Measuring What Matters: User Perceptions of Knowledge Graph Quality Dimensions in Cultural Heritage

Maria Angela Pellegrino[1,*,†], Anisa Rula[2,†], Lisa Ehrlinger[3], András Micsik[4], Blerina Spahiu[5] and Lorena Etcheverry[6]

[1]*Università di Salerno, Italy*
[2]*Università di Brescia, Italy*
[3]*Hasso Plattner Institute, University of Potsdam, Germany*
[4]*Institute for Computer Science and Control (SZTAKI), Hungarian Research Network (HUN-REN), Hungary*
[5]*University of Milano-Bicocca, Italy*
[6]*Universidad de la República, Uruguay*

Abstract

Knowledge graph (KG) quality frameworks are typically based on multiple dimensions, such as accuracy, completeness, or timeliness. Although these dimensions are widely used and discussed, their definitions, categorizations, and, specifically, their weighting during aggregation remain largely unspecified. Despite existing research that defines and proposes data quality (DQ) dimensions, and even ISO standards, their application in practice is often left to expert judgment without empirical grounding. This paper investigates how practitioners perceive the importance of KG quality dimensions and whether these perceptions vary across downstream tasks. We conducted a structured survey. 15 participants from the cultural heritage domain rated the importance of 19 quality dimensions in two scenarios: as data publishers who are responsible to make a KG reusable, and as data consumers who use a KG for research or decision-making. Our results show that accuracy, availability, and verifiability consistently rank as top priorities, while licensing, interoperability, understandability, and performance show high context dependency. A total of 73.3% of participants confirmed that their quality prioritisation changes when the context shifts to a specific task, in line with the DQ *fitness for use* principle. These outcomes provide an empirical basis to understand practitioner-grounded quality priorities, having direct implications for deriving context-sensitive weights for composite KG quality indicators.

Keywords

Knowledge Graph Quality, Data Quality Dimensions, Quality Weights, User Study, Best-Worst Scaling, Scenario-Based Evaluation, Cultural Heritage, Linked Data

1. Introduction

Knowledge graph (KG) quality assessment has reached a level of methodological maturity, with established taxonomies [1], reusable toolchains [2, 3, 4, 5, 6], and standard vocabularies [7, 8] for representing and computing quality metrics. Yet, practical guidance on which data quality (DQ) dimensions actually matter in which context (and to what extent) remains largely unexplored. Many taxonomies, classifications, and definitions coexist without a shared understanding of how dimensions should be prioritised in practice [9, 10]. These frameworks typically support multi-dimensional quality assessment by allowing users to specify weights over quality dimensions when computing composite quality indicators. However, the question of how those weights should be selected and justified remains largely unaddressed. In practice, it is often unclear which dimensions are most relevant to a given use case, and weights for their aggregation are either assumed equal or delegated to expert judgment, with no empirical grounding.

This gap matters because DQ is not a fixed property of a dataset. The *fitness for use* perspective, established in the information systems literature [11], holds that DQ dimensions cannot be evaluated

QKG@ESWC2026: Workshop on Quality of Knowledge Graphs at ESWC 2026, May 10–11, 2026, Dubrovnik , Croatia
*Corresponding author.
†These authors contributed equally.
mapellegrino@unisa.it (M. A. Pellegrino); anisa.rula@unibs.it (A. Rula); Lisa.Ehrlinger@hpi.de (L. Ehrlinger); micsik@sztaki.hu (A. Micsik); blerina.spahiu@unimib.it (B. Spahiu); lorenae@fing.edu.uy (L. Etcheverry)

independently of the task at hand. Building on this foundation, Mohammed et al. [10] argue that DQ assessment depends not only on the task but on five interacting facets: the data itself, its source, the system processing it, the task, and the human involved. This broader view reflects that quality is shaped by the context in which data is produced and consumed. Applied to KGs, this implies that a data publisher and a data consumer may legitimately assign different importance to the same DQ dimension and that weighting schemes derived without reference to user context may systematically misrepresent quality requirements in practice.

This paper is guided by two research questions (RQs):

RQ1 Which KG quality dimensions do practitioners consider essential?

RQ2 To which extent does the importance of DQ dimensions vary across task and roles?

This paper addresses these questions through a structured user study. We designed and administered a survey that allows us to obtain, from domain practitioners, a context-dependent ranking of priorities regarding the quality dimensions of KGs. We used a scenario-based evaluation and best-worst scaling to determine, on an empirical basis, the importance of DQ dimensions. This is a first approach to the problem of establishing predefined weights. The study is conducted in the cultural heritage domain, where KGs integrate heterogeneous institutional sources and serve both expert and public audiences, making the quality trade-off particularly visible and consequential.
The contributions of this paper are:

- A reusable and general purpose survey methodology for eliciting task and role-specific quality dimension priorities from KG practitioners, combining categorical importance ratings with intra-category best-worst scaling.
- DQ dimension priorities informing a non-uniform weighting system are empirically identified for 19 KG quality dimensions under two contrasting task scenarios — data publication and data consumption — derived from 15 participants in the cultural heritage domain.
- Evidence that the importance of quality dimensions is task-dependent: 73.3% of participants confirmed that their prioritisation changes according to the task context, with measurable shifts observed for licensing, interoperability, understandability, and performance across scenarios.

The remainder of the paper is structured as follows. Section 2 reviews related work on DQ assessment in general, KG quality assessment in particular, and user-centred evaluation. Section 3 outlines the design of the survey, while Section 4 reports the empirical results. Section 5 discusses findings and acknowledges limitations. Section 6 concludes the paper and identifies directions for future work.

2. Related Work

We position our work at the intersection of (i) general theories of DQ dimensions, (ii) DQ assessment and DQ dimensions for KGs, and (iii) user-centered evaluation approaches that conceptualize quality as experienced in context. Accordingly, the rest of this section is structured around these three areas.

2.1. General Theories of Data Quality

Data quality is broadly defined as *fitness for use* [11], emphasising that DQ assessment depends on the context in which data is used. Early research framed DQ as a multi-dimensional concept, where dimensions — also referred to as characteristics or attributes — capture distinct aspects of quality such as accuracy, completeness, or timeliness [11]. Wang and Strong [11] derived their influential four-category taxonomy (intrinsic, contextual, representational, and accessibility-related dimensions) empirically, by surveying data consumers about which dimensions they considered relevant. Their approach closely mirrors the methodology of the present study, applied here to the KG domain.

Despite decades of research, no consensus on a standard set of DQ dimensions has emerged. Many taxonomies and definitions coexist [9, 12], and even ISO standards propose different sets of dimensions depending on the perspective: ISO 8000 [13] targets enterprise data exchange, ISO 25012 [14] defines DQ characteristics from a software quality perspective, and ISO 5259 [15] addresses AI-specific needs. DQ dimensions can be quantified through DQ metrics — numerical functions that map a dimension to a measurable value, with ISO 25024 [16] providing a corresponding catalogue. However, none of these standards specify how dimensions should be prioritised or weighted in practice.

The *fitness for use* principle implies that quality cannot be assessed independently of user goals and task context [17], and recent work further highlights the need for context-aware solutions across all stages of DQ management [18]. This provides the theoretical basis for our study: KG quality is a multi-dimensional construct whose relative importance depends on task and role, and — as the lack of standardisation suggests — on the perceptions of those who use the data.

2.2. Knowledge Graph Quality

Quality assessment of KGs and linked data has been extensively studied through surveys, taxonomies, and metric catalogues. The foundational work by Zaveri et al. [1] systematises linked DQ dimensions and organises them into four categories — intrinsic, contextual, representational, and accessibility-related — following the classification by Wang and Strong [11]. This work established a shared conceptual vocabulary and enabled comparability across KG quality assessment approaches.

Building on this taxonomy, several toolchains operationalise quality assessment by computing and reporting metrics. Luzzu [2] provides an extensible architecture for linked DQ assessment, supporting reusable metric definitions and machine-readable quality metadata. SemQuire [6] builds on DQV to assess the quality of linked open data sources through a structured, vocabulary-aligned evaluation process. KGHeartBeat [3] focuses on continuous monitoring of KG availability and technical health, evaluating aspects such as endpoint accessibility and dereferenceability. Pizhuk et al. [5] extend this perspective to domain-specific settings, proposing a DQ dashboard tailored to security KGs. LOD Laundromat [4] offers a complementary perspective by automating the crawling, cleaning, and republishing of linked data at Web scale. At the standards level, the W3C Data Quality Vocabulary (DQV) [7] and the Data Quality Dimension vocabulary (DQD) [8] provide lightweight schemas to represent DQ metadata, including DQ dimensions, metrics, and measurements and can be used to encode DQ assessment results.

While these frameworks provide strong technical foundations, they typically assume that relevant dimensions are predefined and that aggregation can rely on fixed or user-specified weights. How such weights should be selected and justified remains largely underexplored, and weighting is often treated as a configuration parameter rather than an empirically grounded construct.

2.3. User-Centered KG Evaluation

While the frameworks discussed above provide important conceptual and technical foundations, there is increasing recognition that quality assessment should reflect how users experience quality when performing data-related tasks. In HCI and user experience research, scenario-based and task-based evaluation methods are routinely used to elicit judgements under realistic contexts of use [19, 20, 21]. Applied to KGs, these methods motivate studies that examine how users interpret dimensions such as completeness, provenance, or timeliness depending on their goals. Prior user-involved approaches have primarily focused on crowdsourcing data validation or error detection tasks [22, 23, 24, 25].

Despite the maturity of KG quality frameworks and tooling, deriving and justifying weights for quality dimensions remains largely underexplored. This concerns how aggregation choices should be justified, how importance changes with task context, and how weighting decisions can reflect user priorities rather than expert assumptions.

In contrast to prior work, our study does not ask participants to assess the correctness of individual data items, but to articulate and prioritise the quality dimensions themselves. Rather than adopting

predefined weights, we elicit practitioners' relative prioritisation through contextualised scenarios, yielding empirically grounded weighting structures that can inform composite KG quality indicators.

3. Study Design for DQ Dimension Importance Assessment

In this section, we introduce the study design for assessing how practitioners rate the importance of KG quality dimensions across different usage contexts. To address RQ1 and RQ2, the study treats quality not as a fixed technical property but as a multi-dimensional construct whose relative importance depends on task and role. The instrument combines conceptual alignment, scenario-based evaluation, and comparative prioritisation to derive empirically grounded importance ratings for KG quality dimensions.

3.1. Study Design Pipeline

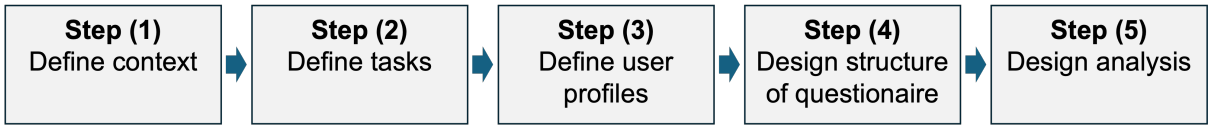

Figure 1: Study design pipeline.

Our experimental setting is organised as a five-step pipeline illustrated in Figure 1.

Step 1: Context. Since quality judgements are meaningful only according to the domain of use, we chose to frame the study in a particular domain, specifically the cultural heritage domain, where KGs integrate heterogeneous institutional sources and support both publication and consumption tasks.

Step 2: Tasks. We operationalise KG quality using a selected subset of dimensions from the Zaveri et al. [1] taxonomy, grouped into four categories. We acknowledge that our categorisation does not fully align with the original classification proposed by Zaveri et al. This is due to a reinterpretation of certain dimensions and a restructuring of categories to better match the goals of our study. We provide participants with concise definitions adapted to the study context. Before performing the main tasks (rating quality dimensions), participants assess the relevance and the clarity of these definitions. This is done to ensure that the subsequent prioritisation judgements are less affected by definitions' ambiguity. The main elicitation tasks then ask participants to express the importance of each dimension under different usage scenarios.

Step 3: User profiles. Since the aim is to elicit practitioner-grounded quality priorities, participants should be familiar with the domain and, ideally, with data or semantic web practices related to the scenarios under this study.

Step 4: Survey structure. The questionnaire is organised into blocks that capture: (i) participant background, (ii) conceptual alignment on the meaning of the selected quality dimensions, and (iii) scenario-based prioritisation tasks. For the latter, participants evaluate the same set of dimensions in two scenarios; data publication and data consumption. The evaluation was performed using a three-level categorical scale (irrelevant, worth having, and must have).

Step 5: Analysis and weight derivation. This study specifically analyses the distribution of the ratings (to indicate agreement), best-worst selection counts, and per-dimension dominance scores. Categorical importance ratings and intra-category comparative selections are then combined to derive scenario-specific profiles, which serve as empirically grounded weighting structures for KG quality dimensions.

3.2. Context Definition

The study is grounded in the cultural heritage domain, which is particularly suitable to answer our research questions, because of the wide heterogeneity of institutions, different professional roles, and

downstream use cases. The basis for our DQ dimension selection is the well-established taxonomy by Zaveri et al. [1], which we adopt as-is to ensure consistency with prior work and to enable future comparability across studies. In particular, the contextual cluster has been refined to group dimensions that are functionally related within our evaluation setting. We select a subset of dimensions based on their integration in available KG quality assessment frameworks, as KGHeartBeat. Table 1 lists all DQ dimensions included in our survey, their definitions as presented to the participants, and the quality category to which they belong within the taxonomy [1].

Table 1
KG quality dimensions included in the user survey along with their definitions as presented to participants, grouped by category according to [1].

Dimension	Definition
Accessibility	
Availability	The extent to which data (or some portion of it) is present, obtainable, and ready for use.
Interlinking	The degree to which entities that represent the same concept are linked to each other, be it within or between two or more data sources.
Licensing	The granting of permission for a consumer to re-use a dataset under defined conditions.
Performance	The efficiency and responsiveness of systems that deliver or process cultural heritage data (e.g., query response time).
Security	The extent to which data is protected against unauthorized access.
Intrinsic	
Accuracy	The degree to which the information about cultural heritage objects or resources is syntactically accurate and correctly reflects reality.
Completeness	The degree to which the information about cultural heritage objects or resources are fully documented.
Consistency	The extent to which cultural heritage data are uniform and free of contradictions.
Conciseness	The degree to which cultural heritage descriptions contain only necessary and relevant information, avoiding redundancy or unnecessary verbosity.
Representational	
Interoperability	The ability of cultural heritage data to be integrated, exchanged, and used across different systems, platforms, and institutions by using shared standards, formats, and vocabularies.
Interpretability	The degree to which cultural heritage data is represented using an appropriate notation and whether the machine is able to process the data.
Representational-Conciseness	The representation of the data, which is compact and well formatted on the one hand and clear and complete on the other hand.
Understandability	The ease with which data can be comprehended without ambiguity and be used by a human information consumer.
Versatility	The availability of the data in different representations and in an internationalized way.
Contextual	
Amount of Data	The volume or quantity of cultural heritage records or digital objects available.
Believability	The degree to which cultural heritage data are perceived as credible and trustworthy by users, based on source authority and identity of the information provider.
Reputation	The perceived trustworthiness and standing of the institution, project, or source that provides the cultural heritage data.
Timeliness	The degree to which cultural heritage data are made available within a time frame that makes them useful for a given purpose.
Verifiability	The extent to which claims about cultural heritage objects or data can be independently checked through references to external vocabularies, consulting publishers, or via digital signatures.

3.3. Task Definition: Scenario-Based Elicitation of Contextual Priorities

Each scenario in this study is defined by a combination of a *role* and a *task*. The *role* refers to the professional perspective adopted by the participant (data publisher or data consumer), and the *task* refers to the activity performed within that role (preparing a KG for external reuse, or consuming a KG for research or decision-making). Participants were asked to rate each dimension definition on *clarity* and *domain relevance*. This preliminary step serves two methodological purposes. First, it reduces the risk that later prioritisation judgements are based on definitional ambiguity, rather than genuine differences in importance. Second, it helps identify dimensions whose meanings may be overlapping or unclear, such as *reputation* and *believability*, or *interpretability* and *understandability*.

To illustrate how quality priorities might change based on user perspective, the survey introduces participants to two specific scenarios. In the first scenario (named scenario A), participants take on the role of a data publisher responsible for preparing a cultural heritage KG for external users. In the second scenario (named scenario B), participants take on the role of a data consumer who relies on the knowledge graph for research or decision-making. Each scenario describes a realistic professional task involving a cultural heritage KG. These tasks are easily generalisable to other domains. Each participant selected the scenarios of interest and then rated the importance of each quality dimension in each scenario. The following is the text included in the survey to describe these scenarios.

Scenario A — Data Publication.

> *You are a "digital heritage data steward" responsible for preparing and publishing a collection of cultural heritage data for public reuse.*

Scenario B — Data Consumption.

> *You are an end-user consuming cultural heritage knowledge graphs.*

The set of dimensions remains fixed across scenarios, varying the role and objective of the task. This approach lets us attribute differences in prioritisation to context rather than to participant background. Within each scenario, participants rated each dimension on a three-level categorical scale: *irrelevant*, *worth having*, and *must have*. This structure adapted the MoSCoW prioritisation framework [26], from which the could have level was excluded to reduce response ambiguity and sharpen the distinction between optional and essential dimensions. This framework is well-established in requirements engineering for eliciting ordinal importance distinctions without imposing artificial numerical precision.

3.4. User profiling

The aim of the study is to elicit practitioner-grounded prioritisation of KG quality dimensions. For this reason, the target participants were individuals with familiarity with the cultural heritage domain and, ideally, with data-related or semantic web practices relevant to KG publication or consumption.

The participants were recruited through the GOBLIN[1] COST Action network and represent six professional domains: museums and collections, archives and libraries, heritage authorities, universities and research institutions, digital humanities, and industry/technology providers. Three participants indicated domains outside this classification (web information extraction, linguistics). A more detailed description of the participants is reported in Section 4.

3.5. Survey Structure

The questionnaire was organised into four main blocks.

The first block collected **participant background information**, including professional domain, level of experience, familiarity with the cultural heritage domain, familiarity with DQ dimensions, and

[1]Global Network on Large-Scale, Cross-domain and Multilingual Open Knowledge Graphs (GOBLIN): https://goblin-cost.eu

familiarity with semantic web technologies. These variables support interpretation of the prioritisation results and provide context for the participant sample.

The second block addressed **conceptual alignment**. In this part of the survey, participants rated the selected quality dimension definitions in terms of clarity and relevance to the cultural heritage domain. This step was included to ensure that later prioritisation judgements were not influenced by uncertainty about the meaning of the dimensions.

The third block contained the **scenario-based prioritisation tasks**. Participants were asked to evaluate the importance of the same set of quality dimensions under the two different scenarios, using a three-level categorical scale. This block captures how dimension importance changes with task context.

The fourth block contained the comparative and **reflective questions**. In addition to the categorical ratings, participants completed a forced best-worst selection task within predefined groups of dimensions. The questionnaire also included final feedback questions on whether quality priorities change with task context and on the perceived generalisability of the instrument to other domains.

3.6. Analysis and Intra-Group Comparative Prioritisation

Participants rated each definition on two axes: relevance to the cultural heritage domain and clarity of formulation. This preliminary step serves two methodological purposes. First, it reduces semantic ambiguity by ensuring a consistent interpretation of each construct before prioritisation judgments are made. Second, it identifies dimensions whose definitions are viewed as overlapping or unclear—like interpretability and understandability, or believability and reputation—which might introduce noise into later comparisons.

We assessed conceptual alignment by examining relevance and clarity ratings for each dimension. Dimensions with mostly low clarity ratings were flagged for further inspection, and patterns of confusion are described in Section 4.2. By explicitly separating agreement from importance, observed prioritisation differences reflect true task-related trade-offs instead of definitional misunderstandings.

Categorical ratings are prone to acquiescence bias, where respondents assign uniformly high scores across all dimensions. To address this, the survey includes a best-worst selection task within predefined clusters of related dimensions, according to DQ categories.

Cluster structure by data quality categories. Dimensions are grouped into four category clusters following the taxonomy by Zaveri et al. [1]: (i) *accessibility*: availability, licensing, interlinking, security, performance; (ii) *intrinsic*: accuracy, consistency, conciseness, completeness; (iii) *representational*: representational-conciseness, interoperability, interpretability, versatility, understandability, (iv) *contextual*: amount of data, timeliness, reputation, believability, verifiability,

Best-worst selection. Within each category and for each scenario, participants select the *most* and *least* relevant dimension. This best-worst scaling approach [27] forces explicit trade-off among conceptually related dimensions, exposing relative dominance relationships that categorical ratings cannot reveal.

Weight derivation. A per-dimension dominance score is computed as the difference between most-relevant and least-relevant selection counts, following standard best-worst scaling conventions [27]. Combining importance tiers from categorical ratings with dominance scores from cluster selection yields a scenario-specific prioritisation profile for each dimension. These profiles form the basis for the empirical weighting structures discussed in Section 5.

A known limitation of the cluster-based design is that it does not capture trade-offs across DQ categories, resembling clusters. A participant who values *accuracy* (intrinsic) and *verifiability* (contextual) equally has no mechanism to express this cross-cluster equivalence. Categorical ratings partially compensate for this, but cross-cluster relative dominance remains an open issue for future work.

4. Results

This section reports the results from 15 completed responses, structured around three aspects: participant background (Section 4.1), shared understanding of dimension definitions (Section 4.2), and scenario-dependent prioritisation patterns (Sections 4.3–4.5). Participation in each scenario was optional; 9 participants completed scenario A and 6 completed scenario B. All 15 participants contributed to the best-worst scaling tasks, which were not scenario-specific.

4.1. Participant Background

The first part of the user study gathers background information about the participants. Table 2 summarises self-reported familiarity across three axes on a 1–5 Likert scale. Familiarity with semantic web technologies is notably skewed towards high values (47% at level 5), reflecting the technical profile of the GOBLIN network. Familiarity with the cultural heritage domain and with DQ dimensions is more evenly distributed, indicating that the sample spans both domain practitioners and technical specialists. Experience levels spanned from early to senior career: fewer than 5 years (20%), 5–10 years (27%), 11–20 years (33%), and more than 20 years (20%). Nine participants (60%) identified primarily as data providers; six (40%) as data consumers.

Table 2
Participant familiarity profiles ($n = 15$).

Axis	1	2	3	4	5
Cultural heritage domain	20%	13%	33%	27%	7%
DQ dimensions	13%	27%	27%	13%	20%
Semantic web technologies	6%	7%	20%	20%	47%

4.2. Shared Understanding of Quality Dimensions

Before the scenario-based tasks, participants rated each dimension definition on a 1–5 agreement scale. Table 3 reports the percentage of responses at levels 4–5 (i.e., high agreement) per dimension and flags dimensions that generated substantive qualitative feedback. Most dimensions achieved high agreement. *Accuracy* (73.33% at level 5, no responses below 3), *licensing* (93.33% at levels 4–5), and *interoperability* (93.33% at levels 4–5) showed the strongest consensus. *Consistency* (86.67% at levels 4–5) and *representational-conciseness* (86.67% at levels 4–5) were also well received, though participants noted partial overlap between the two constructs. Three dimensions generated notable disagreement or qualitative concern:

- *Timeliness* showed the widest rating spread (13.33% at level 1, 40% at level 3, 27% at level 5) and the highest volume of comments from participants questioning its applicability to cultural heritage, where data are often historically stable rather than time-sensitive.
- *Security* exhibited a bimodal distribution (20.0% at level 1, 53.33% at level 5), reflecting a division between participants who questioned its relevance to public datasets and those who considered it essential.
- *Reputation* and *believability* were perceived as conceptually overlapping by multiple participants, with several comments suggesting that the two constructs are difficult to distinguish operationally.

Note that disagreement indicates ambiguity in how these three dimensions are perceived. Consequently, the ratings in the subsequent sections should also be interpreted with caution.

4.3. Scenario A — Data Publication

Figure 2 reports the importance ratings per dimension for the data publication scenario. Figure 3 reports best-worst selection counts per category.

Table 3
Scoring of dimension definitions on a 1–5 agreement scale, reporting the percentage of responses at levels 4–5 (i.e., high agreement) per dimension.

Dimension	% responses at levels 4–5
Accessibility	
Licensing	93.33
Availability	86.67
Interlinking	73.33
Performance	66.67
Security	66.67
Intrinsic	
Consistency	86.67
Conciseness	80.00
Accuracy	73.33
Completeness	73.33
Representational	
Interoperability	93.33
Interpretability	86.67
Understandability	80.00
Representational-Conciseness	86.67
Versatility	66.67
Contextual	
Amount of Data	80.00
Verifiability	80.00
Reputation	73.33
Believability	66.67
Timeliness	40.00

Categorical ratings. *Accuracy* received the highest concentration of "must have" ratings, with no participant marking it as irrelevant. *Availability* and *licensing* followed, confirming that data correctness, access, and legal conditions for reuse are perceived as foundational in a publication context. *Interlinking* and *security* showed greater variability, with a non-trivial share of "worth having" responses.

Best-worst selection. Results per category are summarised in Table 4. Within the accessibility-related cluster, *availability* dominates while *performance* is least critical, suggesting that ensuring data is accessible outweighs system responsiveness concerns at publication time. The selection of *completeness* as least relevant within the intrinsic cluster is noteworthy: it may reflect the practical difficulty of guaranteeing completeness in heterogeneous heritage collections, or a perception that completeness is aspirational rather than enforceable at publication. Within the contextual cluster, *verifiability* dominates over *reputation* and *believability*, suggesting that traceable, independently checkable claims are more operationally valued than perceived source authority.

Table 4
Best-worst selection results, scenario A ($n = 9$).

Cluster/Quality Category	Most relevant	Least relevant
Accessibility	Availability	Performance
Intrinsic	Accuracy	Completeness
Representational	Understandability	Versatility
Contextual	Verifiability	Amount of data

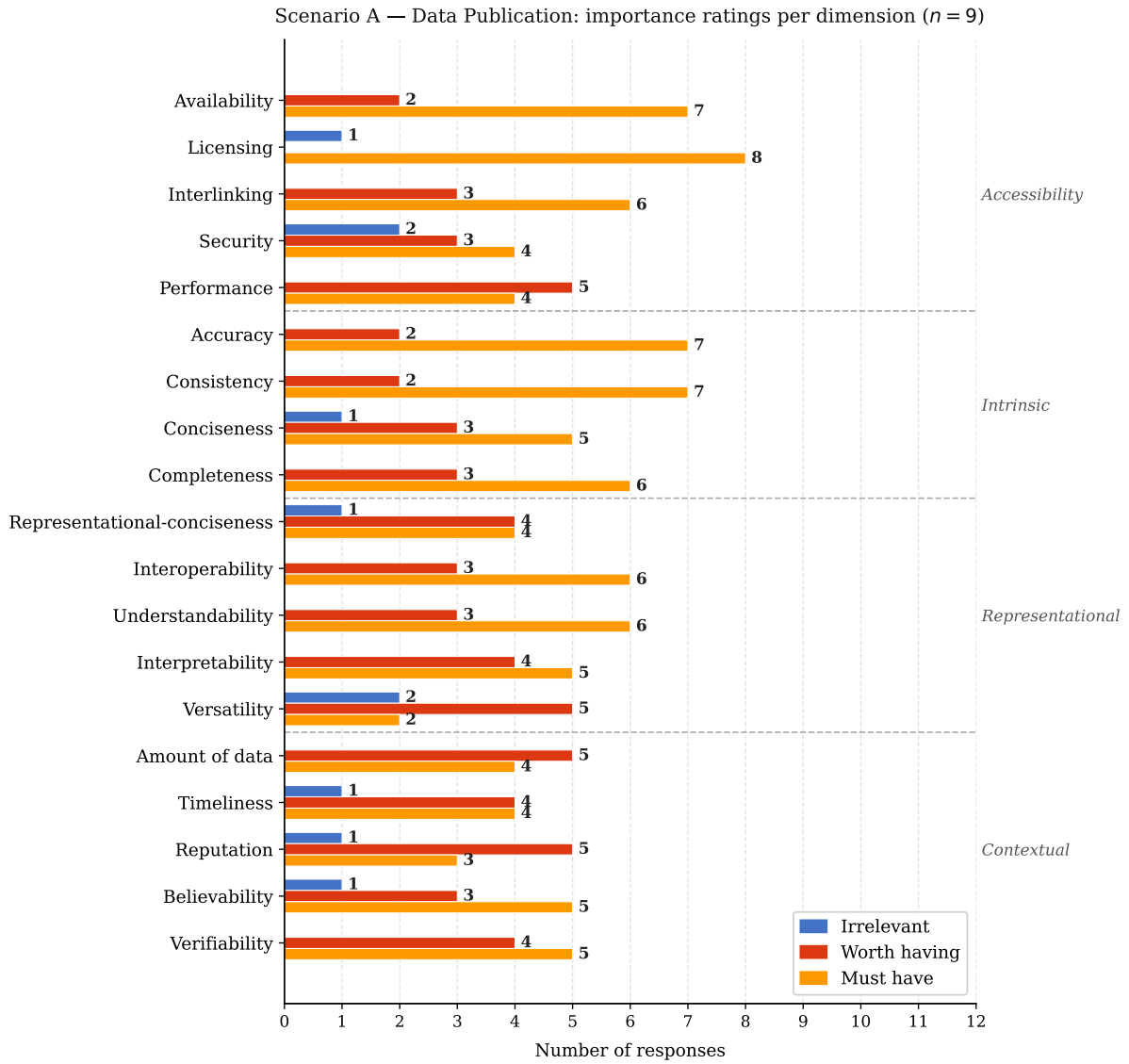

Figure 2: Categorical importance ratings per quality dimension, scenario A (data publication), $n = 9$. Dimensions are grouped by quality family (Zaveri et al. [1]).

4.4. Scenario B — Data Consumption

Figure 4 and Figure 5 report categorical ratings and best-worst selections for scenario B.

Categorical ratings. *Accuracy* and *availability* achieved high "must have" ratings, which confirms cross-scenario stability. *Licensing* shifted towards "worth having", which is consistent with legal compliance being perceived as a publisher's responsibility rather than a consumer concern. *Interlinking* received a more favourable profile than in scenario A, reflecting its importance for users to navigate through data across sources.

Best-worst selection. Results per category are summarised in Table 5. The most pronounced cross-scenario shift occurs in the representational cluster: *understandability* (most relevant in scenario A) is replaced by *interoperability* (most relevant in scenario B), showing that consumers prioritise integration over presentation clarity. Correspondingly, *representational-conciseness* becomes least relevant in scenario B, where it had been higher rated by publishers. Within the intrinsic cluster, *conciseness* is ranked least relevant in contrast to *completeness* in the publication scenario. This suggests that

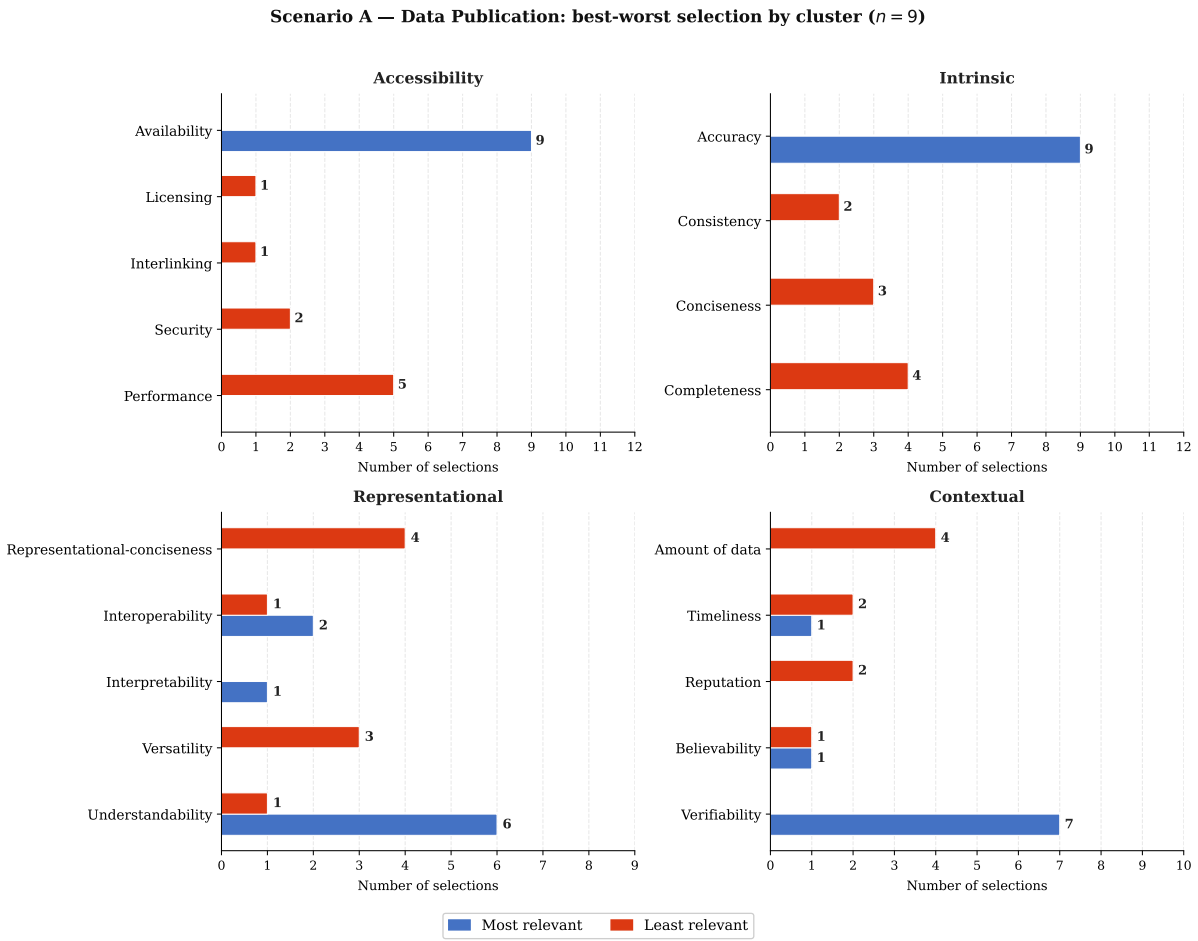

Figure 3: Best-worst selection counts per dimension category, scenario A (data publication), $n = 9$.

Table 5
Best-worst selection results, scenario B ($n = 6$).

Cluster/Quality category	Most relevant	Least relevant
Accessibility	Availability	Security
Intrinsic	Accuracy	Conciseness
Representational	Interoperability	Representational-conciseness
Contextual	Verifiability	Timeliness

consumers accept redundancy as long as the data is correct.

4.5. Cross-Scenario Comparison

Table 6 summarises stable and context-sensitive DQ dimensions across both scenarios. Three dimensions are stable at high priority in both scenarios: *accuracy*, *availability*, and *verifiability*. One dimension, *versatility*, is consistently low in both scenarios. The remaining dimensions are context-sensitive. The most pronounced shifts occur for *licensing* (high priority for publishers, lower for consumers), *understandability* (prioritised by publishers over consumers), and *interoperability* (prioritised by consumers over publishers). When asked whether they prioritise quality dimensions differently depending on the task, 73.3% of participants answered *yes*, confirming that quality importance is task-dependent. Additionally, 80% of the participants rated the generalisability of the instrument to other domains (e.g., life sciences, linguistics) at level 4 or 5 out of 5, which indicates that the approach transfers beyond cultural heritage.

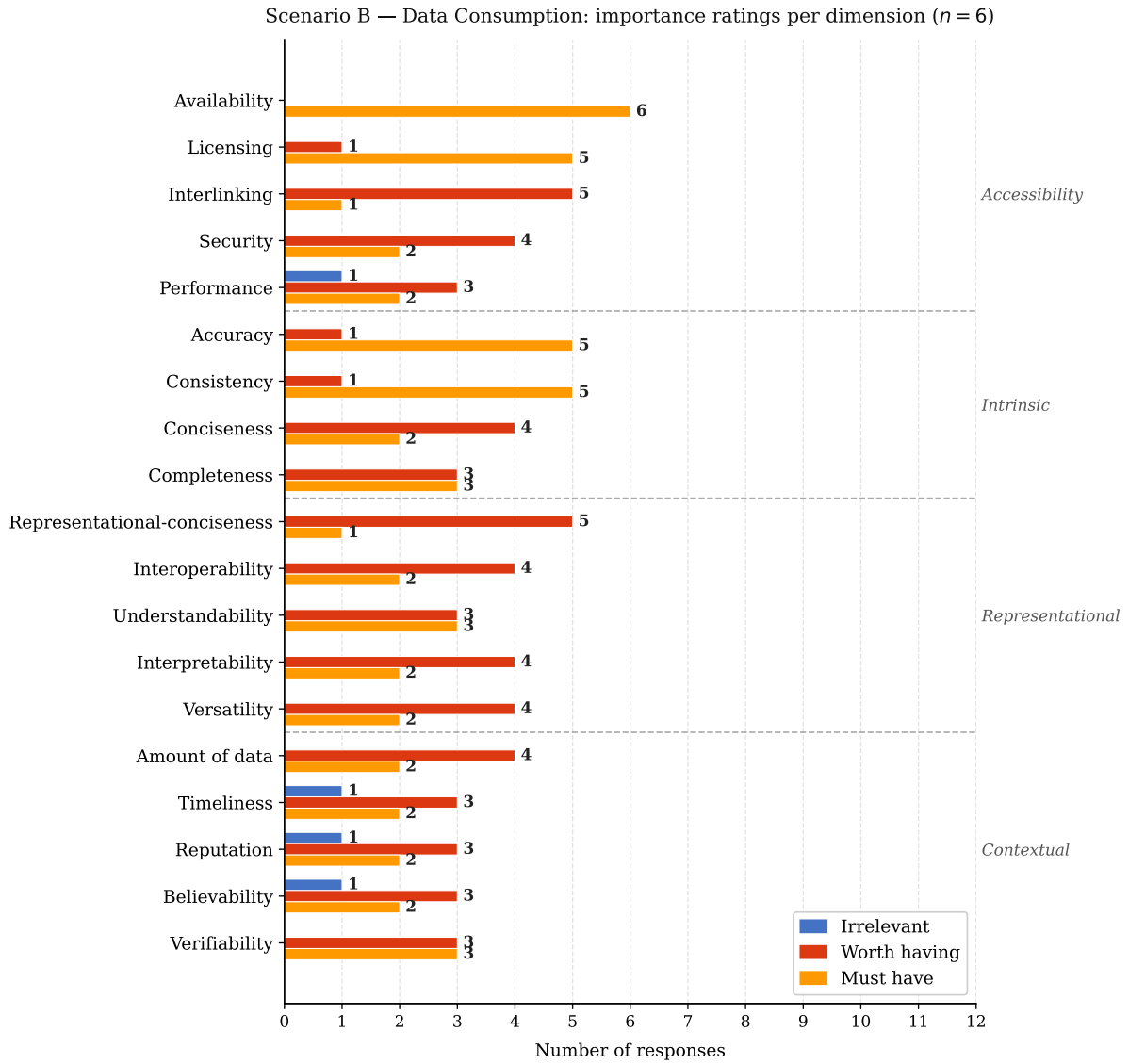

Figure 4: Categorical importance ratings per quality dimension, scenario B (data consumption), $n = 6$. Dimensions are grouped by quality family (Zaveri et al.).

The cross-scenario profiles in Table 6 already support practical guidance without requiring numerical weights. Stable high-priority dimensions (*accuracy*, *availability*, *verifiability*) form a baseline that any quality assessment in this setting should cover. Context-sensitive dimensions (*licensing*, *understandability*, *interoperability*) should be weighted according to the deployment scenario: practitioners can identify which scenario their use case most resembles and adjust emphasis accordingly. Stable low-priority dimensions such as *versatility* are candidates for exclusion in lightweight assessments. However, a full derivation of weighting structures is left as future work.

5. Discussion

In the following, we revisit and discuss the two research questions introduced at the beginning of this paper, and comment on the generalisability and limitations of this survey.

Comparative selection reveals structure hidden by quality categories (RQ1). RQ1 asked which KG quality dimensions practitioners consider essential. The best-worst selection task exposed prioritisation hierarchies that the three-level categorical ratings alone could not distinguish. Within

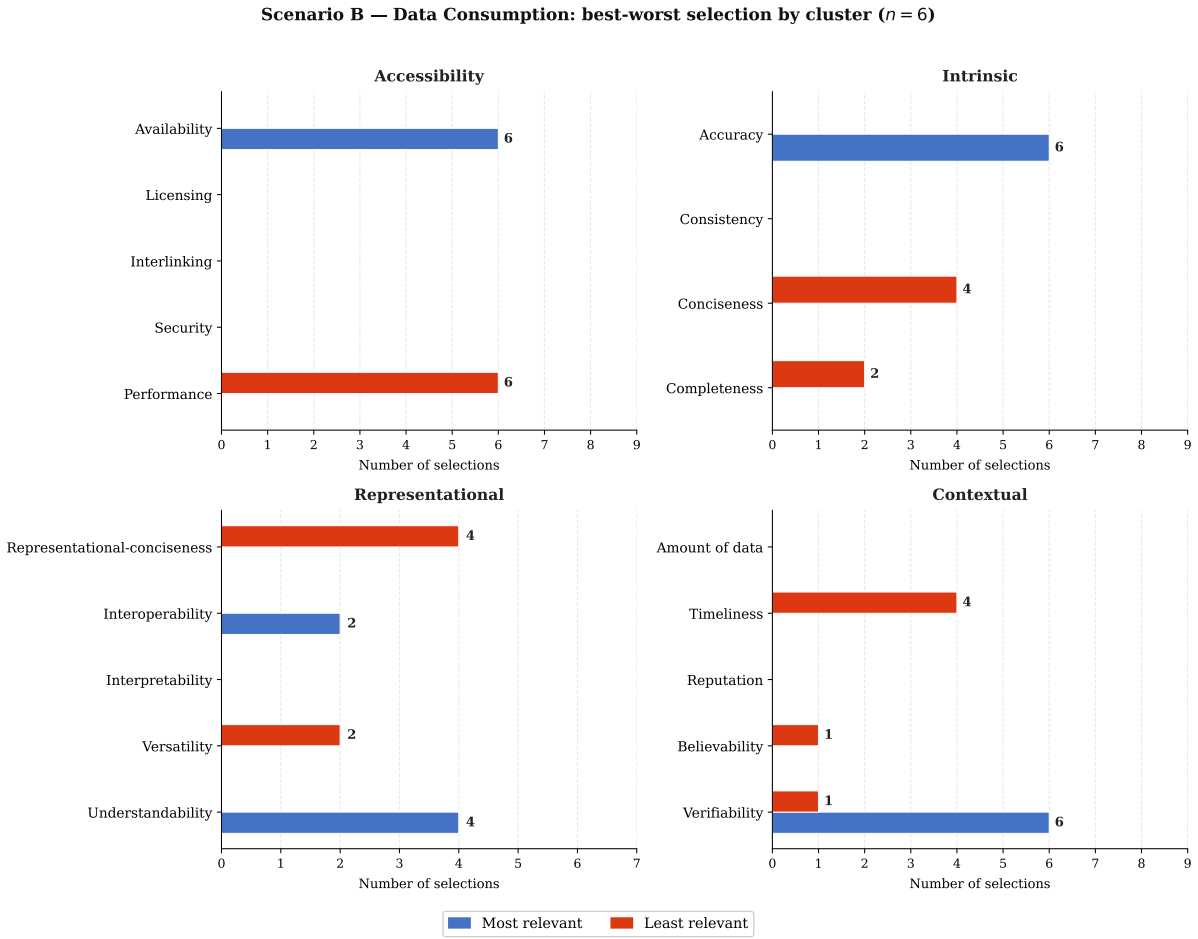

Figure 5: Best-worst selection counts per dimension cluster, scenario B (data consumption), $n = 6$.

Table 6
Cross-scenario prioritisation summary ($n = 15$). Arrows indicate directional shifts between scenario A and scenario B.

Dimension	Scenario A	Scenario B	Status
Accuracy	Must have	Must have	Stable
Availability	Must have	Must have	Stable
Verifiability	Top (BWS)	Top (BWS)	Stable
Licensing	Must have	Worth having	Context-sensitive (↓)
Understandability	Top (BWS)	Mid	Context-sensitive (↓)
Interoperability	Mid	Top (BWS)	Context-sensitive (↑)
Performance	Least (BWS)	Mid	Context-sensitive (↑)
Completeness	Least (BWS)	Mid	Context-sensitive (↑)
Versatility	Least (BWS)	Least (BWS)	Stable (low)

the intrinsic cluster, for example, *accuracy* and *completeness* both attracted high "must have" ratings in scenario A, yet *completeness* was simultaneously the most frequently selected "least relevant" dimension in the forced task. This apparent contradiction reflects the difference between what participants consider generally important and what they would actually prioritise when forced to choose. This pattern is not isolated: similar trade-offs are visible across clusters, for instance between *understandability* and *interoperability* in the representational cluster across scenarios. This result supports the use of best-worst scaling [27] as a complement to Likert-type instruments in quality elicitation studies.

Quality priority is task-dependent (RQ2). RQ2 asked to what extent the importance of DQ dimensions varies across tasks and roles. The central finding is that practitioners do not assign uniform importance to KG quality dimensions. Across both scenarios, *accuracy*, *availability*, and *verifiability* are stable at high priority, while other dimensions shift substantially with role and task. *Licensing* migrates from essential (publication) to secondary (consumption); *interoperability* gains prominence for consumers over publishers; *understandability* follows the reverse pattern. This confirms empirically what the *fitness for use* perspective [11] implies theoretically: a single fixed weighting scheme cannot adequately represent quality requirements across different usage contexts. Existing KG quality frameworks such as KGHeartBeat [3] and Luzzu [2] provide the infrastructure to express dimension weights, but leave their derivation to the user. The prioritisation profiles produced by this study provide an empirical basis for populating those weights in cultural heritage settings.

Limitations and recommendations for DQ dimension definitions. Section 4.2 identified three dimensions where definitional uncertainty may have distorted prioritisation ratings. For *timeliness*, several participants questioned the applicability of the concept in a domain where data is historically stable. For *security*, the ratings were bimodal, which reflects the disagreement about whether the dimension applies to publicly accessible datasets. For *reputation* and *believability*, the perceived conceptual overlap limits the reliability of their individual prioritisation scores. These three dimensions should therefore be interpreted with caution. Based on these observations, we derive three concrete recommendations for future deployments. First, the definitions of the DQ dimensions should be presented to a small domain-specific group before the main study to verify that they are interpreted consistently. Second, *reputation* and *believability* should be merged into a single "source trustworthiness" dimension. Third, *timeliness* should be reformulated in terms that are meaningful for slowly-evolving collections, where temporal currency is less relevant than provenance and stability.

Generalisability. 80% of participants rated the instrument as generalisable to other domains (levels 4–5 out of 5), which suggests that the elicitation methodology transfers beyond cultural heritage, even if dimension definitions require domain-specific adaptation. This is consistent with the taxonomy by Zaveri et al. [1], which was designed to apply across linked data domains. Replication in domains such as life sciences or linguistics, both identified by participants, would test whether the stable dimensions observed here (*accuracy*, *availability*, *verifiability*) reflect a domain-general core, or whether they are specific to the governance and provenance requirements of cultural heritage.

Limitations of the study design. Three limitations should be noted. First, the sample size of 15 participants is appropriate for an exploratory study but insufficient to support statistical generalisation. Second, recruitment through the GOBLIN network introduces a selection bias towards individuals with high semantic web familiarity, which may skew ratings for technically defined dimensions such as *interpretability* and *interlinking*. Third, the two-scenario design does not cover the full range of KG interaction contexts; dimensions such as *timeliness* or *performance* may rank differently in data integration or real-time retrieval scenarios not represented here.

6. Conclusion and Future Work

This paper presents an empirical study on the context-sensitivity of KG quality dimensions and their importance. We designed a survey that combines categorical importance ratings with best-worst scaling across two scenarios (data publication and data consumption) and collected responses from 15 practitioners in the cultural heritage domain. Results show that *accuracy*, *availability*, and *verifiability* have a stable high priority rating across both scenarios, while *licensing*, *interoperability*, *understandability*, and *performance* indicate measurable context-sensitivity. 73.3% of the participants confirmed that their quality prioritisation changes with the task context, which provides direct empirical support for the context-sensitive importance assessment in composite KG quality indicators.

These findings address a gap recognized in existing KG quality frameworks: while tools such as Luzzu [2] or KGHeartBeat [3] and vocabularies such as DQV [7] and DQD [8] support the expression of dimension weights, they provide no mechanism to derive them empirically. The results reported in this paper provide preliminary insights into DQ dimension prioritisation in cultural heritage settings, while adopting a general-purpose and reusable approach that can be transferred to other domains.

Several directions remain open for future work. The most straightforward extension is to apply the instrument in other domains, such as life sciences and linguistics, where participants themselves rated generalisability highly (80% at levels 4–5), to assess whether the observed prioritisation patterns transfer across contexts. Beyond the two scenarios studied here, further roles and tasks should be explored to test whether the stable dimensions identified here generalise more broadly. This includes special cases such as data integration and cross-institutional federation, where dimensions such as *interoperability* and *timeliness* may rank differently than observed here. As a more technical next step, the empirically derived importance profiles should be integrated into an operational quality assessment pipeline and evaluated against expert-assumed or uniform baselines on a concrete KG quality reporting task. Finally, the ambiguity observed in several dimension definitions suggests that refining and sharpening existing definitions [1] is a valuable direction, in particular merging *reputation* and *believability* into a single "source trustworthiness" construct and reformulating *timeliness* for slowly-evolving collections.

Acknowledgments

This publication is based upon work from COST Action CA23147 GOBLIN – Global Network on Large-Scale, Cross-domain and Multilingual Open Knowledge Graphs, supported by COST (European Cooperation in Science and Technology, https://www.cost.eu).

Declaration on Generative AI

During the preparation of this work, the author(s) used GPT-4 in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, S. Auer, Quality assessment for linked data: A survey, Semantic Web 7 (2016) 63–93. doi:10.3233/SW-150175.

[2] J. Debattista, S. Auer, C. Lange, Luzzu—a methodology and framework for linked data quality assessment, Journal of Data and Information Quality (JDIQ) 8 (2016) 1–32. doi:10.1145/2992786.

[3] M. A. Pellegrino, A. Rula, G. Tuozzo, Kgheartbeat: An open source tool for periodically evaluating the quality of knowledge graphs, in: The Semantic Web - ISWC - 23rd International Semantic Web Conference, volume 15233 of *Lecture Notes in Computer Science*, Springer, 2024, pp. 40–58. doi:10.1007/978-3-031-77847-6_3.

[4] W. Beek, L. Rietveld, H. R. Bazoobandi, J. Wielemaker, S. Schlobach, LOD laundromat: A uniform way of publishing other people's dirty data, in: The Semantic Web – ISWC: 13th International Semantic Web Conference. Proceedings, Part I, Springer-Verlag, Berlin, Heidelberg, 2014, p. 213–228. doi:10.1007/978-3-319-11964-9_14.

[5] D. Pizhuk, L. Ehrlinger, G. Denk, V. Geist, A data quality dashboard for (security) knowledge graphs, in: BTW, volume P-361 of *LNI*, Gesellschaft für Informatik e.V., 2025, pp. 803–810. doi:10.18420/BTW2025-45.

[6] A. Langer, V. Siegert, C. Göpfert, M. Gaedke, Semquire - assessing the data quality of linked open data sources based on DQV, in: ICWE Workshops, volume 11153 of *LNCS*, Springer, 2018, pp. 163–175. doi:10.1007/978-3-030-03056-8_14.

[7] R. Albertoni, A. Isaac, Introducing the data quality vocabulary (dqv), Semantic Web 12 (2020) 81–97. doi:10.3233/SW-200382.
[8] J. Schrott, R. Meindl, C. Lettner, W. Wöß, L. Ehrlinger, DQD: The data quality definition ontology, in: MTRS, Springer, 2023, pp. 291–297. doi:10.1007/978-3-031-65990-4_27.
[9] C. Cichy, S. Rass, An overview of data quality frameworks, IEEE Access 7 (2019) 24634–24648.
[10] S. Mohammed, L. Ehrlinger, H. Harmouch, F. Naumann, D. Srivastava, The five facets of data quality assessment, SIGMOD Rec. 54 (2025) 18–27. doi:10.1145/3749116.3749120.
[11] R. Y. Wang, D. M. Strong, Beyond accuracy: What data quality means to data consumers, Journal of Management Information Systems 12 (1996) 5–33. doi:10.1080/07421222.1996.11518099.
[12] V. Papastergios, L. Ehrlinger, A. Gounaris, Unfolding data quality dimensions in practice: A survey, J. Data and Information Quality 18 (2026). doi:10.1145/3786328.
[13] ISO 8000-8:2015(E), Data Quality – Part 8: Information and Data Quality Concepts and Measuring, Standard, International Organization for Standardization, 2015.
[14] International Organization for Standardization, Software engineering – Software product Quality Requirements and Evaluation (SQuaRE) – Data quality model, 2008.
[15] International Organization for Standardization, Artificial intelligence – Data quality for analytics and machine learning (ML) – Part 1: Overview, terminology, and examples, 2024.
[16] International Organization for Standardization, Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Measurement of data quality, 2015.
[17] D. M. Strong, Y. W. Lee, R. Y. Wang, Data quality in context, Communication ACM 40 (1997) 103–110. doi:10.1145/253769.253804.
[18] F. Serra, V. Peralta, A. Marotta, P. Marcel, Use of context in data quality management: A systematic literature review, J. Data and Information Quality 16 (2024). URL: https://doi.org/10.1145/3672082. doi:10.1145/3672082.
[19] B. Spahiu, M. Palmonari, R. A. Alva Principe, A. Rula, Understanding the structure of knowledge graphs with abstat profiles, Semantic Web 15 (2024) 1519–1545.
[20] J. M. Carroll, Making use: scenario-based design of human-computer interactions, MIT press, 2003.
[21] D. Benyon, C. Macaulay, Scenarios and the hci-se design problem, Interacting with Computers 14 (2002) 397–405. doi:10.1016/S0953-5438(02)00007-3.
[22] E. Kapsalis, Wikidata: Recruiting the crowd to power access to digital archives, Journal of Radio & Audio Media 26 (2019) 134–142. doi:10.1080/19376529.2019.1559520.
[23] D. Kontokostas, A. Zaveri, S. Auer, J. Lehmann, Triplecheckmate: A tool for crowdsourcing the quality assessment of linked data, in: International Conference on Knowledge Engineering and the Semantic Web, Springer, 2013, pp. 265–272. doi:10.1007/978-3-642-41360-5_22.
[24] M. Acosta, A. Zaveri, E. Simperl, D. Kontokostas, F. Flöck, J. Lehmann, Detecting linked data quality issues via crowdsourcing: A dbpedia study, Semantic web 9 (2018) 303–335. doi:10.3233/SW-160239.
[25] M. Cao, J. Zhang, S. Xu, Z. Ying, Knowledge graphs meet crowdsourcing: a brief survey, in: International Conference on Cloud Computing, Springer, 2020, pp. 3–17. doi:10.1007/978-3-030-69992-5_1.
[26] D. Clegg, R. Barker, CASE Method Fast-track: A RAD Approach, CASE method, Addison-Wesley Publishing Company, 1994. URL: https://books.google.it/books?id=86ZfQgAACAAJ.
[27] J. J. Louviere, T. N. Flynn, A. A. J. Marley, Best-Worst Scaling: Theory, Methods and Applications, Cambridge University Press, 2015.