RQSS: Referencing Quality Scoring System for Wikidata

Seyed Amir Hosseini Beghaeiraveri a, Alasdair Gray a and Fiona McNeill b

a 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
b School of Informatics, The University of Edinburgh, Edinburgh, EH8 9AB, UK
E-mail: f.j.mcneill@ed.ac.uk

Abstract. Wikidata is a collaborative multi-purpose knowledge graph with the unique feature of adding provenance data to the statements of items as a reference. About 73% of Wikidata statements have provenance metadata, but there are few studies on the referencing quality in this knowledge graph, with existing studies focusing on relevancy and trustworthiness. While there are existing frameworks to assess the quality of Linked Data, there are none focused on reference quality. We define a comprehensive referencing quality assessment framework based on Linked Data quality dimensions. We implement the objective metrics of the assessment framework as the Referencing Quality Scoring System - RQSS. RQSS 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 overall 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 knowledge graphs in general.

Keywords: Reference Quality, Data Quality, Wikidata, Knowledge Graphs, Subetting, Topical Subsets, Random Subsets, Big Data, RQSS

1. Introduction

Approaching its tenth birthday, Wikidata [1] is now the paramount general-purpose user-contributed knowledge graph in research and industry [2]. By August 2022, Wikidata has 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 introduction1, Wikidata is a secondary database that collects statements along with their provenance. Providing provenance in Wikidata is called referencing. More than 73% of Wikidata statements have at least one reference2. Wikidata references can help AI tools to detect errors and make decisions based on the supporting evidence [5]. Having references also make Wikidata a believable and verifiable knowledge base for human users and researchers.

2https://wikidata-todo.toolforge.org/stats.php - accessed 17 August 2022. The page is not updating anymore
Linked Data Quality is a multi-dimensional concept [4, 6–13] in which, providing the source of facts is considered part of believability and verifiability dimensions [4, 11, 13]. Providing the provenance increases the trust in data [4, 11, 13]. Despite the high percentage of referencing in Wikidata and a large portion of metadata 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 [14], extended in 2021 by Amaral et al. [2]. These studies evaluated two subjective data quality dimensions (relevancy and verifiability) on Wikidata references. 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 referencing quality in Linked Data, including Wikidata. We aim to address this gap by defining and implementing a comprehensive framework for assessing referencing.

Although some KGs, e.g. DBPedia, support referencing on the resource (item) level, Wikidata is the only KG that supports referencing at the statement level among open general-purpose knowledge graphs. 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 [11, 15], 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. 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 knowledge graph 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 [16]. We use three topical subsets [15] 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 researches. Our contributions are (i) defining the first comprehensive referencing quality assessment framework for Linked Data based on the Wikidata data model, (ii) developing RQSS that is the referencing quality scoring system to automatically monitor referencing quality of Wikidata-driven datasets, and (iii) providing statistical scores of Wikidata subsets referencing quality during the evaluation of RQSS. This study is the most comprehensive evaluation of Wikidata references in different dimensions and complements previous subjective researches [2, 14]. 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. In Section 6 we discuss the main points of the study and a summary of lessons we learned during this research. Finally, in Section 7 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” [17]. In the literature, quality of data is considered a multidimensional concept. Wang and Strong [18] 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 in which extent data 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. [19] 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. [11] 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 [18] into 11 dimensions and 34 metrics and then evaluated five KGs: Freebase, Wikidata, YAGO, Cyc, and DBPedia. 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. [13] examined near 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. [20] surveyed 28 papers on Wikidata quality mostly published in 2017. They stated that trustworthiness needs to be investigated further in Wikidata. Shenoy et al. [21] proposed a framework to recognize low-quality statements in Wikidata. They created a historical dataset of Wikidata removed statements via connecting 311 weekly dumps and applied the removing pattern to current statements. Abian et al. [22] investigated the imbalances of Wikidata in gender, recency and geological data considering user needs. They used Wikipedia page views information to conclude user needs and applied them on 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 [11]. In the literature, the trust category consists of different dimensions like believability, reputation, objectivity and verifiability. Färber et al. [4] defined the trustworthiness dimension as a combination of the Wand and Strong three [18] dimensions: 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 in terms of their source or content). Trustworthiness at the statement level was a metric in Färber et al. [4] and Debattista et al. [13]. 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 and authoritative [23]. Piscopo et al. [14] examined the authoritative nature 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 authoritative nature 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. [24] 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 [25] 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 [11], which is the most comprehensive collection of Linked Data quality metrics to the best of our knowledge. Table 1 shows these dimensions with those that apply to references shown in bold. We first explain the terms used in the definitions in Section 3.1 and then provide the metric definitions in Section 3.2.

3.1. Terminology

We use the following terms and sets in metrics formal definitions. Since we use Wikidata’s data model [26] 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 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., \(S_D := \{x \mid (x, \text{rdf:type}, \text{wikibase:Statement}) \in D\}\).
- \(\forall s_i \in S_D, predicate(s_i)\) denotes the property that statement \(s_i\) is formed by, \(predicate(s_i) := \{y \mid (x, p:y, x) \in D\}\). In Wikidata, \(predicate(s_i)\) will always have only one member.
- \(D_C\) as the set of classes in \(D\), i.e., \(C_D := \{x \mid (x, \text{wdt:P279}, z) \in D\} \cup \{z \mid (x, \text{wdt:P31}, z) \in D\}\).
- \(D_I\) as the set of instances in \(D\), i.e., \(I_D := \{x \mid (x, \text{wdt:P31}, z) \in D\}\).
- \(D_L\) as the set of literals in \(D\), i.e., \(L_D := \{z \mid (x, y, z) \in D \land \neg \text{isIRI}(z)\}\).
- \(D_R\) as the set of reference nodes in \(D\), i.e., \(R_D := \{x \mid (x, \text{rdf:type}, \text{wikibase:Reference}) \in D\}\).
- \(D_RT\) as the set of triples used in reference nodes (reference triples), i.e., \(R_T := \{(x, y, z) \mid x \in R_D\}\).
- \(D_RP\) as the set of properties (predicates) used in reference triples, i.e., \(R_P := \{y \mid (x, y, z) \in R_T\}\).
- \(D_RO\) as the set of objects used in reference triples, i.e., \(R_O := \{z \mid (x, y, z) \in R_T\}\).
- \(D_RL\) as the set of literals used in reference triples, i.e., \(R_L := \{x \mid x \in R_O \land x \in R_L\}\).
- \(\text{urlDomain}(x)\) denotes the domain part of URI \(x\), \(\forall x \in ROD \land RL_D\).
- \(D_R^E\) as the set of external sources in \(D\), i.e., \(R^E_D := \{x \mid x \in R_O \land \text{urlDomain}(x) \in \text{WikiHosts} \lor (x, \text{wdt:P127}, \text{wd:Q180}) \in D\}\), where \(\text{WikiHosts} := \{"wikipedia.org","wikimedia.org","wikivoyage.org","mediawiki.org","wikiversity.org","wikinews.org","wikisource.org","wikibooks.org","wikiquote.org","wiktionary.org","wikiba.se"\}^3\).

Table 1

<table>
<thead>
<tr>
<th>Category</th>
<th>Accessibility</th>
<th>Intrinsic</th>
<th>Trust</th>
<th>Dynamicity</th>
<th>Contextual</th>
<th>representational</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dimension</td>
<td>Availability</td>
<td>Licensing</td>
<td>Security</td>
<td>Interlinking</td>
<td>Performance</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Accuracy</td>
<td>Consistency</td>
<td>Consiceness</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Reputation</td>
<td>Believability</td>
<td>Verifiability</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Currency</td>
<td>Volatility</td>
<td>Objetivity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Completeness</td>
<td>Amount-of-data</td>
<td>Relevancy</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3 owned by (P127) and Wikimedia Foundation (Q180)
– \( RU_{ext}^D \) as the set of external URIs used as an object in reference triples, i.e., \( RU_{ext}^D := \{ x \mid x \in R_D^D \land urlDomain(x) \neq \text{"wikidata.org"} \} \). \( RU_{ext}^D \) exclude those external sources from \( R_D^D \) that have been added as Wikidata Q-ID items and represent a dataset, a book, magazine, etc.

### 3.2. Referencing Quality Metrics

We now provide the formal definitions of our referencing quality assessment metrics. Formalizing definitions prevents ambiguity in measuring and enable us to implement the metrics for automatic evaluation. Categories and dimensions are sorted based on Zaveri et al. [11] survey as shown in Table 1. 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. At each dimension, we provide a brief survey of the Linked Data definition and metrics. We designed the metrics to return a number between 0 and 1 as the mean result, although in the majority of them, providing the distribution is helpful to analyze the data. For each defined metric, there is a discussion part explaining additional considerations to the definition.

#### Category I. Accessibility

This category includes dimensions that are related to access and retrieval of data. There are five dimensions in this category: availability, licensing, interlinking, security, and performance [11]. In the context of referencing, only performance is not applicable.

**DIMENSION 1. AVAILABILITY**

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

**Metric 1. Availability of External URIs**

Consider function \( deref : RU_{ext}^D \rightarrow \{0, 1\} \) as follows:

\[
deref(x) = \begin{cases} 
1 & \text{if http/https request of } x \text{ responds code 200} \\
0 & \text{otherwise}
\end{cases}
\]

Then, we define metric \( m_{deref} \) as below:

\[
m_{deref} = \frac{\sum_{x \in RU_{ext}^D} deref(x)}{|RU_{ext}^D|}
\]

**Discussion.** In Wikidata, reference-specific properties such as reference URL (P854) and stated in (P248) accept URIs as their objects to show an external source for the fact. These properties have been used repeatedly in active Wikidata projects [15]. These external sources must be available at the time of the user’s request, otherwise validation and confirmation of the reference is not possible.

**DIMENSION 2. LICENSING**

“Licensing is defined as the granting of permission for a consumer to re-use a dataset under defined conditions” [11]. In datasets, the licensing criteria are the existence of human-readable [10, 13] or machine-readable license
In the context of references, we define a metric for the licensing status of external URIs.

Metric 2. External URIs Domain Licensing
Consider $RDS_{ext}^D$ to be the set of domains of the external URIs in $RU_{ext}^D$:

$$RDS_{ext}^D := \{\text{urlDomain}(x) \mid x \in RU_{ext}^D\}$$

We define the function $isDS\ Licensed : RDS_{ext}^D \rightarrow \{0, 1\}$ as follows:

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

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

$$m_{license} = \frac{\sum_{x \in RDS_{ext}^D} isDS\ Licensed(x)}{|RDS_{ext}^D|}$$

Discussion. The Wikidata knowledge base is licensed under Creative Commons Zero (CC0)\(^4\). It means that Wikidata references are available for free or for commercial reuse with no limitation. In the context of references, a reference will be more likely to reuse 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 [11]. For example, assume there are two references in a given statement of a protein: one to UniProt [28] and one to InterPro [29]. 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
“Security is the extent to which access to data can be restricted and hence protected against its illegal alteration and misuse” [11]. Security is not covered as much as other Accessibility dimensions. According to Zaveri et al. [11], Flemming’s study [27] is the only work that includes a definition for this dimension. However, because security is rarely applied in governmental or medical datasets, Flemming does not include any metric in the tool for this dimension. Zaveri et al. (based on Wang and Strong [18]) mentioned secure access to data (e.g. via SSL or login credentials) and proprietary access to data as metrics of security.

In the context of references, secure access to external URIs is important. An unsecured external link decreases the trust in the provenance of data and causes the security threats such as man-in-the-middle. Therefore, we consider a metric for measuring the rate of secure TLS/SSL connections in external URIs.

Metric 3. Security of External URIs
Consider function $isSecure : RU_{ext}^D \rightarrow \{0, 1\}$ 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:

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

\(^4\)https://creativecommons.org/publicdomain/zero/1.0/ - accessed 28 August 2021
DIMENSION 4. INTERLINKING

In Linked Data, “interlinking refers to the degree to which entities that represent the same concept are linked to each other, be it within or between two or more linked data sources” [11]. This dimension is measured by data network parameters like interlinking degree, clustering coefficient, centrality, and sameAs chains [30]. Another metric is \texttt{owl:sameAs} links either to internal entities [4] or external URIs [4, 10, 13]. Färber et al. also considered the validity of external \texttt{owl:sameAs} links as a metric in this dimension [4]. Interlinking is one of the four fundamental principles of Linked Data [31]. We evaluate this dimension by a metric that measures the interconnectedness of reference properties to other ontologies.

Metric 4. Interlinking of Reference Properties

We define function \texttt{interlinkExists} \( : \text{RP}_D \rightarrow \{0, 1\} \) as below:

\[
\text{interlinkExists}(x) = \begin{cases} 
1 & \text{x is connected to equivalence property in another ontology} \\
0 & \text{otherwise}
\end{cases}
\]

Then, we define metric \( m_{\text{refPropInterlinking}} \) as follows:

\[
m_{\text{refPropInterlinking}} = \frac{\sum_{x \in \text{RP}_D} \text{interlinkExists}(x)}{|\text{RP}_D|}
\]

Discussion. Interlinking in reference properties ease adaptation. Using equivalent connections, Wikidata-specific approaches and automatic tools of the reference properties can be generalized to other ontologies. In Wikidata, \textit{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 \( \sum_{x \in \text{RP}_D} \text{interlinkExists}(x) \)) is equal to \(|\{x \in \text{RP}_D | (x, \text{wdt:P1628}, z) \in D\}|\).

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” [11]. The measures of evaluating this dimension are the usage of hash-URIs instead of slash-URIs [27] (as cited in [11]), low latency [6, 13, 27], high throughput [13], and scalability of a data source [27] (as cited in [11]). 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 [11]. Dimensions that belong to this category are accuracy, consistency, conciseness [11].

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
Fig. 1. The RDF model of Wikidata references, derived from [26]. 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.

(ii) semantic accuracy, which refers to the degree to which data values represent the correctness of the values to the actual real world values” [11]. Accuracy is an important aspect of data quality as it is sometimes considered a synonym of quality in the literature [4]. Bizer and Cyganiak [19] suggest outlier detection methods (e.g. distance-based, deviations-based, and distribution-based methods [32]) 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, 8, 13]. By evaluating the quality of five open knowledge graphs, Färber et al. [4], based on Batini et al. [33], 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** Consider PatRefD be the reification pattern for the references in Wikidata. Consider function isReiValid : SD \rightarrow \{0, 1\} as follows:

\[\text{isReiValid}(x) = \begin{cases} 1 & x \text{ matches PatRefD} \\ 0 & \text{otherwise} \end{cases}\]

Then, we define \(m_{\text{synTriple}}\) metric as follows:

\[m_{\text{synTriple}} = \sum_{x \in SD} \text{isReiValid}(x) \frac{1}{|SD|}\]

Discussion. Knowledge graphs have their specific data model for adding references. Figure 1 shows the Wikidata data model for 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) [34]. ShEx is a structural schema language allowing 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\(^5\) and PyShEx\(^6\). The number of mismatches returned by a ShEx validator tool can illustrate the metric.

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

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

---

\(^5\)https://github.com/shexjs/shex.js - accessed 28 August 2021

\(^6\)https://github.com/hsolbrig/PyShEx - accessed 28 August 2021
Then, we define metric $m_{\text{synLiteral}}$ as below:

$$m_{\text{synLiteral}} = \sum_{x \in RL_D} \text{isLitS\text{ynValid}}(x)$$

**Discussion.** Some of the reference-specific properties accept literals as the object. For example, retrieved (P813) is a widely used property in Wikidata references that accepts a date-time value, indicating the date that the fact was retrieved and added to Wikidata from an external online 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 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). Figure 2 shows an example of a regular expression assigned to reference-specific property title (P1476).

**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 $EQ^{RT}_S$ as the set of reference triples in $RT_D$ which there is an equivalent ⟨subject, relation⟩ pair in the gold standard set $GS_i$ for them (subject-relation matches):

$$EQ^{RT}_S := \{(x, y, z) \in RT_D | \exists (a, b, c) \in GS_i : \text{equiv}(x, a) \land \text{equiv}(y, b)\}$$

Also, consider $EQ^{RT/GS}_S$ be the set of triples in $RT_D$ which there is an equivalent triple in the gold standard set $GS_i$ for them (exact matches):

$$EQ^{RT/GS}_S := \{(x, y, z) \in RT_D | \exists (a, b, c) \in GS_i : \text{equiv}(x, a) \land \text{equiv}(y, b) \land \text{equiv}(z, c)\}$$

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

$$m_{\text{semTriple}} = \frac{\sum_{S_i \in SS_D} |EQ^{RT/GS}_S|}{\sum_{S_i \in SS_D} |EQ^{RT}_S|}$$

**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 knowledge graph to a carefully selected dataset as a gold standard. This dataset includes 100 triple 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 need to provide completely accurate references for a topic, which needs human experts.
Fig. 3. 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.” [11]. 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, 8, 13], is one of the common criteria for this dimension. Other common metrics for checking consistency in Linked Data are usage of undefined classes [8, 13], ontology hijacking [8, 13], and OWL inconsistencies [8, 13], the extent of values compliance with the domain/range of data types [4, 13], and misuse of predicates [35]. 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** Consider function \( \text{isRefSpecific} : RP_D \rightarrow \{0, 1\} \) as follows:

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

Then, we define metric \( m_{\text{refPropCon}} \) as below:

\[
m_{\text{refPropCon}} = \frac{\sum_{x \in RP_D} \text{isRefSpecific}(x)}{|RP_D|}
\]

**Discussion.** By this metric, one can assure that the dataset is using 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\(^7\) with properties such as dcterms:provenance and dcterms:source and the W3C PROV-O [36] 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 [15]. 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 3 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 \( TYP_{x_i} \) 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 the datatype of \( x_i \) if \( x_i \) is a literal. Also, \( \forall y_i \in RP_D \) suppose there is function \( \text{range}(y_i) \) that return the range(s) of the given reference predicate \( y_i \). Also consider function \( \text{inRange} : RT_D \rightarrow \{0, 1\} \) defined as below:

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

\(^7\)https://dublincore.org/specifications/dublin-core/dc-md/terms/ - accessed 28 August 2021
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 [13]. In Wikidata, ranges of a property can be fetched from the class (P2308) qualifier of the property constraint (P2302) statements that have value-type constraint (Q21510865). Figure 4 shows the value allowed types for the property stated in (P248).

**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).
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” [11]. Redundancy in both schema and instance levels is covered in the Mendes et al. [7] framework. Debattista et al. [13] considered instance-level redundancy in their investigation over 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 in 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.

We consider redundancy in both schema and instance level. 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 [15]. Figure 5 shows reference sharing in the Wikidata data model.

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

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

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

$$m_{\text{schemaRed}} = 1 - \frac{\sum_{x, y \in \mathcal{RP}_D} \text{arePredsRed}(x, y)}{|\mathcal{RP}_D|}$$

**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 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 $\mathcal{SR}_D$ to be the set of reference nodes that provide provenance for more than one statement:

$$\mathcal{SR}_D := \{ x \in \mathcal{R}_D \mid \exists f_1, f_2 \in \mathcal{S}_D : (f_1, \text{prov:wasDerivedFrom}, x) \in \mathcal{D} \land (f_2, \text{prov:wasDerivedFrom}, x) \in \mathcal{D} \land f_1 \neq f_2 \}$$
We define metric $m_{\text{refSharing}}$ as follows:

$$m_{\text{refSharing}} = \frac{|S_R|}{|R_D|}$$

Discussion. Reference sharing refers to using a set of reference triples for more than one statement in common [15]. Shared references are very usual in Wikidata [15]. 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

This category contains dimensions that illustrate the perceived trustworthiness of the dataset [11]. These dimensions are reputation, believability, verifiability, and objectivity [11]. In knowledge graphs, 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” [11]. Reputation is the social aspect of trust in the semantic web [37], thus, the reputation criteria try to measure users’ opinions about datasets [38, 39]. Investigating users’ opinions can be done explicitly through questionnaires and decentralized voting such as Gil and Artz study [39]. On the other hand, implicit methods like relying on page ranks can be used as a metric for reputation [38, 39]. Golbeck and Hendler [37], proposed an algorithm for computing the reputation of objects considering the incoming links to the object. Our suggestion to measure the reputation of references in a dataset is to average the external URIs page ranks.

**Metric 13. External URIs Reputation** Assume there is the function $\text{srcRanker} : RU_{D}^{\text{ext}} \rightarrow [0, 1]$ such that $\text{srcRanker}(x)$ 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 $m_{\text{srcRank}}$ as follows:

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

Discussion. One of the available methods to determine the rank of web URIs is Google PageRank [40]. However, Google is not providing page rank data anymore. The current benchmarks also belong to late 2016. We consider another metric as a proxy to Metric 13 to approximate the reputation of references by checking that the external URIs are blacklisted. In that case, we define $\text{isSrcBL} : RU_{D}^{\text{ext}} \rightarrow \{0, 1\}$ as below:

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

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

$$m_{\text{srcBL}} = 1 - \frac{\sum_{x \in RU_{D}^{\text{ext}}} \text{isSrcBL}(x)}{|RU_{D}^{\text{ext}}|}$$
D**IMENSION 10. B**ELIEVABILITY

Zaveri et al. define believability as “the degree to which the information is accepted to be correct, true, real and credible” [11]. Believability sometimes is considered as a synonym for trustworthiness [4, 13, 41]. 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 [6].

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 [42, 43]. However, there are different objective metrics to measure believability, e.g., the use of trust ontologies in data [44] and clarifying the provenance of data [4, 13]. In the context of references, we define the metric for believability dimension based on the fact that references are added more by humans or machines.

**Metric 14. Human-added References** We define \( m_{humanRe fs} \) as:

\[
m_{humanRe fs} = \frac{|\{ x \in RD | x \ added \ by \ human \}|}{|RD|}
\]

**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 [45] 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.

D**IMENSION 11. V**ERIFIABILITY

Verifiability is defined as the “degree by which a data consumer can assess the correctness of a dataset” [11]. Verifiability indicates the possibility of verifying the correctness of the data [4]. A dataset is verifiable if there exist concrete means of assessing the correctness of data. Therefore, providing the provenance of facts [4, 13] and use of digital signatures to sign RDF datasets [46] are suggested metrics for this dimension. Subjective methods like using unbiased trusted third-party evaluators are also suggested in the literature [6].

In the context of references, the document type of a reference can be the subject of measurement. We define a metric based on the verifiability of the document type of a reference.

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

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

Then, we define metric \( m_{verif} \) as follows:

\[
m_{verif} = \frac{\sum_{x \in RO_D} typeVerif\ Score(x)}{|RO_D|}
\]
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 trustable scientific, librarian, or political sources (e.g., UniProt\(^8\) in life science). 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 challenging.

**DIMENSION 12. OBJECTIVITY**

Objectivity is defined as “the degree to which the interpretation and usage of data is unbiased, unprejudiced and impartial” [11]. 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 [47]. Bizer [6] 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** 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**

Dimensions of this category monitor the freshness and frequency of data updates [11]. These dimensions, according to Zaveri et al. [11] are currency, volatility, and timeliness. [4, 18] considered dynamicity as the timeliness dimension in the contextual category. Bizer [6] however, considered dynamicity as the timeliness dimension in the intrinsic category. More recently, Ferradji et al. [48] 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 [36] 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\(^9\) is also another way to capture reference modification dates.

**DIMENSION 13. CURRENCY**

According to Zaveri et al. “currency measures how promptly the data is updated” [11]. This dimension is usually measured by computing the distance between the latest time data modified and the observation time [7]. Sometimes the release time of data is also included in the calculation [12]. 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 [11]. For example, the time that Wikidata takes to update a wrestler’s statement for his new Olympic medal is a currency measurement.

\(^8\)https://www.uniprot.org/ - accessed 5 November 2022
\(^9\)https://www.wikidata.org/wiki/Wikidata:History_Query_Service - last edited 4 September 2022
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** \( \forall x \in RT_D \), let \( \text{modifTime}(x) \) be the time of the last modification (or creation if there is no modification after creation), and \( \text{startTime}(x) \) be the origin of time for reference triple \( x \). Also, consider \( t_{\text{now}} \) denotes the observation time. We define metric \( m_{\text{freshTriple}} \) as follows:

\[
m_{\text{freshTriple}} = \frac{\sum_{x \in RT_D} t_{\text{now}} - \text{modifTime}(x)}{t_{\text{now}} - \text{startTime}(x)}
\]

**Discussion.** The origin of time is a point in time from which the metric is measured [12]. 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 it. 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\(^{10}\). This dataset is far more extensive than Wikidata dumps and there is no public endpoint for it.

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

\[
m_{\text{freshExternal}} = \frac{\sum_{x \in RU^{ext}_D} t_{\text{now}} - \text{modifTime}(x)}{t_{\text{now}} - \text{startTime}(x)}
\]

**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 \(<\text{Last-Modified}>\) header to the request time, even when the page was published previously.

**Dimension 14. Volatility**

According to Zaveri et al., “volatility refers to the frequency with which data varies in time” [11]. 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 [11].

\(^{10}\)https://dumps.wikimedia.org/backup-index.html - accessed 22 August 2022
The changefreq attribute of Semantic Sitemap [49] is a suggested metric for volatility [27] (as cited in [11]). Based on the changefreq attribute of the external URIs, we can define a metric for the volatility of external URIs.

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

\[
\text{ssChangeFreq}(x) = \begin{cases} 
1 & \text{value of } <\text{changefreq}>\text{x is always} \\
0.9 & \text{value of } <\text{changefreq}>\text{x is hourly} \\
0.8 & \text{value of } <\text{changefreq}>\text{x is daily} \\
0.6 & \text{value of } <\text{changefreq}>\text{x is weekly} \\
0.4 & \text{value of } <\text{changefreq}>\text{x is monthly} \\
0.1 & \text{value of } <\text{changefreq}>\text{x is yearly} \\
0 & \text{otherwise}
\end{cases}
\]

Then, we define metric \( m_{volat} \) as follows:

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

**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**

“Timeliness measures how up-to-date data is, relative to a specific task” [11]. 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** 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:

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

**Discussion.** Timeliness is the difference between the real-world reference updating frequency (freshness) and the expected reference updating frequency (volatility). The closer the real-world frequency is to the expected frequency, the better score timeliness will have.

**Category V. Contextual**

The contextual category includes dimensions that mostly depend on the context of the task at hand [11]. 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. [11], correctness, amount of data, and relevancy belong to the contextual category.
DIMENSION 16. COMPLETENESS

Completeness indicates the extent to which the dataset covers real-world structures and instances. This is a long dimension that contains several sub-categories in some sources, e.g., Furber et al. [9] and Mendes et al. [7] that considered completeness in schema and data instances. Zaveri et al [11] 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, such that can be called "ontology completeness". (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” [11]. 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 Consider $C_{schema} \subseteq schema(D)$ to be the set classes defined in the schema of $D$, $\forall c_i \in C_{schema}$ let $RPC_{c_i}$ 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 $P_{schema} \subseteq schema(D)$ be the set properties defined in the schema of $D$ and $\forall sp_i \in P_{schema}$ let $RPP_{sp_i}$ 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:

$$m_{classSchemaCom} = \frac{|\{x \in C_D | RPC_{schema} \neq \emptyset\}|}{|C_D|}$$

and metric $m_{propertySchemaCom}$ as following:

$$m_{propertySchemaCom} = \frac{|\{x \in R_D | RPP_{schema} \neq \emptyset\}|}{|R_D|}$$

Discussion. Färber et al. [4] measured schema completeness (in knowledge bases) via 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[11]. The existence of such schemata is a key factor to enhance this metric.

Metric 22. Schema-based Property Completeness of References Consider $P_{schema} \subseteq schema(D)$ be the set properties defined in the schema of $D$ and $\forall sp_i \in P_{schema}$ let $RPP_{sp_i}$ 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 $\langle$referenced statement, reference property$\rangle$ pairs, $H \subseteq (S_D \times RP_D)$ as:

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

Also, we define set $IS \subseteq (P_{schema} \times \bigcup_{sp \in P_{schema}} RPP_{sp})$ as the $\langle$property, reference predicate$\rangle$ pairs in the schema level:

$$IS := \{(sp, r) \mid sp \in P_{schema} \land r \in RPP_{sp}\}$$

11For example, see https://www.wikidata.org/wiki/EntitySchema:E265 - accessed 6 November 2022
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:

\[
\text{comRefPropS}_{sp_i}^{r_j} = \frac{|\{(s, r) \in H \mid \text{predicate}(s) = sp_i \land r = r_j\}|}{|\{(s, r) \in H \mid \text{predicate}(s) = sp_i\}|}
\]

and the metric \(m_{sbRefPropCom}\) as the following average:

\[
m_{sbRefPropCom} = \frac{\sum_{(sp_i, r_j) \in IS} \text{comRefPropS}_{sp_i}^{r_j}}{|IS|}
\]

**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 [50]. 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:

\[
\text{comRefPropS}_{sp_i}^{r_j} = \frac{|\{(s, r) \in H \mid \text{predicate}(s) = sp_i \land r = r_j\}|}{|\{s \in IS \mid \text{predicate}(s) = sp_i\}|}
\]

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

\[
[p_i] := \{s \in IS \mid \text{predicate}(s) = p_i\}
\]

Also consider \(H\) 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 \(I \subseteq (P \times RP)\) as:

\[
I := \{([p_i], r_j) \mid [p_i] \in P \land \exists(s, r) \in H : s \in [p_i]\}
\]

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

\[
\text{comRefProp}_{[p_i]}^{r_j} = \frac{|\{(s, r) \in H \mid s \in [p_i] \land r = r_j\}|}{|\{(s, r) \in H \mid s \in [p_i]\}|}
\]

and the metric \(m_{refPropCom}\) as:

\[
m_{refPropCom} = \frac{\sum_{([p_i], r_j) \in I} \text{comRefProp}_{[p_i]}^{r_j}}{|I|}
\]

**Discussion.** The main difference between this metric and Metric 22 is that Metric 22 computes the completeness of reference-specific properties using the dataset schema, 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 Fährber 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 [51]. 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 a same class have used the same properties [51] 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:

\[
\text{comRefProp}^{\text{ref}}_{r_j} = \frac{|\{(s, r) \in H \mid s \in [p_i] \land r = r_j\}|}{|[p_i]|}
\]

**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 \( m_{\text{comPop}} \) as follows:

\[
m_{\text{comPop}} = \frac{|\{f \in S_D \mid f \in SS_D \land f \text{ has at least one reference}\}|}{|SS_D|}
\]

**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 [11] 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 [52] clarify 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 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**

According to Zaveri et al., “Amount-of-data refers to the quantity and volume of data that is appropriate for a particular task” [11]. 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 [11]. In the context of references, this dimension can include quantitative statistics of references. Beghaeiraveri et al. [15] provided a statistical review over 6 Wikidata subsets that are relevant to this dimension. They investigated the number of reference nodes, the total number of reference triples, distribution of triples per reference node, usage frequency of reference specific properties, and percentage of shared references. For all of these concepts, we formally define a quantitative metric in the Amount-of-data dimension.
Unlike other RQSS metrics, some of these quantitative criteria are not normalized between 0 to 1. However, having quantitative statistics helps users to estimate the coverage of references.

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

\[
m_{\text{refNodesPerFact}} = \frac{|R_D|}{|S_D|}
\]

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

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

\[
m_{\text{refTriplesPerFact}} = \frac{|RT_D|}{|S_D|}
\]

**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 \( m_{\text{refTriplesPerNode}} \) as follows:

\[
m_{\text{refTriplesPerNode}} = 1 - \frac{|R_D|}{|RT_D|}
\]

**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” [15]. By knowing how many triples there are for each reference node on average, we can estimate the detail level of referencing.

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

\[
m_{\text{refLiteralPerTriple}} = \frac{|RL_D|}{|RT_D|}
\]

**Discussion.** The Wikidata data model has three types of reference values: external sources, internal sources, and literals (Figure 6). 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**

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” [11]. In Linked Data, relevancy metrics are checking the existence of meta-information attributes and the extent of using relevant external links and/or relevant \texttt{owl:sameAs} predicates [6]. Farber et al. [4] measured the relevancy of facts in knowledge graphs by looking at whether there is a ranking system on facts in the KG.

Relevancy is one of the main conditions of Wikidata references [23]. According to Wikidata guidelines, references “should point to specific sources that back up the data provided in a statement” [23]. Few efforts are measur-
ing the relevance of references in Wikidata. Judging on the relevance of a reference is highly subjective [4]. Due to the subjective nature of the concept, Piscopo et al. [14] proposed an approach to evaluate the relevance of Wikidata English external sources through microtask crowdsourcing followed up with a machine-learning algorithm. Recently, they developed 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 data is useful to measure 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)** Assume we have function \(\text{isRelevant} : RT_D \rightarrow \{0, 1\}\) as below:

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

Then we define metric \(m_{\text{relTriples}}\) as follows:

\[
\begin{aligned}
m_{\text{relTriples}} &= \frac{\sum_{x \in RT_D} \text{isRelevant}(x)}{|RT_D|} \\
\end{aligned}
\]

**Discussion.** Previous works [2, 14] 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 metrics 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:

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

Then, consider function \(\text{isSharedTripleIrrelevant} : FT \rightarrow \{0, 1\}\) as below:

\[
\text{isSharedTripleIrrelevant}(x) = \begin{cases} 
1 & \text{triple } x. (a, b, c) \text{ is not relevant to the fact } x. f \\
0 & \text{otherwise}
\end{cases}
\]
Then we define metric $m_{rel\text{shared}}$ as follows:

$$m_{rel\text{shared}} = 1 - \frac{\sum_{x \in FT} is\text{sharedTripleIrrelevant}(x)}{|FT|}$$

**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**

Representational dimensions indicate the proper presentation and ease of understanding of data to the user. According to Zaveri et al. [11], 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).

**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” [11]. Literature measures this by keeping URIs short and free of SPARQL parameters [10, 13] and also avoiding the use of RDF reification, containers, and collections [4, 10, 13]. As references are statements about statements, reification is inevitable [4]. However, short URLs in external sources can help machines processing references.

**Metric 31. External Sources URL Length** Assume $\forall x \in RU_{\text{ext}}^D$, function $ASCII\text{len}(x)$ returns the the number of ASCII characters of $x$. Now we define function $URL\text{shortness} : RU_{\text{ext}}^D \to [0, 1]$ as below:

$$URL\text{shortness}(x) = \begin{cases} 
1 & ASCII\text{len}(x) \leq 80 \\
0.75 & 80 < ASCII\text{len}(x) \leq 2083 \\
0.5 & 2083 < ASCII\text{len}(x) \leq 4096 \\
0 & \text{otherwise}
\end{cases}$$

Then, we define metric $m_{url\text{Length}}$ as follows:

$$m_{url\text{Length}} = \frac{\sum_{x \in RU_{\text{ext}}^D} URL\text{shortness}(x)}{|RU_{\text{ext}}^D|}$$

**Discussion.** The Hypertext Transfer Protocol HTTP/1.1 RFC [53] 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\(^\text{12}\)). Old browsers like Microsoft Internet Explorer cannot handle URLs with more than 2083 characters\(^\text{13}\). The best practice for URL


length is less than 80 characters for search engine crawlers. 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 [54]. This method maps non-ASCII characters with a % sign followed by two hexadecimal numbers.

**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” [11]. Representational consistency metrics assess the degree of using existing terms in the context [4] and established terms that already are used in the dataset [13]. In the context of referencing, despite there being no standard vocabulary, there are well-known general ontologies, e.g., Dublin Core Metadata [14] and the W3C PROV-O [36]. In addition, some ontologies use their specific properties for references, e.g., Genealogy [15].

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 [11], we define a metric based on the diversity of properties used in reference triples.

**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 [15]. 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 [15].

**DIMENSION 21. UNDERSTANDABILITY**

Understanding deals with the readability and accessibility of data for humans. According to Zaveri et al. [11], “understandability refers to the ease with which data can be comprehended, without ambiguity, and used by a human information consumer”. 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, 13], the existence of example SPARQL queries for the dataset [27], the existence of a regular expression that expresses the URIs of the dataset [4, 13], the existence of a vocabulary list for the dataset [13], and the getting use of mailing list.

---

15http://gov.genealogy.net/ontology.owl - accessed 20 July 2022
Metric 33. Human-readable Labeling of Reference Properties We define metric $m_{ref\text{HumanLabel}}$ as follows:

$$m_{ref\text{HumanLabel}} = \frac{|\{x \in RPD | \exists z: (x, rdfs:label, z) \in D\}|}{|RP_D|}$$

Discussion. Different predicates are used in Linked Data to express the label of a subject\textsuperscript{16}. 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 $m_{ref\text{HumanComment}}$ as follows:

$$m_{ref\text{HumanComment}} = \frac{|\{x \in RPD | \exists z: (x, schema:description, z) \in D\}|}{|RP_D|}$$

Discussion. Descriptions are effective in removing ambiguity of predicate usage. According to Wikidata\textsuperscript{17}, descriptions have the differentiating role for the entities with the similar labels. Wikidata RDF dump uses schema:description predicate for each description.

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

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

Then, We define metric $m_{\text{handyExt}}$ as follows:

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

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

\textsuperscript{16}For a comprehensive list of labelling predicates see [13, §(U1)]

\textsuperscript{17}https://www.wikidata.org/wiki/Help:Label - accessed 27 July 2022
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 6. We can represent external online-available sources as the set \( \{ x \in R^e_D \mid (x, \text{wdt:P31}/\text{wdt:P279*}, \text{wd:Q7094076}) \in \text{Wikidata} \} \). On the last line, we see 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), automatic investigating online availability is not feasible.

**D I M E N S I O N 2 2 . I N T E R P R E T A B I L I T Y**

According to Zaveri et al., “Interpretability refers to technical aspects of the data, that is, whether information is represented using an appropriate notation and whether it conforms to the technical ability of the consumer” [11]. Interpretable data increases the reusability and facilitates the integration with other datasets [11]. 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 [6, 13], and avoiding the usage of RDF blank nodes [4, 10, 13]. In the context of references, we define a metric based on avoiding blank node usage in references.

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

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

**Discussion.** Blank nodes existing at the populating time, when the dataset expect a reference node or a reference triple which is not available. Serialization errors also can cause this problem. Automatic tools cannot interpret these nodes. Thus in terms of interoperability, having no references is better than having blank nodes. As shown in Figure 1, 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.

**D I M E N S I O N 2 3 . V E R S A T I L I T Y**

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, 13] and multilingualism [4, 13, 55]. 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 Labeling of Reference Properties** We define metric \( m_{\text{ref\_MLLabel}} \) as follows:

\[
m_{\text{ref\_MLLabel}} = \frac{|\{ x \in RP_D \mid \exists z : (x, \text{rdfs:label}, z) \in D \land \text{lang}(z) \neq \text{"en"} \}|}{|RP_D|}
\]

**Metric 38. Multilingual Commenting of Reference Properties** We define metric \( m_{\text{ref\_MLComment}} \) as follows:

\[
m_{\text{ref\_MLComment}} = \frac{|\{ x \in RP_D \mid \exists z : (x, \text{schema:description}, z) \in D \land \text{lang}(z) \neq \text{"en"} \}|}{|RP_D|}
\]

\[\text{online database (Q7094076)}\]
Discussion. Wikidata is one of the multilingual open KGs. 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.

**Metric 39. Multilingual Sources** \(\forall x_i \in RO_D \setminus RL_D\) assume function \(srcLang(x_i)\) returns the ISO 639-1:2002\(^{19}\) language code of the source. We define metric \(m_{\text{refMLSources}}\) as follows:

\[
m_{\text{refMLSources}} = \frac{|\{x \in RO_D \setminus RL_D | langS rc(x) \neq "en"\}|}{|RO_D \setminus RL_D|}
\]

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 function \(srcLang(x_i)\) from Metric 39. Also, consider set \(MS\) to be the set of facts having at least one non-English source as a reference:

\[
MS := \{x \in S_D | \exists c \in RO_D \setminus RL_D : (x, prov: wasDerivedFrom, z) \in D \land (z, b, c) \in D \land langS rc(c) \neq "en"\}
\]

Then, we define metric \(m_{\text{MLFacts}}\) as follows:

\[
m_{\text{MLFacts}} = \frac{|MS|}{|S_D|}
\]

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.

### 3.3. Metrics Targets

Table 2 shows the classification of all defined metrics based on the metric targets and the part of referencing on which the quality review is conducted.

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

The Referencing Quality Scoring System (RQSS) is a data quality assessment methodology\(^{11}\) that aims to measure the referencing quality of the Wikidata and other Wikibase-hosted\(^{20}\) datasets. 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 7 shows these components and (part of) data flow between them. In the following paragraphs, we explain the details of the system.

\(^{19}\)https://www.iso.org/standard/22109.html - accessed 9 August 2022

\(^{20}\)https://wikiba.se/ - accessed 6 November 2022
Table 2

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

<table>
<thead>
<tr>
<th>Target</th>
<th>Metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metadata (schemas, historical metadata, sources metadata)</td>
<td>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</td>
</tr>
<tr>
<td>Source content</td>
<td>Availability of External URIs, Security External URIs, Multiple References Consistency, Verifiable Type of References, Relevance of Reference Triples, Relevance of Shared References, External Sources URL Length, Handy External Sources, Multilingual Sources, Multilingual Referenced Statements</td>
</tr>
</tbody>
</table>

Fig. 7. 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 is based on the Wikidata data model). The Metadata Extractor (component B) independently retrieves information such as Entity-Schema summary and historical data from Wikidata. The extracted data is then given to the Framework Runner (component C), which performs 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).

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. 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.

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 performs 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 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 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.

4.1. 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 [56].

In the current version v1.0.1, all main components of Figure 7 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 established a comparison platform and gain valuable insight into the referencing quality in different topics and also Wikidata as a whole.

5.1. Subsetting Overview

We extract three topical subsets corresponding to three Wikidata WikiProjects: Gene Wiki[57], Music[58], and Ships[15]. 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[24] using WDumper [58]. Our subsetting approach is item-based, i.e., selecting the desired items (Q-IDs) and extracting all statements of those items [16]. For topical subsetting, we use the approach of [15]. For random subsetting, we tweaked the WDumper code to extract items from the dump by Q-IDs [59]. We then deployed a Python script[25] 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. To optimize the subsets size, we ignore metadata irrelevant to referencing, such as item labels, item descriptions, and item qualifiers. The specification files of topical and random subsets can be found in the GitHub repository of the paper [60]. The RDF files for each of the subsets can be found in [61].

---

Table 3

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

Table 4

<table>
<thead>
<tr>
<th></th>
<th>Random 100K #2</th>
<th>Random 500K</th>
<th>Random 1M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random 100K #1</td>
<td>62</td>
<td>372</td>
<td>779</td>
</tr>
<tr>
<td>Random 100K #2</td>
<td>399</td>
<td>802</td>
<td></td>
</tr>
<tr>
<td>Random 500K</td>
<td>3,861</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3 shows for each subset the number of items, statements, references, 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. 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 4 shows the intersection between the random subsets, i.e., the number of overlapped items. Considering the sum-up size of each pair of subsets, the amount of overlap is negligible. However, the uniformity of referencing rate and similarity of missing items percentage in random subsets with variable sizes reveals the need for a deeper look at what is in these subsets. We seek to find the topic coverage of the random subsets, i.e., separating random subset items based on topics similar to Wikidata [3, §(What is in Wikidata)]

Figure 8 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.

---


27Note that the pie chart belongs to December 2019 when Wikidata had about 71 million items.


Fig. 8. Topic coverage of the four random subsets. Note that the colours are consistent across the four charts.

Table 5
The average of RQSS metric scores in each category, the total average, and an example weighted average.

<table>
<thead>
<tr>
<th>Subset</th>
<th>Accessibility</th>
<th>Intrinsic</th>
<th>Trust</th>
<th>Dynamicity</th>
<th>Contextual</th>
<th>Representational</th>
<th>Overall</th>
<th>Weighted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene Wiki</td>
<td>0.5323</td>
<td>0.7901</td>
<td>0.5086</td>
<td>0.0338</td>
<td>0.3211</td>
<td>0.7819</td>
<td>0.5816</td>
<td>0.5349</td>
</tr>
<tr>
<td>Music</td>
<td>0.4606</td>
<td>0.7824</td>
<td>0.3622</td>
<td>0.0758</td>
<td>0.2265</td>
<td>0.8534</td>
<td>0.5569</td>
<td>0.5013</td>
</tr>
<tr>
<td>Ships</td>
<td>0.5592</td>
<td>0.7469</td>
<td>0.3525</td>
<td>0.1239</td>
<td>0.1703</td>
<td>0.8245</td>
<td>0.5406</td>
<td>0.4840</td>
</tr>
<tr>
<td>Random 100K #1</td>
<td>0.5269</td>
<td>0.8918</td>
<td>0.5043</td>
<td>0.1116</td>
<td>0.2960</td>
<td>0.7868</td>
<td>0.5944</td>
<td>0.5476</td>
</tr>
<tr>
<td>Random 100K #2</td>
<td>0.5220</td>
<td>0.8951</td>
<td>0.5027</td>
<td>0.0842</td>
<td>0.2929</td>
<td>0.7871</td>
<td>0.5926</td>
<td>0.5451</td>
</tr>
<tr>
<td>Random 500K</td>
<td>0.5196</td>
<td>0.8849</td>
<td>0.5062</td>
<td>0.1029</td>
<td>0.2936</td>
<td>0.7881</td>
<td>0.5921</td>
<td>0.5451</td>
</tr>
<tr>
<td>Random 1M</td>
<td>0.5170</td>
<td>0.8891</td>
<td>0.5079</td>
<td>0.1116</td>
<td>0.2945</td>
<td>0.7878</td>
<td>0.5930</td>
<td>0.5463</td>
</tr>
</tbody>
</table>

5.2. Subsets Overall Scores

Table 5 shows the overall RQSS metric scores of each subset in different categories, the total average of all scores, and an example of a weighted average. 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 8). 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 the presence of lots of schema definitions 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 [62].

5.2.1. Scores by Dimension

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

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

In the Intrinsic category (accuracy-related metrics), the average score is 0.99 for accuracy, 0.56 for consistency and 0.65 for conciseness. Despite the high accuracy scores, in Metric 6, we observe that the lack of regexes for a few frequently used properties causes many literals not to be checked. The consistency of reference properties is higher than 0.7 in all subsets, and random subsets have better scores than topical subsets. In range consistency, scores vary from 0.2 (Ships) to 0.44 (Gene Wiki), and besides low scores, all subsets suffer from having no specified ranges for reference properties. The reference sharing ratio as the proxy of conciseness varies between 0.3 and 0.7 and is considerably higher in random subsets than topical subsets.

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

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

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

In the Representational category, the average is 0.88 for representational-conciseness, 0.99 for representational-consistency, 0.85 for understandability, 0.99 for interoperability, and 0.59 for versatility. 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 in 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.
Table 6

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

5.2.2. 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 more 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.

We present one possible weighting in the last column of Table 5. Our weighted average aims to make approximated or imprecise metrics less effective and magnify the essential and more meaningful metrics. The weights and the justifications of the importance of metrics are as below:

- Metrics 22 and 23 weights are set to three. It 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 black-listed IPs as the proxy of reputation is not accurate.
- The rest of the metrics are assigned a weight of one.

Our weighted scores are lower than the overall scores. It shows that reference quality weakness in Wikidata subsets appears in metrics that RQSS accurately measures. This can indicate that high scores in the low-weighted metrics (e.g., in the ratio of human-added references and freshness of reference triples) are obtained due to the lack of accurate measuring, and it is likely with having more precise measurement tools, the realistic scores of these metrics will be lower. The weighted average also confirm that Wikidata is weak in completeness and multilingualism of referencing. Note that our weighting is only an example.

5.3. Referencing Quality Analysis

Running RQSS over topical and random subsets, we 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.

5.3.1. Availability, Licensing and Security

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 10 seconds request and 60 seconds 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.30

---

30https://github.com/seyedahbr/RQSSFramework/blob/main/RQSSFramework/Licensing/LicenseExistanceChecking.py

*licensing_keywords* list
Table 7

<table>
<thead>
<tr>
<th>Subset</th>
<th>Reference Properties</th>
<th>Score (Metric 4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene Wiki</td>
<td>855</td>
<td>0.1274</td>
</tr>
<tr>
<td>Music</td>
<td>1,194</td>
<td>0.1122</td>
</tr>
<tr>
<td>Ships</td>
<td>97</td>
<td>0.2886</td>
</tr>
<tr>
<td>Random 100K #1</td>
<td>586</td>
<td>0.0972</td>
</tr>
<tr>
<td>Random 100K #2</td>
<td>607</td>
<td>0.0889</td>
</tr>
<tr>
<td>Random 500K</td>
<td>969</td>
<td>0.0804</td>
</tr>
<tr>
<td>Random 1M</td>
<td>1,159</td>
<td>0.0733</td>
</tr>
</tbody>
</table>

Fig. 9. The distribution of reference properties equivalents (between those with $\geq 1$ equivalents). Red lines are medians, triangles are means, and circles are outliers.

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.

5.3.2. 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 on 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. Figure 9 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.

5.3.3. Accuracy

Syntactic Validity of Reference Triples

RQSS deploys the PyShEx evaluator tool [63] to verify the reification of all referenced statements, reference nodes and reference values. We use a ShEx schema\(^{31}\) that starts from the statement 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 \(\langle\text{reference property, literal}\rangle\) pairs, we matched the literals with the regular expressions obtained from the format as a regular expression (P1793) qualifiers of each

Table 8
RQSS results for reference triple syntax accuracy.

<table>
<thead>
<tr>
<th>Subset</th>
<th>Statement Nodes</th>
<th>Failures</th>
<th>Score (Metric 5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene Wiki</td>
<td>97,062,660</td>
<td>124,783</td>
<td>0.9987</td>
</tr>
<tr>
<td>Music</td>
<td>12,743,480</td>
<td>2,798</td>
<td>0.9997</td>
</tr>
<tr>
<td>Ships</td>
<td>1,116,976</td>
<td>51</td>
<td>0.9999</td>
</tr>
<tr>
<td>Random 100K #1</td>
<td>1,225,313</td>
<td>580</td>
<td>0.9995</td>
</tr>
<tr>
<td>Random 100K #2</td>
<td>1,226,097</td>
<td>624</td>
<td>0.9994</td>
</tr>
<tr>
<td>Random 500K</td>
<td>6,117,915</td>
<td>2,482</td>
<td>0.9995</td>
</tr>
<tr>
<td>Random 1M</td>
<td>12,231,380</td>
<td>4,945</td>
<td>0.9995</td>
</tr>
</tbody>
</table>

Table 9
RQSS results for reference literal syntax accuracy.

<table>
<thead>
<tr>
<th>Subset</th>
<th>Reference Properties</th>
<th>Literals</th>
<th>Regexes</th>
<th>Invalid</th>
<th>Failures</th>
<th>Score (Metric 6)</th>
<th>No Regex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene Wiki</td>
<td>705</td>
<td>4,668,209</td>
<td>684 (0.97)</td>
<td>5</td>
<td>0</td>
<td>1.0</td>
<td>70,751 (2%)</td>
</tr>
<tr>
<td>Music</td>
<td>1,036</td>
<td>704,514</td>
<td>1,049 (1.01)</td>
<td>15</td>
<td>0</td>
<td>1.0</td>
<td>95,533 (13%)</td>
</tr>
<tr>
<td>Ships</td>
<td>69</td>
<td>2,128</td>
<td>63 (0.91)</td>
<td>1</td>
<td>0</td>
<td>1.0</td>
<td>968 (45%)</td>
</tr>
<tr>
<td>Random 100K #1</td>
<td>543</td>
<td>51,004</td>
<td>590 (1.08)</td>
<td>6</td>
<td>0</td>
<td>1.0</td>
<td>5,334 (10%)</td>
</tr>
<tr>
<td>Random 100K #2</td>
<td>569</td>
<td>50,449</td>
<td>589 (1.03)</td>
<td>8</td>
<td>0</td>
<td>1.0</td>
<td>5,212 (10%)</td>
</tr>
<tr>
<td>Random 500K</td>
<td>902</td>
<td>243,147</td>
<td>939 (1.04)</td>
<td>10</td>
<td>0</td>
<td>1.0</td>
<td>15,472 (6%)</td>
</tr>
<tr>
<td>Random 1M</td>
<td>1,082</td>
<td>479,231</td>
<td>1,132 (1.04)</td>
<td>16</td>
<td>0</td>
<td>1.0</td>
<td>27,085 (5%)</td>
</tr>
</tbody>
</table>

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 12 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 belongs to PubMed ID (P698) values.

5.3.4. 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 5.3.10). 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.
Fig. 10. The top three reference properties with the highest percentage of literals in each subset.

Table 10
RQSS results for consistency of reference properties.

<table>
<thead>
<tr>
<th>Subset</th>
<th>Reference Properties</th>
<th>Score (Metric 8)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene Wiki</td>
<td>855</td>
<td>0.7298</td>
</tr>
<tr>
<td>Music</td>
<td>1,194</td>
<td>0.8072</td>
</tr>
<tr>
<td>Ships</td>
<td>97</td>
<td>0.7319</td>
</tr>
<tr>
<td>Random 100K #1</td>
<td>586</td>
<td>0.8788</td>
</tr>
<tr>
<td>Random 100K #2</td>
<td>607</td>
<td>0.8896</td>
</tr>
<tr>
<td>Random 500K</td>
<td>969</td>
<td>0.8627</td>
</tr>
<tr>
<td>Random 1M</td>
<td>1,159</td>
<td>0.8714</td>
</tr>
</tbody>
</table>

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 on 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 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. 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.
Table 11
RQSS results for range consistency of reference triples.

<table>
<thead>
<tr>
<th>Subset</th>
<th>Reference Properties</th>
<th>Reference Values</th>
<th>Ranges</th>
<th>Failures</th>
<th>Score (Metric 9)</th>
<th>No Ranges</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene Wiki</td>
<td>122</td>
<td>14,528,575</td>
<td>462</td>
<td>8,150,998</td>
<td>0.4389</td>
<td>1,571,716 (11%)</td>
</tr>
<tr>
<td>Music</td>
<td>134</td>
<td>1,475,080</td>
<td>689</td>
<td>1,170,486</td>
<td>0.2064</td>
<td>45,740 (3%)</td>
</tr>
<tr>
<td>Ships</td>
<td>20</td>
<td>55,083</td>
<td>140</td>
<td>38,181</td>
<td>0.3068</td>
<td>678 (1%)</td>
</tr>
<tr>
<td>Random 100K #1</td>
<td>33</td>
<td>96,581</td>
<td>154</td>
<td>63,066</td>
<td>0.3470</td>
<td>2,670 (3%)</td>
</tr>
<tr>
<td>Random 100K #2</td>
<td>28</td>
<td>97,352</td>
<td>140</td>
<td>63,034</td>
<td>0.3525</td>
<td>3,263 (3%)</td>
</tr>
<tr>
<td>Random 500K</td>
<td>53</td>
<td>464,968</td>
<td>241</td>
<td>302,038</td>
<td>0.3504</td>
<td>13,032 (3%)</td>
</tr>
<tr>
<td>Random 1M</td>
<td>63</td>
<td>917,746</td>
<td>306</td>
<td>595,109</td>
<td>0.3515</td>
<td>251,50 (3%)</td>
</tr>
</tbody>
</table>

Table 12
RQSS results for reference sharing.

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

5.3.5. Conciseness: Ratio of Reference Sharing

Similar to [15], 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.

5.3.6. Reputation: External URIs

We use pydnsbl\(^3\) to check whether URI domains are among the public black-listed domains on the web. 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.

5.3.7. 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 Wikidata software. Figure 11 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. Note that there is an upper date limit set to 3 January 2022 (the

\(^3\)https://pypi.org/project/pydnsbl/0.5.4/ - accessed 19 August 2022
Table 13
RQSS results for the reputation of external URIs (Pydnsbl).

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

Fig. 11. ‘View History’ tab of **Albert Einstein (Q937)** on 20 September 2022. The second record shows an addition of a reference to a claim.

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.

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

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 is no historical metadata for them. The results of Gene Wiki were not available after three unsuccessful attempts and more than 90 days of processing. The scores vary between random and topical subsets. Due to the presence of active bots in the Gene Wiki WikiProject, such as Pathwaybot33 and ProteinBoxBot34, 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

34https://www.wikidata.org/wiki/User:ProteinBoxBot - accessed 14 November 2022
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 on Wikidata revision history which is not the scope of this paper.

5.3.8. 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\(^\text{35}\). Table 15 shows the disaggregated statistics of sources 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. Note that many of external links can be blog posts, encyclopedic articles, or even scholarly articles, but investigating the content of the external links is subjective. The ‘Unclassified’ column shows the number and percentage of external sources for which RQSS can not classify their type. Music and Ships contains a large number of such external sources, which explains the reason for their low score.

5.3.9. Objectivity: Multiple References for Statements

RQSS counts the number of reference nodes connected to each statement node via $\text{prov:wasDerivedFrom}$ links (Figure 1). 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 12 shows the distribution of references in statements having two or more references. Gene Wiki has the best average, and most of its multiple-referenced statements have between 2 and 4 references. Note that there are statements in Gene Wiki having more than 100 references.

5.3.10. Currency

Freshness of Reference Triples

As mentioned in Section 5.3.7, we do not have access to the historical metadata of a single triple. Using HTML history pages of items (Figure 11) and for all $\langle$item, referenced fact$\rangle$ pairs, RQSS extracts the first creation time of each fact as its $\text{startTime}$, and the latest reference creation or revision of the fact as the $\text{modi}f\text{Time}$. The upper date limit is set to 3 January 2022. The results of fact-reference freshness are shown in Table 17. The results of Gene Wiki were not computable after 90 days of processing. The percentage of missing historical data is similar to Section 5.3.7 (Table 14). The freshness scores, which includes only found referenced facts, are high, and there is not much difference between different subsets.

---

Table 15

RQSS results for type of sources and the verifiability.

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

---

Table 16
RQSS results for having multiple references for statements.

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

Fig. 12. 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.

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.

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

**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 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 have Last-Modified header. However, if we consider using of Last-Modified header as sign of dataset professionalism, Gene Wiki uses the most credible external sources.
Table 18
RQSS results for freshness of external URIs.

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

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

<table>
<thead>
<tr>
<th>Subset</th>
<th>Classes</th>
<th>Fact Properties</th>
<th>Score (Metric 21)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene Wiki</td>
<td>17,184</td>
<td>4,206</td>
<td>0.0004</td>
</tr>
<tr>
<td>Music</td>
<td>1,381</td>
<td>3,506</td>
<td>0.0014</td>
</tr>
<tr>
<td>Ships</td>
<td>1,133</td>
<td>701</td>
<td>0.0008</td>
</tr>
<tr>
<td>Random 100K #1</td>
<td>3,484</td>
<td>4,141</td>
<td>0.0025</td>
</tr>
<tr>
<td>Random 100K #2</td>
<td>3,498</td>
<td>4,191</td>
<td>0.0022</td>
</tr>
<tr>
<td>Random 500K</td>
<td>8,299</td>
<td>5,917</td>
<td>0.0010</td>
</tr>
<tr>
<td>Random 1M</td>
<td>11,908</td>
<td>6,630</td>
<td>0.0007</td>
</tr>
</tbody>
</table>

5.3.11. Volatility and Timeliness
To compute Metric 19, RQSS uses the Ultimate Sitemap Parser\textsuperscript{36} Python package. This package searches the root domain for XML sitemap files and finds the \texttt{<changefreq>} tag of the given URI. However, downloading, decompressing, and searching XML sitemaps is very time-consuming, even for the smallest subset. Thus we are not able to present volatility results. As Metric 20 is the distance between freshness and volatility, timeliness results are also not computed.

5.3.12. Completeness

Class/Property Schema Completeness of References
RQSS deploys PyShEx schema loader to parse Wikidata Entity Schema ShEx-C raw texts and create a summary of the schema-level referenced classes, referenced fact 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 class/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 5.3.12) RQSS extracts all \langle statement, reference property \rangle pairs from subsets and checks each pair over E-ID summaries. There are total of 193 \langle fact property, reference property \rangle 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 \langle statement, reference property \rangle 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 \langle fact property, reference property \rangle (\texttt{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.

\textsuperscript{36}https://pypi.org/project/ultimate-sitemap-parser/ - accessed 25 September 2022
Table 20
RQSS results for schema-based property completeness of references.

<table>
<thead>
<tr>
<th>Subset</th>
<th>⟨statement, reference property⟩ pairs</th>
<th>Non-Referenced Facts</th>
<th>Score (Metric 22)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Without</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Non-Referenced Facts</td>
<td>With</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene Wiki</td>
<td>180,955,497</td>
<td>33,540,964</td>
<td>0.6098</td>
</tr>
<tr>
<td>Music</td>
<td>12,148,520</td>
<td>6,395,340</td>
<td>0.1203</td>
</tr>
<tr>
<td>Ships</td>
<td>490,748</td>
<td>815,686</td>
<td>0.1177</td>
</tr>
<tr>
<td>Random 100K #1</td>
<td>2,754,858</td>
<td>278,790</td>
<td>0.4331</td>
</tr>
<tr>
<td>Random 100K #2</td>
<td>2,722,602</td>
<td>285,545</td>
<td>0.4252</td>
</tr>
<tr>
<td>Random 500K</td>
<td>13,681,074</td>
<td>1,413,017</td>
<td>0.4946</td>
</tr>
<tr>
<td>Random 1M</td>
<td>27,304,697</td>
<td>2,838,831</td>
<td>0.5369</td>
</tr>
</tbody>
</table>

Fig. 13. The distribution of completeness ratios of the 193 schema-level ⟨fact property, reference property⟩ (comRefProp values). Red lines are medians, triangles are means.

The Majority of the current Entity-Schemas belong to Gene Wiki classes. 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 it helps 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 comRefProp values increase and due to a fixed 193 pairs, the total score raises. Figure 13 shows the distribution of all 193 comRefProp values. In all subsets, there are a variety of comRefProp values between 0 and 1. The details of comRefProp values can be found at [62].

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. Table 21 shows the result of property completeness of references. The fourth column shows the number of ⟨statement, reference property⟩ pairs (comRefProp values in Metric 23), which are the averaging factors. Comparing the results with Section 5.3.12, 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 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.

Table 21

RQSS results for property completeness of references.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene Wiki</td>
<td>180,955,497</td>
<td>33,540,964</td>
<td>0.2942</td>
<td>0.1587</td>
</tr>
<tr>
<td>Music</td>
<td>12,148,520</td>
<td>6,395,340</td>
<td>0.2196</td>
<td>0.0975</td>
</tr>
<tr>
<td>Ships</td>
<td>490,748</td>
<td>815,686</td>
<td>0.3243</td>
<td>0.1673</td>
</tr>
<tr>
<td>Random 100K #1</td>
<td>2,754,858</td>
<td>278,790</td>
<td>0.4711</td>
<td>0.3318</td>
</tr>
<tr>
<td>Random 100K #2</td>
<td>2,722,602</td>
<td>285,545</td>
<td>0.4597</td>
<td>0.3214</td>
</tr>
<tr>
<td>Random 500K</td>
<td>13,681,074</td>
<td>1,413,017</td>
<td>0.3945</td>
<td>0.2429</td>
</tr>
<tr>
<td>Random 1M</td>
<td>27,304,697</td>
<td>2,838,831</td>
<td>0.3616</td>
<td>0.2128</td>
</tr>
</tbody>
</table>

Fig. 14. The distribution of completeness ratios (fact property, reference property) \( \text{comRefProp} \) values at instance-level. Red lines are medians, triangles are means. Circles on the Music bar are outliers.

completeness. Figure 14 shows the distribution of averaging factors \( \text{comRefProp} \) values. The distribution shows topical subset \( \text{comRefProp} \) values are less scattered. Detailed statistics of (fact property, reference property) pairs can be found on [1].

5.3.13. 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 15 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 inverses the ratio and subtracts it from one to map the ratio into a number between 0 and 1. Figure 16 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 Amount-Of-Data.

5.3.14. 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.
Fig. 15. 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. 16. The distribution of triples and literals per reference node. Red lines are medians and triangles are means. Outliers are ignored due to readability.

5.3.15. 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 lesser 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 17 shows the top five properties with the highest frequency of use in each subset. The
Table 22
RQSS results for URI length of external sources.

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

Table 23
RQSS results for diversity of reference properties.

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

Fig. 17. Five properties with the highest frequency of use in each subset.

The frequency of property usage in topical subsets is similar to [15] 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 ID* (P698), which is one the most frequent literal accepting properties in the random subsets, is also the fourth most frequent property in general.
Table 24

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

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

Fig. 18. The distribution of the number of labels and comments in reference properties. Red lines are medians, triangles are means, and circles are outliers.

5.3.16. Understandability

Human-readable Labeling/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 18 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 anchor in the middle of 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.
Table 25

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

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 no anchors). Figure 19 shows the share of each handy external source type in the final score. As Figure 19 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.

5.3.17. Interpretability: Usage of Blank Nodes in References

RQSS checks the number of blank nodes amongst reference nodes and reference value nodes (Figure 1). 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 reference nodes, showing that some reference values are shared between reference nodes.
Table 26
RQSS results for blank nodes in referencing reification.

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

Table 27
RQSS results for multilingual labelling and commenting of reference properties.

<table>
<thead>
<tr>
<th>Subset</th>
<th>Reference Properties</th>
<th>Labelling Score (Metric 37)</th>
<th>commenting Score (Metric 38)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene Wiki</td>
<td>835</td>
<td>1.0</td>
<td>0.9988</td>
</tr>
<tr>
<td>Music</td>
<td>1,194</td>
<td>0.9983</td>
<td>0.9958</td>
</tr>
<tr>
<td>Ships</td>
<td>97</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Random 100K #1</td>
<td>586</td>
<td>0.9965</td>
<td>0.9931</td>
</tr>
<tr>
<td>Random 100K #2</td>
<td>607</td>
<td>0.9967</td>
<td>0.9934</td>
</tr>
<tr>
<td>Random 500K</td>
<td>969</td>
<td>0.9979</td>
<td>0.9938</td>
</tr>
<tr>
<td>Random 1M</td>
<td>1,159</td>
<td>0.9974</td>
<td>0.9948</td>
</tr>
</tbody>
</table>

5.3.18. 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 labelling and commenting on reference properties. Compared to Section 5.3.16, the scores of multilingual metadata are lower. However, high scores show that Wikidata is rich in non-English labelling/commenting. Figure 20 shows the distribution of the number of non-English labels and comments in reference properties. However, high scores show that Wikidata is rich in non-English labelling/commenting. Figure 20 shows the distribution of the number of non-English labels and comments in reference properties, which is identical to Figure 18.

**Multilingual Sources**
RQSS retrieve 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 multilinguality 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, signs, 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, the score of Gene Wiki has the lowest score as it widely uses well-known biomedical datasets IDs/sources in references, which are published in English. Figure 21 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 5.3.18 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 uses non-English references.
Fig. 20. 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.

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

Table 29  
RQSS results for multilingual referenced statements.

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

In addition to the statistical analytics and referencing scores, there are other lessons learned in this comprehensive and in-depth study of Wikidata references. 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. We could not establish the Wikibase docker containers due to the lack of root privileges on the server. 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 issues with Wikibase Docker 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 query the 03 January 2022 full dump locally. The first solution is not possible with current subsetting tools. The second solution, however, requires expensive infrastructure.

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 missed 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 studies, 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 [23]. For example, the claim for the power of a nuclear 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.

Despite the limitations discussed above, this research has positive results. The most important achievement of this research is that statistical analysis can identify data quality weaknesses in the context of referencing. 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. RQSS 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 knowledge graphs hosted in Wikibase instances that follow the Wikidata model, e.g., the EU Knowledge Graph [64]. In this research, we devised innovative methods to deal with the limitations. For example, subsetting was used to deal with the large size of Wikidata.

7. Conclusions

In this study, we investigated the referencing quality of a collaborative knowledge graph, 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 evaluated 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-consenseness, 0.99 for representational-consenseness, 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.

Our evaluation had multiple challenges: the large volume of the Wikidata dump and the lack of proper documentation to establish local copies of data, 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, would further strengthen the RQSS framework. In this regard, the next steps in referencing quality assessment are overcoming the challenges of big data and historical metadata, defining reusable frameworks for subjective criteria (a combination of convolutional networks learning and human opinions), and generalizing the RQSS assessment framework for all semantic web knowledge graphs.
Acknowledgement. We appreciate the helpful suggestions and fruitful discussions of the Shape Expressions (ShEx) community 38 and the ELIXIR BioHackathon Europe Subsetting Project [65]: Dan Brickley, Katherine Thornton, Eric Prud’hommeaux, and Andrea Waagmeester.

References


38https://www.w3.org/community/shex/ - accessed 9 October 2022
REFERENCES


[34] 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.


[53] Help:Sources/Items not needing sources - Wikidata. https://www.wikidata.org/wiki/Help:Sources/Items_not_needing_sources#When_the_item_has_a_statement_that_refers_to_an_external_source.


