JLIS.it vol. 15, no. 1 (January 2024)
ISSN: 2038-1026 online

DOI: 10.36253/jlis.it-547

JLIS.it

Quality Assessment of Research Comparisons in the Open Research Knowledge Graph: A Case Study

Jennifer D'Souza[a], Hassan Hussein[b], Julia Evans[c], Lars Vogt[d], Oliver Karras[e], Vinodh Ilangovan[f], Anna-Lena Lorenz[g], Sören Auer[h]

a-h) TIB Leibniz Information Centre for Science and Technology

Contact: Jennifer D'Souza, jennifer.dsouza@tib.eu; Hassan Hussein, hassan.hussein@tib.eu; Julia Evans, julia.evans@tib.eu; Lars Vogt, lars.vogt@tib.eu; Oliver Karras, oliver.karras@tib.eu; Vinodh Ilangovan, vinodh.ilangovan@tib.eu; Anna-Lena Lorenz, anna.lorenz@tib.eu; Sören Auer, auer@tib.eu
Received: 04 April 2023; **Accepted:** 21 June 2023; **First Published:** 15 January 2024

ABSTRACT
The Open Research Knowledge Graph (ORKG) is a digital library for machine-actionable scholarly knowledge, with a focus on structured research comparisons obtained through expert crowdsourcing. While the ORKG has attracted a community of more than 1,000 users, the curated data has not been subject to an in-depth quality assessment so far. Here, proposed as a first exemplary step, within a team of domain experts, we evaluate the quality of six selected ORKG Comparisons based on three criteria, namely: 1) the quality of semantic modelling, 2) the maturity of the Comparisons in terms of their completeness, syntactic representation, identifier stability, and their linkability mechanisms ensuring the interoperability and discoverability. Finally, 3) the informative usefulness of the Comparisons to expert and lay users. We have found that each criterion addresses a unique and independent aspect of quality. Backed by the observations of our quality evaluations presented in this paper, a fitting model of knowledge graph quality appears one that is indeed multidimensional as ours.

KEYWORDS
Knowledge Graph; Open Research Knowledge Graph; Linked Open Data (LOD); Human-Computer Interaction; Survey.

JLIS.it vol. 15, no. 1 (January 2024)
ISSN: 2038-1026 online

DOI: 10.36253/jlis.it-547

1. Introduction

Scholarly communication has long relied on discourse-based methods, but the sheer volume of new publications each year can be overwhelming for researchers. The Open Research Knowledge Graph (ORKG) project (Auer et al. 2020) advocates for a structured format to represent scholarly contributions as a knowledge graph (KG) readable by machines and humans. With its next-generation semantic publishing model (Berners-Lee, Hendler, and Lassila 2001; Shotton 2009), it claims that only with a combination of human- and machine-actionability can we allow for novel exploration and assistance services for scholars, thereby strengthening research even in the face of voluminous publication cycles worldwide (Bornmann and Mutz 2015, 2217; Johnson, Watkinson, and Mabe 2018, 6).

The ORKG simplifies access to scholarly knowledge by explicitly representing a machine-actionable set of property-value pairs to describe scholarly contributions. With such a structured description format, several works of research can be compared by their similar properties using the ORKG Research Comparisons feature (Oelen et al. 2019; 2020). In keeping with standards, it adheres to the FAIR (Findable, Accessible, Interoperable, and Reusable) Data Principles (Wilkinson et al. 2016). ORKG Comparisons are a way of making science more accessible to all researchers, thereby a concrete step toward improved scientific communication and also encouraging scientific interdisciplinarity. However, given that the ORKG employs crowdsourcing to populate the KG, data quality unavoidably varies. Maintaining high standards thus necessitates effective evaluation and management of the information quality of the graph.

Information quality is generally defined as "fitness for use" (Zaveri et al. 2016, 2). Embedded within this definition is a question: use by *whom* and for *what purpose*? As alluded to earlier, the ORKG is intended to be both human-readable and machine-actionable – two very disparate user groups with equally disparate (and sometimes competing) desiderata. A graph with high machine-actionability may be overly complex and confusing for human readers, while a graph with high human-readability may lack the explicit conceptual relationships necessary for machines. In other words, there is no singular definition of quality suitable to all users and use cases.

The ORKG knowledge curation is based on the wisdom-of-the-crowd philosophy[1] concretely implemented via the crowdsourcing methodology. Crowdsourcing is "the outsourcing of a piece of work to a crowd of people via an open call for contributions" rather than hiring or contracting employees for the task (Daniel et al. 2018, 2). It has become an indispensable tool for data acquisition in many fields (Zhang 2022, 749), but ensuring the quality of the obtained data remains a complex open issue (Daniel et al. 2018, 2).

This work presents a qualitative case study evaluating six crowdsourced ORKG Comparisons based on three criteria: 1) semantic modelling quality; 2) the maturity of the Comparisons in terms of their completeness, syntactic representation, identifier stability, and their linkability mechanisms ensuring the interoperability and discoverability; and finally 3) the informative usefulness of the Comparisons to expert and lay users. Each of these quality dimensions corresponds to an ORKG usage goal for humans and machines – as already noted, these goals are not always complementary. Therefore, each criterion is addressed with its own quality evaluation approach.

[1] https://en.wikipedia.org/wiki/Wisdom_of_the_crowd

JLIS.it vol. 15, no. 1 (January 2024)
ISSN: 2038-1026 online

DOI: 10.36253/jlis.it-547

The paper is structured as follows. Section 2 discusses related work on KG quality evaluations and introduces the Knowledge Graph Maturity Model (KGMM) (Hussein et al. 2022) used for criterion 2. Section 3 introduces the ORKG and its features, in particular the ORKG Comparisons analysed in this work. Section 4 details the evaluation criteria, namely semantic modelling, maturity, and informativeness. Section 5 presents the quality evaluations of the six selected ORKG Comparisons for each criterion. Section 6 discusses the obtained results, possible conclusions, and limitations of the study, and Section 7 concludes the paper.

2. Related Work

Data quality is a multidimensional, inherently contextual, and task-dependent concept. To evaluate data quality, researchers advise combining multiple metrics adapted to each individual use case, such as surveys of stakeholders' perceptions, assessments of ease of access, and statistics of accuracy and completeness (Pipino, Lee, and Wang 2002; Bizer and Cyganiak 2009; Zaveri et al. 2016; Färber et al. 2018). The following subsections present two perspectives on knowledge graph quality evaluation that are particularly relevant to the ORKG: the generic frameworks of Linked Open Data and FAIR Guiding Principles, and the more specific Knowledge Graph Maturity Model (KGMM).

2.1 Linked Open Data and FAIR Guiding Principles

The Five-Star Linked Open Data (LOD) system was introduced by Tim Berners-Lee in 2010 to encourage publishing data on the Web in machine-readable, non-proprietary formats following the Resource Description Framework (RDF) standards and containing links to other data.[2] Another prominent approach to linked data management is FAIR – Findable, Accessible, Interoperable, and Reusable – which defines the "characteristics that contemporary data resources, tools, vocabularies and infrastructures should exhibit to assist discovery and reuse" by both humans and machines (Wilkinson et al. 2016). Zaveri et al. (2016) surveyed thirty quality measures for linked data quality. They compiled eighteen "core" data quality dimensions divided amongst four classifications and encompassing both qualitative and quantitative metrics for assessment. However, they find that most metrics were not explicitly defined or were not statistically precise, and none were formally validated. The resulting framework they present is a starting point for further development of more defined metrics.

2.2 Knowledge Graph Maturity Model

The Knowledge Graph Maturity Model (KGMM) offers an assessment of data in KGs with the goal of offering it in "the most mature, complete, representable, stable, and linkable shape" (Hussein et al. 2022). The model itself distinguishes five maturity levels, with each level possessing its own set of quality measures that are ranked in priority as either essential, important, or useful.

[2] https://www.w3.org/DesignIssues/LinkedData.html

JLIS.it vol. 15, no. 1 (January 2024)
ISSN: 2038-1026 online

DOI: 10.36253/jlis.it-547

To pass to the next maturity level, all measures rated essential and at least half of those rated as important must be satisfied. Whether a measure is satisfied or not is a binary choice. Below, we briefly describe the five levels of the KGMM in the context of the specific application scenario of ORKG Comparisons.

- **Level 1: Published.** Passing the quality measures defined for this maturity level indicates that the KG is accessible on the web with an open licence. ORKG Comparisons satisfy this level by default, due to the ORKG publishing model.
- **Level 2: Completeness.** Passing this level implies that the ORKG Comparison has sufficient information scope and density to meet the requirements of the given task (Wang and Strong 1996). This encompasses the accuracy, timeliness, and trustworthiness of the data, as well as a clear record of its provenance. It additionally includes whether there are missing values, whether all necessary statements are included, and whether there is documentation of the KG's contents.
- **Level 3: Representation.** Clearing this level assures that the Comparison is free from redundant or duplicated entities; its format and metadata ensure it is reusable by others. Where possible, data might be presented in different formats (e.g., tables, charts, etc.) while still maintaining full backward compatibility.
- **Level 4: Stability.** Stability implies that the KG's data is trackable w.r.t. provenance and version history. Furthermore, it enables mechanisms to leverage globally unique IDs such as DOIs and ORCID. It provides a means for querying the data in the KG, such as SPARQL or an API endpoint. Based on the features of the ORKG, Comparisons satisfy this level by default.
- **Level 5: Linkability.** The KG can be dereferenced using URIs and an HTTP call retrieves the graph in RDF syntax. The KG is well-linked to internal and external resources. The ORKG satisfies dereferenceability by default even though they may or may not be well-linked. Based on the minimal dereferenceability criteria, Comparisons automatically pass this level.

In general, the KGMM can be applied in a crowdsourcing setting involving multiple reviewers of the quality of the KG. Survey questions addressing many of the quality measures have been developed to appear alongside a KG, and crowdsourced reviewers are asked to reply with their opinion.

3. Open Research Knowledge Graph

In this section, we introduce two relevant facets of the Open Research Knowledge Graph (ORKG).

Paper Contributions. Figure 1 illustrates a structured description of a paper's contribution in the ORKG. The ORKG represents papers as research contribution(s) in a structured and semantic way using (subject, predicate, object) triple statements,[3] where the predicates are salient properties of a given work's contribution. The subject and object positions are generally filled by resources.[4]

[3] https://www.w3.org/TR/rdf11-concepts/\#data-model
[4] https://www.w3.org/TR/rdf11-concepts/\#resources-and-statements

JLIS.it vol. 15, no. 1 (January 2024)
ISSN: 2038-1026 online

DOI: 10.36253/jlis.it-547

ORKG resources are resolved to the ORKG namespace and constitute the ORKG scholarly knowledge vocabulary.[5] Mapping between an ORKG resource and an equivalent concept resource in another ontology is done using the OWL sameAs property.[6] ORKG objects may also be filled by a literal such as a text, number, or date. The ORKG also stores basic metadata of the original paper including the title, publication date, and list of authors. Papers may also be assigned to a research field within the ORKG's 700-fields taxonomy.[7]

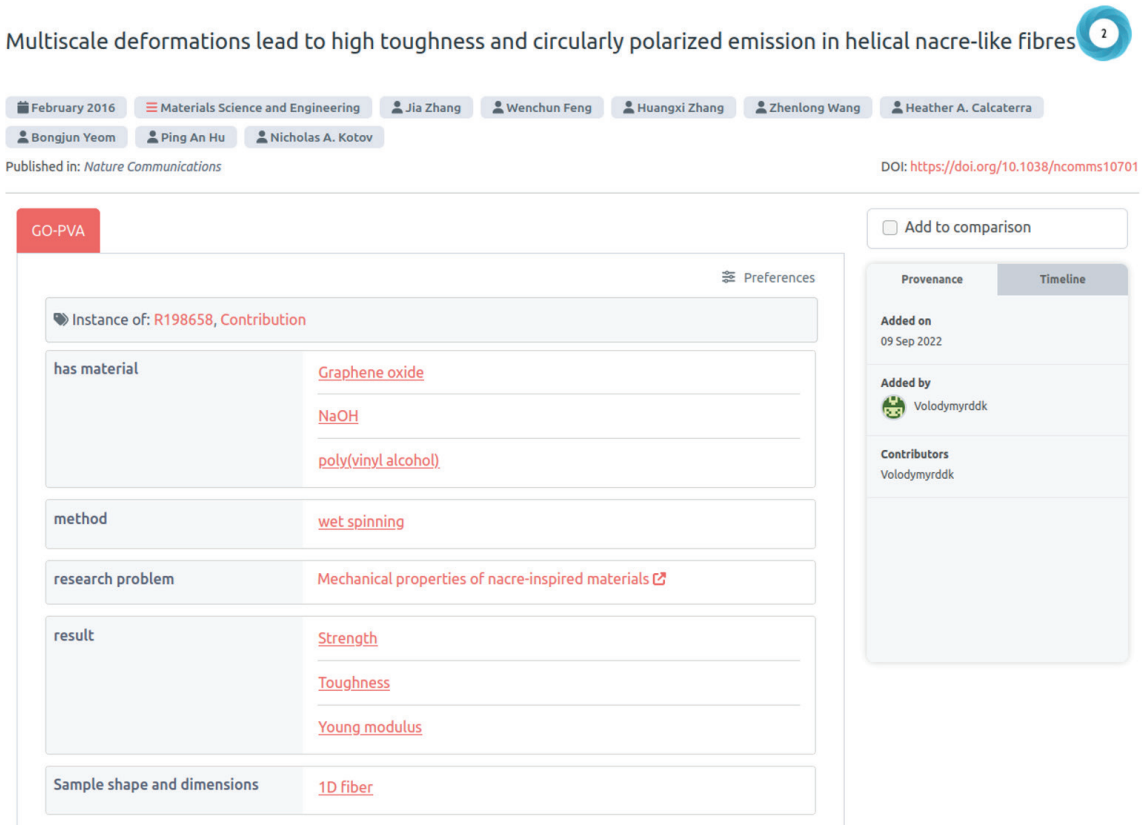

Figure 1. The ORKG user interface of an example scholarly paper w.r.t. its structured contribution content as (property, value) pairs, where the properties are the contribution relevant aspects.

ORKG Research Comparisons. The ORKG platform supports downstream services such as the creation of tabular Comparisons of contributions which constitute ORKG subgraphs. Figure 2 shows an example of one such Comparison. The ORKG Comparisons feature creates one aggregated view of the values of several structured contributions with more-or-less similar sets of predicates. Comparisons can encompass multiple contributions from the same article (e.g., a Comparison of AI benchmark characteristics introduced as several contributions in a single paper); or contributions with similar predicates from different articles (e.g., a Comparison of the Covid-19

[5] https://www.w3.org/TR/rdf11-concepts/\#vocabularies
[6] https://orkg.org/property/SAME_AS
[7] https://orkg.org/help-center/article/20/ORKG_Research_fields_taxonomy

JLIS.it vol. 15, no. 1 (January 2024)
ISSN: 2038-1026 online

DOI: 10.36253/jlis.it-547

reproductive number (R0) estimate studies undertaken by different research groups in different countries).

1D fiber based on graphene oxide

September 2022 | Volodymyrddk

Mechanical properties of nacre-inspired 1D fiber based on graphene oxide.

Properties	Scalable One-Step Wet-Spinning of Graphene Fibers and Yarns from Liquid Crystalline Dispersions of Graphene Oxide: Towards Multifunctional Textiles *GO fibers coagulated by CaCl2 - 2013*	Scalable One-Step Wet-Spinning of Graphene Fibers and Yarns from Liquid Crystalline Dispersions of Graphene Oxide: Towards Multifunctional Textiles *GO fibers coagulated by chitosan - 2013*	Liquid crystal self-templating approach to ultrastrong and tough biomimic composites *GGO-HPG - 2013*	Ultrastrong Fibers Assembled from Giant Graphene Oxide Sheets *GO-ca2+ - 2012*
has material	calcium(2+) Graphene oxide	chitosan Graphene oxide	Ca2+ Graphene oxide hyperbranched polyglycerol	Ca2+ Graphene oxide
has method	wet spinning	wet spinning	wet spinning	wet spinning
has result	Strength Toughness Young modulus	Strength Toughness Young modulus	Strength Toughness Young modulus	Strength Toughness Young modulus
has result/strength				
has measurement value*	412	442	652	364.4
has unit*	megapascal	megapascal	megapascal	megapascal
has unit/megapascal/same as*	https://www.wikidata.org/entity/Q21062777	https://www.wikidata.org/entity/Q21062777	https://www.wikidata.org/entity/Q21062777	https://www.wikidata.org/entity/Q21062777
same as*	strength	strength	strength	strength
same as/strength/same as*	https://www.wikidata.org/entity/Q605035	https://www.wikidata.org/entity/Q605035	https://www.wikidata.org/entity/Q605035	https://www.wikidata.org/entity/Q605035
has result/toughness				
has measurement value*	4.8	4.8	14	6.8
has unit*	megajoule per cubic metre	MJ/m3	MJ/m3	%
has unit/mj/m3/same as*		megajoule per cubic metre	megajoule per cubic metre	
same as*	toughness	toughness	toughness	toughness

Figure 2. An example ORKG Comparison based on a common set of properties (depicted in the grey column) used to semantically describe four different contributions of scholarly articles (depicted in the columns with the red header).

3.1 Six ORKG Research Comparisons – the Qualitative Analysis Corpus of this Work

The ORKG partners with researchers (senior PhD students, postdocs, and other advanced scholars) from diverse domains to curate contributions and Comparisons that address key research problems in their field. Grantees receive up to €400 per month as an incentive as well as mentoring from the ORKG team. Of the 1,134 total Comparisons in the ORKG, 348 were created by curation grantees. Six of these Comparisons were selected based on the quality of their semantic modelling, as defined by experts (see Section 4.1 for a detailed explanation of the semantic modelling criteria). Table 1 provides overview statistics of these Comparisons.

JLIS.it vol. 15, no. 1 (January 2024)
ISSN: 2038-1026 online

DOI: 10.36253/jlis.it-547

Comparison name	Contributions (Papers)	Research field	Research problem
1D Fiber[8]	8 (5)	Materials Science and Engineering	Mechanical properties of nacre-inspired materials
Microneedle Tech.[9]	7 (7)	Medicinal Chemistry and Pharmaceutics	Microneedle technology
Supply Chain[10]	16 (16)	Operations Research, Systems Engineering and Industrial Engineering	Supply chain
CO_2 Gas Flux[11]	11 (11)	Oceanography	CO_2 flux estimation of the ocean
Glucose Sensors[12]	7 (7)	Nanoscience and Nanotechnology	Development of Glucose Sensors Based on Nanomaterials
Smart Cities[13]	10 (9)	Information Systems, Process and Knowledge Management	Smart cities

Table 1. Overview of the six ORKG Comparisons qualitatively analysed in this work. A research paper can have more than one contribution hence the count of papers and contributions are not always equal.

4. Six ORKG Research Comparisons – a Qualitative Analysis

The six selected ORKG Comparisons were qualitatively analysed by the following three criteria: 1) semantic modelling, 2) maturity ratings based on the KGMM, and 3) the usefulness of their content for researchers and laypeople. These criteria address different user groups and use cases, with semantic modelling focusing on machine-actionability, content usefulness for human readers, and the maturity model blending both with the principles of LOD and FAIR.

4.1 Qualitative Evaluation Criterion 1 – Semantic Modelling

The evaluation dataset in this study was initially selected based on the semantic modelling quality determined by two semantic modelling experts in the ORKG team. Good semantic modelling criteria, especially for crowdsourced knowledge, are defined based on observations made on the data in terms of which types of modelling decisions create a human-readable and machine-actionable graph. However, the ORKG recommended guidelines suggest best practices, but are not mandated. For this evaluation, we selected three Comparisons which satisfy all or many of the semantic modelling guidelines and can be deemed good examples to learn from; and three Comparisons that are examples of pitfalls to avoid in semantic modelling. It should be noted that *only* the quality of the modelling is judged here – the content itself is not considered.

[8] https://orkg.org/comparison/R215963/
[9] https://orkg.org/comparison/R161079/
[10] https://orkg.org/comparison/R212576/
[11] https://orkg.org/comparison/R160742/
[12] https://orkg.org/comparison/R143853/
[13] https://orkg.org/comparison/R140131/

JLIS.it vol. 15, no. 1 (January 2024)
ISSN: 2038-1026 online

DOI: 10.36253/jlis.it-547

4.1.1 Semantic Modelling Principles

I. *Use of ontology resources.* A marker of a high-quality ORKG Comparison is in its use of resources rather than literals for object nodes wherever possible. Resources can be reused across statements, thereby establishing connections between papers. At best, resources should have a user-supplied definition and additional description statements, which we identify as *ontology resources*. However, this requires more input work on the part of the creator, and it is often the case that resources are not described or reused within the ORKG. We identify these as *ad hoc resources*. Use of ontology resources is always preferable to use of ad hoc resources.

II. *Establishing mappings between resources.* Where appropriate, ORKG resources may be linked to external ontologies, such as Wikidata and GeoNames. Such resources are also considered ontology resources. This results in ORKG resources situated in the linked data cloud as well.

III. *Modelling best practices for resources that are constants.* A prerequisite to following these recommended practices is to determine whether a triple object should indeed be modelled as a literal value and not a resource. A hint for which values to model as literals can be gleaned from the available RDF datatypes for literals – e.g., boolean, string, integer, decimal, date, etc.[14] However, there still remains significant subjective decision making on the part of the modeller. To support this, the following sub-list offers some recommendations for common use cases, although it is by no means exhaustive.

   A. *Use the QUDT specification to model measurements.* A resource representing a measurement may be used in the object position in one statement and then reused in the subject position of additional statements for each facet of information such as measurement value (specified with a literal) and measurement unit (specified with a resource).

   B. *Use the international standard YYYY-MM-DD date format.* Building upon XML Schema, RDF specification supports the date datatype for literals and specifies the format.

   C. *Express time period nodes as a two-statement resource in the graph, one for start and one for end.* The literals for date and time must not constitute more than one date or time reference point. Ranges can be expressed as a resource node in the graph which is further specified via begin and end predicates leading to literal value specifications of the begin and end date/time.

   D. *Maintaining consistency in modelling.* Maintaining consistency – of terms, labels, modelling strategies, and so on – is also a hallmark of high-quality modelling. The modelling choices should follow a consistent pattern of expression.

IV. *Select relevant properties which have values across multiple contributions.* An ORKG Comparison with many empty cells affects human-readability. It is best to minimise the number of empty cells as far as possible by careful selection of properties with values across most if not all contributions in the Comparison.

V. *Use phrase-based resource names.* Avoid communicating too much information (e.g., complete sentences) in individual literals or labels of resources. This hinders machine-actionability. However, the use of sentences is encouraged for the description predicate.

[14] https://www.w3.org/TR/rdf11-concepts/\#section-Datatypes

JLIS.it vol. 15, no. 1 (January 2024)
ISSN: 2038-1026 online

DOI: 10.36253/jlis.it-547

4.2 Qualitative Evaluation Criterion 2 – Knowledge Graph Maturity Model

KGMM (Hussein et al. 2022) was discussed in theory to apply to KGs generally, but also included a proposed implementation specifically for ORKG Comparisons. All measures that were not either directly satisfied by the ORKG publishing model or which did not have a mechanism in the ORKG based on which they could be satisfied, were proposed as a survey questionnaire. A survey question asked the user to offer a yes-or-no answer based on whether the measure was deemed satisfied or not. The cumulative responses to the survey, in combination with those measures which are automatically satisfied, determine the maturity of the ORKG Comparison.

4.2.1 Maturity Evaluation as a Questionnaire

- **Pilot evaluations.** The proposed KGMM questionnaire was pilot tested to assess the clarity and difficulty of the questions. Specialists from Biology, Physics, and Meteorology were involved in the pilot study, and based on their feedback, four questions were rephrased. The final survey questions and the corresponding KGMM maturity measure are presented in Table 2, and the survey itself is published online.
- **Final evaluations.** The survey was conducted with six subject librarians from TIB. They were selected based on their one-to-one domain expertise mapping against the selected Comparisons, where in one case the subject librarian was a specialist in the broader domain of the Comparison but not the specific research problem addressed.[15] Each subject librarian was presented with a Comparison from their domain to rate. A summary of their responses is provided in Table 4 (see Section 5).

A requirement of the current KGMM is that the participant taking it would also need to have a basic understanding of semantic modelling to accurately address its questions. E.g., to correctly respond to the usefulness of the Linkability measure in the KGMM Level 5 would mean the participant knows what ontologies are – a concept from the field of semantic modelling – and has an idea of the ones that are suitable. Here, we make the following disclaimer: our six participants, while they are domain experts, may not necessarily be aware of the field of semantic modelling. Thus, to obtain a balanced view, the maturity evaluation survey was also run with one additional participant: the developer of the KGMM model. His responses are recorded in parentheses in Table 4 alongside the maturity evaluations from the subject librarians. Thus, to get an accurate estimation of the maturity ratings, the reader of the table can see the response from the domain-expert who may not be aware of semantic modelling juxtaposed against the response of one with sufficient proficiency in semantic modelling and advanced proficiency in science but who is not a domain-expert.

[15] Specifically, *Several works on the dynamic redesign of a closed-loop supply chain network using accelerated Benders decomposition and robust optimisation models.*

JLIS.it vol. 15, no. 1 (January 2024)
ISSN: 2038-1026 online

DOI: 10.36253/jlis.it-547

	Measure	Survey Question
Level 1: Published	Responsiveness** Licence** Syntactic Accuracy* Easiness*	Satisfied by ORKG publishing model.
Level 2: Completeness	Provenance**	Satisfied by ORKG publishing model.
	Timeliness*	Do you think that the comparison is outdated?
	Instance Completeness*	Do you think the properties of this comparison can generalize as a template to capture new contributions from additional scholarly articles on this research theme?
	Property Completeness*	Do you think that the comparison has enough properties?
	Documentation Completeness*	Do you think the comparison's description is descriptive?
Level 3: Representation	Reusability**	Satisfied by ORKG publishing model.
	Conciseness**	Do you think that the comparison has adequate resource/property values?
	Data Representation	If the comparison doesn't have a visualisation, do you think it needs one? If the comparison has a visualisation, do you think it is a fitting representation of the results?
Level 4: Stability	Trackability** Queryability*	Satisfied by ORKG publishing model.
	Identifier Stability*	Mechanism for measure already present in ORKG.
Level 5: Linkability	Dereferencability	Satisfied by ORKG publishing model.
	Linkability	Would you have suggestions for ontologies to link/resolve the predicates and resources of the selected comparison?

Table 2. Questionnaire methodology used to evaluate the ORKG Comparisons' maturity based on KGMM. Measures marked with double asterisks are essential, those with one asterisk are important, and those without an asterisk are useful. The evaluation for each measure is formulated as a corresponding question to which the responses are recorded as a yes/no.

4.3 Qualitative Evaluation Criterion 3 – Usefulness

Another measure of quality is how well a given KG serves its intended information purpose. For the ORKG, one intended purpose of Comparisons is to offer readers an overview of the scientific progress on a research problem or similar. To assess this, we asked subject librarians from each domain: *Is this comparison useful to a specialist in this field/layperson trying to understand this topic?* The evaluations are described in Section 5 and summarised in Table 5.

JLIS.it vol. 15, no. 1 (January 2024)
ISSN: 2038-1026 online

DOI: 10.36253/jlis.it-547

5. Evaluations

5.1 1D fiber based on graphene oxide

This Comparison contains 9 predicates, 50 resources, and 32 literals and has a maximum subgraph depth of 4 nodes. It presents studies measuring the strength, toughness, and compressive stiffness of composite fibres made of graphene oxide in combination with different materials.

- **Semantic modelling.** This Comparison makes good use of ontology resources, including establishing mappings from ORKG resources to external ontologies. Modelling best practice standards are followed and the representations are clear and consistent. The selected properties contain values across multiple contributions. Resource names are succinct.
- **Maturity rating.** This Comparison has a maturity level of 2. It satisfies Timeliness and Instance Completeness, but not Property Completeness or Documentation Completeness (Level 2). It does not satisfy Conciseness or Data Representation (Level 3). It also does not satisfy Linkability (Level 5).
- **Usefulness.** In the opinion of the subject librarian, this Comparison is lacking critical information for specialists, namely details about the process such as sample size and mean variation. Moreover, the names used for materials are inconsistent (and therefore confusing) and in some cases unclear. As for the suitability of the Comparison for laypeople, the subject librarian was unsure, believing that more information about the samples and materials, as well as the testing method, would probably make it more useful for laypeople.

5.2 Microneedle Technology as Insulin Delivery Systems

This Comparison contains 19 predicates, 99 resources, and 133 literals and has a maximum subgraph depth of 3. It surveys contributions describing microneedles constructed from six different materials and of varying sizes for the purpose of subcutaneous insulin injection.

- **Semantic modelling.** This Comparison uses ontology resources very well, although at the time it was created the ORKG did not have the functionality for linking to external ontologies. Modelling best practices are followed and the overall structure is consistent. Almost all properties have values across all contributions. Resources have short phrase-based names.
- **Maturity rating.** This Comparison is still Level 1. It satisfies Property Completeness, but not Timeliness, Instance Completeness, or Documentation Completeness (Level 2). It satisfies Conciseness and Data Representation (Level 3). It does not satisfy Linkability (Level 5).
- **Usefulness.** This Comparison was deemed not useful to a specialist because it only evaluates one review article, despite a considerable output of additional research on this topic. It is also not suitable for laypeople because it is too niche.

5.3. Several works on the dynamic redesign of a closed-loop supply chain network using accelerated Benders decomposition and robust optimisation models

This Comparison contains 15 predicates, 45 resources, and 83 literals and has a maximum subgraph depth of 2 nodes. It presents research on problems arising in supply chain reconfiguration

JLIS.it vol. 15, no. 1 (January 2024)
ISSN: 2038-1026 online

DOI: 10.36253/jlis.it-547

according to a number of variables, such as whether the entire network is expanding, the future facility plans, and the number of products.

- **Semantic modelling.** While this Comparison does use resources, most are ad hoc rather than ontology resources. The Comparison contains no measurements but still follows good modelling practices and maintains consistency. Almost all properties have values across all Comparisons. Resource names are succinct.
- **Maturity rating.** This Comparison is also still Level 1. It satisfies Property Completeness, but not Timeliness, Instance Completeness, or Documentation Completeness (Level 2). It satisfies Conciseness and Data Representation (Level 3). It also satisfies Linkability (Level 5).
- **Usefulness.** The subject librarian described the Comparison as providing an expert-level overview of the differences and similarities of the contributions. While it serves this purpose well, it was unclear how useful this overview would be for those already experts in the field. On the other hand, there is too little pedagogical exposition for a layperson. A particular stumbling point for this Comparison is that the topic has connections to multiple fields, none of which are explained. Additionally, the property names are confusing and lacking context.

5.4 CO_2 gas flux assessment of the various oceanic regimes

This Comparison contains 20 predicates, 12 resources, and 116 literals and has a maximum subgraph depth of 1 node. It gives an overview of studies measuring changes in atmospheric CO_2 concentration over bodies of water depending on the time and location.

- **Semantic modelling.** It has an overabundance of literals, and only uses ad hoc resources. The modelling does not follow best practices (e.g. dates are not expressed in the proper format, time periods are expressed using a single statement) and is inconsistent. Many of the selected properties are not relevant to all contributions. However, it does use appropriate phrase-based naming.
- **Maturity rating.** This Comparison has a maturity level of 5. It satisfies all measures except for Data Representation (Level 3).
- **Usefulness.** The subject librarian was unable to judge if this Comparison would be useful for specialists. The Comparison was judged not suitable for laypeople.

5.5 Comparison among glucose sensors based on different nanomaterials

This Comparison contains 9 predicates, 8 resources, and 80 literals and has a maximum subgraph depth of 1 node. It includes the sensitivity and measurements of glucose sensors produced from seven different nanomaterial constructions.

- **Semantic modelling.** Similar to the previous Comparison, there are a multitude of literals which would be more appropriate as resources, and only a single ad hoc resource. The modelling does not follow best practices (e.g., measurements are modelled as single statements, and include text preceding the numeric value), nor is it consistent. While some of the values follow phrase-based naming conventions, the values for sensor measurements are given as full sentences.

JLIS.it vol. 15, no. 1 (January 2024)
ISSN: 2038-1026 online

DOI: 10.36253/jlis.it-547

- **Maturity rating.** This Comparison also has a maturity level of 5. All measures are met, except for Instance Completeness (Level 2).
- **Usefulness.** The subject librarian determined that this Comparison would be useful for a specialist in the field, but would not be suitable for a layperson as it is too specific.

5.6 Smart cities and cultural heritage

This Comparison contains 26 predicates, 104 resources, and 5 literals and has a maximum subgraph depth of 2 nodes. It contains surveys, frameworks, and methodology publications reporting broad insights on modelling cultural heritage in smart cities, according to varied definitions of smart urban development.

- **Semantic modelling.** This Comparison contains many ad hoc resources instead of ontology resources. There are no measurements to evaluate, but the modelling is inconsistent and confusing. Several properties apply to only a handful of contributions, leading to a Comparison with many empty values. Full-sentences are given as values for multiple properties, including those describing the research problem and results.
- **Maturity rating.** This Comparison has a maturity level of 2. It satisfies Timeliness and Documentation Completeness, but not Instance Completeness or Property Completeness (Level 2). It does not satisfy Conciseness, but it does satisfy Data Representation (Level 3). It satisfies Linkability (Level 5).
- **Usefulness.** For specialists in the field, the subject librarian was unable to judge if this Comparison would be useful. They deemed it useful for laypeople.

	1D Fiber	Micro-needle Tech.	Supply Chain	CO_2 Gas Flux	Glucose Sensors	Smart Cities
Ontology resources	✔	✔	✔	✘	✘	✘
External links	✔	✘	✔	✘	✘	✘
Modelling best practices	✔	✔	❍	✘	✘	❍
Consistency	✔	✔	✔	✘	✘	✘
Relevant properties	✔	✔	✔	✘	✔	✘
Phrase-based names	✔	✔	✔	✔	✘	✘

Table 3: The results of the semantic modelling evaluation. The ✔ symbol indicates the presence of the given element while the ✘ symbol indicates its absence. The ❍ symbol indicates this element is not applicable to the given Comparison.

JLIS.it vol. 15, no. 1 (January 2024)
ISSN: 2038-1026 online

DOI: 10.36253/jlis.it-547

	1D Fiber	Micro-needle Tech.	Supply Chain	CO_2 Gas Flux	Glucose Sensors	Smart Cities
Responsiveness**	★	★	★	★	★	★
Licence**	★	★	★	★	★	★
Syntactic Accuracy*	★	★	★	★	★	★
Easiness*	★	★	★	★	★	★
Provenance**	★	★	★	★	★	★
Timeliness*	✔ (✔)	✘ (✘)	✘ (✔)	✔ (✘)	✔ (✔)	✔ (✔)
Instance Completeness*	✔ (✔)	✘ (✔)	✘ (✔)	✔ (✔)	✘ (✔)	✘ (✘)
Property Completeness*	✘ (✔)	✔ (✔)	✔ (✔)	✔ (✔)	✔ (✔)	✘ (✔)
Documentation Comp.*	✘ (✘)	✘ (✘)	✘ (✔)	✔ (✔)	✔ (✘)	✔ (✔)
Reusability**	★	★	★	★	★	★
Conciseness**	✘ (✔)	✔ (✘)	✔ (✔)	✔ (✘)	✔ (✔)	✘ (✘)
Data Representation	✘ (✘)	✔ (✔)	✔ (✔)	✘ (✔)	✔ (✔)	✔ (✔)
Trackability**	★	★	★	★	★	★
Queryability*	★	★	★	★	★	★
Identifier Stability*	☆	☆	☆	☆	☆	☆
Dereferencability	★	★	★	★	★	★
Linkability	✘ (✘)	✘ (✘)	✔ (✘)	✔ (✘)	✔ (✔)	✔ (✘)
Maturity Level	2 (5)	1 (2)	1 (5)	5 (2)	5 (5)	2 (2)

Table 4. Whether a Comparison satisfies the surveyed measures, according to a subject librarian (or in the opinion of the KGMM developer), and the maturity level resulting from these judgments. The ✔ symbol indicates that it satisfies the given measure, while the ✘ symbol indicates that it does not. The ★ symbol indicates measures which are automatically satisfied by the ORKG publishing model, while ☆ means the mechanism for satisfying this measure is present but not guaranteed to be applied. Measures marked with double asterisks are essential, those with one asterisk are important, and those without an asterisk are useful.

	1D Fiber	Micro-needle Tech.	Supply Chain	CO_2 Gas Flux	Glucose Sensors	Smart Cities
Expert	✘	✘	❍	❍	✔	❍
Layperson	❍	✘	✘	✘	✘	✔

Table 5. Whether a Comparison is useful to an expert or layperson, in the opinion of a subject librarian. The ✔ symbol indicates it is useful, the ✘ symbol indicates not useful, and the ❍ symbol indicates the subject librarian was unsure or felt unable to judge.

6. Discussion

6.1 Evaluation Results

The KGMM is still in its early development stages. Many of the measures are inherently difficult to evaluate, in that they are both subjective and challenging for even domain experts to assess. As some measures lack proposed survey questions and are not addressed by the ORKG, it was necessary to exclude four measures from Level 2 – of which two were classified as essential – from our implementation. Additionally, we classify Timeliness as important and not essential because we observe it is encumbered by the research field productivity artefact. In a highly productive field,

JLIS.it vol. 15, no. 1 (January 2024)
ISSN: 2038-1026 online

DOI: 10.36253/jlis.it-547

e.g., Computer Science, where the publication cycle is rapid, if Timeliness were essential a Comparison for that field would never cross the next maturity stage; on the other hand, a Comparison for a low productive field, e.g., Mathematics, would face no such hindrance. This insight about an essential measure halting progression through the KGMM relates more broadly to one of our findings. Though the KGMM is a hierarchical framework, its stagewise measures can be entirely independent between stages. Therefore, it is possible for a KG to receive a low maturity rating, even while satisfying the essential measures of higher levels. This may give the impression that the amount of improvement or development necessary for a KG to attain a high maturity level is much greater than it actually is.

The KGMM takes a holistic approach to quality evaluation which incorporates a broad range of criteria and standards, some of which are also influenced by external factors such as UI. While care has been taken in the development of the KGMM to ensure a minimum level of clean and accurate data modelling, it cannot replace an assessment focused on and specific to semantic modelling. Thus, we observe no correspondence between the semantic modelling evaluation and the KGMM rankings.

With regard to the usefulness evaluation, we note that the results are heavily influenced by the subjectivity of the domain experts consulted. Moreover, ORKG Comparisons may be used for a variety of independent purposes (e.g., understanding a research problem, finding datasets, identifying potential collaborators, etc), and whether a Comparison is useful for one purpose does not determine its usefulness for another. Nonetheless, two interesting points of consideration emerge from this assessment. First, none of the Comparisons were judged to be useful for both specialists and laypeople, and in the case of those that were not useful for either, it was for different reasons. Second, semantic modelling which is highly machine-actionable is often not highly human-actionable and may be difficult for readers to interpret. Similarly, flawed semantic modelling often simplifies the KG and presents information in a more intuitive structure, so reduced machine-actionability may result in increased human-actionability.

6.2 Differing Information Objectives of ORKG Comparisons

A good quality evaluation of ORKG Comparisons should consider the Comparison's information objective as seen by its creator. For instance, Comparison of Water Domain Ontologies provides an overview of multiple ontologies developed for describing water and water-related systems, such as water quality management and solar-powered water heating. The Comparison shows generic ontology aspects such as number of classes, number of data properties, number of object properties, and links to the ontologies themselves. However, if someone were seeking qualitative information on the theme of water domain ontologies showing the reader how the knowledge capture objective of each of the ontologies differ from or are similar to each other, this Comparison would not suffice. While it is briefly modelled in terms of the 'ontology component' property in the given Comparison, a new Comparison that directly addresses the new information objective sought would be ideally befitting.

JLIS.it vol. 15, no. 1 (January 2024)
ISSN: 2038-1026 online

DOI: 10.36253/jlis.it-547

6.3 Limitations of Our Survey Methodology

Using a survey to solicit feedback from KG users has two notable advantages: it is easy to source the feedback, and the relevant users are consulted directly. However, this approach also has some limitations. One limitation is that the survey takers must be knowledgeable about both semantic modelling and the relevant field of study. As shown in Table 4, the results of the KGMM survey are dramatically different when taken by the developer of the KGMM, who is not a domain expert in any of the Comparison fields but is an expert in KGs, compared with the subject librarians. In five cases the developer identified additional external ontologies which ought to be linked within the given Comparison, which were only identified by subject librarians in two cases. Conversely, when asked if the properties generalise well enough for additional contributions to be added, the developer answered affirmatively in five cases, where subject librarians answered affirmatively in only two cases. Determining the generalizability of the properties requires an understanding of the associated research landscape, which a domain expert is best equipped to assess. Another limitation is that a binary response does not befit all question types. For instance, a question asking about the descriptiveness of the Comparison documentation would best be addressed by maybe a response continuum bar rather than a simple yes or no.

7. Conclusion

This work presents a case study evaluating ORKG Comparisons based on three quality criteria. First, the quality of the semantic modelling in each KG was assessed by an expert in this field. Second, subject librarians completed a survey evaluating the KGs on a research topic from within their domain of expertise. This survey elicited feedback relevant to establishing a KG's maturity level, as defined in Hussein et al. (2022). Third, the same subject librarians assessed whether they believed the given KG would be useful for informing either a domain expert or a layperson on the research theme. No Comparison was highly rated in all approaches, nor were any judged useful for both domain experts and laypeople. While the study's small sample size of only six subgraphs from a single KG and the subjective nature of the evaluation criteria limit the statistical rigour and generalizability of the conclusions, they are nonetheless suggestive. These results highlight that quality criteria must be defined independently according to each use case, and assessing multiple quality dimensions in aggregate measures is challenging.

Some avenues for future work include the possibility of incorporating explicit semantic modelling evaluation into the maturity model, reformulating the maturity model to allow for answers other than a binary yes or no, and developing objective measures for assessing maturity. More generally, for any quality evaluation framework, it is necessary to assess quality by identifying independent quality criteria and reflecting this independence in the overall quality assessment.

JLIS.it vol. 15, no. 1 (January 2024)
ISSN: 2038-1026 online

DOI: 10.36253/jlis.it-547

References

Auer, Sören, Allard Oelen, Muhammad Haris, Markus Stocker, Jennifer D'Souza, Kheir Eddine Farfar, Lars Vogt, Manuel Prinz, Vitalis Wiens, and Mohamad Yaser Jaradeh. 2020. "Improving Access to Scientific Literature with Knowledge Graphs." *Bibliothek Forschung Und Praxis* 44 (3): 516–29. https://doi.org/10.1515/bfp-2020-2042.

Berners-Lee, Tim, James Hendler, and Ora Lassila. 2001. "The Semantic Web." *Scientific American* 284 (5): 34–43. https://www.lassila.org/publications/2001/SciAm.pdf.

Bizer, Christian, and Richard Cyganiak. 2009. "Quality-Driven Information Filtering Using the WIQA Policy Framework." *Journal of Web Semantics* 7 (1): 1–10. https://doi.org/10.1016/j.websem.2008.02.005.

Bornmann, Lutz, and Rüdiger Mutz. 2015. "Growth Rates of Modern Science: A Bibliometric Analysis Based on the Number of Publications and Cited References: Growth Rates of Modern Science: A Bibliometric Analysis Based on the Number of Publications and Cited References." *Journal of the Association for Information Science and Technology* 66 (11): 2215–22. https://doi.org/10.1002/asi.23329.

Daniel, Florian, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. 2018. "Quality Control in Crowdsourcing: A Survey of Quality Attributes, Assessment Techniques, and Assurance Actions." *ACM Computing Surveys* 51 (1). https://doi.org/10.1145/3148148.

Färber, Michael, Basil Ell, Carsten Menne, Achim Rettinger, and Frederic Bartscherer. 2018. "Linked Data Quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO." *Semantic Web Journal* 9 (2): 77–129. Accessed March 15, 2023. https://doi.org/10.3233/SW-170275.

Hussein, Hassan, Allard Oelen, Oliver Karras, and Sören Auer. 2022. "KGMM - A Maturity Model for Scholarly Knowledge Graphs Based on Intertwined Human-Machine Collaboration." In *From Born-Physical to Born-Virtual: Augmenting Intelligence in Digital Libraries: 24th International Conference on Asian Digital Libraries, ICADL 2022*, 253–69. Hanoi, Vietnam: Springer-Verlag. https://doi.org/10.1007/978-3-031-21756-2_21.

Johnson, Rob, Anthony Watkinson, and Michael Mabe. 2018. "The STM Report: An Overview of Scientific and Scholarly Publishing. 5th Edition." https://www.stm-assoc.org/2018_10_04_STM_Report_2018.pdf.

Oelen, Allard, Mohamad Yaser Jaradeh, Kheir Eddine Farfar, Markus Stocker, and Sören Auer. 2019. "Comparing Research Contributions in a Scholarly Knowledge Graph." In *SciKnow 2019: Third International Workshop on Capturing Scientific Knowledge*, 21-26. https://doi.org/10.15488/9388.

Oelen, Allard, Mohamad Yaser Jaradeh, Markus Stocker, and Sören Auer. 2020. "Generate FAIR Literature Surveys with Scholarly Knowledge Graphs." In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, 97–106. https://doi.org/10.34657/5212.

Pipino, Leo L., Yang W. Lee, and Richard Y. Wang. 2002. "Data Quality Assessment." *Communications of the ACM* 45 (4): 211–18. https://doi.org/10.1145/505248.506010.

JLIS.it vol. 15, no. 1 (January 2024)
ISSN: 2038-1026 online

DOI: 10.36253/jlis.it-547

Shotton, David. 2009. "Semantic Publishing: The Coming Revolution in Scientific Journal Publishing." *Learned Publishing* 22 (2): 85–94. https://doi.org/10.1087/2009202.

Wang, Richard Y., and Diane M. Strong. 1996. "Beyond Accuracy: What Data Quality Means to Data Consumers." *Journal of Management Information Systems* 12 (4): 5–33. https://doi.org/10.1080/07421222.1996.11518099.

Wilkinson, Mark D, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3 (1): 1–9. https://doi.org/10.1038/sdata.2016.18.

Zaveri, Amrapali, Anisa Rula, Andrea Maurino, Ricardo Pietrobon, Jens Lehmann, and Sören Auer. 2016. "Quality Assessment for Linked Data: A Survey." *Semantic Web* 7 (1): 63–93. https://doi.org/10.3233/SW-150175.

Zhang, Jing. 2022. "Knowledge Learning With Crowdsourcing: A Brief Review and Systematic Perspective." *IEEE/CAA Journal of Automatica Sinica* 9 (5): 749–62. https://doi.org/10.1109/jas.2022.105434.