitor Wikidata schema-based referencing quality and other Wikibase-derived datasets.
3. The External URIs Reputation metric, which uses deny-listed URIs as a proxy to measure URLs reputation (instead of using page ranks). Until finding a reliable measurement, this metric can be ignored in referencing quality assessments, unless end users want to find those deny-listed URIs to achieve a 100% score.

Our evaluation had multiple challenges: the large volume of the Wikidata dump and the lack of proper documentation to establish local copies of data namely, regarding the Docker images, the lack of a feasible approach to access Wikidata revision history, and the impact of the subjective quality issues on objective metrics. RQSS is the first reusable comprehensive referencing quality investigation and gives us valuable insights into referencing quality strengths and weaknesses. Adding support for subjective criteria in relevancy, authoritativeness and consistency, by deploying a combination of convolutional networks learned over human opinions would further strengthen the RQSS framework. Another important future step is to overcome the challenges of massive data and historical metadata. Although RQSS can effectively calculate referencing quality scores and the analysis of scores provided valuable information about Wikidata, RQSS scores should be evaluated by human experts to ensure their usefulness. Finally, the RQSS assessment framework should be generalized to all RDF KGs. In the current version, RQSS and its assessment framework are based on the Wikidata data model. This means that the Python implementation and the formal definitions are made using Wikidata terminology, vocabulary, and RDF model. In addition, several necessary metadata for computing the metrics come directly from Wikidata, e.g., schemata and historical information. The good news is that the nature of the referencing quality metrics and dimensions can be reproduced for any other KGs. In all KGs that support referencing, references must be available, complete, reputable, etc. Even the type of calculation can be generalized with few changes. For example, in the Amount-of-data dimension, for KGs that references are bound to the items (instead of statements), one can change the ratios per item (instead of statements). The current implementation can be applied to any Wikibase-derived dataset with minor changes in prefixes and namespaces. Generalizing RQSS for any RDF KG enables data quality researchers to compare provenance quality across different KGs.

*Acknowledgement.* We appreciate the helpful suggestions and fruitful discussions of the Shape Expressions (ShEx) community and the ELIXIR BioHackathon Europe Subsetting Project [66]: Dan Brickley, Katherine Thornton, Eric Prud'hommeaux, and Andra Waagmeester.<sup>24</sup> The first author would like to acknowledge the EPSRC grant EP/T022124/1.

## References

- [1] D. Vrandečić and M. Krötzsch, Wikidata: a free collaborative knowledgebase, *Communications of the ACM* **57**(10) (2014), 78–85. doi:10.1145/2629489.
- [2] G. Amaral, A. Piscopo, L.-A. Kaffee, O. Rodrigues and E. Simperl, Assessing the quality of sources in Wikidata across languages: a hybrid approach, *arXiv preprint arXiv:2109.09405* (2021).
- [3] Wikidata:Statistics, 2022. <https://www.wikidata.org/wiki/Wikidata:Statistics>.
- [4] M. Färber, F. Bartscherer, C. Menne and A. Rettinger, Linked data quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO, *Semantic Web* **9**(1) (2017), 77–129. doi:10.3233/SW-170275.
- [5] Wikidata:Introduction - Wikidata, 2023, visited on 27 November 2023. <https://www.wikidata.org/wiki/Wikidata:Introduction>.
- [6] Help:Sources - Wikidata, visited on 27 November 2023. <https://www.wikidata.org/wiki/Help:Sources>.
- [7] S.A. Hosseini Beghaeiraveri, Towards Automated Technologies in the Referencing Quality of Wikidata, in: *Companion Proceedings of the Web Conference 2022*, WWW '22 Companion, Association for Computing Machinery, New York, NY, USA, 2022, pp. 324–328. ISBN 9781450391306. doi:10.1145/3487553.3524192.
- [8] C. Bizer, Quality-driven information filtering in the context of web-based information systems, PhD Thesis, Freie Universität Berlin, 2007.
- [9] P.N. Mendes, H. Mühleisen and C. Bizer, Sieve: linked data quality assessment and fusion, in: *Proceedings of the 2012 Joint EDBT/ICDT Workshops*, 2012, pp. 116–123.
- [10] A. Hogan, A. Harth, A. Passant, S. Decker and A. Polleres, Weaving the pedantic web, in: *LDOW*, 2010.
- [11] C. Fürber and M. Hepp, SWIQA – A SEMANTIC WEB INFORMATION QUALITY ASSESSMENT FRAMEWORK, *ECIS 2011 Proceedings* (2011). <https://aisel.aisnet.org/ecis2011/76>.
- [12] A. Hogan, J. Umbrich, A. Harth, R. Cyganiak, A. Polleres and S. Decker, An empirical survey of Linked Data conformance, *Journal of Web Semantics* **14** (2012), 14–44. doi:10.1016/j.websem.2012.02.001. <https://www.sciencedirect.com/science/article/pii/S1570826812000352>.
- [13] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann and S. Auer, Quality assessment for Linked Data: A Survey, *Semantic web* **7**(1) (2016), 63–93, Publisher: IOS Press.
- [14] A. Rula, M. Palmonari and A. Maurino, Capturing the age of linked open data: Towards a dataset-independent framework, in: *2012 IEEE Sixth International Conference on Semantic Computing*, IEEE, 2012, pp. 218–225.
- [15] J. Debattista, C. Lange, S. Auer and D. Cortis, Evaluating the quality of the LOD cloud: An empirical investigation, *Semantic Web* **9**(6) (2018), 859–901, Publisher: IOS Press.
- [16] A. Piscopo, L.-A. Kaffee, C. Phethean and E. Simperl, Provenance information in a collaborative knowledge graph: an evaluation of Wikidata external references, in: *International semantic web conference*, Springer, 2017, pp. 542–558.
- [17] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak and Z. Ives, DBpedia: A Nucleus for a Web of Open Data, in: *The Semantic Web*, Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2007, pp. 722–735. ISBN 978-3-540-76298-0. doi:10.1007/978-3-540-76298-0\_52.
- [18] S.A.H. Beghaeiraveri, A. Gray and F. McNeill, Reference Statistics in Wikidata Topical Subsets, in: *Proceedings of the 2nd Wikidata Workshop (Wikidata 2021)*, CEUR Workshop Proceedings, Vol. 2982, CEUR, Virtual Conference, October, 2021, ISSN: 1613-0073. <http://ceur-ws.org/Vol-2982/#paper-3>.
- [19] S.A.H. Beghaeiraveri, A.J.G. Gray and F.J. McNeill, Experiences of Using WDumpster to Create Topical Subsets from Wikidata, in: *CEUR Workshop Proceedings*, Vol. 2873, CEUR-WS, 2021, p. 13, ISSN: 1613-0073. <https://researchportal.hw.ac.uk/en/publications/experiences-of-using-wdumper-to-create-topical-subsets-from-wikid>.
- [20] J.M. Juran, *Quality control handbook*, McGraw Hill, 1962, Issue: 658.562 Q-1q. ISBN 0-07-033175-8.
- [21] R.Y. Wang and D.M. Strong, Beyond accuracy: What data quality means to data consumers, *Journal of management information systems* **12**(4) (1996), 5–33, ISBN: 0742-1222 Publisher: Taylor & Francis.
- [22] C. Bizer and R. Cyganiak, Quality-driven information filtering using the WIQA policy framework, *Journal of Web Semantics* **7**(1) (2009), 1–10. doi:10.1016/j.websem.2008.02.005. <https://www.sciencedirect.com/science/article/pii/S157082680800019X>.
- [23] G. Developer, Basic Concepts | Freebase API (Deprecated), 2019, visited on 27 November 2023. [https://developers.google.com/freebase/guide/basic\\_concepts](https://developers.google.com/freebase/guide/basic_concepts).
- [24] M. Fabian, K. Gjergji and W. Gerhard, Yago: A core of semantic knowledge unifying wordnet and wikipedia, in: *16th International world wide web conference*, WWW, 2007, pp. 697–706.
- [25] D. Foxvog, Cyc, in: *Theory and Applications of Ontology: Computer Applications*, Springer Netherlands, Dordrecht, 2010, pp. 259–278. ISBN 978-90-481-8847-5. doi:10.1007/978-90-481-8847-5\_12.

---

<sup>24</sup><https://www.w3.org/community/shex/> - accessed 15 April 2024

- [26] A. Piscopo and E. Simperl, What we talk about when we talk about Wikidata quality: a literature survey, in: *Proceedings of the 15th International Symposium on Open Collaboration*, 2019, pp. 1–11.
- [27] K. Shenoy, F. Ilievski, D. Garijo, D. Schwabe and P. Szekely, A Study of the Quality of Wikidata, *Journal of Web Semantics* **72** (2022), 100679, Publisher: Elsevier.
- [28] D. Abián, A.M. Penuela and E. Simperl, An Analysis of Content Gaps versus User Needs in the Wikidata Knowledge Graph, in: *The Semantic Web—ISWC 2022 21st International Semantic Web Conference, ISWC 2022, Virtual Event, October 23–27, 2022, Proceedings, 2022*.
- [29] Wikidata:Verifiability - Wikidata, <https://www.wikidata.org/wiki/Wikidata:Verifiability> - accessed 28 July 2020. <https://www.wikidata.org/wiki/Wikidata:Verifiability>.
- [30] A. Piscopo, P. Vougiouklis, L.-A. Kaffee, C. Phethean, J. Hare and E. Simperl, 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 - OpenSym '17*, ACM Press, Galway, Ireland, 2017, pp. 1–10. ISBN 978-1-4503-5187-4. doi:10.1145/3125433.3125445. <http://dl.acm.org/citation.cfm?doid=3125433.3125445>.
- [31] P. Curotto and A. Hogan, Suggesting Citations for Wikidata Claims based on Wikipedia's External References., in: *Wikidata@ ISWC*, 2020.
- [32] A. Flemming, Quality characteristics of linked data publishing datasources, *Master's thesis, Humboldt-Universität of Berlin* (2010).
- [33] F. Callegati, W. Cerroni and M. Ramilli, Man-in-the-Middle Attack to the HTTPS Protocol, *IEEE Security & Privacy* **7**(1) (2009), 78–81, Conference Name: IEEE Security & Privacy. doi:10.1109/MSP.2009.12. [https://ieeexplore.ieee.org/abstract/document/4768661?casa\\_token=HW9jWgKX8dwAAAAA:TfTVxthWZ\\_9EisEwtndEkKmtYWaeVqtJav67DFsmcZAK0WRfotzX8RjclLjKnxF4xQCQYYY](https://ieeexplore.ieee.org/abstract/document/4768661?casa_token=HW9jWgKX8dwAAAAA:TfTVxthWZ_9EisEwtndEkKmtYWaeVqtJav67DFsmcZAK0WRfotzX8RjclLjKnxF4xQCQYYY).
- [34] S.A. Thomas, *SSL & TLS essentials: securing the Web*, Wiley, New York, 2000. ISBN 978-0-471-38354-3.
- [35] C. Guéret, P. Groth, C. Stadler and J. Lehmann, Assessing linked data mappings using network measures, in: *Extended semantic web conference*, Springer, 2012, pp. 87–102.
- [36] T. Berners-Lee, Linked Data - Design Issues, 2006, visited on 27 November 2023. <https://www.w3.org/DesignIssues/LinkedData>.
- [37] E.M. Knorr, R.T. Ng and V. Tucakov, Distance-based outliers: algorithms and applications, *The VLDB Journal* **8**(3) (2000), 237–253. doi:10.1007/s007780050006.
- [38] C. Batini, C. Cappiello, C. Francalanci and A. Maurino, Methodologies for data quality assessment and improvement, *ACM computing surveys (CSUR)* **41**(3) (2009), 1–52, Publisher: ACM New York, NY, USA.
- [39] Wikibase/Indexing/RDF Dump Format - MediaWiki, visited on 27 November 2023. [https://www.mediawiki.org/wiki/Wikibase/Indexing/RDF\\_Dump\\_Format](https://www.mediawiki.org/wiki/Wikibase/Indexing/RDF_Dump_Format).
- [40] C. Böhm, F. Naumann, Z. Abedjan, D. Fenz, T. Grütze, D. Hefenbrock, M. Pohl and D. Sonnabend, Profiling linked open data with ProLOD, in: *2010 IEEE 26th International Conference on Data Engineering Workshops (ICDEW 2010)*, 2010, pp. 175–178. doi:10.1109/ICDEW.2010.5452762.
- [41] J. Golbeck, Inferring Reputation on the Semantic Web, in: *In Proceedings of the 13th International World Wide Web Conference*, 2004.
- [42] C. Batini and M. Scannapieco, *Data and Information Quality: Dimensions, Principles and Techniques*, Data-Centric Systems and Applications, Springer International Publishing, Cham, 2016. ISBN 978-3-319-24104-3 978-3-319-24106-7. doi:10.1007/978-3-319-24106-7.
- [43] Y. Gil and D. Artz, Towards content trust of web resources, *Journal of Web Semantics* **5**(4) (2007), 227–239. doi:10.1016/j.websem.2007.09.005. <https://www.sciencedirect.com/science/article/pii/S1570826807000376>.
- [44] I. Jacobi, L. Kagal and A. Khandelwal, Rule-based trust assessment on the semantic web, in: *International Workshop on Rules and Rule Markup Languages for the Semantic Web*, Springer, 2011, pp. 227–241.
- [45] O. Hartig, Trustworthiness of data on the web, in: *Proceedings of the STI Berlin & CSW PhD Workshop*, Citeseer, 2008.
- [46] J. Golbeck and A. Mannes, Using Trust and Provenance for Content Filtering on the Semantic Web., in: *MTW*, 2006, pp. 3–4.
- [47] I. Jacobi, L. Kagal and A. Khandelwal, Rule-based trust assessment on the semantic web, in: *International Workshop on Rules and Rule Markup Languages for the Semantic Web*, Springer, 2011, pp. 227–241.
- [48] J.J. Carroll, Signing RDF graphs, in: *International Semantic Web Conference*, Springer, 2003, pp. 369–384.
- [49] F. Naumann, *Quality-driven query answering for integrated information systems*, Vol. 2261, Springer, 2003.
- [50] M.A. Ferradji and F. Benchikha, Enhanced metrics for temporal dimensions toward assessing Linked Data: A case study of Wikidata, *Journal of King Saud University. Computer and information sciences* (2021), Publisher: Elsevier BV. doi:10.1016/j.jksuci.2021.05.010.
- [51] T. Lebo, S. Sahoo and D. McGuinness, PROV-O: The PROV Ontology, 2021, visited on 27 November 2023. <https://www.w3.org/TR/prov-o/#generatedAtTime>.
- [52] R. Cyganiak, H. Stenzhorn, R. Delbru, S. Decker and G. Tummarello, Semantic sitemaps: Efficient and flexible access to datasets on the semantic web, in: *European Semantic Web Conference*, Springer, 2008, pp. 690–704.
- [53] S. Weibel, J. Kunze, C. Lagoze and M. Wolf, Dublin core metadata for resource discovery. <https://www.rfc-editor.org/rfc/rfc2413>.
- [54] J.E.L. Gayo, D. Kontokostas and S. Auer, Multilingual linked open data patterns, *Semantic Web journal* (2013), Publisher: Citeseer.
- [55] S.A.H. Beghaeiraveri, RQSSFramework. <https://github.com/seyedahbr/RQSSFramework/releases/tag/v1.0.2>.
- [56] S. Burgstaller-Muehlbacher, A. Waagmeester, E. Mitraka, J. Turner, T. Putman, J. Leong, C. Naik, P. Pavlidis, L. Schriml and B.M. Good, Wikidata as a semantic framework for the Gene Wiki initiative, *Database* **2016** (2016), Publisher: Oxford Academic.
- [57] B. Fünfstück, WDump, 2019. <https://github.com/bennofs/wdumper>.
- [58] S.A.H. Beghaeiraveri, WDump, 2021. <https://github.com/seyedahbr/wdumper>.
- [59] S.A.H. Beghaeiraveri, RQSS\_Evaluation. [https://github.com/seyedahbr/RQSS\\_Evaluation/releases/tag/v1.0.2](https://github.com/seyedahbr/RQSS_Evaluation/releases/tag/v1.0.2).
- [60] S.A.H. Beghaeiraveri, Wikidata 3 Topical Subsets (Gene Wiki, Music, Ships) and 4 Random Subsets, *Zenodo*, 2022, <https://doi.org/10.5281/zenodo.7332161>. doi:10.5281/zenodo.7332161.



- [61] H. Solbrig, PyShEx, 2018. <https://github.com/hsolbrig/PyShEx>.
- [62] D. Olteanu, H. Meuss, T. Furche and F. Bry, XPath: Looking Forward, in: *XML-Based Data Management and Multimedia Engineering — EDBT 2002 Workshops*, Vol. 2490, Springer Berlin Heidelberg, Berlin, Heidelberg, 2002, pp. 109–127, Series Title: Lecture Notes in Computer Science. ISBN 978-3-540-00130-0 978-3-540-36128-2. doi:10.1007/3-540-36128-6\_7.
- [63] J.E.L. Gayo, E. Prud'hommeaux, I. Boneva and D. Kontokostas, Validating RDF Data, *Synthesis Lectures on the Semantic Web: Theory and Technology* 7(1) (2017), 1–328. doi:10.2200/S00786ED1V01Y201707WBE016.
- [64] S.A.H. Beghaeiraveri, Output files of the RQSS extractor and framework on 3 Topical (Gene Wiki, Music, Ships) subsets and 4 Random Subsets, Zenodo, 2022, <https://doi.org/10.5281/zenodo.7336208>. doi:10.5281/zenodo.7336208.
- [65] D. Diefenbach, M.D. Wilde and S. Alipio, Wikibase as an Infrastructure for Knowledge Graphs: The EU Knowledge Graph, in: *The Semantic Web – ISWC 2021*, Lecture Notes in Computer Science, Springer International Publishing, Cham, 2021, pp. 631–647. ISBN 978-3-030-88361-4. doi:10.1007/978-3-030-88361-4\_37.
- [66] J.E. Labra-Gayo et al., Project 21 - Biohackathon 2021 - KG subsets, kg-subsetting, 2021, original-date: 2021-11-08T13:27:08Z. <https://github.com/kg-subsetting/biohackathon2021>.
- [67] U. Consortium, UniProt: a hub for protein information, *Nucleic acids research* 43(D1) (2015), D204–D212, Publisher: Oxford University Press.
- [68] S. Hunter, R. Apweiler, T.K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bork, U. Das, L. Daugherty and L. Duquenne, InterPro: the integrative protein signature database, *Nucleic acids research* 37(suppl\_1) (2009), D211–D215, Publisher: Oxford University Press.
- [69] I. Rogers, The Google Pagerank algorithm and how it works (2002). <http://ianrogers.uk/google-page-rank/>.
- [70] Wikidata:Bots - Wikidata, 2021-08-22, visited on 27 November 2023. <https://www.wikidata.org/wiki/Wikidata:Bots>.
- [71] Wikidata:Database reports/EntitySchema directory, 2023, visited on 27 November 2023. [https://www.wikidata.org/wiki/Wikidata:Database\\_reports/EntitySchema\\_directory](https://www.wikidata.org/wiki/Wikidata:Database_reports/EntitySchema_directory).
- [72] Wikidata:Schemas - Wikidata, visited on 27 November 2023. <https://www.wikidata.org/wiki/Wikidata:Schemas>.
- [73] L.L. Pipino, Y.W. Lee and R.Y. Wang, Data quality assessment, *Communications of the ACM* 45(4) (2002), 211–218. doi:10.1145/505248.506010.
- [74] Help:Sources/Items not needing sources - Wikidata, visited on 27 November 2023. [https://www.wikidata.org/wiki/Help:Sources/Items\\_not\\_needing\\_sources#When\\_the\\_item\\_has\\_a\\_statement\\_that\\_refers\\_to\\_an\\_external\\_source](https://www.wikidata.org/wiki/Help:Sources/Items_not_needing_sources#When_the_item_has_a_statement_that_refers_to_an_external_source).
- [75] H. Nielsen, J. Mogul, L.M. Masinter, R.T. Fielding, J. Gettys, P.J. Leach and T. Berners-Lee, Hypertext Transfer Protocol – HTTP/1.1, Request for Comments, RFC 2616, Internet Engineering Task Force, 1999, Num Pages: 176. doi:10.17487/RFC2616. <https://datatracker.ietf.org/doc/rfc2616>.
- [76] HTML URL Encoding Reference, visited on 27 November 2023. [https://www.w3schools.com/tags/ref\\_urlencode.ASP](https://www.w3schools.com/tags/ref_urlencode.ASP).

## Appendix A. Formal Definitions of Metrics

This appendix provides the formal definitions of our referencing quality assessment metrics. Formalizing definitions prevents ambiguity in measuring and enables us to implement the metrics for automatic evaluation.

### A.1. Terminology

We use the following terms and sets in the metrics formal definitions. Since we use Wikidata data model [39] as the base, the reification model and prefixes used are the same as Wikidata:

- *Wikidata* as the set of all Wikidata RDF dump triples in the form of  $(x, y, z)$  which  $x$  is the subject,  $y$  is the predicate, and  $z$  is the object part of the triple.
- $D \subseteq \text{Wikidata}$  as the input dataset.
- $P_D$  as the set of properties (predicate) used in statements of  $D$ , i.e.,  $P_D := \{y \mid (x, p: y, z) \in D\}$
- $S_D$  as the set of statements in  $D$ , i.e.,  $S_D := \{x \mid (x, \text{rdf:type}, \text{wikibase:Statement}) \in D\}$ .  $\forall s_i \in S_D$ ,  $\text{predicate}(s_i)$  denotes the property that statement  $s_i$  is formed by,  $\text{predicate}(s_i) = \{y \mid (x, p: y, s_i) \in D\}$ . In Wikidata,  $\text{predicate}(s_i)$  will always have only one member.
- $C_D$  as the set of classes in  $D$ , i.e.,  $C_D := \{x \mid (x, \text{wdt:P279}, z) \in D\} \cup \{z \mid (x, \text{wdt:P31}, z) \in D\}$
- $I_D$  as the set of instances in  $D$ , i.e.,  $I_D := \{x \mid (x, \text{wdt:P31}, z) \in D\}$
- $L_D$  as the set of literals in  $D$ , i.e.,  $L_D := \{z \mid (x, y, z) \in D \wedge \neg \text{isIRI}(z)\}$
- $R_D$  as the set of reference nodes in  $D$ , i.e.,  $R_D := \{x \mid (x, \text{rdf:type}, \text{wikibase:Reference}) \in D\}$
- $RT_D$  as the set of triples used in reference nodes (reference triples), i.e.,  $RT_D := \{(x, y, z) \mid x \in R_D\}$
- $RP_D$  as the set of properties (predicates) used in reference triples, i.e.,  $RP_D := \{y \mid (x, y, z) \in RT_D\}$
- $RO_D$  as the set of objects used in reference triples, i.e.,  $RO_D := \{z \mid (x, y, z) \in RT_D\}$

- 1 –  $RL_D$  as the set of literals used in reference triples, i.e.,  $RL_D := \{x \mid x \in RO_D \wedge x \in L_D\}$  1
- 2 –  $urlDomain(x)$  denotes the domain part of URI  $x \forall x \in RO_D \setminus RL_D$ . 2
- 3 –  $R_D^{ext}$  as the set of external sources in  $D$ , i.e.,  $R_D^{ext} := \{x \mid x \in RO_D \setminus RL_D \wedge \neg(urlDomain(x) \in WikiHosts \vee$  3  
4  $(x, wdt:P127, wd:Q180) \in D)\}$ ,<sup>25</sup> where  $WikiHosts := \{"wikipedia.org", "wikimedia.org",$  4  
5  $"wikivoyage.org", "mediawiki.org", "wikiversity.org", "wikinews.org",$  5  
6  $"wikisource.org", "wikibooks.org", "wikiquote.org", "wiktionary.org",$  6  
7  $"wikiba.se"\}$  7
- 8 –  $RU_D^{ext}$  as the set of external URIs used as an object in reference triples, i.e.,  $RU_D^{ext} := \{x \mid x \in R_D^{ext} \wedge$  8  
9  $urlDomain(x) \neq "wikidata.org"\}$ .  $RU_D^{ext}$  exclude those external sources from  $R_D^{ext}$  that have been 9  
10 added as Wikidata Q-ID items and represent a dataset, a book, a magazine, etc. 10

## 11 A.2. Formal Definitions and Discussions 11

### 12 Category I. Accessibility 12

#### 13 DIMENSION 1. AVAILABILITY 13

14 **Category I. Accessibility** 14

15 DIMENSION 1. AVAILABILITY 15

16 DIMENSION 1. AVAILABILITY 16

17 *Metric 1. Availability of External URIs* Consider function  $deref : RU_D^{ext} \rightarrow \{0, 1\}$  as follows: 17

$$18 \quad deref(x) = \begin{cases} 1 & \text{if http/https request of } x \text{ responds with status code 200} \\ 0 & \text{otherwise} \end{cases} \quad 18$$

19 Then, we define metric  $m_{deref}$  as below: 19

$$20 \quad m_{deref} = \frac{\sum_{x \in RU_D^{ext}} deref(x)}{|RU_D^{ext}|} \quad 20$$

21 *Discussion.* In Wikidata, reference-specific properties such as *reference URL (P854)* and *stated in (P248)* accept 21  
22 URIs as their objects to show an external source for the fact. These properties have been used repeatedly in ac- 22  
23 tive Wikidata projects [18]. These external sources must be available at the time of the user's request, otherwise, 23  
24 validation and confirmation of the reference is not possible. 24

#### 25 DIMENSION 2. LICENSING 25

26 *Metric 2. External URIs Domain Licensing* Consider  $RDS_D^{ext}$  to be the set of domains of the external URIs in  $RU_D^{ext}$ : 26

$$27 \quad RDS_D^{ext} := \{urlDomain(x) \mid x \in RU_D^{ext}\} \quad 27$$

28 We define the function  $isDSLicensed : RDS_D^{ext} \rightarrow \{0, 1\}$  as follows: 28

$$29 \quad isDSLicensed(x) = \begin{cases} 1 & \text{if } x \text{ has a human or machine-readable license} \\ 0 & \text{otherwise} \end{cases} \quad 29$$

30 Then, we define  $m_{license}$  as: 30

$$31 \quad m_{license} = \frac{\sum_{x \in RDS_D^{ext}} isDSLicensed(x)}{|RDS_D^{ext}|} \quad 31$$

---

32 <sup>25</sup> owned by (P127) and Wikimedia Foundation (Q180) 32

*Discussion.* The Wikidata knowledge base is licensed under Creative Commons Zero (CC0).<sup>26</sup> It means that Wikidata references are available for free or for commercial reuse with no limitations. In the context of references, a reference will be more likely to be reused if the external dataset has a license. A clear license makes the users and third parties aware of legal rights and permission to use the data [13]. For example, assume there are two references in a given statement of a protein: one to Uniprot [67] and one to InterPro [68]. The former is more likely to be reused as the UniProt dataset has a CC BY 4.0 license, while InterPro has no clear license as of this writing.

### DIMENSION 3. SECURITY

*Metric 3. Security of External URIs* Consider function  $isSecure : RU_D^{ext} \rightarrow \{0, 1\}$  as follows:

$$isSecure(x) = \begin{cases} 1 & \text{if } x \text{ supports TLS/SSL requests} \\ 0 & \text{otherwise} \end{cases}$$

Then, we define metric  $m_{secure}$  as below:

$$m_{secure} = \frac{\sum_{x \in RU_D^{ext}} isSecure(x)}{|RU_D^{ext}|}$$

### DIMENSION 4. INTERLINKING

*Metric 4. Interlinking of Reference Properties* We define function  $interlinkExists : RP_D \rightarrow \{0, 1\}$  as below:

$$interlinkExists(x) = \begin{cases} 1 & \text{if } x \text{ is connected to an equivalent property in another ontology} \\ 0 & \text{otherwise} \end{cases}$$

Then, we define metric  $m_{refPropInterlinking}$  as follows:

$$m_{refPropInterlinking} = \frac{\sum_{x \in RP_D} interlinkExists(x)}{|RP_D|}$$

*Discussion.* Interlinking in reference properties eases adaptation. Using equivalent connections, Wikidata-specific approaches and automatic tools of the reference properties can be generalized to other ontologies. In Wikidata, *equivalent property* (P1628) indicates the similarity of a Wikidata property to a fellow property in another ontology. Considering this, the numerator of the metric fraction (the amount of  $\sum_{x \in RP_D} interlinkExists(x)$ ) is equal to  $|\{x \in RP_D \mid (x, \text{wdt} : \text{P1628}, z) \in D\}|$ .

### DIMENSION 5. PERFORMANCE

Performance is not applicable in the context of references (see Section 5).

## Category II. Intrinsic

### DIMENSION 6. ACCURACY

<sup>26</sup><https://creativecommons.org/publicdomain/zero/1.0/> - accessed 15 April 2024

*Metric 5. Syntactic Validity of Reference Triples* Consider  $PatRef_D$  be the reification pattern for the references in Wikidata. Consider function  $isReifValid : S_D \rightarrow \{0, 1\}$  as follows:

$$isReifValid(x) = \begin{cases} 1 & x \text{ matches } PatRef_D \\ 0 & \text{otherwise} \end{cases}$$

Then, we define  $m_{synTriple}$  metric as follows:

$$m_{synTriple} = \frac{\sum_{x \in S_D} isReifValid(x)}{|S_D|}$$

*Discussion.* KGs have their specific data model for adding references. An accurate reference should follow this data model. Failure to follow the right pattern makes the reference unavailable for the user and causes inaccuracy in data. The patterns can be defined using Shape Expressions (ShEx) [63]. ShEx is a structural schema language that allows validation, traversal and transformation of RDF graphs. ShEx is well-organized to describe RDF patterns. Evaluation of references over patterns can then be done with validator tools like shex.js and PyShEx.<sup>27</sup> The number of mismatches returned by a ShEx validator tool can illustrate the metric.

*Metric 6. Syntactic Validity of Reference Literals* Consider function  $isLitSynValid : RL_D \rightarrow \{0, 1\}$  as follows:

$$isLitSynValid(x) = \begin{cases} 1 & x \text{ matches the specified literal rule} \\ 0 & \text{otherwise} \end{cases}$$

Then, we define metric  $m_{synLiteral}$  as below:

$$m_{synLiteral} = \frac{\sum_{x \in RL_D} isLitSynValid(x)}{|RL_D|}$$

*Discussion.* Some of the reference-specific properties accept literals as the object. For example, *title* (P1476) is a widely used property in Wikidata references that accepts a string value, indicating the published name of a source. This metric assesses that the literals are syntactically compatible with their specified data type. The compatibility can be checked by regular expressions that are specified to properties by the Wikidata community. In Wikidata, *property constraint* (P2302) carries metadata about how the property should be used. One of the values that *property constraint* (P2302) can have is *format constraint* (Q21502404) in which this statement can have a qualifier (a piece of metadata attached to the statements to explain more context) with the property *format as a regular expression* (P1793).

*Metric 7. Semantic Validity of Reference Triples (Subjective)* Let  $SS_D \subseteq S_D$  be a finite set of selected statements from  $S_D$ . For each  $S_i \in SS_D$ , let  $GS_i$  be the gold standard reference triples for the statement  $S_i$ . We define  $EQ_{S_i}^{RT_D}$  as the set of reference triples in  $RT_D$  for which an equivalent  $\langle \text{subject, relation} \rangle$  pair in the gold standard set  $GS_i$  exists (subject-relation matches):

$$EQ_{S_i}^{RT_D} := \{(x, y, z) \in RT_D \mid \exists(a, b, c) \in GS_i : equiv(x, a) \wedge equiv(y, b)\}$$

<sup>27</sup>ShEx.js: <https://github.com/shexjs/shex.js> and PyShEx: <https://github.com/hsolbrig/PyShEx> - accessed 15 April 2024

Also, consider  $EQ_{S_i}^{RT_D|GS_D}$  to be the set of triples in  $RT_D$  for which an equivalent triple in the gold standard set  $GS_i$  exists (exact matches):

$$EQ_{S_i}^{RT_D|GS_D} := \{(x, y, z) \in RT_D \mid \exists(a, b, c) \in GS_i : equiv(x, a) \wedge equiv(y, b) \wedge equiv(z, c)\}$$

Then, we define  $m_{semTriple}$  as the ratio of all exact matches to all (subject, relation) pair matches:

$$m_{semTriple} = \frac{\sum_{S_i \in SS_D} |EQ_{S_i}^{RT_D|GS_D}|}{\sum_{S_i \in SS_D} |EQ_{S_i}^{RT_D}|}$$

*Discussion.* Färber et al. [4] used the ‘semantic validity of triples’ metric to evaluate whether the statements presented by the triples are true. They compared 100 samples from each KG to a carefully selected dataset as a gold standard. This dataset includes 100 triples about persons gathered from a trusted source (Integrated Authority File from German National Library). We take the same approach of comparing with a gold standard set. The evaluation of this metric is highly dependent on the trustworthiness of the gold standard set [4]. To form such a gold standard set, one needs to provide completely accurate references for a topic, which needs human experts. To reflect the entire Wikidata in a relatively small set, the provided gold standard set should be unbiased and complete through sampling diversly and involving multiple domain experts.

## DIMENSION 7. CONSISTENCY

*Metric 8. Consistency of Reference Properties* Consider function *isRefS pecific* :  $RP_D \rightarrow \{0, 1\}$  as follows:

$$isRefS pecific(x) = \begin{cases} 1 & x \text{ is a reference-specific property} \\ 0 & \text{otherwise} \end{cases}$$

Then, we define metric  $m_{refPropCon}$  as below:

$$m_{refPropCon} = \frac{\sum_{x \in RP_D} isRefS pecific(x)}{|RP_D|}$$

*Discussion.* By this metric, one can ensure that the dataset uses reference-specific properties in the reference triples as much as possible. It will be difficult for humans and machines to track references that do not use reference-specific predicates. There is no standard for reference-specific predicates. Dublin Core Metadata terms [53] with properties such as `dcterms:provenance` and `dcterms:source` and the W3C PROV-O [51] with properties such as `prov:wasDerivedFrom` are examples of widely used provenance ontologies in Linked Data [4]. Wikidata has its own ontology to keep the provenance information. Predicates like *reference URL* (P854) and *stated in* (P248) are widely used in Wikidata references [18]. In Wikidata *property constraint* (P2302) carries another metadata about where the property should be used. This metadata is placed under the *property scope* (P5314) qualifier of the *property scope constraint* (Q53869507) values. Figure 4 shows the scope constraints of the property *stated in* (P248).

*Metric 9. Range Consistency of Reference Triples*  $\forall x_i \in RO_D$  let set  $TYP_{x_i}$  to be all types that  $x_i$  can be an instance of them, i.e., the classes that  $x_i$  belongs to if  $x_i$  is an item or the datatype of  $x_i$  if  $x_i$  is literal. Also,  $\forall y_i \in RP_D$  suppose there is a function *range*( $y_i$ ) that returns the range(s) of the given reference predicate  $y_i$ . Also consider function *inRange* :  $RT_D \rightarrow \{0, 1\}$  defined as below:

$$inRange(x := (a, b, c)) = \begin{cases} 1 & range(b) \in TYP_c \\ 0 & \text{otherwise} \end{cases}$$

Then we define the metric  $m_{trpRangeCon}$  as follows:

$$m_{trpRangeCon} = \frac{\sum_{x \in RT_D} inRange(x)}{|RT_D|}$$

*Discussion.* Nonconformity of domains (expected type of the subject of a triple) and ranges (expected type of the object of a triple) in triples can lead to inconsistencies in queries and make information retrieval hard [15].

*Metric 10. Multiple References Consistency (Subjective)* Let  $MRS_D \subseteq S_D$  be the set of those statements that have more than one reference. We define function  $isRefCon : MRS_D \rightarrow \{0, 1\}$  as follows:

$$isRefCon(x) = \begin{cases} 1 & \text{if references of the statement } x \text{ are compatible pairwise} \\ 0 & \text{otherwise} \end{cases}$$

Then the metric  $m_{multiRefCon}$  will be:

$$m_{multiRefCon} = \frac{\sum_{x \in MRS_D} isRefCon(x)}{|MRS_D|}$$

*Discussion.* Mentioning multiple separate references for a statement is usual in Wikidata. In cases where there are several separate references for a fact, these references need to be consistent. Assessing the consistency of two references is not doable without human opinions as it needs checking the relevancy and the equivalence of the content of the two references. Thus, this metric is subjective. This metric should be considered along with the other subjective dimension Relevancy (DIMENSION 18).

#### DIMENSION 8. CONCISENESS

*Metric 11. Schema-level Consiciencies of Reference Properties (Subjective)* Suppose there is function  $arePredsRed : RP_D \rightarrow \{0, 1\}$  as follows:

$$arePredsRed(x, y) = \begin{cases} 1 & \text{if } x \text{ and } y \text{ are equivalent} \\ 0 & \text{otherwise} \end{cases}$$

Then we define metric  $m_{schemaRed}$  as follows:

$$m_{schemaRed} = 1 - \frac{\sum_{\substack{x, y \in RP_D \\ x \neq y}} arePredsRed(x, y)}{|RP_D|}$$

*Discussion.* An example of redundancy in schema level in Wikidata is *reference URL (P854)* versus *URL (P2699)*. The former is a reference-specific property that presents the Internet URL of a source. The latter is a regular property (not reference-specific) used for the same reason. If a dataset uses both properties for referencing, a schema-level redundancy occurs. The same situation can be considered for *stated in (P248)* and *published in (P1433)*. However, these judgments are quite subjective.

*Metric 12. Ratio of Reference Sharing* Consider the set  $SR_D$  to be the set of reference nodes that provide provenance for more than one statement:

$$SR_D := \{x \in R_D \mid \exists f_1, f_2 \in S_D : (f_1, \text{prov:wasDerivedFrom}, x) \in D \wedge (f_2, \text{prov:wasDerivedFrom}, x) \in D \wedge f_1 \neq f_2\}$$

We define metric  $m_{refSharing}$  as follows:

$$m_{refSharing} = \frac{|SR_D|}{|R_D|}$$

*Discussion.* Reference sharing refers to using a set of reference triples for more than one statement in common [18]. Shared references are very usual in Wikidata [18]. Shared references are assumed to be created by bots where they add references for a bunch of statements at once. Using shared references can reduce the redundancy of data in reference triples. However, sharing a reference between statements can violate the relevancy condition. Together with Metric 30, a balance can be made to the relevancy and redundancy of reference triples.

### Category III. Trust

#### DIMENSION 9. REPUTATION

*Metric 13. External URIs Reputation* Assume there is the function  $srcRanker : RU_D^{ext} \rightarrow [0, 1]$  such that  $srcRanker(x)$  returns the page rank (a real number between 0 and 1) of the external source  $x$  based on the number of incoming web links to  $x$ . We define metric  $m_{srcRank}$  as follows:

$$m_{srcRank} = \frac{\sum_{x \in RU_D^{ext}} srcRanker(x)}{|RU_D^{ext}|}$$

*Discussion.* One of the available methods to determine the rank of web URIs is Google PageRank [69]. However, Google is not providing page rank data anymore. The current benchmarks belong to late 2016. We consider another metric as a proxy to *Metric 13* to approximate the reputation of references by checking if the external URIs are blacklisted. In that case, we define  $isSrcBL : RU_D^{ext} \rightarrow \{0, 1\}$  as below:

$$isSrcBL(x) = \begin{cases} 1 & \text{if } x \text{ is blacklisted} \\ 0 & \text{otherwise} \end{cases}$$

then, the proxy metric  $m_{srcBL}$  as follows:

$$m_{srcBL} = 1 - \frac{\sum_{x \in RU_D^{ext}} isSrcBL(x)}{|RU_D^{ext}|}$$

Online datasets such as PydnsI can be used to identify where a URL is blacklisted.<sup>28</sup> Please note that such a proxy is a very weak approximation of the real score.

#### DIMENSION 10. BELIEVABILITY

*Metric 14. Human-added References* We define  $m_{humanRefs}$  as:

$$m_{humanRefs} = \frac{|\{x \in R_D \mid x \text{ added by human}\}|}{|R_D|}$$

<sup>28</sup>PydnsI: <https://pypi.org/project/pydnsbl/0.5.4/> - accessed 15 April 2024

*Discussion.* Data users trust datasets more if data is added and curated by humans (especially experts) instead of automated tools [4]. Automated tools are widely used to provide the provenance of statements. YAGO uses `yago:extractionTechnique` predicate to indicate the extraction method of a statement. Wikidata uses *bots* [70] for adding references -however, distinguishing bot activities from humans is challenging. This task requires querying Wikidata revision history which is not hosted anywhere. Furthermore, there is no differentiating method for detecting bots and humans: the activity of some human user accounts is similar to bots in terms of adding bulk data at once and detecting this needs pattern recognition over data.

#### DIMENSION 11. VERIFIABILITY

*Metric 15. Verifiable Type of References* Assume there is function  $typeVerifScore : RO_D \setminus RL_D \rightarrow [0, 1]$  as follows:

$$typeVerifScore(x) = \begin{cases} 1 & \text{if type of } x \text{ is scholarly article} \\ 0.75 & \text{if type of } x \text{ is well-known trusted knowledge base} \\ 0.5 & \text{if type of } x \text{ is book, encyclopedia, or encyclopedic article} \\ 0.25 & \text{if type of } x \text{ is magazine, blog, or blog post} \\ 0 & \text{otherwise} \end{cases}$$

Then, we define metric  $m_{verif}$  as follows:

$$m_{verif} = \frac{\sum_{x \in RO_D} typeVerifScore(x)}{|RO_D|}$$

*Discussion.* Once it comes to verifying a reference, a peer-reviewed article is more verifiable than a book, and a book is more verifiable than a web URI. Well-known knowledge bases gather and structurize data in their focus topic from a trustable scientific, librarian, or political sources (e.g., UniProt in life science).<sup>29</sup> We consider such datasets more verifiable than books and less than scholarly articles. For reference values that are Wikidata items, we can check the *instance of (P31)* property of the reference value. However, detecting the value type of external URIs is a challenging task, requiring involving human judgment and machine-learning methods.

#### DIMENSION 12. OBJECTIVITY

*Metric 16. Multiple References for Statements* Let  $RS_D \subseteq S_D$  be the set of referenced statements in  $D$ , i.e., statements that have at least one reference, and let  $MRS_D \subseteq RS_D$  be the set of those statements that have two or more references. Then we define metric  $m_{multi}$  as follows:

$$m_{multi} = \frac{|MRS_D|}{|RS_D|}$$

*Discussion.* A fact with multiple references is more verifiable and reliable. Considering objectivity as the data provider's effort to increase quality, we check whether the dataset provides more than one reference for a single fact.

### Category IV. Dynamicity

#### DIMENSION 13. CURRENCY

<sup>29</sup><https://www.uniprot.org/> - accessed 15 April 2024



*Metric 17. Freshness of Reference Triples*  $\forall x \in RT_D$ , let  $modifTime(x)$  be the time of the last modification (or creation if there is no modification after creation), and  $startTime(x)$  be the origin of time for reference triple  $x$ . Also, consider  $t_{now}$  denotes the observation time. We define metric  $m_{freshTriple}$  as follows:

$$m_{freshTriple} = \frac{\sum_{x \in RT_D} \frac{t_{now} - modifTime(x)}{t_{now} - startTime(x)}}{|RT_D|}$$

*Discussion.* The origin of time is a point in time from which the metric is measured [14]. One option for time origin is the publish time of the entire dataset  $D$ . A more accurate time origin for reference triple  $x$  is the creation time of  $S_x$ , which  $S_x$  is the statement that  $x$  is a reference for. Finding freshness data for Wikidata triples is challenging. The metadata of addition, deletion and changes of the Wikidata statements, including times and editors, is called Wikidata Revision History.<sup>30</sup> This dataset is far more extensive than Wikidata dumps and there is no public endpoint for it.

*Metric 18. Freshness of External URIs*  $\forall x \in RU_D^{ext}$ , let  $modifTime(x)$  be the time of the last modification (or creation if there is no modification after creation), and  $startTime(x)$  be the origin of time for external URI  $x$  (see Metric 17). Also, consider  $t_{now}$  denotes the observation time. We define metric  $m_{freshExternal}$  as follows:

$$m_{freshExternal} = \frac{\sum_{x \in RU_D^{ext}} \frac{t_{now} - modifTime(x)}{t_{now} - startTime(x)}}{|RU_D^{ext}|}$$

*Discussion.* The creation or last modification time of a URI can be fetched by the HTTP response headers, or via Google Cache. HTTP headers can be inaccurate as some servers set the <Last-Modified> header to the request time, even when the page was published previously.

#### DIMENSION 14. VOLATILITY

*Metric 19. Volatility of External URIs* Assume there is a function  $ssChangeFreq : RU_D^{ext} \rightarrow [0, 1]$  that maps the value of <changeFreq> attribute to numbers between 0 and 1, as follows:

$$ssChangeFreq(x) = \begin{cases} 1 & \text{value of } \langle \text{changeFreq} \rangle x \text{ is always} \\ 0.9 & \text{value of } \langle \text{changeFreq} \rangle x \text{ is hourly} \\ 0.8 & \text{value of } \langle \text{changeFreq} \rangle x \text{ is daily} \\ 0.6 & \text{value of } \langle \text{changeFreq} \rangle x \text{ is weekly} \\ 0.4 & \text{value of } \langle \text{changeFreq} \rangle x \text{ is monthly} \\ 0.1 & \text{value of } \langle \text{changeFreq} \rangle x \text{ is yearly} \\ 0 & \text{otherwise} \end{cases}$$

Then, we define metric  $m_{volat}$  as follows:

$$m_{volat} = \frac{\sum_{x \in RU_D^{ext}} ssChangeFreq(x)}{|RU_D^{ext}|}$$

<sup>30</sup>The revision history can be downloaded from <https://dumps.wikimedia.org/backup-index.html> - accessed 15 April 2024

*Discussion.* A highly volatile reference means the user can expect the source to be regularly edited, updated, and curated in short periods. Volatility is a way to measure how the provenance data provider manages its content.

#### DIMENSION 15. TIMELINESS

*Metric 20. Timeliness of External URIs* Let  $m_{freshExternal}$  and  $m_{volat}$  be the measurements for *freshness of external URIs* and *volatility of external URIs* for a given dataset. Then we define metric  $m_{timeliness}$  of the dataset as follows:

$$m_{timeliness} = \begin{cases} \frac{m_{freshExternal}}{m_{volat}} & m_{volat} > 0 \text{ and } m_{volat} > m_{freshExternal} \\ 1 & \text{otherwise} \end{cases}$$

*Discussion.* Timeliness is the fraction of the real-world reference updating frequency (freshness) on the expected reference updating frequency (volatility). The closer the real-world frequency is to the expected frequency, the better the score timeliness will be.

### Category V. Contextual

#### DIMENSION 16. COMPLETENESS

*Metric 21. Class/Property Schema Completeness of References* Consider  $C_D^{schema} \subset schema(D)$  to be the set classes defined in the schema of  $D$ .  $\forall c_i \in C_D^{schema}$  let  $RPC_{c_i}^{schema}$  be the set of reference-specific properties defined in  $schema(D)$  to be used as a reference predicate for instances of class  $c_i$ . Likewise, consider  $P_D^{schema} \subset schema(D)$  be the set properties defined in the schema of  $D$  and  $\forall sp_i \in P_D^{schema}$  let  $RPP_{sp_i}^{schema}$  be the set of reference-specific properties defined in  $schema(D)$  to be used as a reference predicate for property  $sp_i$ . We define metric  $m_{classSchemaCom}$  as below:

$$m_{classSchemaCom} = \frac{|\{x \in C_D \mid RPC_x^{schema} \neq \emptyset\}|}{|C_D|}$$

and metric  $m_{propertySchemaCom}$  as following:

$$m_{propertySchemaCom} = \frac{|\{x \in P_D \mid RPP_x^{schema} \neq \emptyset\}|}{|P_D|}$$

*Discussion.* Färber et al. [4] measured schema completeness (in knowledge bases) by comparing the dataset to a gold standard set containing real-world classes. We do not compare reference schemata of  $D$  with a gold standard. Instead, we count classes that have a reference schema. Wikidata uses Entity-Schemas (based on ShEx) in which the shape of references for each class and properties of that class can be specifically determined<sup>31</sup>. The existence of such schemata is a key factor to enhance this metric. The full list of Wikidata EntitySchemas can be found in [71].

*Metric 22. Schema-based Property Completeness of References*

Consider  $P_D^{schema} \subset schema(D)$  be the set properties defined in the schema of  $D$  and  $\forall sp_i \in P_D^{schema}$  let  $RPP_{sp_i}^{schema}$  be the set of reference-specific properties defined in  $schema(D)$  to be used as a reference predicate for property  $sp_i$ . Consider the set of all (referenced statement, reference property) pairs,  $H \subseteq (S_D \times RP_D)$  as:

$$H := \{(s, r) \mid s \in S_D \wedge r \in RP_D \wedge \exists o \in R_D : (s, \text{prov:wasDerivedFrom}, o) \in D \wedge (o, r, x) \in RT_D\}$$

<sup>31</sup>For example, see <https://www.wikidata.org/wiki/EntitySchema:E265> - accessed 15 April 2024

Also, we define set  $IS \subseteq (P_D^{schema} \times \bigcup_{sp_i \in D^{schema}} RPP_{sp_i}^{schema})$  as the ⟨property, reference predicate⟩ pairs in the schema level:

$$IS := \{(sp, r) \mid sp \in P_D^{schema} \wedge r \in RPP_{sp}^{schema}\}$$

Then,  $\forall (sp_i, r_j) \in IS$ , we define the completeness ratio of reference property  $r_j$  w.r.t. its references schema property  $ps_i$  as follows:

$$comRefPropS_{r_j}^{sp_i} = \frac{|\{(s, r) \in H \mid predicate(s) = sp_i \wedge r = r_j\}|}{|\{(s, r) \in H \mid predicate(s) = sp_i\}|}$$

and the metric  $m_{sbRefPropCom}$  as the following average:

$$m_{sbRefPropCom} = \frac{\sum_{(sp_i, r_j) \in IS} comRefPropS_{r_j}^{sp_i}}{|IS|}$$

*Discussion.* Although having a data schema is not mandatory in semi-structured datasets, Wikidata encourages users to define schemata to improve the quality of data [72]. As a complement to Metric 21, this metric is an indicator of the richness of the input dataset schema in references. Note that the set  $H$  contains only referenced statements. There might be statements at the instance level with no references. These statements are not included in calculating the completeness metrics as we assume that non-referenced statements do not need to be referenced according to Wikidata policies. However, we can calculate completeness metrics by taking both cases into account. In that case, the completeness ratio of reference property  $r_j$  w.r.t. its references schema  $ps_i$  would be as follows:

$$comRefPropS_{r_j}^{sp_i} = \frac{|\{(s, r) \in H \mid predicate(s) = sp_i \wedge r = r_j\}|}{|\{s \in S_D \mid predicate(s) = sp_i\}|}$$

**Metric 23. Property Completeness of References** Assume we partition the set  $S_D$  into the family of fact class sets  $P = \{[p_1], \dots, [p_n]\}$ , based on an equivalence relation  $X = \{(s_i, s_j) \mid predicate(s_i) = predicate(s_j)\}$  as follows:

$$[p_i] := \{s \in S_D \mid predicate(s) = p_i\}$$

Also, consider  $H$  to be the set of all combinations of the referenced facts and their reference as defined in Metric 22 and consider ⟨fact class, reference property⟩ pairs set  $I \subseteq (P \times RP_D)$  as:

$$I := \{\langle [p], r \rangle \mid [p] \in P \wedge \exists (s, r) \in H : s \in [p]\}$$

Then,  $\forall \langle [p_i], r_j \rangle \in I$ , we define completeness ratio of reference property  $r_j$  w.r.t. fact class  $[p_i]$  as follows:

$$comRefProp_{r_j}^{[p_i]} = \frac{|\{(s, r) \in H \mid s \in [p_i] \wedge r = r_j\}|}{|\{(s, r) \in H \mid s \in [p_i]\}|}$$

and the metric  $m_{refPropCom}$  as:

$$m_{refPropCom} = \frac{\sum_{\langle [p_i], r_j \rangle \in I} comRefProp_{r_j}^{[p_i]}}{|I|}$$

*Discussion.* The main difference between this metric and Metric 22 is that Metric 22 computes the completeness of reference-specific properties using the dataset schemata, while this metric computes the completeness of reference-specific properties by comparing the current status of similar data instances, regardless of any schema. The logic is similar to Färber et al. *column completeness* metric [4]. In traditional relational datasets that have a fixed schema, property (aka relation or column) completeness is the degree by which a defined property in schema level is used in the instance records [73]. In semi-structured datasets (like RDF), there is no fixed schema. Therefore, one can measure the column completeness as the extent to which instances of the same class have used the same properties [73] in instance level. In the context of references, we expect facts that are formed by the same property to have similar references using the same reference-specific properties. For example, if there is a fact about a wrestler’s mass (e.g., using *mass (P2067)* property), and the fact has a reference using *reference URL (P854)*, then we expect all equivalent mass-facts to have a reference using *reference URL (P854)* property. This metric is the average of this expectation. Similar to Metric 22, this metric can be calculated by taking non-referenced statements into account. In that case, the completeness ratio of reference property  $r_j$  w.r.t. fact class  $[p_i]$  will be as follows:

$$comRefProp_{r_j}^{[p_i]} = \frac{|\{(s, r) \in H \mid s \in [p_i] \wedge r = r_j\}|}{|[p_i]|}$$

*Metric 24. Population Completeness of References (Subjective)* Let  $SS_D \subseteq S_D$  be a finite set of selected facts from  $S_D$  that need referencing. We define the metric  $m_{comPop}$  as follows:

$$m_{comPop} = \frac{|\{f \in S_D \mid f \in SS_D \wedge f \text{ has at least one reference}\}|}{|SS_D|}$$

*Discussion.* In Linked Data, the population completeness is measured by using the ratio of the number of represented real-world objects to the total number of real-world objects [13] in a gold standard set (for example, see [4]). In the context of references, we redefine the ratio as the number of referenced statements to all statements that need referencing. We use the “need for referencing” concept according to Wikidata. The Wikidata Help [74] clarifies that all statements need references except:

- When the value of the statement is common human knowledge. This usually happens with properties like *instance of*, *subclass of*, and *occupation* (just for well-known Items). For example, “Earth is an instance of an inner planet” does not need a source.
- When the item has a statement that refers to an external source. For example, the Douglas Adams item’s statement *Amazon author ID* does not need a reference because the external source of information allows easy verification of the statement.
- When the item itself is a source for a statement. For example, consider a statement about a book that has some authors. In this case, the authors do not need to include their book as a source of this statement.

Removing the above list of candidates from  $S_D$  will create the  $SS_D$  in this metric. However, Some parts of the exception list are subjective. Alternatively, the  $SS_D$  from Metric 7 (Semantic Validity of Reference Triples) is also suitable for this metric.

#### DIMENSION 17. AMOUNT-OF-DATA

*Metric 25. Ratio of Reference Nodes per Statement* We define metric  $m_{refNodesPerFact}$  as follows:

$$m_{refNodesPerFact} = \frac{|R_D|}{|S_D|}$$

*Discussion.* The ratio of distinct reference nodes per fact can show the richness of reference metadata in the dataset.

**Metric 26. Ratio of Reference Triples per Statement** We define metric  $m_{refTriplesPerFact}$  as follows:

$$m_{refTriplesPerFact} = \frac{|RT_D|}{|S_D|}$$

*Discussion.* Like Metric 25, this metric can give an overview of the richness in referencing.

**Metric 27. Ratio of Reference Triples per Reference Node** We define metric  $m_{refTriplesPerNode}$  as follows:

$$m_{refTriplesPerNode} = 1 - \frac{|R_D|}{|RT_D|}$$

*Discussion.* In the Wikidata data model, reference nodes collect a set of reference triples for facts. “Having more triples in a reference node provides more details about the source which is likely to increase the accuracy” [18]. By knowing how many triples there are for each reference node on average, we can estimate the detail level of referencing. As the number of reference triples is always equal or greater than reference nodes, to normalize the score between 0 and 1, we use complementary of the reverse fraction. For example, consider dataset  $D_1$  has two facts, each has been referenced using three reference triples and dataset  $D_2$  has four facts, each has been referenced with one triple. Then the  $D_1$  score is 0.66 and  $D_2$  is 0.

**Metric 28. Ratio of Reference Literals per Reference Triple** We define metric  $m_{refLiteralPerTriple}$  as follows:

$$m_{refLiteralPerTriple} = \frac{|RL_D|}{|RT_D|}$$

*Discussion.* This metric helps users to know to what extent reference values consist of literals. Literal value amongst reference triples can increase human readability. However, a high ratio of literal can affect the external referencing and decrease the trust in data.

## DIMENSION 18. RELEVANCY

**Metric 29. Relevance of Reference Triples (Subjective)** Assume we have function  $isRelevant : RT_D \rightarrow \{0, 1\}$  as below:

$$isRelevant(x) = \begin{cases} 1 & x \text{ is relevant to the fact to which it belongs} \\ 0 & \text{otherwise} \end{cases}$$

Then we define metric  $m_{relTriples}$  as follows:

$$m_{relTriples} = \frac{\sum_{x \in RT_D} isRelevant(x)}{|RT_D|}$$

*Discussion.* Previous works [2, 16] consider only external sources as the subject of relevancy evaluation. We believe that the entire reference triples, including the reference property and reference value (either external or internal source), should be evaluated for relevance. However, computing this metric needs aggregating human opinions, which makes it subjective.

**Metric 30. Relevance of Shared References (Subjective)** Consider shared references set  $SR_D$  as defined in Metric 12. Now consider  $SRT_D \subseteq RT_D$  as the set of all shared reference triples, i.e.  $SRT_D := \{(a, b, c) \mid a \in SR_D\}$ , and set  $FT$  as the set of all ⟨shared triple, fact⟩ pairs:

$$FT := \{\langle f, t : (a, b, c) \rangle \in SR_D \times SRT_D \mid (f, \text{prov:wasDerivedFrom}, a) \in D\}$$

Then, consider function  $isSharedTripleIrrelevant : FT \rightarrow \{0, 1\}$  as below:

$$isSharedTripleIrrelevant(x) = \begin{cases} 1 & \text{triple } x.(a, b, c) \text{ is not relevant to the fact } x.f \\ 0 & \text{otherwise} \end{cases}$$

Then we define metric  $m_{relShared}$  as follows:

$$m_{relShared} = 1 - \frac{\sum_{x \in FT} isSharedTripleIrrelevant(x)}{|FT|}$$

*Discussion.* The metric aims to measure whether the shared references are relevant to all of their connected statements. Reference sharing is considered a positive point in Metric 29. However, a high reference-sharing ratio can potentially decrease the relevancy of the facts connected to them.

## Category VI. Representational

### DIMENSION 19. REPRESENTATIONAL-CONCISENESS

*Metric 31. External Sources URL Length* Assume  $\forall x \in RU_D^{ext}$ , function  $ASCIILen(x)$  returns the number of ASCII characters of  $x$ . Now we define function  $URLShortness : RU_D^{ext} \rightarrow [0, 1]$  as below:

$$URLShortness(x) = \begin{cases} 1 & ASCIILen(x) \leq 80 \\ 0.75 & 80 < ASCIILen(x) \leq 2083 \\ 0.5 & 2083 < ASCIILen(x) \leq 4096 \\ 0 & \text{otherwise} \end{cases}$$

Then, we define metric  $m_{urlLength}$  as follows:

$$m_{urlLength} = \frac{\sum_{x \in RU_D^{ext}} URLShortness(x)}{|RU_D^{ext}|}$$

*Discussion.* The Hypertext Transfer Protocol HTTP/1.1 RFC [75] does not recommend an upper limit for the length of URLs. However, short URLs are easier for machines to parse and more efficient for datasets or servers to store. Web software applies different limitations on the length of URLs. Popular web server management software can handle URLs with 4096 characters (the lowest belongs to NGINX with 4098 characters).<sup>32</sup> Old browsers like Microsoft Internet Explorer cannot handle URLs with more than 2083 characters.<sup>33</sup> Traditional practice for characters per line is 80 characters.<sup>34</sup> Based on these different recommendations, we tried to define multi-level scoring. Since URLs can contain unsafe ASCII characters, counting the characters of the raw URL string does not work. The standard URL encoding on the web is Percent-encoding [76]. This method maps non-ASCII characters with a % sign followed by two hexadecimal numbers.

### DIMENSION 20. REPRESENTATIONAL-CONSISTENCY

<sup>32</sup>[https://nginx.org/en/docs/http/nginx\\_http\\_core\\_module.html#large\\_client\\_header\\_buffers](https://nginx.org/en/docs/http/nginx_http_core_module.html#large_client_header_buffers) - accessed 15 April 2024

<sup>33</sup><https://support.microsoft.com/en-us/topic/maximum-url-length-is-2-083-characters-in-internet-explorer-174e7c8a-6666-f4e0-6fd6-908b53c12246> - accessed 14 April 2024

<sup>34</sup>[https://en.wikipedia.org/wiki/Characters\\_per\\_line](https://en.wikipedia.org/wiki/Characters_per_line) - accessed 15 April 2024

*Metric 32. Diversity of Reference Properties* We define metric  $m_{refPropDiversity}$  as follows:

$$m_{refPropDiversity} = 1 - \frac{|RP_D|}{|RT_D|}$$

*Discussion.* The metric returns a lower score for input with a greater variety of properties, considering the number of total reference triples. The Wikidata reference properties are limited. Subsets may use similar numbers and types of properties. For a better insight into diversity, we can compute the usage frequency of reference properties [18]. In this case,  $\forall rp_i \in RP_D$  we define  $m_{refPropUse}^{rp_i}$  as follows:

$$m_{refPropUse}^{rp_i} = \frac{|\{(x, y, z) \in RT_D \mid y = rp_i\}|}{|RT_D|}$$

The above fraction shows how much the property  $rp_i$  is used for referencing in  $D$ . Such a distribution helps users to understand the usage balance of internal sitelinks against external sources and which external dataset is used more in references [18].

#### DIMENSION 21. UNDERSTANDABILITY

*Metric 33. Human-readable labelling of Reference Properties* We define metric  $m_{refHumanLabel}$  as follows:

$$m_{refHumanLabel} = \frac{|\{x \in RP_D \mid \exists z : (x, rdfs:label, z) \in D\}|}{|RP_D|}$$

*Discussion.* Different predicates are used in Linked Data to express the label of a subject.<sup>35</sup> In KGs like Wikidata, entities -including reference predicates- are named using Q, P, S, E, etc. IDs. Every entity in Wikidata needs to have a human-readable label. Without labels, using the entity within the user interface would be very ambiguous for human users. Wikidata RDF dump uses `rdfs:label`, `skos:prefLabel`, and `schema:name` predicates for each label of subjects. The essential labelling predicate that every Wikidata item should have is `rdfs:label`. Wikidata entities might have also different “Also known as” labels using `skos:altLabel` predicates.

*Metric 34. Human-readable Commenting of Reference Properties* We define metric  $m_{refHumanComment}$  as follows:

$$m_{refHumanComment} = \frac{|\{x \in RP_D \mid \exists z : (x, schema:description, z) \in D\}|}{|RP_D|}$$

*Discussion.* Descriptions are effective in removing the ambiguity of predicate usage. According to Wikidata, descriptions have a differentiating role for entities with similar labels.<sup>36</sup> Wikidata RDF dump uses `schema:description` predicate for each description.

*Metric 35. Handy External Sources* Assume function  $handyExt : R_D^{ext} \rightarrow [0, 1]$  as below:

$$handyExt(x) = \begin{cases} 1 & x \text{ is an online-available URL with anchor} \\ 0.75 & x \text{ is an online-available URL} \\ 0.5 & x \text{ is an online-available source} \\ 0.25 & x \text{ is an offline sources} \\ 0 & \text{otherwise} \end{cases}$$

<sup>35</sup>For a comprehensive list of labelling predicates see [15, §(U1)]

<sup>36</sup><https://www.wikidata.org/wiki/Help:Label> - accessed 15 April 2024

Then, we define metric  $m_{handyExt}$  as follows:

$$m_{handyExt} = \frac{\sum_{x \in R_D^{ext}} handyExt(x)}{|R_D^{ext}|}$$

*Discussion.* This metric measures to what extent external sources are easy to access for human users. In the first line, there are URLs with anchor; a # character in the path part of the URL. Anchors refer to a specific section or header in a long HTML page and direct the web browser to a particular point in the destination HTML page. Therefore, anchors can help human users save time verifying an online external source. In the next step, there are online-available URLs. These URLs have no anchor but point to a specific page. Those can be external dataset items' HTML pages, CSV files, PDF documents, etc. The next level is external online-available sources. These sources have not been added as a specific URL but are datasets which users can investigate online. Those have been added as Wikidata Q-IDs corresponding to a third-party dataset, e.g., *Integrated Authority File (Q36578)* in Figure 7. We can represent external online-available sources as the set  $\{x \in R_D^{ext} \mid (x, wdt:P31/wdt:P279*, wd:Q7094076) \in Wikidata\}$ .<sup>37</sup> The last category is the Wikidata items that point to offline sources such as books, magazines, compact Disks, etc. While some of these sources may be available online (free or by fee), automatically investigating online availability is not feasible as finding the web page that provides these sources is challenging.

#### DIMENSION 22. INTERPRETABILITY

*Metric 36. Usage of Blank Nodes in References* Consider set  $UN := R_D \cup RP_D \cup RO_D$ . We define metric  $m_{blankNode}$  as follows:

$$m_{blankNode} = 1 - \frac{|\{x \in UN \mid isBlank(x)\}|}{|UN|}$$

*Discussion.* Blank nodes occur at the population time when the dataset expects a reference node or a reference triple which is not available. Serialization errors also can cause this problem. Automatic tools can not interpret these nodes. Thus in terms of interoperability, having no references is better than having blank nodes. As shown in Figure 2, reference nodes, reference predicates, and reference values are the main parts of referencing in the Wikidata RDF model. This metric examines all those IRIs to find blank nodes in each.

#### DIMENSION 23. VERSATILITY

*Metric 37. Multilingual labelling of Reference Properties* We define metric  $m_{refMLLabel}$  as follows:

$$m_{refMLLabel} = \frac{|\{x \in RP_D \mid \exists z : (x, rdfs:label, z) \in D \wedge lang(z) \neq "en"\}|}{|RP_D|}$$

*Metric 38. Multilingual Commenting of Reference Properties* We define metric  $m_{refMLComment}$  as follows:

$$m_{refMLComment} = \frac{|\{x \in RP_D \mid \exists z : (x, schema:description, z) \in D \wedge lang(z) \neq "en"\}|}{|RP_D|}$$

*Discussion.* Wikidata is a multilingual open KG. Almost all entities in Wikidata (including reference properties) have labels and descriptions for multiple languages. Besides Metric 37 and Metric 38 definitions above, we investigate how many languages are added for each property.

<sup>37</sup>online database (Q7094076)



*Metric 39. Multilingual Sources*  $\forall x_i \in RO_D \setminus RL_D$  assume function  $srcLang(x_i)$  returns the ISO 639-1:2002 language code of the source.<sup>38</sup> We define metric  $m_{refMLSources}$  as follows:

$$m_{refMLSources} = \frac{|\{x \in RO_D \setminus RL_D \mid langSrc(x) \neq "en"\}|}{|RO_D \setminus RL_D|}$$

*Discussion.* This metric returns the ratio of non-English sources, considering both internal and external. We hypothesise that most of the non-English references in Wikidata are Wikimedia Foundation sources such as Wikipedia. For sources that are Wikidata items, *language of work or name (P407)* property indicates the language of the source as another Wikidata item. Language items have *ISO 639-1 code (P218)* item that returns the Alpha 2 code of the language. For other URLs, we check the `lang` attribute of the `<html>` tag.

*Metric 40. Multilingual Referenced Statements* Assume the function  $srcLang(x_i)$  from Metric 39. Also, consider setting  $MS$  to be the set of facts having at least one non-English source as a reference:

$$MS := \{x \in S_D \mid \exists c \in RO_D \setminus RL_D : (x, prov:wasDerivedFrom, z) \in D \wedge (z, b, c) \in D \wedge langSrc(c) \neq "en"\}$$

Then, we define metric  $m_{MLFacts}$  as follows:

$$m_{MLFacts} = \frac{|MS|}{|S_D|}$$

*Discussion.* Having multilingual references ease verification of the reference for non-English users. For some facts, e.g., contemporary facts related to closed non-English speaking countries, it is necessary to refer to the sources of the same language.

## Appendix B. Subsets Overall Scores

Although each metric has been measured in its specific method, the quality scores from different metrics of a given dimension can be combined and averaged to show the overall quality score of the dataset in the dimension (see, for example, Färber et al. [4, §6]). In this appendix, we present the discussions around a simple average and an exemplary weighted average of the RQSS quality scores.

Table 30 shows the overall RQSS scores of each subset in different categories, the total average of all scores, and an example of a weighted average (we explain the weights and the scenario in Section B.2). Despite waiting for more than 90 days and having three unsuccessful attempts, Metrics 14 and 18 scores were not obtained for Gene Wiki due to the large size of this subset. Metric 19, and therefore Metric 20 scores were not obtained due to the lack of an efficient tool for fetching `<changeFreq>` tags. We ignore these metrics in all averages. Considering the Overall Average column, the four random subsets have a higher score than the topical subsets. The scores of random subsets differ by less than 2%. This is most likely due to the similarity of their topic coverage (Figure 9). Gene Wiki has the highest score of the topical subsets and is only 1% less than the random subsets. This is most likely due to having a high amount of corresponding EntitySchemas (E-IDs) and the use of bots to populate the data. The Extractor and the Framework Runner outputs of performing RQSS on the topical and random subsets can be found in [64].

<sup>38</sup><https://www.iso.org/standard/22109.html> - accessed 15 April 2024

### 1 *B.1. Scores by Dimension* 1

2  
3 To investigate the quality of referencing by dimension, we calculate the average scores of all subsets in each 3  
4 dimension. At a summary level, we observe that all subsets have good scores in Intrinsic (accuracy-related metrics) 4  
5 and Representational dimensions but weak scores in Dynamicity (freshness-related) and Contextual (completeness 5  
6 and amount of data) categories. Contextual and Representation is where topical subsets have better scores than 6  
7 random subsets. 7

8 In the Accessibility category, the average of subsets is 0.95 for availability (obtained from the availability of 8  
9 external URIs results) and 0.92 for security (obtained from the security of external URIs results), but 0.06 for 9  
10 licensing (obtained from the external URIs domain licensing results) and 0.12 for interlinking (obtained from the 10  
11 interlinking of reference properties results). Regarding licensing, we have been expecting low scores due to the 11  
12 lack of explicit licenses in many external sources. However, in the case of interlinking, the low score means a high 12  
13 number of reference properties have no link to their equivalents in external vocabularies. In such cases, only curating 13  
14 reference properties can improve quality scores. 14

15 In the Intrinsic category (accuracy-related metrics), the average score is 0.99 for accuracy (obtained from the 15  
16 syntactic validity of reference triples and syntactic validity of reference literals results), 0.56 for consistency (ob- 16  
17 tained from the consistency of reference properties and range consistency of reference triples results), and 0.65 for 17  
18 conciseness (obtained from the ratio of reference sharing results). Despite the high accuracy scores, in the syntactic 18  
19 validity of reference literals metric, we observe that the lack of regexes for a few frequently used properties causes 19  
20 many literals not to be checked. The consistency of reference properties is higher than 0.7 in all subsets, and random 20  
21 subsets have better scores than topical subsets. In range consistency, scores vary from 0.2 (Ships) to 0.44 (Gene 21  
22 Wiki), and besides low scores, all subsets suffer from having no specified ranges for reference properties. The ref- 22  
23 erence sharing ratio as the proxy of conciseness varies between 0.3 and 0.7 and is considerably higher in random 23  
24 subsets than topical subsets. 24

25 In the Trust category, the average for reputation is 0.99 (obtained from the external URIs reputation results), for 25  
26 believability is 0.5 (obtained from the human-added references results), for verifiability is 0.35 (obtained from the 26  
27 verifiable type of references results), but for objectivity is 0.02 (obtained from the multiple references for statements 27  
28 results). In reputation, we investigated the blacklisted domains only, so having a small number of blacklisted URLs 28  
29 was expected. The blacklisted domain datasets identify highly malicious URLs, which are unlikely to be used as an 29  
30 external source in Wikidata. In believability, for which we use added-by-humans as the proxy, scores vary from 0.43 30  
31 to 0.78, and topical subsets have considerably higher scores than random subsets. The computation of Gene Wiki 31  
32 results timed out, but we think the scores should be close to random subsets due to active bots in its WikiProject. 32  
33 The added-by-human ratio is essential to explain the reasons behind other quality metrics. In the verifiable type of 33  
34 sources, random subsets and Gene Wiki have similar scores around average, but Ships and Music have notably low 34  
35 scores. In objectivity, for which we use having multiple references as the proxy, scores are less than 0.07 in topical 35  
36 subsets and even less than 0.01 in random subsets. 36

37 In the Dynamicity category, the average is 0.94 for the freshness of facts-reference pairs but 0.09 for the freshness 37  
38 of external URIs. In the fact-reference freshness, Ships has the highest scores. It was not expected because Ships has 38  
39 the highest percentage of human-added references and we hypothesized bots perform better in constantly updating 39  
40 reference information, but we observe the opposite. The freshness of external URIs is notably lower than reference- 40  
41 fact pairs, and Ships has the highest scores. It shows that the Ships WikiProject community uses up-to-date sources 41  
42 more than other subsets. In both metrics, there are many records that RQSS cannot find historical metadata for them. 42

43 In the Contextual category, the average of schema completeness is less than 0.01. As there are many Enti- 43  
44 tySchemas (E-IDs) in Wikidata related to life science, we expected Gene Wiki to score high in class/property schema 44  
45 completeness, but it has low scores. Instead, Ships and Music E-IDs provide more information about references 45  
46 despite being fewer in number. In schema-based property completeness, the average is 0.39. Here Gene Wiki has the 46  
47 highest score, and Music and Ships score notably low. It shows that Gene Wiki references comply with schemata 47  
48 better than other subsets. In instance-based property completeness, the average is 0.35, and random subset scores 48  
49 are higher than topical subsets. In the amount-of-data, the average is 0.34. 49

50 In the Representational category, the average is 0.88 for representational-conciseness (obtained from the external 50  
51 sources URL length results), 0.99 for representational-consistency (obtained from the diversity of reference proper- 51

Table 30

The average of RQSS metric scores in each category, the total average, and an example weighted average.

Subset	Accessibility	Intrinsic	Trust	Dynamicity	Contextual	Representational	Overall	Weighted
Gene Wiki	0.5332	0.7901	0.5086	0.0338	0.3211	0.7819	0.5816	0.5349
Music	0.4606	0.7824	0.3622	0.0758	0.2265	0.8534	0.5569	0.5013
Ships	0.5592	0.7469	0.3525	0.1239	0.1703	0.8245	0.5406	0.4840
Random 100K #1	0.5269	0.8918	0.5043	0.1116	0.2960	0.7868	0.5944	0.5476
Random 100K #2	0.5220	0.8951	0.5027	0.0842	0.2929	0.7871	0.5926	0.5451
Random 500K	0.5196	0.8849	0.5062	0.1029	0.2936	0.7881	0.5921	0.5451
Random 1M	0.5170	0.8891	0.5079	0.1116	0.2945	0.7878	0.5930	0.5463

ties results), 0.85 for understandability (obtained from the human-readable labelling and commenting of reference properties and handy external sources results), 0.99 for interoperability (obtained from the usage of blank nodes in references results), and 0.59 for versatility (obtained from the multilingual labelling and commenting of reference properties, multilingual sources, and multilingual referenced statements results). In having handy (easily accessible) external sources, topical subsets have higher scores than random subsets, and Music has the highest scores as it uses URLs with anchors more than other subsets. In multilingualism of reference properties, all subsets score 0.99 to 1. However, the use of multilingual sources for facts is notably low in all subsets. Music uses multilingual sources as references most frequently and Gene Wiki less than all subsets.

From the framework, many interrelations can be found between dimensions. Verifiability and objectivity are affected by human-added references. It can be concluded by the similarity of Gene Wiki scores to the random subsets scores. Multilingualism is affected by human-added references, but it is also affected by having multiple references for statements. We also observe that curating reference-specific properties and adding proper equivalents, regular expressions, ranges and schema metadata can increase referencing quality efficiently. Although referencing completeness and having multiple references are essential, they are time-consuming to improve; currently Wikidata scores low in these metrics.

## B.2. Weighted Average Score

It is possible to apply weights to the metrics to emphasise the perceived relative importance of the different scores. Assigning weight to the metrics is subjective and depends on the task at hand and users' qualitative requirements [4]. Data consumers can assign a higher weight to those quality metrics that are more important to their use case. For example, in the case of having a better schema in referencing, Metrics 21 and 22 weights should be higher, or if the understandability for humans is a matter of importance, Metrics 33 and 34 should be higher. The weighting strategy is up to the user as well. The provided example weights are coefficients of one. Another approach can be using normalized weights where the sum of weights is one.

We present one hypothetical weighting scenario in the last column of Table 30. Suppose our referencing quality investigation firstly cares about decision-making, which highly depends on the completeness of references, and secondly cares about understandability. Also, suppose we care about certainty in metric computations; thus, we cannot accept proxies in computing metrics. Then, the weights and the justifications for the importance of metrics are as below:

- Metrics 22 and 23 weights are set to three. This indicates the importance of completeness in references, as incomplete referencing can decrease the trust in data and make it hard for machines to perform decisions based on references.
- Metrics 35, 39, and 40 weights are set to two. That is because of the importance of online access to the provenance, and the existence of references for non-English users, which is also one of Wikidata's intentions.
- Metric 13 weight is set to zero as the current RQSS approach to use deny-listed IPs as the proxy of reputation is not accurate.
- The rest of the metrics are assigned a weight of one.

Note that our weighting scenario is only one example of many. The above scenario's weighted scores are lower than the overall scores. It can be a sign of Wikidata (subsets) reference quality weaknesses in completeness and multilingualism of referencing. It also shows that ignoring proxy-based metrics in computation can decrease the score, and therefore, it is likely that current proxies can produce unrealistic high scores. This phenomenon indicates the need for calibrating the metric set against a gold standard dataset to determine their effectiveness and expected outcomes. This calibration process typically involves establishing a correlation with human judgments to validate the accuracy of metrics. Such calibration outcomes aligned metrics, making the metric results understandable to the end users. We discuss this as a required future work, entitled The Human Evaluation of RQSS Scores in Section ??.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51