A reproducible framework to assess the quality of linked open data in GLAM datasets

Gustavo Candela [a,*], Meltem Dişli [b], Milena Dobreva [c] and Sally Chambers [d]
[a] *Department of Software and Computing Systems, University of Alicante, Spain*
E-mail: gcandela@ua.es
[b] *Department of Information Management, Hacettepe University, Turkey*
E-mail: meltem.disli@hacettepe.edu.tr
[c] *Computer and Information Sciences Department, University of Strathclyde, Scotland*
E-mail: milena.dobreva@strath.ac.uk
[d] *Ghent Centre for Digital Humanities, Ghent University, Belgium*
E-mail: sally.chambers@ugent.be

Abstract. Over the last decade GLAM (Galleries, Libraries, Archives and Museums) organizations have been exploring new ways to make their content available using the Semantic Web and Linked Open Data (LOD). The growing number of GLAM institutions converting their collections to LOD, coupled with the increasing demand for high-quality data, has made the assessment of LOD quality a critical concern. In addition, there has been a significant increase in the global interest among researchers in reproducible research, a cornerstone of Open Science, requiring code to generate the experimental results. This study aims to present a reproducible framework to assess LOD quality within the GLAM sector. Based on the literature, a number of data quality criteria were established including 4 dimensions and 32 criteria. Subsequently, six LOD datasets were assessed according to these criteria. The assessments revealed that the LOD datasets performed well on accessibility, while they did not yield satisfactory results for other criteria, such as contextual information. These results can serve as a benchmark for other LOD datasets. Additionally, the study provides a detailed analysis that could be beneficial for other organizations and researchers interested in making their digital collections accessible and reusable as LOD. The study concludes by identifying further research based on the implementation in real practice for data spaces, the Cultural Heritage Cloud (ECCCH) and FAIR advancement in GLAMs.

Keywords: Libraries, Linked Open Data, Collections as data, Reproducible code, Data quality, Semantic web, Open Science

1. Introduction

GLAM (Galleries, Libraries, Archives, and Museums) institutions host rich collections including a digitized variety of analogue materials resulting in texts, images, audio, video or 3D objects supplied with metadata. While digitization of these collections has significantly enhanced their accessibility, GLAM institutions are continually exploring new technological avenues to enrich, publish, and facilitate the reuse of their collections. New initiatives have recently emerged in order to publish digital collections suitable for computational use, enabling the application of Artificial Intelligence and Machine Learning techniques [38, 43]. In parallel, during the last decade, there has been a significant increase in the global interest among researchers in reproducible research, a cornerstone of Open Science [3]. Reproducible research results such as code are crucial for the community to generate the experimental results [26]. In this sense, Jupyter notebooks have emerged as a popular tool in the GLAM context for

*Corresponding author. E-mail: gcandela@ua.es.

data exploration and analysis in the cloud, providing a combination of scripts and documentation, and an interactive interface [9].

The Semantic Web and Linked Open Data (LOD) was introduced as an extension of the traditional web to provide machine-readable content based on standards such as the Resource Description Framework (RDF) [5, 63]. Numerous institutions have utilized these standards to publish and enrich their catalogs in the form of LOD, employing interoperable and standardized vocabularies and external datasets. In this context, collaborative-edition approaches, such as Wikidata, have played a leading role in the enrichment of datasets by defining particular properties per institution to link resources to, such as geographic locations, subjects, authors and works [14]. Examples of LOD datasets can be found in various archives, libraries, and museums [22, 55].

Data quality is crucial in order to reuse the content to its full potential [56]. Previous research has identified issues concerning data quality across several domains, such as Natural Language Processing and Semantic Web [6, 59]. In this sense, several approaches have focused on the assessment of the data quality of the content provided by the digital collections, providing data quality criteria covering several aspects, such as *understandability*, *accuracy* or *timeliness*, with which to assess the data [11, 27]. However, despite all these efforts to assess data quality in LOD datasets, none of the previous work provides a reproducible approach including scripts, therefore enabling users to replicate the results.

The purpose of this work is to provide a reproducible framework to assess the quality of LOD in GLAM institutions. To achieve this goal, data quality criteria were first established based on the literature. Subsequently, the quality of selected LOD datasets was evaluated according to these criteria. The intended audience of this work is library managers, digital curators and Digital Humanities researchers interested in curating, exploring and reusing the LOD provided by GLAM institutions for research projects. In terms of recent work, this research: i) builds capacity of academic librarians in understanding data quality issues and assessment [37]; promotes the publication of high-quality data required for new data infrastructures such as the common European data space for cultural heritage[1]; and assists in understanding how data quality issues such as inconsistency and lack of provenance documentation hinders integration and searching accross datasets as well as diminish research value [2, 56].

The main contributions of this work are: i) a reproducible framework for LOD quality assessment in GLAM; and ii) the provision of reproducible scripts and examples in the form of Jupyter Notebooks. These contributions are intended to foster the reproducibility in research, leverage the use of LOD in GLAM institutions and to encourage Digital Humanities researchers to reuse the LOD datasets for their research projects. It is important to notice that there are plenty of LOD developments but a relatively modest exploration of the quality aspects which allow upscaling the use of resources in the context of data spaces [25], Cultural Heritage Cloud (ECCCH),[2] which are also vital for FAIR [60] advancement in GLAM.

The paper is structured as follows: following a brief overview of the current state of LOD in the GLAM domain and approaches to data quality, we introduce the framework employed for assessing LOD datasets and subsequently detail our findings. Finally, the paper concludes with a summary of the adopted framework and outlines potential avenues for future research.

2. Related work

This section introduces a diverse list of institutions that have made their digital collections available as LOD. It also provides a literature review on the methods and techniques used to assess and document data quality in the context of LOD, as well as emerging initiatives to support reproducible research.

2.1. Publishing data as LOD

Publishing data as LOD enhances data accessibility by making it interconnected, reusable, and machine-readable. In today's globalized world, the importance of LOD is increasing, resulting in a rise in the publication of LOD.

[1] https://www.dataspace-culturalheritage.eu/en
[2] https://www.echoes-eccch.eu/

Many institutions from different domains such as government, agriculture, cultural heritage and environment have explored the benefits of the application of the Semantic Web to their datasets [48]. These institutions have explored the use of the Semantic Web to publish content in the form of RDF datasets.

In line with the FAIR principles [60], LOD is particularly essential for GLAM because it enhances the discoverability, interoperability, and reusability of their digital collections [60]. LOD supports computational research on collections, enabling deeper understanding and the discovery of new insights. In this regard, GLAM initiatives have promoted the reuse of their digital collections employing the Semantic Web as a basis, in libraries, museums and archives [11, 18, 22, 31, 35, 54]. Initiatives like Europeana and LUX: Yale Collections Discovery,[3] aggregate content from several cultural heritage datasets, displaying it through web interfaces or making it available via APIs. The Golden Agents project[4] is another notable example of LOD initiative focused on the Dutch Golden Age. In these initiatives, resources from different Cultural Heritage datasets or external datasets are brought together. The use of external datasets to enrich information is increasingly prevalent in GLAM [14]. For instance, Wikidata, a collaboratively edited cross-domain dataset, has gained significant popularity in the GLAM sector. Additional examples include geographical databases such as GeoNames[5] or authority files such as Virtual International Authority File (VIAF).[6]

In this context, data modeling plays a vital role in ensuring the quality and interoperability of LOD. Best practices recommend reusing vocabularies, preferably standardized ones, in order to increase interoperability, reduce redundancies and encourage reuse of the data [62, 64]. Table 1 shows an overview of LOD datasets made available to the public by organizations obtained as part of a comprehensive collection of scholarly articles retrieved from three major academic databases: Scopus, Web of Science and ACM Digital Library. The search was conducted in September 2024 using strings such as Semantic Web, Linked Open Data, digital library, GLAM and Cultural Heritage. The screening of the datasets was conducted manually by the authors, removing all the work not focused on the CH and GLAM sectors. The ontologies most used to describe the metadata are Bibliographic Framework (BIBFRAME) and CIDOC Conceptual Reference Model (CIDOC-CRM) [4, 17], with 5 and 4 selected datasets respectively.

Table 1
Overview of LOD datasets published by organizations.

Institution	Vocabulary	URL
Austrian National Library	EDM, BIBFRAME, RDA	https://labs.onb.ac.at/en/dataset/lod
Biblioteca Nacional del Congreso Nacional de Chile	Dublin Core	https://datos.bcn.cl/es/endpoint-sparql/
Biblioteca Virtual Miguel de Cervantes	RDA	https://data.cervantesvirtual.com
Bibliothèque Nationale de France	FRBR	https://data.bnf.fr
Database of Chinese Rare Books	BIBFRAME	https://data.ascdc.tw/en/sparql.php
Europeana	EDM	https://pro.europeana.eu/page/sparql
German National Library	BIBFRAME	https://www.dnb.de/EN/lds
Getty Research Institute	CIDOC-CRM	https://data.getty.edu/provenance
LetterSampo Correspondence	CIDOC-CRM	http://ldf.fi/corresp/sparql
Library of Congress	BIBFRAME	https://id.loc.gov
Linked Open Data at SAAM	CIDOC-CRM	http://edan.si.edu/saam/sparql
National Digital Data Archive of Hungary	DC, FOAF, schema.org, DBpedia	https://lod.sztaki.hu/sparql
National Library of Finland	schema.org, BIBFRAME	https://data.nationallibrary.fi
National Library of the Netherlands	schema.org, LRM	https://data.bibliotheken.nl
National Library of Sweden	KB Base Vocabulary	https://libris.kb.se/sparql
Rijksmuseum	EDM, SKOS	https://data.rijksmuseum.nl/controlled-vocabularies
The Nobel Prize	nobel, FOAF	https://data.nobelprize.org/sparql
World War I as LOD	CIDOC-CRM, ww1lod	https://www.ldf.fi/dataset/ww1lod
Zeri Photo Archive	CIDOC-CRM	http://data.fondazionezeri.unibo.it/sparql

[3] https://lux.collections.yale.edu
[4] https://www.goldenagents.org/
[5] https://www.geonames.org
[6] https://viaf.org

2.2. Assessing data quality in LOD

Best practices and guidelines to publish data are provided in several initiatives such as FAIR [60] and Collections as data [12, 43]. Previous work based on digital libraries, in particular DELOS DLRM (Digital Library Reference Model), identified quality as one of the 6 core concepts [16]. While this work was applicable to digital libraries and not focused on datasets, there is no comprehensive reference model on the quality of datasets. These initiatives recommend enabling high-quality access to data. As the demand of high-quality data continues to grow, there is a greater need for accurate and reliable data to support the requirements of the advances in technology [47]. Ensuring data accuracy and reliability can be achieved through data quality assessments. Data quality is a wide concept that covers various aspects, including: *consistency*, *conformity*, *provenance*, *reproducibility*, *completeness* or *documentation* [27, 68]. To maintain effective data quality, it is recommended to conduct automated quality assessments at regular intervals [70]. A recent survey found that most data quality assessment work relays on a single criterion and often use DBpedia as the dataset [41]. Previous work has assessed the data quality of LOD in the Cultural Heritage sector by using several dimensions [11, 53]. Other work has focused on assessing the quality of Linked Data in European Open Government Data [32]. Larger datasets, such as Wikidata, have also been assessed in terms of data quality [27, 51]. More general approaches to assess and improve the quality of LOD are based on profiling statistics, synonym relationships between predicates, and SPARQL [49]. Other approaches have addressed the assessment of metadata, particularly regarding special information such as uncertain information, competing hypotheses, and temporally evolving data, as provided in environments like Wikidata [44]. Additional initiatives are based on the use of languages such as Shape Expressions (ShEx) and Shapes Constraint Language (SHACL) to define node constraints in RDF datasets and assess its quality [65, 66]. ShEx is human-readable and widely used for different purposes in some initiatives including assessing data quality in Cultural Heritage [6, 13], validating items in Wikidata [50], and sharing RDF data models [57]. In this sense, accessibility is a key aspect when considering the FAIR principles in order to be aligned with existing infrastructures such as data spaces and the ECCCH.

Several approaches have focused on the definition of machine-readable and interoperable vocabularies to describe the content in terms of *provenance*, *description*, *access*, *data quality* and *visualization* [64]. For instance, Vocabulary of Interlinked Datasets (VoID) provides terms and patterns for describing RDF datasets [61] and Data Catalog Vocabulary (DCAT) was designed to facilitate interoperability between data catalogs [67]. Some of them enable the detailed description of data quality assessments through machine-readable metadata. In particular, Data Quality Vocabulary (DQV) is an extension of Data Catalog Vocabulary (DCAT) promoted by the W3C. It provides a list of properties and classes suitable for expressing the quality of a dataset such as *dataset*, *dimension*, *category* and *metric* [1]. Other examples of data quality ontologies define common and reusable concepts in terms of data quality [20, 28, 33]. Further approaches have explored the use of controlled vocabularies to describe datasets in terms of metadata and data quality using the Dataset Characteristics and Quality (DCQ) ontology [42]. Some of the benefits of using ontologies to describe data quality are: i) interoperability thanks to the use of machine-readable vocabularies; ii) knowledge sharing by providing additional documentation; and iii) potential reuse by the research community [58]. Table 2 provides an overview of existing vocabularies to describe data access and quality.

Existing data quality assessment tools based on LOD datasets, such as KGHeartBeat [45], are based on a web interface that provides as a result charts and tables. Others are focused on specific data quality criteria such as accessibility and consistency [6, 39]. Most of the projects provide a GitHub repository with different content including web tools or Python scripts. Others provide a language to define the constraints that can be used to assess the quality of a dataset [19]. However, such projects have the following limitations hindering reuse and adoption: i) they need to be installed in a server or a local environment; ii) they are executed in a linear fashion, which makes it difficult for users to understand how the code works; and iii) they can not be directly executed in cloud services such as EOSC or Binder. In contrast, a reproducible approach would benefit from: a) the use of reproducible scripts opening up new opportunities in terms of extension and adaptation to new datasets and data quality criteria; b) the use of ontologies to describe the data quality assessment (e.g., how criteria can be assessed as well as the results obtained) facilitating a transparent and interoperable method; and c) the ability to be executed in cloud environments enabling scalability. Table 3 shows an analysis of the main features provided by previous work concerning data quality.

Reproducibility is the ability to redo an experiment and get the same results [26]. Giving the dynamic nature of LOD datasets in terms of updates and modifications, it is crucial to maintain up-to-date data quality assessments in

Table 2

Vocabularies promoted by relevant institutions and widely used by the community to describe datasets and data quality. Note that the classes and properties columns provide a summary of the classes focused on data access and quality provided by each vocabulary.

Vocabulary	Description	Classes	Properties
Data Catalog Vocabulary (DCAT)	Description of datasets and data services in a catalog	dcat:Dataset, dcat:Distribution	dcat:downloadURL, dcat:landingPage
Dataset Characteristics and Quality (DCQ)	Vocabulary for attaching the results of quality benchmarking	dac:Dimension, dac:Category, dac:Metric	dac:hasValue
Data Quality Vocabulary (DQV)	Description of the data quality assessment of datasets	dqv:Dimension, dqv:Category, dqv:QualityMeasurement, dqv:Metric	dqv:isMeasurementOf, dqv:hasQualityMeasurement
Dataset Usage Vocabulary (DUV)	Description of metadata about dataset usage	duv:RatingFeedback, duv:Usage	duv:hasUsage, duv:hasFeedback, duv:hasRating
Schema.org	Description of structured data on the Internet	schema:Dataset	schema:query, schema:contributor
Vocabulary of Interlinked Datasets (VoID)	Description of linked datasets	void:Dataset	void:exampleResource, void:vocabulary
Wikidata	Wikidata vocabulary to describe resources	wd:Q1172284	wdt:P1659, wdt:P5305

Table 3

Analysis of previous work concerning data quality according to different aspects. Extensibility refers to the possibility to include new dimensions and criteria. Reproducibility refers to the possibility to reproduce the results with the scripts provided. Implementation support refers to the provision of code or instructions to enhance the project.

Reference	Dimensions	Extensibility	Reproducibility	Implementation support	Output format	Last release
[6]	Consistency	Yes	Yes	ShEx, SPARQL	Text	2023
[39]	Accessibility	Yes	Yes	Yes	RDF	2016
[45]	Availability, Licensing, Interlinking, Security, Performance, Semantic Accuracy, Consistency, Conciseness, Reputation, Believability, Verifiability, Currency, Volatility, Completeness, Amount of data, Representational-conciseness, Interoperability, Understandability, Interpretability, Versatility	Yes	Yes	Yes (Python scripts)	CSV	2024
[29]	Accuracy, Trustworthiness, Consistency, Relevancy, Completeness, Timeliness, Ease of understanding, Interoperability, Accessibility, License, Interlinking	Yes	No	No code provided	No	2018
[19]	Intrinsic, Representational, Contextual	Yes	Yes	Yes	Java	2016
[36]	Consistency	Yes	Yes	Yes	HTML, RDF	2015

order to meet reproducibility's aims. Various criteria for ensuring reproducibility have been outlined in the literature such as the accessibility of methods, data, and models used [40]. Another study identifies three dimensions of reproducibility: methods, results, and inferential [30]. Additionally, the availability of datasets and codes has been highlighted as key criteria for reproducibility [71]. In this context, ensuring reproducibility necessitates the detailed documentation of all stages and methodologies of the research. Moreover, it is imperative that the data, as well as the code, are openly accessible and designed for reusability. In the last years, Jupyter Notebooks have emerged as a powerful tool to provide reproducible code [46, 52]. A Jupyter Notebook combines code, textual documentation, visualizations, widgets, and charts, allowing researchers to iteratively develop their analysis [34]. In addition, best practices and guidelines have been made available to publish reproducible code that can be run in cloud environments such as Binder[7] and more advanced infrastructures such as supercomputing centers [8–10].

These efforts provide an extensive demonstration of how LOD made available by GLAM institutions can be assessed using different techniques, methods and criteria. Nevertheless, to the best of our knowledge, none of this research employs a reproducible framework for assessing the quality of LOD in the GLAM sector. This research aims to fill that gap by providing a machine-readable, reproducible and extensible method built on previous work. This approach can encourage GLAM institutions to assess the quality of their datasets and provide the results as additional documentation, as recommended by the GLAM community [12, 43].

[7] https://mybinder.org/

3. A framework to assess the quality of LOD

The framework proposed in this approach to assess the quality of LOD in GLAM institutions works in three steps: i) defining the data quality criteria to assess the LOD based on previous work and literature; ii) data modeling of the data quality assessment using existing vocabularies; and iii) assessing the LOD datasets. These steps are described in detail below. This approach aims to provide a reproducible method that can be applied to LOD datasets regardless of the vocabularies used to describe the metadata.

The framework was defined considering the following key features:

- it considers the dynamic nature of LOD, where datasets are regularly updated or modified, the framework could be used periodically in order to provide up-to-date data quality assessments.
- it employs widely used vocabularies and ontologies based on previous work [21].
- it enables its extension and adaptation by using a machine-readable format to define the data quality criteria and how it criterion can be assessed by means of SPARQL.
- its implementation follows best practices to take advantage of cloud services and infrastructure in order to overcome potential issues concerning resources and scalability.

Note that this is an initial attempt of a subselection of potential methods to evaluate the various criteria available. Subsequently, this work can be refined and enhanced.

3.1. Data quality criteria

The first step involves establishing the data quality criteria to assess the LOD. These criteria were selected based on a comprehensive literature review. Table 4 shows the data quality criteria based on previous work to assess the quality of LOD datasets. Additional details are provided in the Appendix B.

Note that the data quality criteria provided in this work have been adapted to the GLAM sector when possible. For instance, the adaptations include:

- the use of specific properties focused on the GLAM sector such as the ISNI and VIAF identifiers for the interpretability and conciseness criteria.
- a selection of properties, classes and resources used for the completeness criterion (see Table 7).
- a new criterion interlinking-Wikidata was defined in order to identify whether Wikidata provides a dedicated property.
- the semantic accuracy was measured using GLAM content, by checking a random sample of authors.
- the interlinking criterion was redefined in order to measure the connection of resources typed as any class related to authors such as *foaf:Person* or *def:C1005*.

In addition, most of the data quality criteria provided in this work can be measured by means of SPARQL. Our goal was to provide a reproducible means. The SPARQL scripts provided in this work cover in a general way the GLAM sector in terms of classes and properties. If required, they can be adapted to new datasets and vocabularies.

3.2. Use of formal models and vocabularies

The second step involves the employment of formal models and vocabularies for the quality assessment. This ensures that the assessment is machine-readable, reusable and interoperable.

An analysis of the vocabularies described in Table 2 was conducted, identifying the classes and properties to describe data quality in terms of LOD provided by GLAM institutions. Vocabularies were prioritized according to their suitability to describe data quality in terms of categories and data quality criteria. A summary of the selection of classes and properties is provided below.

- Organizations are described by means of the class `schema:Organization`.
- Datasets are mainly described using the vocabularies DCAT and VoID but also including properties from Wikidata to describe particular features of datasets such as the SPARQL endpoint (`wdt:P5305`) and the properties used to link resources (`wdt:P1659`).

Table 4

Summary of the data quality criteria including the dimensions, criteria, previous work, a combination of possible values including Boolean operators, ranges and ratios, and the issue addressed in the form of a question.

Dimension	Criteria	References	Values	Description
Accessibility	Availability	[11, 27, 39, 45, 68]	{0..1}	Is the public SPARQL endpoint available?
	Licensing	[11, 27, 68]	{0, 0.5, 1}	Is there information about the license available?
	Interlinking-Person	[11, 27, 45, 68]	{0..1}	To what the resources describing authors are linked to external datasets?
	Interlinking-Work	[11, 27, 45, 68]	{0..1}	To what the resources describing works are linked to external datasets?
	Interlinking-Wikidata	[11, 14, 24, 69]	{0, 1}	Is there one or more dedicated properties in Wikidata?
	Security	[27, 45, 68]	{0, 1}	Does the SPARQL provide HTTPS support?
	Performance	[11, 27, 45, 68]	{0, 1}	Is the dataset efficient when processing the requests?
Intrinsic	Syntactic validity	[11, 27, 45, 68]	{0..1}	Do RDF documents conform to the specification of the serialization format?
	Semantic accuracy	[11, 27, 68]	{0..1}	Are the metadata provided correct?
	Consistency	[6, 11, 27, 36, 45, 68]	{0, 1}	Is the consistency of statements with respect to class constraints correct?
	Conciseness	[11, 27, 45, 68]	{0..1}	Does the dataset contain redundant resources?
	Conciseness-Wikidata	[11, 24, 27, 45, 68]	{0..1}	Are external resources linking to duplicated records in the dataset?
	Completeness-Person	[11, 27, 68]	{0..1}	All required metadata about authors is present?
	Completeness-Work	[11, 27, 45, 68]	{0..1}	All required metadata about works is present?
	Completeness-Place	[11, 27, 45, 68]	{0..1}	All required metadata about places is present?
	Completeness-Population	[11, 27, 45, 68]	{0..1}	Are all the resources expected in the dataset?
Contextual	Relevancy	[11, 27, 45, 68]	{0, 1}	Does the dataset support a ranking of statements?
	Trustworthiness-Dataset	[11, 27, 45, 68]	{0, 0.5, 1}	How the dataset is curated?
	Trustworthiness-Unknown	[11, 27, 45, 68]	{0, 0.5, 1}	Does the dataset support unknown and empty values?
	Understandability-Labels	[11, 27, 45, 68]	{0, 1}	Do the resources include a label?
	Understandability-Vocabularies	[11, 27, 45, 68]	{0, 1}	Does the dataset provide information about the vocabularies employed?
	Understandability-URL Patterns	[11, 27, 45, 68]	{0, 1}	Does the dataset provide information about the URL patterns?
	Understandability-Examples	[11, 27, 45, 68]	{0, 1}	Does the dataset provide information about SPARQL examples?
	Timeliness	[11, 27, 45, 68]	{0, 0.5, 1}	Does the dataset provide information concerning the publication of the dataset?
Representational	Representational conciseness-URIs Length	[11, 20, 27, 45, 68]	{0, 1}	Does the dataset use long URIs?
	Representational conciseness-Containers	[11, 20, 27, 45, 68]	{0, 1}	Does the dataset use RDF primitives?
	Interoperability	[11, 20, 27, 45, 68]	{0, 1}	Were existing vocabularies reused to describe the metadata?
	Interpretability-VIAF	[11, 20, 27, 45, 68]	{0, 1}	Are unique identifiers based on VIAF used?
	Interpretability-ISNI	[11, 20, 27, 45, 68]	{0, 1}	Are unique identifiers based on ISNI used?
	Interpretability-Labels	[11, 20, 27, 45, 68]	{0, 1}	Is the language used when providing labels?
	Versatility-Serialization	[11, 20, 27, 45, 68]	{0, 1}	Are the data available in different serialization formats?
	Versatility-Multilingual	[11, 20, 27, 45, 68]	{0, 1}	Are the data available in different languages?

– Metadata about dataset usage is described by DUV and it uses DCAT vocabulary `dcat:Dataset` class and all properties associated with this class.
– Data quality is described by means of the Data Quality Vocabulary including dimensions and metrics as well as the query (`schema:query`) and its description (`schema:description`) with which to assess each of them.

By utilizing these vocabularies, the framework can achieve a high level of interoperability and reusability, ensuring that the quality assessments can be easily integrated with other datasets and systems. Figure 1 describes the main classes and properties used in the data modeling process. A resource with the type `void:Dataset` is related to a resource with the type `dqv:QualityMeasurement`, by means of a `dqv:hasQualityMeasurement` property. It represents the evaluation of a given dataset against a specific data quality criteria. The resource with the type `dqv:QualityMeasurement` is related to several resources with the type `dqv:Metric`, describing the different criteria used to assess the dataset. Each of them include additional information such as the `schema:query` and `schema:description` properties. As an example, Listing 1 shows an overview of the metadata based on the Getty Provenance Index (GPI) dataset using the classes and properties proposed in this work.

```
:myGettyDataset a void:Dataset;
    dcterms:title "Getty Provenance Index";
    dcterms:creator :org_GETTY;
    void:sparqlEndpoint <https://data.getty.edu/provenance/sparql>;
    wdt:P5305 <https://data.getty.edu/provenance/sparql>;
    void:dataDump <https://github.com/thegetty/provenance-index-csv>;
    void:vocabulary <http://www.cidoc-crm.org/cidoc-crm/>;
    dqv:hasQualityMeasurement :myGettyMeasurementAvailability ;
```

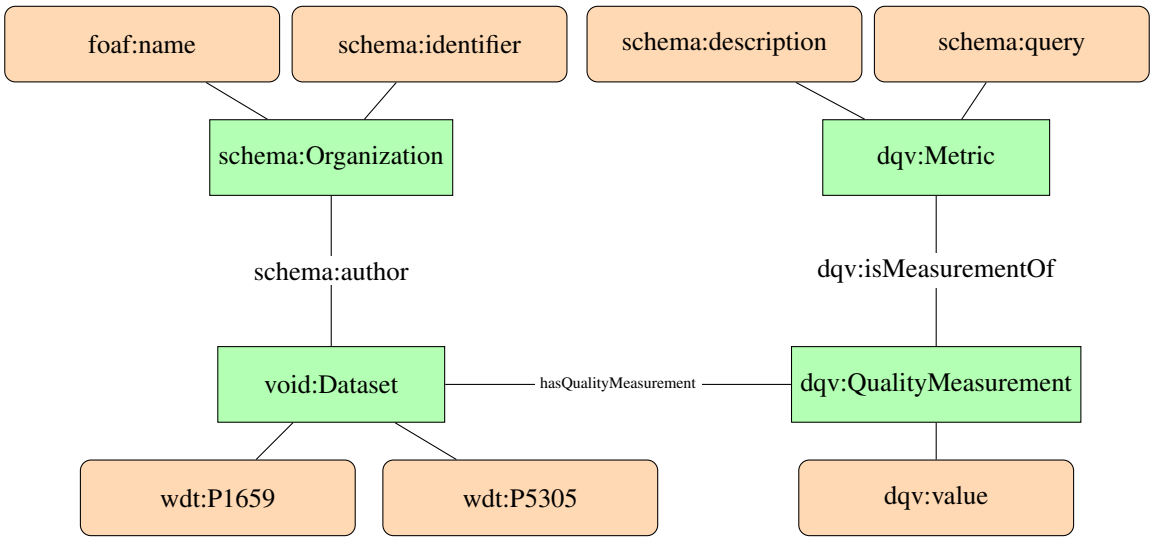

Fig. 1. Overview of classes and properties used to describe the data quality assessment of a given dataset against a specific data quality criteria. A resource with the type `void:Dataset` is related to a resource with the type `dqv:QualityMeasurement`, by means of a `dqv:hasQualityMeasurement` property. A resource with the type `dqv:QualityMeasurement` is related to several resources with the type `dqv:Metric`, describing the different criteria used to assess the dataset. Each of them include a `schema:query` representing the SPARQL to be employed. The properties `wdt:P5305` and `wdt:P1659` correspond to the SPARQL endpoint and related property in the Wikidata vocabulary, respectively.

```
    dqv:hasQualityMeasurement :myGettyMeasurementLicensing ;
    dqv:hasQualityMeasurement :myGettyMeasurementInterlinkingWikidata ;
    dqv:hasQualityMeasurement :myGettyMeasurementPerformance ;
    .

:org_GETTY a foaf:Organization;
    rdfs:label "Getty" ;
    foaf:homepage <https://www.getty.edu/> ;
    .

:availabilityMetric
    a dqv:Metric ;
    skos:prefLabel "Availability" ;
    skos:definition "It checks that the SPARQL endpoint is available"@en ;
    dqv:inDimension ldqd:availability ;
    dqv:expectedDataType xsd:decimal ;
    schema:query "SELECT * WHERE {{?s ?p ?o}} LIMIT {0}" ;
    schema:description "It checks that the SPARQL endpoint is available"
    .

:myGettyMeasurementAvailability
    a dqv:QualityMeasurement ;
    dqv:computedOn :myGettyDataset ;
    dqv:isMeasurementOf :availabilityMetric ;
    dqv:value "1"^^xsd:decimal
    .
```

Listing 1: An example of metadata generated based on the Getty Provenance Index dataset and using the vocabularies and data modeling provided in this work.

3.3. Assessment

The last step involves the actual assessment of the LOD datasets based on the defined data quality criteria and the proposed data modeling. The results obtained from this assessment can be used as additional documentation for datasets, describing essential features for potential reusers [2, 12].

As the research community fosters reproducibility, it encourages the reuse and replication of the obtained results [9, 46]. Jupyter Notebooks have emerged as a powerful tool to combine text, code, charts, interactive widgets

and documentation [9]. In addition, recent initiatives have been created to improve software in research such as the Software Sustainability Institute to promote best practices and guidelines.[8]

In addition, and as more LOD datasets are created as larger-scale datasets, cloud services and infrastructures can be used to perform the data quality assessment. Depending on the size of the datasets and the requirements, some examples are Binder, the European Open Science Cloud (EOSC) or supercomputing center providing services such as JupyterHub.[9]

4. Results

The framework proposed in this work was applied to a selection of LOD datasets made available by relevant GLAM institutions. The criteria to select the datasets were country of origin, language, availability of a SPARQL endpoint and license enabling reuse.

Data quality of LOD datasets has been assessed according to four dimensions and 32 criteria. Table 5 shows the assessment results based on data quality criteria proposed in this work. The assessments were conducted in November 2025.

The information related to the LOD datasets assessed concerning quality criteria within the scope of the research is as follows:

- **datos.bne.es**,[10] is a LOD platform based on bibliographic data published by the National Library of Spain (BNE).
- **data.bnf.fr**,[11] is a LOD project published by the National Library of France (BnF).
- **LetterSampo:correspSearch**,[12] is a project that provides approximately 130,000 historical letters as LOD involving around 17,000 actors, showcasing Finnish language, culture, and history (LDFI).
- **Biblioteca Nacional del Congreso Nacional de Chile (BNC)** is a LOD project.[13]
- **Database of Chinese Rare Books (CRB)** is a LOD dataset describing 900 Chinese rare books and manuscripts and created by the Academia Sinica Center for Digital Cultures (ASCDC) and based on the metadata from the online database of the "Digital Library of Chinese Rare Books".[14]
- **Getty Provenance Index (GPI)** is a LOD project providing access to resources from dealer stock books, sales catalogs, and archival inventories.[15]

A GitHub project [15] was created that includes: i) a README file describing the content and a brief installation guide; ii) the RDF data generated as part of the data modeling step described in the framework; and iii) an interface based on Jupyter Notebooks to enable the reproducibility of the results that reuses the RDF data generated.[16] Figure 2 shows the interface created as part of this work to enable users to reproduce the data quality assessment results. The SPARQL scripts provided in the data quality criteria were adjusted to the selection of datasets employed in this work. To do so, each dataset was analyzed in order to identify the classes and properties employed to describe the metadata and adjust the SPARQL scripts. As a result, a configuration file per dataset describing the data quality criteria was generated.

In order to employ the scripts provided in this work with other datasets, a public SPARQL endpoint is required. A new configuration file containing the data quality criteria should be created using the examples provided as a basis. If required due to the employment of different vocabularies, the SPARQL queries can be refined in order to cover additional properties and restrictions. Figure 3 shows a workflow diagram describing how the scripts work.

[8] https://www.software.ac.uk/
[9] https://jupyterhub.readthedocs.io
[10] https://datos.bne.es/inicio.html
[11] https://data.bnf.fr/
[12] https://www.ldf.fi/dataset/corresp
[13] https://datos.bcn.cl/es/endpoint-sparql
[14] https://lod-cloud.net/dataset/ASCDC-CRB
[15] https://data.getty.edu/provenance
[16] https://github.com/hibernator11/lod-quality-reproducible

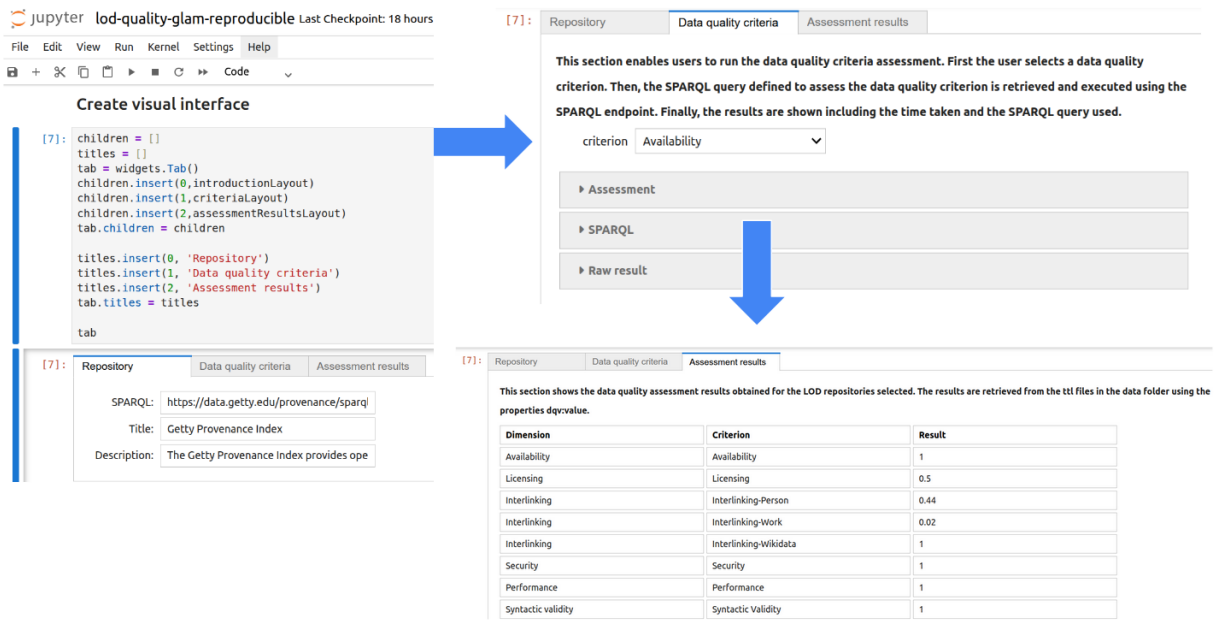

Fig. 2. Interface implemented based on Jupyter Notebooks to reproduce the data quality assessment results. The user is able to consult the information about the dataset and run in real time each data quality criterion. The Assessment results tab retrieves the results obtained by the authors of this work that are described in the RDF data generated.

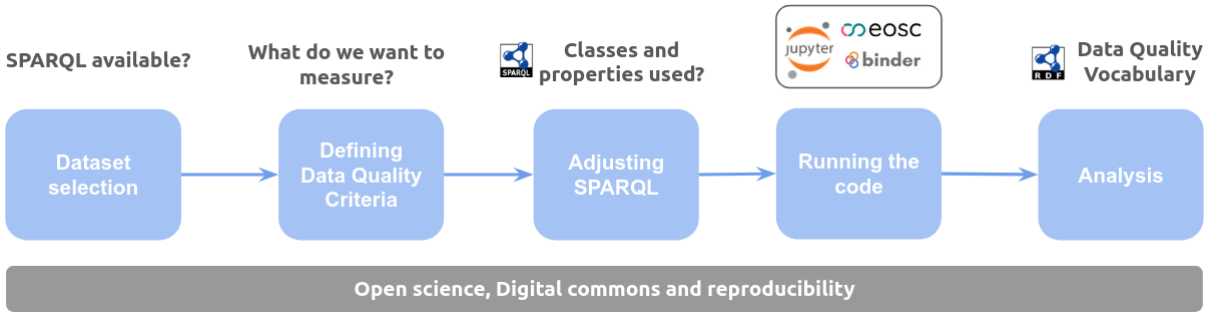

Fig. 3. Workflow diagram describing how the scripts work and can be adapted to new datasets.

Based on these assessments, it has been observed that LOD datasets performed best in the intrinsic dimension, but show the lowest performance in the contextual dimension.

In the accessibility dimension, all LOD datasets successfully met the availability and security criteria. This indicates that all LOD datasets respond to SPARQL queries and make use of HTTPS support. However, only BNE and BnF fully satisfied the criterion regarding the use of properties in Wikidata. The majority of LOD datasets do not have machine-readable licenses, which is a requirement of the Licensing criterion, although textual licensing documentation is available. BNC and CRB did not achieve satisfactory results in the Performance criteria.

All LOD datasets performed well in the intrinsic quality criteria, including syntactic validity, semantic accuracy, and consistency. These results indicate that the RDF documents in the LOD datasets contain no syntax errors and that author names and dates are consistent with those in external datasets. The datasets were also mostly successful in the completeness criterion. However, the data in LOD datasets contain unnecessary duplicates in the conciseness criterion.

In contrast, all LOD datasets failed to satisfy the majority of the criteria in the Contextual dimension. This means that LOD datasets do not provide ranking support (relevancy), human-readable labels of entities (understandability-labels), information about the vocabularies used (understandability-vocabularies), information about URI patterns (understandability-URL patterns) or examples of SPARQL queries (understandability-examples). Only BNC provides machine-readable metadata regarding the publication date, while the others provide textual documentation (timeliness). None of the datasets specify any information about unknown or empty values (trustworthiness–unknown). The BNE, BnF and BNC datasets are curated manually, extracted using automated methods, and enriched by the community, whereas the others include manual curation and automated extraction but lack community-based enrichment (trustworthiness–dataset).

In the Representational dimension, the representational conciseness criterion, which measures URI length, resulted in unsuccessful outcomes for all LOD datasets. However, all datasets performed well in the representational conciseness (containers) criterion, which assesses the use of RDF containers. All LOD datasets were also successful in the interoperability criterion, indicating that the vocabularies used to describe metadata are reusable. Furthermore, although LOD datasets include unique identifiers such as ISNI and VIAF in their metadata, labels are not provided in the appropriate languages. Additionally, the data are not available in different serialization formats or languages.

Considering all assessments, the BNE datasets show the best overall performance, followed respectively by the LDFI and BnF datasets. CRB, on the other hand, demonstrates a lower performance compared to the others.

4.1. Discussion

While this work is focused on the GLAM sector, the framework can be extended for different purposes. For instance, in order to assess new datasets, the examples provided can be employed as a basis and adjusted to the classes and properties used. New data quality criteria can be integrated by adjusting the configuration file. Other options can be focused on the user interface to provide additional functionalities.

A limitation of this study is the manual measurement of the semantic accuracy criterion due to the absence of federation among LOD datasets against known existing datasets. Instead of maintaining datasets as isolated data silos, there is a need to interconnect them, enabling reusers to search across multiple datasets and thereby enhancing the possibilities of analysis. Federated SPARQL queries could serve as a viable solution in this context[23].

The assessment of completeness is limited in scope to the classes and properties provided as an example of how this criterion can be assessed. The population completeness assessment is limited in scope to a small selection of artists provided by a collaborative-edition platform such as Wikidata. In general, the completeness assessment can be enhanced by refining the classes, properties and resources used to assess the dataset.

The relevancy data quality criterion is based on a property defined by Wikidata. GLAM institutions need to work further in this matter in order to provide this type of information. Additional vocabularies could be explored in order to provide metadata concerning relevancy.

The results of the performance criterion are contingent upon response delays and may vary based on several factors, including the dataset used, the SPARQL query, and the system executing the code. In some cases, SPARQL queries (e.g., trustworthiness criterion) provide as a result a time out. This could be solved by refining the SPARQL queries or loading a dump file of the LOD dataset in a local RDF storage system such as the Apache Jena TDB [7]. Other approaches could be based on cloud infrastructures adapted to larger-scale datasets that could be used to load the datasets and run the data quality assessment.

The scripts provided in this work consist of several resources including Jupyter Notebooks, SPARQL queries and RDF files. The Jupyter Notebooks combine textual documentation describing the work and Python code. They can be used by non-technical users (e.g., curators) to reproduce the results but also to assess new datasets by creating a new file using the examples provided. Advanced users (e.g., data scientist or developers) could extend the scripts in order to refine the SPARQL queries and to create new functionalities such as new data quality criteria, widgets or visualizations.

The framework has been applied to a wide diversity of datasets in terms of size and ontologies used demonstrating its adoption. Some issues were identified concerning the use of the SPARQL which could be solved by refining the queries or using local environments.

Table 5
Assessment results of LOD datasets based on the data quality criteria defined.

Dimension	Criteria	BNE	BnF	LDFI	GPI	BNC	CRB
Accessibility	Availability	1	1	1	1	1	1
	Licensing	1	0.5	1	0.5	0.5	0.5
	Interlinking-Person	0.25	0.44	0.67	0.21	0	0
	Interlinking-Work	0.02	0.02	0	0	0	0.02
	Interlinking-Wikidata	1	1	0	0	0.5	0
	Security	1	1	1	1	1	1
	Performance	1	1	1	1	0	0
Intrinsic	Syntactic validity	1	1	1	1	1	1
	Semantic accuracy	1	1	1	1	1	1
	Consistency	1	1	1	1	1	1
	Conciseness	1	1	1	1	1	1
	Conciseness-Wikidata	0.99	0.99	-	-	0.93	-
	Completeness-Person	1	1	1	1	1	0
	Completeness-Work	1	1	1	1	1	1
	Completeness-Place	0	0	1	1	0	1
	Completeness-Population	1	1	1	0	0	0
Contextual	Relevancy	0	0	0	0	0	0
	Trustworthiness-Dataset	1	1	0.5	0.5	1	0.5
	Trustworthiness-Unknown	0	0	0	0	0	0
	Understandability-Labels	0	0	0	0	0	0
	Understandability-Vocabularies	0	0	0	0	0	0
	Understandability-URL Patterns	0	0	0	0	0	0
	Understandability-Examples	0	0	0	0	0	0
	Timeliness	0.5	0.5	0.5	0.5	1	0.5
Representational	Representational conciseness-URIs Length	0	0	0	0	0	0
	Representational conciseness-Containers	1	1	1	1	1	1
	Interoperability	1	1	1	1	1	1
	Interpretability-VIAF	1	1	1	1	1	1
	Interpretability-ISNI	1	1	1	1	1	1
	Interpretability-Labels	0	0	0	0	0	0
	Versatility-Serialization	0	0	0	0	0	0
	Versatility-Multilingual	0	0	0	0	0	0
	Total	18.76	17.45	17.67	15.71	15.93	13.52

The developed Jupyter notebook can be utilized in various studies and by GLAM institutions to assess their metadata quality. For instance, the conciseness criterion, which was found lacking in the assessed LOD datasets, could be improved by reducing data duplication. The conciseness criterion was assessed using the `owl:sameAs` property in order to identify duplicates. Additional approaches could be based on the Wikidata properties defined to link GLAM datasets. However, in some cases, these duplicates might be necessary in the catalog for some reason such as resources with different classes. For example, the BNE provides an anonymous author used in certain works.[17] Conversely, known anonymous works such as *El Lazarillo de Tormes* do not include this type of metadata. Further utilization of this information is required to enhance metadata quality. The metadata can also be refined by including multilingual labels for the description of the data quality criteria. Generally, the assessed LOD datasets do

[17] https://datos.bne.es/persona/XX1193997.html

not provide metadata about the vocabularies and ontologies used, complicating the assessment and measurement of criteria such as consistency. Hence, LOD datasets should consider including this information as part of the dataset.

While the metadata provided by the assessed LOD datasets is generally of high quality, adopting emerging standards such as DCAT to describe the dataset itself could improve their reuse in new environments, such as data spaces [25]. GLAM institutions should align with these requirements to maintain high-quality standards.

The figures in Table 5 are useful to select the dataset that best fits a specific purpose. For instance, if the most relevant aspects for an institution are machine-readable licensing and interlinking, then the LDFI and BNE might be the first choice in order to enrich a collection.

The LOD datasets assessed showed that in some cases the information is only provided in the form of textual documentation (e.g., license). In other cases, such as the timeliness criterion, the lack of metadata can pose difficulties when identifying the novelty or the version of the data. In addition, different properties are used for the same purpose (e.g., `olw:sameAs` and `skos:exactMatch`) hindering the assessment and requiring to adjust the SPARQL scripts. Additional work is required in terms of the richness of the data published by institutions.

We also noticed that recent datasets tend to obtain a higher value for interlinking. This relation could be caused to different reasons such as the maintenance of the dataset. In addition, and in terms of reproducibility and obtaining the same results, LOD can change over time and results can vary when running the system at different times.

5. Conclusions

GLAM are increasingly sharing their data as LOD to enhance accessibility and achieve high-quality standards. It is crucial to assess the quality of these data in a reproducible manner. While various studies have explored data quality assessment methodologies, this study uniquely provides a reproducible and machine-readable framework to explain assessment results using existing vocabularies that can be a valuable resource for new and emergent data infrastructures such as data spaces.

Based on prior research, data quality criteria were identified and applied to LOD datasets. The assessment revealed that all three LOD datasets exhibited strong performance in similar domains (e.g., interlinking, security, interoperability), while they demonstrated less satisfactory outcomes in other domains (e.g., conciseness, relevancy, trustworthiness). Furthermore, the developed framework and GitHub project can be adapted to different LOD datasets to guide data quality enhancement.

Future research includes the integration of additional data quality criteria such as diversity and source reliability, the FAIR advancement in GLAM as well as the implementation in real practice for data spaces and the ECCCH.

Acknowledgements

This work was inspired by the International GLAM Labs Community, Collections as Data, DARIAH-EU, The common European data space for cultural heritage, the GLAM Workbench and the IMPACT Centre of Competence in Digitization. A preliminary version of this work was presented at the 2024 LD4 Conference, *Building Community for Linked Open Data*, and the Archives as data event, at Columbia University.

References

[1] R. Albertoni and A. Isaac, Introducing the Data Quality Vocabulary (DQV), *Semantic Web* **12**(1) (2021), 81–97. doi:10.3233/SW-200382.

[2] H. Alkemade, S. Claeyssens, G. Colavizza, N. Freire, J. Lehmann, C. Neudecker, G. Osti and D. van Strien, Datasheets for Digital Cultural Heritage Datasets, *Journal of Open Humanities Data* (2023). doi:10.5334/johd.124.

[3] B. Antunes and D.R.C. Hill, Reproducibility, Replicability and Repeatability: A survey of reproducible research with a focus on high performance computing, *Comput. Sci. Rev.* **53** (2024), 100655. doi:10.1016/J.COSREV.2024.100655. https://doi.org/10.1016/j.cosrev.2024.100655.

[4] T. Baker, K. Coyle and S. Petiya, Multi-entity models of resource description in the Semantic Web: A comparison of FRBR, RDA and BIBFRAME, *Libr. Hi Tech* **32**(4) (2014), 562–582. doi:10.1108/LHT-08-2014-0081.

[5] T. Berners-Lee, J. Hendler and O. Lassila, The Semantic Web in Scientific American, *Scientific American Magazine* **284** (2001).

[6] G. Candela, An automatic data quality approach to assess semantic data from cultural heritage institutions, *J. Assoc. Inf. Sci. Technol.* **74**(7) (2023), 866–878. doi:10.1002/ASI.24761. https://doi.org/10.1002/asi.24761.

[7] G. Candela, Towards a semantic approach in GLAM Labs: The case of the Data Foundry at the National Library of Scotland, *Journal of Information Science* **0**(0) (2023). doi:10.1177/01655515231174386.

[8] G. Candela, Browsing Linked Open Data in Cultural Heritage: a shareable visual configuration approach, *J. Comput. Cult. Herit.* (2024), Just Accepted. doi:10.1145/3707647.

[9] G. Candela, S. Chambers and T. Sherratt, An approach to assess the quality of Jupyter projects published by GLAM institutions, *J. Assoc. Inf. Sci. Technol.* **74**(13) (2023), 1550–1564. doi:10.1002/ASI.24835. https://doi.org/10.1002/asi.24835.

[10] G. Candela, C. Rosiński and A. Margraf, A reproducible framework to publish and reuse Collections as data: the case of the European Literary Bibliography, Zenodo, 2024. doi:10.5281/zenodo.14106707.

[11] G. Candela, P. Escobar, R.C. Carrasco and M. Marco-Such, Evaluating the quality of linked open data in digital libraries, *J. Inf. Sci.* **48**(1) (2022), 21–43. doi:10.1177/0165551520930951.

[12] G. Candela, N. Gabriëls, S. Chambers, M. Dobreva, S. Ames, M. Ferriter, N. Fitzgerald, V. Harbo, K. Hofmann, O. Holownia, A. Irollo, M. Mahey, E. Manchester, T.-A. Pham, A. Potter and E. Van Keer, A checklist to publish collections as data in GLAM institutions, *Global Knowledge, Memory and Communication* (2023). doi:10.1108/gkmc-06-2023-0195. http://dx.doi.org/10.1108/GKMC-06-2023-0195.

[13] G. Candela, J. Pereda, D. Sáez, P. Escobar, A. Sánchez, A.V. Torres, A.A. Palacios, K. McDonough and P. Murrieta-Flores, An Ontological Approach for Unlocking the Colonial Archive, *J. Comput. Cult. Herit.* **16**(4) (2023). doi:10.1145/3594727.

[14] G. Candela, M. Cuper, O. Holownia, N. Gabriëls, M. Dobreva and M. Mahey, A Systematic Review of Wikidata in GLAM Institutions: a Labs Approach, in: *Linking Theory and Practice of Digital Libraries - 28th International Conference on Theory and Practice of Digital Libraries, TPDL 2024, Ljubljana, Slovenia, September 24-27, 2024, Proceedings, Part II*, A. Antonacopoulos, A. Hinze, B. Piwowarski, M. Coustaty, G.M.D. Nunzio, F. Gelati and N. Vanderschantz, eds, Lecture Notes in Computer Science, Vol. 15178, Springer, 2024, pp. 34–50. doi:10.1007/978-3-031-72440-4_4.

[15] G. Candela, M. Dişli, M. Dobreva and S. Chambers, hibernator11/lod-quality-reproducible: v1.1, Zenodo, 2025. doi:10.5281/zenodo.18002668.

[16] L. Candela, D. Castelli, N. Ferro, Y. Ioannidis, G. Koutrika, C. Meghini, P. Pagano, S. Ross, D. Soergel, M. Agosti, M. Dobreva, V. Katifori and H. Schuldt, *The DELOS Digital Library Reference Model. Foundations for Digital Libraries*, 2007, DELOS Network of Excellence on Digital Libraries Project no. 507618.

[17] K. Coyle, Works, Expressions, Manifestations, Items: An Ontology, *The Code4Lib Journal* **53** (2022). https://journal.code4lib.org/articles/16491.

[18] M. Daquino, F. Mambelli, S. Peroni, F. Tomasi and F. Vitali, Enhancing Semantic Expressivity in the Cultural Heritage Domain: Exposing the Zeri Photo Archive as Linked Open Data, *ACM Journal on Computing and Cultural Heritage* **10**(4) (2017), 21:1–21:21. doi:10.1145/3051487.

[19] J. Debattista, S. Auer and C. Lange, Luzzu - A Methodology and Framework for Linked Data Quality Assessment, *ACM J. Data Inf. Qual.* **8**(1) (2016), 4:1–4:32. doi:10.1145/2992786.

[20] J. Debattista, C. Lange and S. Auer, daQ, an Ontology for Dataset Quality Information, in: *Proceedings of the Workshop on Linked Data on the Web co-located with the 23rd International World Wide Web Conference (WWW 2014), Seoul, Korea, April 8, 2014*, C. Bizer, T. Heath, S. Auer and T. Berners-Lee, eds, CEUR Workshop Proceedings, Vol. 1184, CEUR-WS.org, 2014. https://ceur-ws.org/Vol-1184/ldow2014_paper_09.pdf.

[21] J. Debattista, C. Lange, S. Auer and D. Cortis, Evaluating the quality of the LOD cloud: An empirical investigation, *Semantic Web* **9**(6) (2018), 859–901. doi:10.3233/SW-180306.

[22] C. Dijkshoorn, L. Jongma, L. Aroyo, J. van Ossenbruggen, G. Schreiber, W. ter Weele and J. Wielemaker, The Rijksmuseum collection as Linked Data, *Semantic Web* **9**(2) (2018), 221–230. doi:10.3233/SW-170257.

[23] M. Dişli, G. Osti, G. Candela and R. Zijdeman, From Linked Open Data to Collections as Data: A Reproducible Framework Using Federated Queries, *Information Technology and Libraries* **44**(4) (2025). doi:10.5860/ital.v44i4.17432. https://ital.corejournals.org/index.php/ital/article/view/17432.

[24] M. Dişli, G. Candela, S. Gutiérrez and G. Fontenelle, Open Data Practices of Art Museums in Wikidata: A Compliance Assessment, *Journal of Open Humanities Data* (2025). doi:10.5334/johd.438.

[25] M. Dobreva, K. Stefanov and K. Ivanova, Data Spaces for Cultural Heritage: Insights from GLAM Innovation Labs, in: *From Born-Physical to Born-Virtual: Augmenting Intelligence in Digital Libraries - 24th International Conference on Asian Digital Libraries, ICADL 2022, Hanoi, Vietnam, November 30 - December 2, 2022, Proceedings*, Y. Tseng, M. Katsurai and H.N. Nguyen, eds, Lecture Notes in Computer Science, Vol. 13636, Springer, 2022, pp. 492–500. doi:10.1007/978-3-031-21756-2_41.

[26] C. Drummond, Reproducible research: a minority opinion, *J. Exp. Theor. Artif. Intell.* **30**(1) (2018), 1–11. doi:10.1080/0952813X.2017.1413140.

[27] M. Färber, F. Bartscherer, C. Menne and A. Rettinger, Linked data quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO, *Semantic Web* **9**(1) (2018), 77–129. doi:10.3233/SW-170275.

[28] A.U. Frank, Data Quality Ontology: An Ontology for Imperfect Knowledge, in: *Spatial Information Theory, 8th International Conference, COSIT 2007, Melbourne, Australia, September 19-23, 2007, Proceedings*, S. Winter, M. Duckham, L. Kulik and B. Kuipers, eds, Lecture Notes in Computer Science, Vol. 4736, Springer, 2007, pp. 406–420. doi:10.1007/978-3-540-74788-8_25.

[29] M. Färber and L. Ao, The Microsoft Academic Knowledge Graph enhanced: Author name disambiguation, publication classification, and embeddings, *Quant. Sci. Stud.* **3**(1) (2022), 51–98. doi:10.1162/qss_a_00183.

[30] S.N. Goodman, D. Fanelli and J.P.A. Ioannidis, What does research reproducibility mean?, *Science Translational Medicine* (2016). doi:10.1126/scitranslmed.aaf5027.

[31] E. Hyvönen, P. Leskinen and J. Tuominen, LetterSampo–Historical Letters on the Semantic Web: A Framework and Its Application to Publishing and Using Epistolary Data, *J. Comput. Cult. Herit.* **16**(1) (2023). doi:10.1145/3569372.

[32] L. Ibáñez, I. Millard, H. Glaser and E. Simperl, An Assessment of Adoption and Quality of Linked Data in European Open Government Data, in: *The Semantic Web - ISWC 2019 - 18th International Semantic Web Conference, Auckland, New Zealand, October 26-30, 2019, Proceedings, Part II*, 2019, pp. 436–453. doi:10.1007/978-3-030-30796-7_27.

[33] S.G. Johnson, S.M. Speedie, G.J. Simon, V. Kumar and B.L. Westra, A Data Quality Ontology for the Secondary Use of EHR Data, in: *AMIA 2015, American Medical Informatics Association Annual Symposium, San Francisco, CA, USA, November 14-18, 2015*, AMIA, 2015. https://knowledge.amia.org/59310-amia-1.2741865/t007-1.2744224/f007-1.2744225/2248523-1.2744287/2246427-1.2744284.

[34] P. Jupyter, M. Bussonnier, J. Forde, J. Freeman, B.E. Granger, T. Head, C. Holdgraf, K. Kelley, G. Nalvarte, A. Osheroff, M. Pacer, Y. Panda, F. Perez, B. Ragan-Kelley and C. Willing, Binder 2.0 - Reproducible, interactive, sharable environments for science at scale, in: *Proceedings of the 17th Python in Science Conference 2018 (SciPy 2018), Austin, Texas, July 9 - July 15, 2018*, 2018, pp. 113–120. doi:10.25080/Majora-4af1f417-011.

[35] M. Koho, E. Ikkala, P. Leskinen, M. Tamper, J. Tuominen and E. Hyvönen, WarSampo knowledge graph: Finland in the Second World War as Linked Open Data, *Semantic Web* **12**(2) (2021), 265–278. doi:10.3233/SW-200392.

[36] D. Kontokostas, P. Westphal, S. Auer, S. Hellmann, J. Lehmann, R. Cornelissen and A. Zaveri, Test-driven evaluation of linked data quality, in: *23rd International World Wide Web Conference, WWW '14, Seoul, Republic of Korea, April 7-11, 2014*, C. Chung, A.Z. Broder, K. Shim and T. Suel, eds, ACM, 2014, pp. 747–758. doi:10.1145/2566486.2568002.

[37] G. Liu, B. Bordelon, R. Nagar, U. Nguyen, J. Sarti and J. Boettcher, *Data Quality Literacy: A Guidebook*, 2024. ISBN 9798218417475. https://doi.org/10.31219/osf.io/ruawm.

[38] M. Mahey, A. Al-Abdulla, S. Ames, P. Bray, G. Candela, C. Derven, M. Dobreva-McPherson, K. Gasser, S. Chambers, S. Karner, K. Kokegei, D. Laursen, A. Potter, A. Straube, S.-C. Wagner and L. Wilms, *Open a GLAM lab*, International GLAM Labs Community, Book Sprint, Doha, Qatar, 2019, p. 164. ISBN 978-9927-139-07-9. doi:10.21428/16ac48ec.f54af6ae.

[39] N. Mihindukulasooriya, R. García-Castro and A. Gómez-Pérez, LD Sniffer: a quality assessment tool for measuring the accessibility of Linked Data, in: *Knowledge Engineering and Knowledge Management*, Vol. 1, Springer, Cham, Suiza, 2016, pp. 149–152. ISBN 978-3-319-49004-5. doi:10.1007/978-3-319-58694-6_20. https://link.springer.com/chapter/10.1007/978-3-319-58694-6%5f20.

[40] J. Navarro, A. Deruyver and P. Parrend, A systematic survey on multi-step attack detection, *Computers & Security* (2018). https://www.sciencedirect.com/science/article/pii/S0167404818302141.

[41] A. Nayak, B. Bozic and L. Longo, Linked Data Quality Assessment: A Survey, in: *Web Services - ICWS 2021 - 28th International Conference, Held as Part of the Services Conference Federation, SCF 2021, Virtual Event, December 10-14, 2021, Proceedings*, 2021, pp. 63–76. doi:10.1007/978-3-030-96140-4_5.

[42] A. Nayak, B. Bozic and L. Longo, Data Quality Assessment and Recommendation of Feature Selection Algorithms: An Ontological Approach, *J. Web Eng.* **22**(1) (2023), 175–196. doi:10.13052/JWE1540-9589.2219. https://doi.org/10.13052/jwe1540-9589.2219.

[43] T. Padilla, H. Scates Kettler, S. Varner and Y. Shorish, Vancouver Statement on Collections as Data, Zenodo, 2023. doi:10.5281/zenodo.8342171.

[44] A.D. Pasquale, V. Pasqual, F. Tomasi and F. Vitali, On assessing weaker logical status claims in Wikidata Cultural Heritage records, *Semantic Web* (2024).

[45] M.A. Pellegrino, A. Rula and G. Tuozzo, KGHeartBeat: An Open Source Tool for Periodically Evaluating the Quality of Knowledge Graphs, in: *The Semantic Web – ISWC 2024: 23rd International Semantic Web Conference, Baltimore, MD, USA, November 11–15, 2024, Proceedings, Part III*, Springer-Verlag, Berlin, Heidelberg, 2024, pp. 40–58–. ISBN 978-3-031-77846-9. doi:10.1007/978-3-031-77847-6_3.

[46] J.F. Pimentel, L. Murta, V. Braganholo and J. Freire, Understanding and improving the quality and reproducibility of Jupyter notebooks, *Empir. Softw. Eng.* **26**(4) (2021), 65. doi:10.1007/S10664-021-09961-9. https://doi.org/10.1007/s10664-021-09961-9.

[47] Publications Office of the European Union, Data.europa.eu data quality guidelines, 2022. https://data.europa.eu/doi/10.2830/333095.

[48] J. Raad, W. Beek, F. van Harmelen, J. Wielemaker, N. Pernelle and F. Saïs, Constructing and Cleaning Identity Graphs in the LOD Cloud, *Data Intell.* **2**(3) (2020), 323–352. doi:10.1162/dint_a_00057.

[49] S. Salem and F. Benchikha, LODQuMa: A Free-ontology process for Linked (Open) Data quality management, *J. King Saud Univ. Comput. Inf. Sci.* **34**(8 Part A) (2022), 5552–5563. doi:10.1016/j.jksuci.2021.06.001.

[50] J. Samuel, ShExStatements: Simplifying Shape Expressions for Wikidata, in: *Companion of The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, J. Leskovec, M. Grobelnik, M. Najork, J. Tang and L. Zia, eds, ACM / IW3C2, 2021, pp. 610–615. doi:10.1145/3442442.3452349.

[51] K. Shenoy, F. Ilievski, D. Garijo, D. Schwabe and P.A. Szekely, A study of the quality of Wikidata, *J. Web Semant.* **72** (2022), 100679. doi:10.1016/j.websem.2021.100679.

[52] T. Sherratt, GLAM-Workbench/recordsearch, Zenodo, 2023. doi:10.5281/zenodo.7553047.

[53] G. Sugimoto, Missing Links: Investigating the Quality of Linked Data and its Tools in Cultural Heritage and Digital Humanities, PhD thesis, Vrije Universiteit Amsterdam, 2025.

[54] P.A. Szekely, C.A. Knoblock, F. Yang, E.E. Fink, S. Gupta, R. Allen and G. Goodlander, Publishing the Data of the Smithsonian American Art Museum to the Linked Data Cloud, *Int. J. Humanit. Arts Comput.* **8**(supplement) (2014), 152–166. doi:10.3366/IJHAC.2014.0104. https://doi.org/10.3366/ijhac.2014.0104.

[55] K. Tallerås, Quality of Linked Bibliographic Data: The Models, Vocabularies, and Links of Data Sets Published by Four National Libraries, *Journal of Library Metadata* **17**(2) (2017), 126–155. doi:10.1080/19386389.2017.1355166.
[56] T. Tasovac, S. Chambers and E. Tóth-Czifra, Cultural Heritage Data from a Humanities Research Perspective: A DARIAH Position Paper, 2020, DARIAH's response to European Commission's evaluation and possible revision of the Commission Recommendation of 27 October 2011 on Digitisation and Online Accessibility of Cultural Material and Digital Preservation (REC 2011/711/EU). https://hal.science/hal-02961317.
[57] K. Thornton, H. Solbrig, G.S. Stupp, J.E.L. Gayo, D. Mietchen, E. Prud'hommeaux and A. Waagmeester, Using Shape Expressions (ShEx) to Share RDF Data Models and to Guide Curation with Rigorous Validation, in: *The Semantic Web - 16th International Conference, ESWC 2019, Portorož, Slovenia, June 2-6, 2019, Proceedings*, 2019, pp. 606–620. doi:10.1007/978-3-030-21348-0_39.
[58] M. Uschold and M. Gruninger, Ontologies: principles, methods and applications, *Knowl. Eng. Rev.* **11**(2) (1996), 93–136. doi:10.1017/S0269888900007797.
[59] D. van Strien, K. Beelen, M.C. Ardanuy, K. Hosseini, B. McGillivray and G. Colavizza, Assessing the Impact of OCR Quality on Downstream NLP Tasks, in: *Proceedings of the 12th International Conference on Agents and Artificial Intelligence, ICAART 2020, Volume 1, Valletta, Malta, February 22-24, 2020*, A.P. Rocha, L. Steels and H.J. van den Herik, eds, SCITEPRESS, 2020, pp. 484–496. doi:10.5220/0009169004840496.
[60] M.D. Wilkinson, M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L.B. da Silva Santos, P.E. Bourne et al., The FAIR Guiding Principles for scientific data management and stewardship, *Scientific data* **3** (2016).
[61] World Wide Web Consortium, Describing Linked Datasets with the VoID Vocabulary, 2011. https://www.w3.org/TR/void/.
[62] World Wide Web Consortium, Best Practices for Publishing Linked Data, 2014. https://www.w3.org/TR/ld-bp/#MODEL.
[63] World Wide Web Consortium, RDF 1.1 Concepts and Abstract Syntax, 2014. https://www.w3.org/TR/rdf11-concepts/.
[64] World Wide Web Consortium, Data on the Web Best Practices, 2017. https://www.w3.org/TR/dwbp/.
[65] World Wide Web Consortium, Shapes Constraint Language (SHACL), 2017. https://www.w3.org/TR/shacl/.
[66] World Wide Web Consortium, Shape Expressions (ShEx) Primer, 2022. http://shexspec.github.io/primer/.
[67] World Wide Web Consortium, Data Catalogue Vocabulary, 2024. https://www.w3.org/TR/vocab-dcat-3/.
[68] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann and S. Auer, Quality assessment for Linked Data: A Survey, *Semantic Web* **7**(1) (2016), 63–93. doi:10.3233/SW-150175.
[69] B. Zhang, F. Ilievski and P.A. Szekely, Enriching Wikidata with Linked Open Data, in: *Proceedings of the 3rd Wikidata Workshop 2022 co-located with the 21st International Semantic Web Conference (ISWC2022), Virtual Event, Hanghzou, China, October 2022*, L. Kaffee, S. Razniewski, G. Amaral and K.S. Alghamdi, eds, CEUR Workshop Proceedings, Vol. 3262, CEUR-WS.org, 2022. https://ceur-ws.org/Vol-3262/paper8.pdf.
[70] T. Çakmak, Metadata Quality Assessment in Libraries, in: *Current approaches in information and records management technology in the 100th anniversary of the Republic*, İstanbul University Press, 2024, pp. 987–1010. doi:10.26650/B/SS53.2024.015.39.
[71] A.H. Çetin, E. Doğan and E. Tüzün, A review of code reviewer recommendation studies: Challenges and future directions, *Science of Computer Programming* (2021). doi:10.1016/j.scico.2021.102652.

Appendix A. Namespaces used

The prefixes in Table 6 are used to abbreviate namespaces throughout this article.

Table 6
Common prefixes used to designate RDF vocabularies.

prefix	URI
cidoc	http://www.cidoc-crm.org/cidoc-crm/
dcat	http://www.w3.org/ns/dcat#
daq	http://purl.org/eis/vocab/daq
dcterms	http://purl.org/dc/terms/
dqv	http://www.w3.org/ns/dqv#
foaf	http://xmlns.com/foaf/0.1/
frbr	http://iflastandards.info/ns/fr/frbr/frbrer/
frbr-rda	http://rdvocab.info/uri/schema/FRBRentitiesRDA
isbd	http://iflastandards.info/ns/isbd/elements/
kbv	https://id.kb.se/vocab/
ldqd	https://www.w3.org/2016/05/ldqd#
loc	http://id.loc.gov/ontologies/bibframe/
lssc	http://ldf.fi/schema/lssc/
nobel	http://data.nobelprize.org/terms/
owl	http://www.w3.org/2002/07/owl#
rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#
rdfs	http://www.w3.org/2000/01/rdf-schema#
schema	http://schema.org/
skos	http://www.w3.org/2004/02/skos/core#
void	http://rdfs.org/ns/void#
wdt	http://www.wikidata.org/prop/direct/
wd	http://www.wikidata.org/entity/

Appendix B. Additional details concerning the definition of the data quality criteria

This section describes the definition of each data quality criterion.

Availability: *refers to the extent to which data (or some portion of it) is present, obtainable and ready for use* [11, 27, 68]. For instance, it can be assessed by using a URI checking whether the server responds to a SPARQL query over a period of time.

$$m_{\text{availability}} = \frac{\text{Number of successful requests}}{\text{Total number of requests}} \tag{1}$$

Licensing: *is defined as the granting of permission for a consumer to re-use a dataset under defined conditions* [11, 27, 68]. It can be measured by identifying metadata properties such as `dcterms:license` and `schema:license`.

$$m_{\text{licensing}} = \begin{cases} 1 & \text{machine-readable license} \\ 0.5 & \text{text documentation license} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Interlinking: *refers to the degree to which entities that represent the same concept are linked to each other* [11, 27, 68]. It can be measured by identifying the existence and usage of external URIs (e.g., using the property `owl:sameAs`). In addition, external datasets such as Wikidata can be used in order to retrieve the number of resources linked (see Listing 2). Many institutions have created a dedicated property in Wikidata to establish links and connect their resources [14, 24].

Let g be a dataset, d be the list of values of each property p included in $prop$, v each of the different values included in d, r_i each of the resources typed as a particular class (e.g., `foaf:Person` or `bibframe:Work`) included in g and $checkExternal$ a function that assesses that the value of v is an external URI:

$$\begin{aligned} isExternal(r_i) = (r_i, p, d) \mid (r_i, p, d) \in g \\ \wedge\, (r_i, \text{owl:sameAs}, v) \;\wedge\; p \in prop \wedge v \in d \mid \\ (p, v) \in g \wedge checkExternal(v) \end{aligned} \tag{3}$$

$$m_{\text{interlinking-Class}} = \frac{1}{n} \sum_{i=1}^{n} isExternal(r_i) \tag{4}$$

```
SELECT (COUNT(?subject) AS ?total)
WHERE {
    ?subject wdt:P268 ?object
}
```

Listing 2: An example of SPARQL query to retrieve the number of resources linked in Wikidata according to the property `wdt:P2799` used to link authors in the Biblioteca Virtual Miguel de Cervantes.

$$m_{\text{interlinking-Wikidata}} = \begin{cases} 1 & \text{two or more dedicated property available in Wikidata} \\ 0.5 & \text{one dedicated property available in Wikidata} \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

Security: *is the extent to which data is protected against alteration and misuse* [68]. Given the open nature of some of the data available, in some cases security was not considered [29]. In order cases, security was measured by verifying the support of HTTPS [45].

$$m_{\text{security}} = \begin{cases} 1 & \text{HTTPS support} \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

Performance: *refers to the efficiency of a system that binds to a large dataset* [68]. This can be measured in terms of scalability by means of the detection of whether the time to answer ten requests divided by ten is not longer than the time it takes to answer one request [68].

$$m_{\text{performance}} = \begin{cases} 1 & \text{answering ten requests divided by ten is not longer than} \\ & \text{the time it takes to answer one request} \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

Note that queries have different levels of complexity. In our approach, a general SPARQL query was used in order to facilitate its employment in different datasets.

Syntactic validity: *is defined as the degree to which an RDF document conforms to the specification of the serialization format* [11, 27, 68]. It can be measured by identifying syntax errors by using, for instance, existing RDF validators.[18] Listing 3 shows an example to retrieve 100 resources with the type `cidoc:Person` that can be used with the RDF validator.

```
SELECT ?subject
WHERE {
    ?subject rdf:type cidoc:E21_Person
}
LIMIT 100
```

Listing 3: An example of SPARQL query to retrieve 100 resources with the type `cidoc:E21_Person`.

Note that this criterion can be improved by using a larger number of resources provided by the dataset. When using a small portion of the resources, the confidence interval can be obtained as additional statistical information [6].

[18] See, for example, the W3C RDF Validator available at https://www.w3.org/RDF/Validator/

$$m_{\text{syntactic}} = \begin{cases} 1 \text{ RDF syntactically correct} \\ 0 \text{ otherwise} \end{cases} \tag{8}$$

Semantic accuracy: *is defined as the extent to which data are correct, reliable, and certified free of error* [6, 11, 68]. This can be measured, for instance, by checking a random sample of authors against external datasets in order to identify whether the dates and names provided are correct. This can be automatically performed using federated queries. For example, Wikidata provides a list of federated datasets that can be used in combination with Wikidata.[19] Listing 4 shows an example of a federated SPARQL query to retrieve the titles of works from Wikidata and the Biblioteca Virtual Miguel de Cervantes. This approach can be applied in the future when specific steps are taken towards a federation of resources which are not federated currently.

```
SELECT ?workLabel WHERE {
  wd:Q165257 wdt:P2799 ?id
  BIND(
    uri(concat("https://data.cervantesvirtual.com/person/",
      ?id)) as ?bvmcID)
  SERVICE
    <http://data.cervantesvirtual.com/openrdf-sesame/\note{datasets}/data>
  {
    ?bvmcID <http://rdaregistry.info/Elements/a/authorOf> ?work .
    ?work rdfs:label ?workLabel
  }
}
LIMIT 100
```

Listing 4: An example of federated SPARQL query to retrieve the titles of the works of the author Lope de Vega from Biblioteca Virtual Miguel de Cervantes by employing the Wikidata SPARQL endpoint.

$$m_{\text{accuracy}} = \frac{1}{n} \sum_{i=1}^{n} isAccurate(r_i) \tag{9}$$

Consistency: *means that a knowledge base is free of (logical/formal) contradictions with respect to particular knowledge representation and inference mechanisms* [6, 11, 27, 68]. This can be measured, for instance, by checking the consistency of statements with respect to class constraints. Following previous approaches [11], we limit ourselves to the class constraint `owl:disjointWith`. Other initiatives have also focused on the validation of properties [27]. Listing 5 shows an example of SPARQL query to measure the consistency according to the criterion proposed in this work.

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>

SELECT ?s ?class1 ?class2
WHERE {
    ?s a ?class1 .
    ?s a ?class2 .
```

[19] https://www.wikidata.org/wiki/Wikidata:Lists/SPARQL_endpoints

```
    ?class1 owl:disjointWith ?class2
}
```

Listing 5: An example of SPARQL query to measure the consistency in LOD by means of the property `owl:disjointWith`.

$$m_{\text{consistency}} = \begin{cases} 1 \text{ RDF consistent} \\ 0 \text{ otherwise} \end{cases} \tag{10}$$

Conciseness: *refers to the minimization of redundancy of entities at the schema and the data level* [68]. It refers to the case when the data does not contain redundant resources. This can be measured by identifying duplicates included in the datasets [11]. It can be measured using identifiers such as VIAF and ISNI. Other approaches are based on the use of external repositories. Listing 6 shows and example of SPARQL query to retrieve duplicate links in Wikidata based on resources featured in BnF.

```
SELECT ?wdItem (COUNT(?bnfResource) AS ?total)
WHERE {
    ?wdItem wdt:P268 ?bnfResource
}
GROUP BY ?wdItem
HAVING (COUNT(?bnfResource) > 1)
```

Listing 6: An example of SPARQL query to retrieve the number of duplicated resources in Wikidata according to the property `wdt:P268` used to link resources in BnF.

$$m_{\text{conciseness}} = \frac{1}{n} \sum_{i=1}^{n} isDuplicated(r_i) \tag{11}$$

$$m_{\text{conciseness-Wikidata}} = \frac{1}{n} \sum_{i=1}^{n} isDuplicated(r_i) \tag{12}$$

Completeness: *refers to the degree to which all required information is present in a particular dataset* [11, 27, 68]. It can be measured using different metrics and including: i) schema completeness by using a selection of classes and properties; and ii) population completeness in terms of the resources available in the dataset. For the population completeness, and since there is no gold standard to test the dataset, we defined a gold standard including classes, properties and resources that could be available in a GLAM dataset. The classes and properties were selected according to the main vocabularies used to describe the resources. Concerning the population, a gold standard was defined based on ten recognized artists included in Wikidata since it is one the most relevant datasets used to enrich the resources in GLAM institutions. Listing 7 shows the SPARQL query used to extract the artists from Wikidata to create the gold standard.

```
SELECT DISTINCT ?artist ?viaf
WHERE {
    VALUES ?artist {
```

Table 7
Classes and properties used to measure the schema completeness.

Category	Classes	Properties
Person	foaf:Person	foaf:givenName
	bne:C1005	rdfs:label
	cidoc:E21_Person	skos:prefLabel
Work	lssc:Letter	rdfs:label
	frbr-rda:Work	skos:prefLabel
	cidoc:E73_Information_Object	lssc:has_time
	bne:C1001	bne:P1001
		bne:P1004
		dc:subject
		dc:date
		frbr-rda:dateOfWork
Place	cidoc:E53_Place	rdfs:label
		skos:prefLabel

```
    wd:Q5682 wd:Q1512 wd:Q692 wd:Q165257
    wd:Q9068 wd:Q535 wd:Q16867 wd:Q762
    wd:Q5597 wd:Q5592
    }.
    ?artist wdt:P735 ?name .
    ?artist wdt:P214 ?viaf
}
```

Listing 7: SPARQL query used to extract the artists from Wikidata to create the gold standard. The values instruction is used to provide the identifiers of the resources selected.

Based on the classes and properties defined in the vocabularies proposed in our literature review, we created a selection based on the main classes and properties identified. Note that the class Work represents any type of document (e.g., `lssc:Letter` and `frbr-rda:Work`).

$$m_{\text{schema\_completeness}} = \begin{cases} 1 \text{ classes and properties used to describe the resources} \\ 0 \text{ otherwise} \end{cases} \tag{13}$$

$$m_{\text{population\_completeness}} = \begin{cases} 1 \text{ resources available in the dataset} \\ 0 \text{ otherwise} \end{cases} \tag{14}$$

Relevancy: *refers to the provision of information which is in accordance with the task at hand and important to the users' queries* [11, 68]. It assesses whether the dataset supports a ranking of statements in order to express the relative relevance of statements. For example, Wikidata enables users to define the ranking of the properties [44].[20] In order to assess this criterion, we used the property `wikibase:rank` provided by Wikidata.

[20]For example, a city may include the list of its mayors in which the current mayor would receive the preferred rank.

$$m_{\text{ranking}} = \begin{cases} 1 \text{ ranking of statements supported} \\ 0 \text{ otherwise} \end{cases} \tag{15}$$

Trustworthiness: *is defined as the degree to which the information is accepted to be correct, true, real and credible* [11, 27, 68]. It can be measured by identifying how the dataset is curated. To do so, we distinguish between manual and automated curation methods, giving relevancy to the enrichment by the community such as the case of Wikidata.

$$m_{\text{trustworthiness-Dataset}} = \begin{cases} 1 & \text{manual curation, automatic extraction and enrichment by a community} \\ 0.5 & \text{manual curation and automatic extraction} \\ 0 & \text{otherwise} \end{cases} \tag{16}$$

Trustworthiness can be increased by supporting unknown and empty values [44]. In the GLAM context, the typical case is when a creator of a book or an sculpture is an anonymous person. For example, Wikidata provides the unknown values using the entity `wd:Q4233718`.[21] Simpler approaches can be based on the use of the literals such as "Unknown", "Unknown value" or "Anonymous".

$$m_{\text{trustworthiness-Unknown}} = \begin{cases} 1 & \text{using machine-readable unknown and empty values} \\ 0.5 & \text{using unknown and empty literals} \\ 0 & \text{otherwise} \end{cases} \tag{17}$$

Note that this is an initial metric and there is more to be done to establish how to use annotations, compute trustworthiness values, compute trustworthiness of information providers, etc.

Understandability: *refers to the ease with which data can be comprehended without ambiguity and be used by a human information consumer* [11, 27, 68]. This can be measured following different approaches: i) detection of human-readable labelling of entities (e.g., use of the property `rdfs:label`); ii) provision of information of the vocabularies used in the dataset (e.g., use of the property `void:vocabulary`); iii) provision of information concerning the URL patterns used (e.g., use of the properties `void:uriRegexPattern` and `void:uriSpace`); and iv) providing examples of SPARQL queries (e.g., use of the property `schema:query`).

$$m_{\text{labels}} = \begin{cases} 1 \text{ Using labels to describe the resources} \\ 0 \text{ otherwise} \end{cases} \tag{18}$$

[21] See, for example, the Spanish Literature book El Lazarillo de Tormes in Wikidata available at https://www.wikidata.org/wiki/Q4233718

$$m_{\text{vocabularies}} = \begin{cases} 1 \text{ Providing information about the vocabularies used} \\ 0 \text{ otherwise} \end{cases} \tag{19}$$

$$m_{\text{urlpatterns}} = \begin{cases} 1 \text{ Using URL patterns} \\ 0 \text{ otherwise} \end{cases} \tag{20}$$

$$m_{\text{queries}} = \begin{cases} 1 \text{ Providing SPARQL queries} \\ 0 \text{ otherwise} \end{cases} \tag{21}$$

Timeliness: *measures how up-to-date data is relative to a specific task* [27, 68]. It can be measured by identifying machine-readable information (e.g., `dcterms:modified`) used by known vocabularies such as DCAT, schema.org and VoID concerning the publication of the dataset.

$$m_{\text{timeliness}} = \begin{cases} 1 & \text{provision of a machine-readable publication date} \\ 0.5 & \text{provision of text documentation} \\ 0 & \text{otherwise} \end{cases} \tag{22}$$

Note that even if something is published recently it does not really mean it offers the newest view on something. This criterion provides a first formalization that may require further work in the future.

Representational conciseness: *refers to the representation of the data, which is compact and well formatted on the one hand and clear and complete on the other hand* [11, 27, 68]. It can be measured by checking the use of long URIs and the use of RDF primitives such as RDF reification, RDF containers and RDF collections (e.g., `rdf:Bag` or `rdf:Seq`) [11]. We have considered the length of URIs of 60 characters as recommended for search engine optimization purposes.

$$m_{\text{long\_uris}} = \begin{cases} 1 \text{ otherwise} \\ 0 \text{ use of long URIs} \end{cases} \tag{23}$$

$$m_{\text{rdf\_primitive}} = \begin{cases} 1 \text{ otherwise} \\ 0 \text{ use of RDF primitives} \end{cases} \tag{24}$$

Note that LOD is based on the provision of permanent identifiers as URIs and this might not be the case for other type of datasets.

Interoperability: *refers to the degree to which the format and structure of the information conforms to previously returned information as well as data from other sources* [11, 27, 68]. It can be measured by identifying whether existing vocabularies have been reused to describe the metadata, as recommended by the W3C [62]. Properties such as `void:vocabulary` provides information about the vocabularies used to describe the metadata. A selection of the vocabularies presented in Table 1 were used as a basis to measure this criterion.

$$m_{\text{interoperability}} = \begin{cases} 1 \text{ using existing vocabularies} \\ 0 \text{ otherwise} \end{cases} \tag{25}$$

Interpretability: *refers to technical aspects of the data, that is, whether information is represented using an appropriate notation and whether the machine is able to process the data* [11, 27, 68]. This includes identifying objects and terms used to define these objects with globally unique identifiers (e.g., ISNI and VIAF) as well as detecting the use of appropriate language (e.g., use of the language when providing labels).

$$m_{\text{interpretability\_property}} = \begin{cases} 1 \text{ information is represented using an appropriate notation} \\ 0 \text{ otherwise} \end{cases} \tag{26}$$

$$m_{\text{interpretability\_language}} = \begin{cases} 1 \text{ use of language for labels} \\ 0 \text{ otherwise} \end{cases} \tag{27}$$

Versatility: *refers to the availability of the data in different representations and in an internationalised way*. It can be measured by checking whether data is available in different serialization formats (e.g., RDF/XML and Turtle) and whether data is available in different languages (e.g., when using the property `rdfs:label`). In order to identify the serialization formats, this criterion uses the property `void:feature` and as a value the Unique URIs for File Formats provided by the W3C.[22]

$$m_{\text{serialization}} = \begin{cases} 1 \text{ different serialization formats} \\ 0 \text{ otherwise} \end{cases} \tag{28}$$

$$m_{\text{multilingual}} = \begin{cases} 1 \text{ use of multilingual labels} \\ 0 \text{ otherwise} \end{cases} \tag{29}$$

[22] https://www.w3.org/ns/formats/

Appendix C. Installation guide

The code provided as a result of this work can be installed in a computer by following the steps outlined below.

- Download the code from GitHub: https://github.com/hibernator11/lod-quality-reproducible.
- Open the folder with the code.
- Run the command: pip install -r requirements.txt
- Open Jupyter.
- Open and run the notebook provided by this work: lod-quality-glam-reproducible.ipynb