A Shape Expression approach for assessing the quality of Linked Open Data in Libraries

Gustavo Candela [a,*], Pilar Escobar [a], María Dolores Sáez [a] and Manuel Marco-Such [a]
[a] *Department of Software and Computing systems, University of Alicante, Spain*
E-mail: gcandela@ua.es
Victor de Boer, Vrije Universiteit Amsterdam, the Netherlands. Enrico Daga, The Open University, United Kingdom. Marieke van Erp, KNAW Humanities Cluster, the Netherlands . Eero Hyvönen, University of Helsinki, Aalto University, Finland. Albert Meroño Peñuela, Vrije Universiteit Amsterdam, the Netherlands. Harald Sack, FIZ Karlsruhe - Leibniz Institute for Information Infrastructure, Germany.

Editors: Mehwish Alam, FIZ Karlsruhe, Germany; Victor de Boer, Vrije Universiteit Amsterdam, Netherlands; Eero Hyvönen, University of Helsinki, Aalto University, Finland; Albert Meroño Peñuela, Vrije Universiteit Amsterdam, Netherlands; Harald Sack, FIZ Karlsruhe, Germany
Solicited reviews: Jouni Tuominen, Aalto University, Finland; Katherine Thornton, Yale University, USA; Marilena Daquino, University of Bologna, Italy

Abstract. Cultural heritage institutions are exploring Semantic Web technologies to publish and enrich their catalogues. Several initiatives, such as Labs, are based on the creative and innovative reuse of the materials published by cultural heritage institutions. In this way, quality has become a crucial aspect to identify and reuse a dataset for research. In this article, we propose a methodology to create Shape Expressions definitions in order to validate LOD datasets published by libraries. The methodology was then applied to four use cases based on datasets published by relevant institutions. It intends to encourage institutions to use ShEx to validate LOD datasets as well as to promote the reuse of LOD, made openly available by libraries.

Keywords: Linked Open Data, Data Quality, Libraries, Cultural Heritage

1. Introduction

Galleries, Libraries, Archives and Museums (GLAM institutions) have traditionally provided access to digital collections. The wide range of material formats include text, image, video, audio or maps.

As technologies have evolved over the years, GLAM organisations have adapted to the new environments in terms of new skills, service design or digital research [1]. Institutions have started to make their collections accessible for computational uses such as data science, Machine Learning, and Artificial Intelligence [2, 3]. Recently, the Lab concept has emerged as a means to publish digital collections as datasets amenable to computational use as well as to identify innovative and creative ways of reusing them [4]. GLAM institutions are engaging users and researchers to conduct research on the digital collections.

The *Semantic Web* was presented by Tim Berners-Lee in 2001 as an extension of the current Web in which information is structured in a way that is readable by computers [5]. The Semantic Web is based on a range of technologies that enable the connection of resources, known as *Linked Open Data* (LOD).

Applying the LOD concepts to the digital collections provided by libraries has become highly popular in the research community. Many institutions have adopted Resource Description Framework (RDF) to describe their content. In addition, collaborative editing approaches have been proposed using Wikidata and Wikibase to highlight research opportunities in community-based collections, as well as community-

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

owned infrastructure, to facilitate open scholarship practices [6, 7]. The use of LOD enhances the discoverability and impact of digital collections by transforming isolated repositories (data silos) into valuable datasets that are connected to external repositories.

However, the use of Semantic Web technologies requires complex technical skills and professional knowledge in different fields, hindering their adoption. Many aspects must be taken into account such as the vocabulary to describe the resources, the identification of external repositories to create the links and the system to store the final dataset. As in the case of other types of structured data, LOD suffers from quality problems such as inaccuracy, inconsistency, and incompleteness, impeding its full potential of reuse and exploitation.

Bibliographic records published as LOD by libraries have gained value in contexts outside of the library domain in order to connect and reuse resources [8, 9]. It is crucial to provide libraries with higher-quality and richer metadata for reuse in a linked data environment, not only to create contextual information, but also to facilitate the work of the library staff [10].

Recent studies have focused on the next generation of metadata in libraries to develop quality assurance practices [11]. Some approaches have assessed the quality of LOD using several methods and techniques [12, 13]. A preliminary query-based approach assesses the quality of the LOD published by four relevant libraries [14]. Shape Expressions (ShEx) have emerged as a concise, formal, modelling and validation language for RDF structures, addressing the Semantic Web community's need to ensure data quality for RDF graphs [15, 16]. However, to the best of our knowledge, none of these previous approaches provide a user-friendly syntax, systematic and reproducible method to assess the quality of LOD published by libraries based on ShEx as a main component.

The objective of the present study was to introduce a systematic and reproducible approach to analyse the data quality of LOD published by libraries. The methodology was then applied to four LOD repositories issued by relevant institutions. The collection of ShEx schemas provided as a result of this study is publicly available and can be used to reproduce the results and extend the examples provided using new rules based on additional properties and vocabularies.

The main contributions of this paper are as follow: (a) a methodology to assess the quality of the LOD published by libraries using ShEx; (b) the results obtained after the quality assessment; and (c) the ShEx definitions to assess LOD published by libraries.

The paper is organised as follows: after a brief description of the state of the art in Section 2, Section 3 describes the methodology employed to evaluate LOD in libraries using ShEx. Section 4 shows the results of the methodology's application. The paper concludes with an outline of the adopted methodology, and general guidelines on how to use the results and future work.

2. Related work

2.1. Background

The Semantic Web is a web of data that is machine-readable and includes a collection of technologies to describe and query the data, as well as to define standard vocabularies. Linked Data was introduced by Tim Berners-Lee [17] as a essential component of the Semantic Web to create relationships between datasets. Thus, the Resource Description Framework (RDF) [18] lies at the heart of the Semantic Web as it provides a standard model for data interchange on the Web and extends the Web's linking structure by means of URIs. In addition, SPARQL provides a standardised query language for data represented as RDF in which a query can include a list of triple patterns, conjunctions, disjunctions, and optional patterns [19].

Libraries have traditionally provided the descriptive metadata of bibliographic records using standards such as MARC.[1] While MARC is the most common formats used by libraries to publish bibliographic information, it presents limitations regarding its use as RDF, since MARC was not defined for a Web environment [20].

In this sense, several initiatives provide a more expressive and modern framework for bibliographic information based on Semantic Web technologies. Some examples include: Functional Requirements for Bibliographic Records (FRBR), the family of conceptual models [21], and Resource Description and Access (RDA) specification [22], the IFLA Library Reference Model (LRM) [23], the Bibliographic Ontology (BIBO) [24], the Bibliographic Framework Initiative (BIBFRAME) [25] and FRBRoo [26]. However, translating the old records into the new format is not an easy task [27], since libraries usually host large catalogues,

[1] https://www.loc.gov/marc/

including many types of resources that often require a manual revision to transform the data with accuracy.

Several major libraries (e.g., the OCLC, the British Library, the National Library of France), publishers, and library catalogue vendors have applied LOD to their catalogues in an effort to make these records more useful to users. For instance, the *Library of Congress Linked Data Service* (id.loc.gov) provides access to authority data. The Bibliothèque nationale de France (BnF) has published data.bnf.fr by aggregating information on works, authors and subjects. The Biblioteca Nacional de España (BNE) has transformed its catalogue to RDF and is available at datos.bne.es [28]. The Biblioteca Virtual Miguel de Cervantes (BVMC) catalogue has been transformed to RDF based on the RDA vocabulary to describe the resources [29]. The British National Bibliography (BNB) LOD platform provides access to the British National Bibliography published as LOD and made available through SPARQL.[2]

LOD promotes cultural heritage discovery and access by providing a resource context through the linking of bibliographic catalogue records to external repositories such as Wikidata, GeoNames and the Virtual International Authority File (VIAF). GLAM institutions are increasingly embracing the value of contributing information to open knowledge and collaborative projects such as Wikidata. In this sense, many institutions have linked their collections to Wikidata by means of dedicated properties. For instance, the property *BNB person ID* (P5361) at Wikidata links to the BNB LOD platform. The linking and enrichment of entities enables combining information from datasets stored in different places and with varying SPARQL endpoints [30].

2.2. Validating LOD

For researchers, data quality is a key factor when choosing a dataset for reuse [13, 31]. In this way, several methods and tools have recently emerged allowing to assess the quality of datasets built using Semantic Web technologies. In addition, the research community has highlighted the need for reproducible research by providing articles, data and code [32].

The development of tools to support data validation has accelerated over the past few years [33]. SemQuire consists of a quality assessment tool for analysing aspects of quality of particular LOD. It recommends a series of 55 intrinsic, representational, contextual and accessibility quality metrics [34]. Stardog Integrity Constraint Validation (ICV) allows to write constraints that are translated to SPARQL in order to assess RDF triples in a repository [29]. DistQualityAssessment is an open source implementation of quality assessment of large RDF datasets using Apache Spark [35]. Luzzu is a platform that assesses Linked Data quality using a library of generic and user-provided domain specific quality metrics [36].

Shapes Constraint Language (SHACL) is a World Wide Web Consortium (W3C) specification for validating graph-based data against a set of conditions [37]. As a result of the validation process, SHACL provides a validation report described with the SHACL Validation Report Vocabulary that reports the conformance and the set of all validation results. It provides advanced features such as SHACL-SPARQL that can be used to express restrictions based on a SPARQL query.

ShEx enables RDF validation through the declaration of constraints on the RDF model [38]. ShEx schemas are defined using terms from RDF semantics such as node which corresponds to one IRI, a blank node or a literal, and graph as a set of triples described as subject, predicate, object. ShEx enables defining of node constraints to determine the set of a node's allowed values, including their associated cardinalities and datatypes.

ShEx also enables the definition of constraints on the allowed neighbourhood of a node called Shape, in terms of the allowed triples that contain this node as subject or object. Listing 1 shows an example of ShEx to validate entities of type person described using FOAF.

```
<PersonShape> {
    foaf:givenName    xsd:string+,
    foaf:familyName   xsd:string ,
    foaf:phone        IRI*,
    foaf:mbox         IRI
}

# Person1 matches PersonShape
<http://domainexample/Person1>
    foaf:givenName    "Gustave" ;
    foaf:familyName   "Eiffel" ;
    # no phone number needed
```

[2] https://bnb.data.bl.uk/

Table 1
Comparison of data quality assessment tools for LOD.

Tool	Open source	Available	Grammar-oriented	Installation required
DistQualityAssessment	yes	yes	no	yes
Luzzu	yes	yes	no	yes
RDFUnit	yes	yes	no	yes
RDF Validator Service	yes	yes	no	no
SHACL	yes	yes	no	no
SemQuire	yes	yes	no	no
ShEx	yes	yes	yes	no
Stardog ICV	no	yes	no	yes

```
foaf:mbox <mailto:ge@example.com>
```

Listing 1: A ShEx Shape to validate a person described using the FOAF ontology. Person1 matches PersonShape including the required properties.

There are several implementations of ShEx including shex.js for Javascript,[3] Shaclex for Scala[4] and Java ShEx for Java.[5] In particular, shex.js includes a Simple Online Validator[6] to provide a configuration file called manifest that can load a schema, load and execute a query against a particular SPARQL endpoint, and validate the nodes selected by the query. The combination of the validation tool, the ShEx definitions and the manifests offers a reproducible environment with which to replicate the research result. Work on a collection of ShEx schemas has also begun for several vocabularies .[7]

The international research community has become increasingly interested in applying and using ShEx for the validation of RDF data. One example is the description and validation of Fast Healthcare Interoperability Resources (FHIR) for RDF transformations by means of ShEx [39]. Moreover, ShEx is employed in several Wikidata projects to ensure data quality by developing quality-control pipelines [40]. ShEx is also used to facilitate the creation of RDF resources that are validated upon creation [41]. Another approach proposes a set of mappings that can be used to convert from XML Schema to ShEx [42].

While ShEx and SHACL behave similarly with simple examples, ShEx is more grammar-oriented (shapes look like grammar rules) and SHACL is more constraint-oriented. ShEx provides an abstract syntax that can be easily serialized to several formats. SHACL uses inference (e.g. checking `rdfs:subClassOf` relationships) while ShEx focuses on RDF nodes.

Table 1 contrasts all the tools mentioned above to assess LOD by using the following features: (i) published as open source; (ii) available for use or download; (iii) using a grammar-oriented and friendly syntax; and (iv) installation required to start using it.

2.3. Data quality criteria

The definition of LOD quality criteria has been attracting ever more interest. A LOD quality model specifies a set of quality characteristics and quality measures related to Linked Data, together with formulas to calculate measures [43]. A data quality criteria according to which large-scale cross-domain LOD repositories can be analysed provides 34 data quality dimensions grouped into 4 data quality categories [13].

With regard to libraries, a methodology for assessing the quality of linked data resources based on SPARQL query templates has been presented together with an extensive evaluation of five LOD datasets, including the BNE [44]. Another example is based on Europeana; it describes an approach for capturing multilinguality as part of data quality dimensions, including completeness, consistency and accessibility [45]. A new method and the validation results of several catalogues using MARC as a metadata format identifies the structural features of the records and most frequent issues [46]. Moreover, an extensible quality assessment framework which supports multiple metadata schemas describes the requirements that must be considered during the design of such software [47]. A previous computational analysis is based on art historical linked data to assess the authoritativeness of secondary sources recording artwork attributions [48].

[3] https://github.com/shexSpec/shex.js/
[4] https://github.com/labra/shaclex/
[5] https://gforge.inria.fr/projects/shex-impl/
[6] https://rawgit.com/shexSpec/shex.js/master/packages/shex-webapp/doc/shex-simple.html
[7] https://github.com/shexSpec/schemas

A recent methodology provides the dimensions and data quality criteria to assess the LOD published by libraries (see Table 2) [14]. In particular, the dimension category includes the criterion *Consistency of statements with respect to relation constraints*. Let g be the dataset to assess and *prop* the list of properties, this criterion measures the extent to which the instance data is consistent by averaging the scores obtained from the single metrics $m_{conRelat,i}$:

$$m_{\text{conRelatRp}} = \frac{1}{n}\sum_{i=1}^{n} m_{conRelatRp,i}(g) \tag{1}$$

Let R_p be the set of all property constraints defined in *prop*,

$$R_{\text{p}} = (p,d) \mid p \,\epsilon\, prop \mid (p,d) \,\epsilon\, g \wedge isDatatype(d) \tag{2}$$

Then we can define the metrics $m_{conRelatRp}(g)$ as follows:

$$m_{\text{conRelatRp}}(g) = \frac{|\{(s,p,o) \,\epsilon\, g \mid \exists\ (p,d) \,\epsilon\, R_{\text{p}} : datatype(o) = d|\}}{|\{(s,p,o) \,\epsilon\, g \mid \exists\ (p,d) \,\epsilon\, R_{\text{p}}\}|} \tag{3}$$

In the case of an empty set of relation constraints (R_p), the respective metric should evaluate to 1.

However, in these previous works this criterion is used to assess only the properties `rdfs:range` and `owl:FunctionalProperty`. A more complete and detailed assessment based on the properties could be beneficial for potential institutions willing to publish and validate their LOD data.

These efforts provide an extensive demonstration of how LOD can be assessed, specifying how each criterion can be evaluated. Nevertheless, to the best of our knowledge, no evaluation has been conducted regarding the consistency of statements with regard to LOD relation constraints published by libraries using ShEx.

3. Methodology

This section introduces the methodology to assess the data quality of LOD published by libraries using ShEx. The procedure is described in Figure 1 and is based on 3 steps, which are detailed in the following subsections: (i) identification of resources; (ii) definition of ShEx schema; and (iii) validation. The validation step's output is a report describing the results of the evaluation.

Prior works to assess LOD are based on query-based methodologies that can be complex to reproduce for non-expert users. We used ShEx in this approach because: i) it provides a grammar-based language –similar to regular expressions– to define the rules with which to assess the data; ii) ShEx schemas can be reused to reproduce the results; and iii) ShEx schemas can be used as a starting point to be extended with additional classes and properties. In addition, the use of ShEx does not require installing software to use it.

Although LOD repository publications have recently been on the rise, in some cases and for a number of reasons, the URL is no longer available, making its reuse difficult. In this sense, their exploitation and analysis requires specific knowledge about Semantic Web technologies. Nevertheless, promoting them by way of prototypes and reuse examples may help to lower the barriers to entry.

In addition to the publication of the LOD repository, metadata can be enriched using external repositories. This information can also be assessed in order to identify duplicates as well as to validate the number of external links.

3.1. Identification of resources

The identification of resources is a crucial step when analysing the elements and properties to be assessed by means of ShEx.

Publication workflows in libraries are becoming ever more complicated as metadata maintenance is a dynamic and evolving process [49]. In this sense, bibliographic information is stored as metadata using common entities (e.g. author, work, date). Metadata comes in an increasing number of options, including FRBR, BIBFRAME, RDA, Dublin Core (DC), schema.org, Europeana Data Model (EDM) and FRBRoo. In addition, the vocabulary used to describe the contents can be complex, as in the particular case of FRBR based

Table 2
The data quality criteria to assess LOD classified by category and dimension.

Category	Dimension	Criterion
Intrinsic category	Accuracy	Syntactic validity of RDF documents
		Syntactic validity of literals
		Syntactic validity of triples
		Check of duplicate entities
	Trustworthiness	Trustworthiness on KG level
		Trustworthiness on statement level
		Using unknown and empty values
	Consistency	Check of schema restrictions during insertion of new statements
		Consistency of statements w.r.t. class constraints
		Consistency of statements w.r.t. relation constraints
Contextual category	Relevancy	Creating a ranking of statements
	Completeness	Schema completeness
		Column completeness
		Population completeness
	Timeliness	Timeliness frequency of the KG
		Specification of the validity period of statements
		Specification of the modification date of statements
Representational data-quality	Ease of understanding	Description of resources
		Labels in multiple languages
		Understandable RDF serialization
		Self-describing URIs
	Interoperability	Avoiding blank nodes and RDF reification
		Provisioning of several serialization formats
		Using external vocabulary
		Interoperability of proprietary vocabulary
Accessibility category	Accessibility	Dereferencing possibility of resources
		Availability of the KG
		Provisioning of public SPARQL endpoint
		Provisioning of an RDF export
		Support of content negotiation
		Linking HTML sites to RDF serializations
		Provisioning of KG metadata
	License	Provisioning machine-readable licensing information
	Interlinking	Interlinking via owl:sameAs
		Validity of external URIs

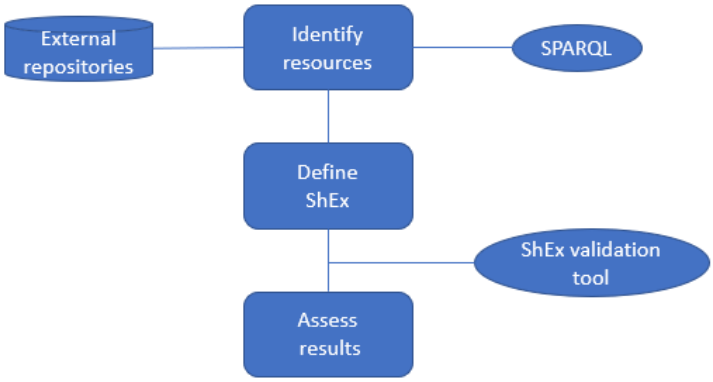

Fig. 1. Methodology for assessing the data quality of LOD repositories published by libraries using ShEx.

Table 3
Main entities described by LOD vocabularies used by libraries to publish bibliographic information.

Vocabulary	Entities
BIBFRAME	Work, Person
BIBO	Event, Agent, Book, Newspaper
FRBRoo	Work, Expression, Person, Event, Place
Dublin Core	BibliographicResource, Agent, Location
EDM	ProvidedCHO, WebResource, Aggregation, Agent, Concept
FRBR	Work, Expression, Manifestation, Agent
LRM	Work, Expression, Manifestation, Agent
RDA	Work, Expression, Manifestation, Agent
schema.org	CreativeWork, Book, Event, Person, Organization

vocabularies in which entities typed as Work follow a hierarchical organisation that includes several layers.

Table 3 shows an overview of the main entities in LOD vocabularies used by libraries to publish their catalogues.[8] Each entity may include several properties with different levels of granularity depending on the vocabulary. For instance, DC includes the properties `dc:author` and `dc:contributor`, while RDA provides a much more expressive vocabulary to describe the roles with works, expressions and manifestations.

The resources are identified by means of the SPARQL query in Listing 2 that shows an example of how to retrieve the different classes stored in an RDF repository. Several classes can be used to type the same resource. For instance, a book can be typed as `dc-terms:BibliographicResource`, as well as `bibo:Book` and `schema:Book`. Resources may be identified by descriptive URIs providing a readable description of the entity (e.g. person or work) or using opaque URIs based on identifiers that are non human-readable.[9]

```
SELECT DISTINCT ?type
WHERE {
  ?s a ?type.
}
```

Listing 2: A SPARQL query to retrieve the different classes stored in an RDF repository.

Once we extracted the main resources described in the repository and identified their type, we extracted the properties of each class using SPARQL queries. For instance, Listing 3 shows a SPARQL query to retrieve the different properties used by the class `bibo:Book`. In some cases, the fact of matching text against a regular expression pattern by means of the SPARQL instruction `regex` can facilitate the filtering and exclusion of properties that are not necessary for a particular purpose. For example, libraries may host a long list of distinct roles to describe how authors contributed to works (e.g. illustrator, transcriber or writer) though users may not be interested in these properties.

```
SELECT distinct ?p WHERE {
   ?s a bibo:Book.
   ?s ?p ?o
}
```

Listing 3: A SPARQL query to retrieve the different properties used by the class `bibo:Book`.

3.2. Definition of ShEx

A collection of RDF triples can be assessed by means of a ShEx definition to determine whether the collection meets the requirements defined in the schema.

According to the entities and its properties identified in the previous step, ShEx schemas are defined to assess RDF data. ShEx can be represented in JSON structures (ShExJ) –intended for human consumption– or a compact syntax (ShExC) –for machine processing– [15].

ShEx has several serialization formats [50]:

- a concise, human-readable compact syntax (ShExC);
- a JSON-LD syntax (ShExJ) which serves as an abstract syntax;
- an RDF representation (ShExR) derived from the JSON-LD syntax.

Following other approaches [16], the ShEx-based validation workflow for libraries consists of:

1. writing a schema for the data type in question;

[8] The prefixes used to abbreviate RDF vocabularies can be found in the appendix (Table 11).

[9] For an overview of URI patterns see, https://www.w3.org/community/bpmlod/wiki/Best_practises_-_previous_notes, requested on 9 April 2021

2. transferring that schema into the library model of items, statements, qualifiers and references;
3. writing a ShEx manifest for the library-based schema.

A manifest file includes several properties: (i) a label for the schema; (ii) a ShEx schema; (iii) a data label describing the dataset; (iv) a data property including a SPARQL endpoint; (v) the SPARQL query to retrieve the data; and (vi) a status property with the value *conformant*. Listing 4 shows an example of manifest file to test entities typed as Person in the LOD repository published by the BNE. ShEx manifests can be hosted on GitHub so they can be used by online services.

When defining the ShEx schemas to assess a dataset, previous examples of ShEx can be reused as a starting point. For instance, if a dataset is based on FOAF, we could use previous examples based on this vocabulary to define the new ShEx schema.

In addition, the definition of ShEx constraints for an existing dataset and its validation can be performed by means of graphical tools aimed at novices and experts; they enable combination and modification functionalities allowing the building of complex ShEx schema [51].

3.3. Validation results

The last step consists of the conformance of the entity data from the library with the ShEx manifest defined in the previous step.

The ShEx2 Simple Online Validator[10] can be used to test and experiment with ShEx. The prototype provides several examples showing how to use ShEx. The validator requires a ShEx expression and a SPARQL endpoint and query to retrieve the entities in order to assess them. The results are shown item by item which allows fixing possible issues in the definition of the ShEx rules.

The ShEx2 Simple Online Validator allows users to select a manifest using a button in the left-hand list. Once a manifest is selected, a query can be chosen using a button in the right-hand list. The validate button then produces a list of results according to the items retrieved by the query selected in the previous step. Users are allowed to edit the schema and query inputs in order to re-execute the query and the validation. The list of results may include errors detailing the resource and the property involved, together with a textual description. Some examples of errors and their interpretation are:

- mismatched datatype: indicates that the tool cannot match an input value with the data type it expects for the value. A common problem is related to the class `rdf:langString` used for language-tagged string values (e.g. "Pain d'épice"@fr) that are validated against `xsd:string` providing a mismatched datatype error.
- missing property: a cardinality indicates that a property requires at least one value for the property.
- exceeds cardinality: a cardinality indicates that a property requires a specific number of values for the property.

Prototypes and tools as illustrated in the example enable the reproducibility of the results. Researchers may thus replicate, reuse and extend findings, and thereby drive scientific progress. Nevertheless, there are some aspects to consider when using a LOD dataset published by a library for assessment. For instance, in order to use the ShEx2 Simple Online Validator, the DL must provide a SPARQL endpoint via the secure HTTPS protocol.

4. Assessing the quality of LOD published by libraries

This section introduces the application of the methodology introduced in Section 3 to three uses cases based on relevant libraries. An additional use case is provided to show how the methodology can be adapted to other contexts.

After having identified the main classes and properties for each LOD repository, a file including the ShEx schema was manually created for each class (e.g. bnf-manifestation.shex), detailing the prefixes used and including the constraints. In order to use the schemas, we created manifests, based on each LOD repository, containing a list of items described as follows: i) the SPARQL query as well as the SPARQL endpoint that gathers the items to be tested; ii) a label describing the schema and the data used; and iii) the ShEx schema used to assess the data. The ShEx definitions are grouped by library in a manifest file (see Table 5). Since ShEx2 Simple Online Validator can process a manifest file by adding the parameter `manifestURL`

[10] https://rawgit.com/shexSpec/shex.js/wikidata/packages/shex-webapp/doc/shex-simple.html

```
{
    "schemaLabel": "BNE person entities",
    "schemaURL": "https://raw.githubusercontent.com/hibernator11/
    ShEx-DLs/master/bne-person.shex",
    "dataLabel": "Get 20 items typed as person from datos.bne.es" ,
    "data": "Endpoint: http://datos.bne.es/sparql",
    "queryMap": "SPARQL '''select ?item where
                {\r\n ?item rdf:type ns2:C1005\r\n}\r\n limit 20'''@START",
    "status": "conformant"
}
```

Listing 4: An example of manifest file to test entities typed as Person – that corresponds to the class `ns2:C10005`– in the LOD repository published by the BNE.

to the URL to load a schema, execute a query against a SPARQL endpoint and validate the nodes retrieved by the query, the examples provided in this study are working examples that can be tested and reproduced online.[11] The project is available on GitHub[12] and has been made citable via Zenodo.[13]

When creating the ShEx schemas, preliminary tests were performed to pass the validation and after several iterations, we succeeded in addressing all the issues. For some classes that included a large number of resources, the properties were extracted manually, since the SPARQL endpoint produced some errors due to the complexity and time of the query. For instance, when using many properties in a ShEx definition, we may receive a 414 HTTP error (URI Too Long).

4.1. Selecting datasets

The selection of a LOD repository is a critical factor as well as a complicated task since many institutions are publishing their metadata as LOD. Choosing the right subject ensures the possibility of replicating existing results as well as presenting new challenges to researchers.

In this sense, benchmarks provide an experimental basis for evaluating and comparing the performance of computer systems [52, 53] as well as the possibility of replicating existing results [54]. Previous research has focused on four LOD repositories published by libraries –BVMC, BnF, BNB and BNE– that serve as a benchmark and has discussed the methodology employed to evaluate linked data in libraries [14]. Other approaches provide a list of potential LOD datasets for reuse such as the LOD Cloud[14] and the list of SPARQL endpoints provided by Wikidata.[15]

There are many aspects to consider when using a LOD repository. For instance, open licenses and clear terms of use and conditions are key when reusing datasets. Depending on the requirements, a SPARQL endpoint may be necessary in order to assess the information provided by the repository. Table 4 shows an overview of LOD repositories published by libraries and the vocabulary used.

In some cases, organisations provide a dump file instead of having a public SPARQL endpoint available. The Library of Congress, for example, suggests to download the bulk metadata and use a SPARQL engine to create custom queries such as RDF4J.[16]

The ShEx Online Validator requires a public SPARQL endpoint that uses HTTPS to test the entities. However, some organisations do not provide this protocol in their services such as the BNE and BVMC. To showcase the re-usability of our methodology to assess LOD, we identified datasets in Wikidata and the current LOD Cloud whose description contains terms such as *library*, or are included in Sections 2-3, providing a SPARQL endpoint via the secure HTTPS pro-

[11] See, for instance, https://rawgit.com/shexSpec/shex.js/wikidata/packages/shex-webapp/doc/shex-simple.html?manifestURL=https://raw.githubusercontent.com/hibernator11/ShEx-DLs/master/bnb.manifest.json

[12] https://github.com/hibernator11/ShEx-DLs

[13] https://doi.org/10.5281/zenodo.4732774

[14] https://lod-cloud.net

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

[16] https://id.loc.gov/techcenter/searching.html

Table 4
Overview of LOD repositories published by libraries.

Institution	Vocabulary	URL
Biblioteca Nacional de España	FRBR	http://datos.bne.es
Biblioteca Virtual Miguel de Cervantes	RDA	http://data.cervantesvirtual.com
Bibliothèque nationale de France	FRBR	https://data.bnf.fr
BNB Linked Data Platform	BIBO	https://bnb.data.bl.uk/
Europeana	EDM	https://pro.europeana.eu/page/sparql
Library of Congress	BIBFRAME	https://id.loc.gov/
National Library of Finland	Schema, BIBFRAME	https://data.nationallibrary.fi
National Library of Netherlands	LRM	https://data.bibliotheken.nl

tocol. As a result, we chose the following datasets for evaluation:

- BNB Linked Data platform
- BnF
- National Library of Finland (NLF)

In order to show how our methodology can be adapted to other domains, we selected an additional dataset from LOD Cloud, the Linked Open Vocabularies (LOV).[17]

The SPARQL endpoints publicly available are used to assess the LOD datasets. The main difference between the repositories is the vocabulary used to describe the information, in particular the entities and properties.

BnF and BNB are linked to Wikidata by means of specific properties. In this way, and in addition to the ShEx definitions created according to the vocabularies used by the libraries, we have created a ShEx schema per library to assess whether the resources linked to Wikidata were typed as human (`wd:Q5`) via the public Wikidata SPARQL endpoint provided by the Wikidata infrastructure.

4.2. The BNB Linked Data Platform

The BNB Linked Data Platform provides access to the British National Bibliography[18] published as LOD and made available through a SPARQL endpoint. The Linked Open BNB is a subset of the full BNB including published books, serial publications and new and forthcoming books, representing approximately 4.4 million records. It is available under a Creative Commons CC0 licence.[19]

The dataset is accessible through different interfaces: (i) a SPARQL online editor; (ii) a SPARQL endpoint for remote access; and (iii) a web interfaces providing a search box to enter a plain text term.

The BNB dataset has been modelled and represented in RDF using a number of standard schemas including the British Library Terms,[20] BIBO, the unconstrained version of the RDA element sets, schema.org and DC, amongst others. In addition, the BNB dataset has been enriched by means of the creation of links to several external datasets such as Wikidata, GeoNames and VIAF. The Book Data model provides an overview of the main classes and properties involved in the data model.[21] Figure 2 shows a summary of the main classes used in the BNB repository.

A ShEx definition was created for each class to perform the assessment. As an example, the definition corresponding to `bibo:Book` can be found in Listing 5. All the definitions were included in a manifest file that can be consumed by the ShEx validation tool [55].[22]

In addition, we created a further ShEx schema to assess the resources linked to the BNB Linked Data platform of Wikidata by means of the property *BNB person ID* (P5361) were typed as human (`wd:Q5`).

4.3. Bibliothèque nationale de France as LOD: data.bnf.fr

The data.bnf.fr project endeavours to make the data produced by Bibliothèque nationale de France more useful on the Web using Semantic Web technologies.

[17] https://lov.linkeddata.es/dataset/lov/sparql
[18] https://www.bl.uk/collection-metadata/metadata-services
[19] http://creativecommons.org/publicdomain/zero/1.0/
[20] http://www.bl.uk/schemas/bibliographic/blterms
[21] http://www.bl.uk/bibliographic/pdfs/bldatamodelbook.pdf
[22] See, for instance, https://rawgit.com/shexSpec/shex.js/wikidata/packages/shex-webapp/doc/shex-simple.html?manifestURL=https://raw.githubusercontent.com/hibernator11/ShEx-DLs/master/bnb.manifest.json

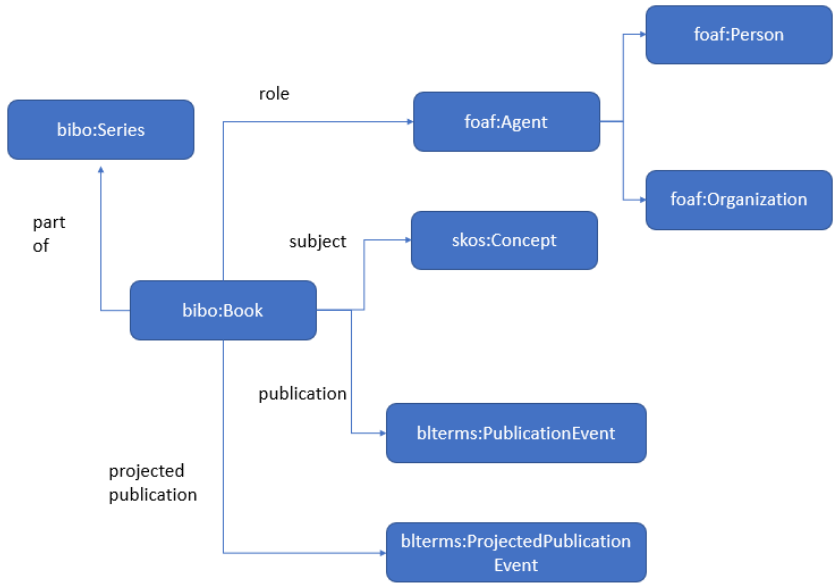

Fig. 2. Main classes retrieved from BNB LOD platform based on BIBO, SKOS and FOAF controlled vocabularies, and how they interact to create meaning.

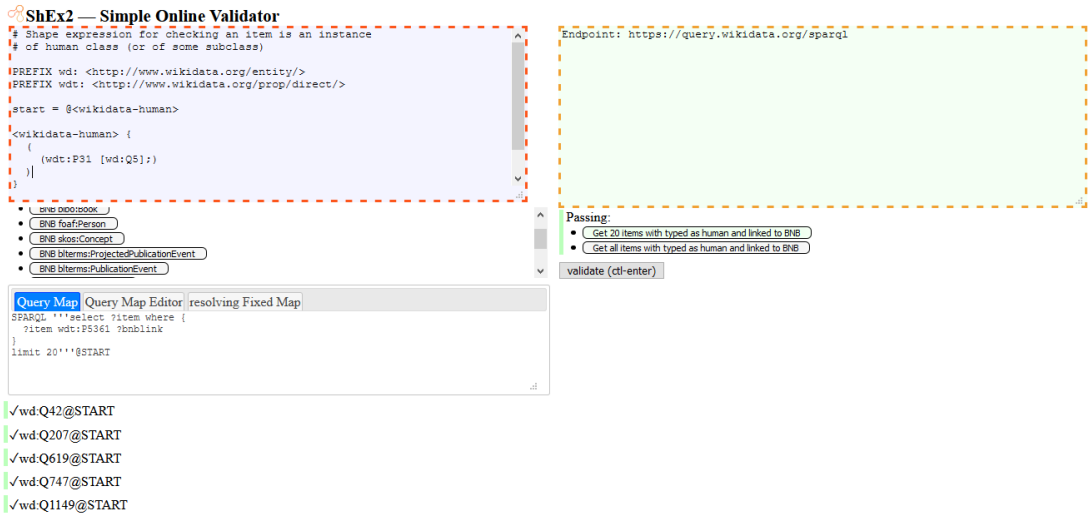

Fig. 3. The ShEx validator interface that uses the manifest file provided for the BNB Linked Data platform to assess each of the ShEx definitions showing the results. Online access to run the examples is available at https://github.com/hibernator11/ShEx-DLs/.

The dataset integrates several databases including the BnF main catalogue, BnF archives and manuscripts, and Gallica. The data model is based on FRBR, FOAF and SKOS as main vocabularies and provides links to external repositories such as GeoNames, Library of Congress and VIAF.[23] The dataset can be used via a public SPARQL endpoint or as a dump file.[24]

An overview of the main classes stored in the LOD repository has been extracted and is shown in Figure 4. A new vocabulary has been defined to describe roles in

[23] https://data.bnf.fr/en/opendata
[24] http://api.bnf.fr/dumps-de-databnffr

```
start = @<book>
<book> {
 rdfs:label  xsd:string ;
 dct:title xsd:string ;
 blterms:bnb xsd:string ;
 dct:language IRI+ ;
 dct:subject  IRI* ;
 dct:alternative xsd:string* ;
 dct:description rdf:langString* ;
 dct:abstract rdf:langString* ;
 dct:spatial IRI* ;
 bibo:isbn13 xsd:string* ;
 bibo:isbn10 xsd:string* ;
 schema:isbn xsd:string* ;
 schema:identifier IRI ;
 schema:datePublished xsd:string ;
 schema:name xsd:string ;
 schema:author IRI* ;
 schema:contributor IRI* ;
 blterms:projectedPublication IRI? ;
 blterms:publication IRI? ;
 owl:sameAs IRI* ;
 isbd:P1053 rdf:langString* ;
 isbd:P1042 rdf:langString* ;
 isbd:P1073 rdf:langString* ;
 rdau:P60048 IRI? ;
 rdau:P60049 IRI* ;
 rdau:P60050 IRI? ;
}
```

Listing 5: A ShEx Shape to validate the resources typed as `bibo:Book` at BNB Linked Data platform. Each line corresponds to a property based on a particular vocabulary used to describe the resources.

which resources are linked to the Library of Congress subject headings (LCSH).[25]

Once we extracted the main resources described in the repository and identified their type, we extracted the properties for each class using SPARQL queries. For instance, Listing 6 shows a SPARQL query to retrieve the different properties used by the class `frbr-rda:Work`. In this case, 394 unique properties where identified to define the ShEx schema including a long list of roles.

```
SELECT DISTINCT ?p
WHERE {
  ?s a frbr-rda:Work .
  ?s ?p ?o
  FILTER (!regex(?p, "roles")) .
  FILTER (!regex(?p, "relators")) .
}
```

Listing 6: A SPARQL query to retrieve the different properties used by the class `frbr-rda:Work`. The `FILTER` instructions exclude the roles and relators properties.

A ShEx schema was defined for each class to perform the validation as is shown in Listing 7. As in the previous use case, all the ShEx schemas were included

[25] https://data.bnf.fr/vocabulary/roles/

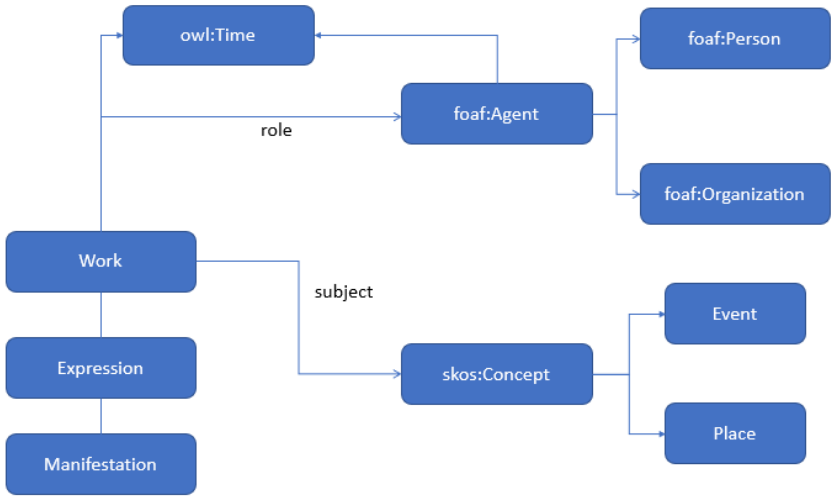

Fig. 4. Overview of the main classes based on FRBR, FOAF and SKOS retrieved from data.bnf.fr and how they interact to create meaning.

in a manifest file that is used by the validation tool as an input.

Moreover, we created an additional ShEx schema to check whether the resources linked from Wikidata to data.bnf.fr by means of the property *Bibliothèque nationale de France ID (P268)* were typed as human (`wd:Q5`).

4.4. National Library of Finland as LOD

The Finnish National Bibliography was published as LOD in 2017. The dataset containing about 40 million of RDF triples and based on schema.org and BIBFRAME, was extracted from MARC bibliographic records. The dataset contains a wide range of materials, including books, maps, journals and digitized documents.

A ShEx schema was defined for each class based on schema.org (e.g. CreativeWork, Periodical, Person and Place) to perform the validation. Properties are mainly based on schema.org as well as additional vocabularies such as the agent and the unconstrained version of the RDA element sets. Groups of related items such as Periodicals and CreativeWorkSeries were assessed by means of the properties `schema:isPartOf` and `schema:has_part` (e.g. nlf-periodical.shex). The collection of ShEx schemas were collected in a manifest file that is used by the validation tool as an input.

4.5. Linked Open Vocabularies

The purpose of LOV is to promote and facilitate the use of well documented vocabularies in the Linked Data environment [56]. The vocabulary collection is maintained by the LOV team of curators and is constantly growing (749 as of April 2021).

The data model is based on specifications to describe vocabularies including Vocabulary of a Friend (VOAF) and VANN –a vocabulary for annotating vocabulary descriptions–, and additional vocabularies such as FOAF, schema.org and DC.

A ShEx schema for the most relevant classes is provided such as `voaf:Vocabulary`, `foaf:Person`, `foaf:Organization` and `lexvo:Language`. Since some of the main classes used in LOV such as `foaf:Person` and `foaf:Organization` are used in the rest of use cases of this work, ShEx schemas were reused. In addition, a manifest file was created that includes all the ShEx schemas to be assessed.

4.6. Results and discussion

In order to assess the datasets, the main resources were identified and validated using a random sample of 1000 items retrieved per entity and library from their SPARQL endpoints. A total of 37 ShEx definitions were created to validate the LOD published by libraries. Table 5 shows the description of the classes, ShEx and manifest files used to assess the BNB and the BnF. Figure 3 shows the ShEx validation interface consuming the manifest file and presenting the evaluation results for the BNB Linked Data platform.

```
start = @<work>
<work> {
 rdfs:label  xsd:string ;
 owl:sameAs IRI* ;
 dct:title rdf:langString ;
 dct:subject  IRI* ;
 dct:creator  IRI+ ;
 dct:contributor  IRI* ;
 dct:language  IRI+ ;
 dct:description rdf:langString* ;
 dct:created xsd:string* ;
 dct:publisher xsd:string? ;
 dct:frequency xsd:string* ;
 bnf-onto:subject IRI* ;
 bnf-onto:translation IRI* ;
 bnf-onto:electronicEdition IRI* ;
 bibo:issn xsd:string* ;
 rdam:P30135 IRI? ;
 rdam:P30086 IRI? ;
 rdam:P30016 IRI? ;
 rdam:P30088 IRI? ;
 rdam:P30176 IRI? ;
 wemi:workManifested IRI? ;
 wemi:expressionOfWork IRI+ ;
 wemi:electronicReproduction IRI* ;
 }
```

Listing 7: A ShEx to validate the resources typed as `rda-frbr:Work` at data.bnf.fr. Each line corresponds to a property used to describe the resources.

Table 6 provides an overview of the data quality evaluation. All the assessed repositories obtained a high score, notably the BNB, the NLF and the BnF.

The BNB reached the highest score. We applied constraints to it based on several properties including `dct:title`, `dct:language` or `schema:author`. However, in some cases, the work's author is missing and we recommend setting them as anonymous.[26] The 414 resources typed as `foaf:Person` and `foaf:Organization` fail due to the lack of the property `schema:name`.

The BNF obtains a high score, even though that some constraints are violated. For instance, resources typed as `frbr-rda:Expression` should provide a language and there are 12 cases in which the property `dcterms:language` is missing. The Wikidata property score for the BnF can be attributed to the fact that the property is used to link all types of content such as countries, persons and organisations.

Regarding the NLF dataset, we identified properties such as `rdaa:P50025` for variant names of corporate body entities that provide a list of literals typed as `rdf:langString` and `xsd:string`, producing a data type mismatch error.[27] The node constraint can be updated (e.g., `Literal*`) to solve this issue in order to obtain a better performance. There are a number of errors due to the lack of the property `schema:inLanguage`, since the ShEx schema requires at least one value for the property. In addition, a more expressive vocabulary such as FRBR or RDA could be used to describe the relationships between aggregated works in order to differentiate the items in-

[26] See, for example, https://www.wikidata.org/wiki/Q4233718

[27] See, for example, http://finto.fi/cn/en/page/2841A.

Table 5
Description of the classes, ShEx definitions and manifest files used to assess the BNB and the BnF provided in the GitHub project.

Library	Manifest file	Class	ShEx file
BNB	bnb.manifest.json	foaf:Agent	bnb-agent.shex
		bibo:Book	bnb-book.shex
		skos:Concept	bnb-concept.shex
		blterms:PublicationEvent	bnb-publication.shex
		blterms:ProjectPublicationEvent	bnb-publication.shex
		wd:Q5	bnb-wikidata.shex
BnF	bnf.manifest.json	foaf:Agent	bnf-agent.shex
		frbr:Work	bnf-work1.shex
		frbr-rda:Work	bnf-work2.shex
		frbr-rda:Expression	bnf-expression.shex
		frbr-rda:Manifestation	bnf-manifestation.shex
		skos:Concept	bnf-concept.shex
		foaf:Organization	bnf-organization.shex
		foaf:Person	bnf-person.shex
		geo:SpatialThing	bnf-place.shex
		wd:Q5	bnf-wikidata.shex

volved (e.g. article, work and periodical publications) instead of using the relationships `schema:hasPart` and `schema:isPartOf`.

The lower score for LOV can be attributed to the lack of values for several properties. Among the LOV dataset errors are the following:

- 422 resources typed as `voaf:Vocabulary` without the properties `dct:language` and `dct:creator`.
- 162 persons without a property `foaf:name`.
- 162 organisations without a `foaf:name`.

In general, the ShEx schemas could be improved by setting the constraint `owl:sameAs` to a more restricted cardinality (e.g. 1 or more) in order to identify resources that are not linked to external repositories. In some cases, the number of resources to test were below 1000 given the random selection, such as the case of the class `voaf:Vocabulary` from LOV.

Moreover, some resources may not include sufficient information to be assessed. For instance, the resources typed as `skos:Concept` only includes a `rdfs:label` in the BNB Linked Data platform. In some cases, the same schema can be used for different classes such as `blterms:PublicationEvent` and `blterms:ProjectedPublicationEvent` because they are based on the same properties and vocabularies. The ShEx definitions defined for the BnF are more detailed since the FRBR model provides additional classes to describe the resources compared to the BIBO vocabulary. Tables 7, 8, 9 and 10 illustrate the total errors per dataset aggregated by class.

The results of the assessment are useful for librarians in several ways since they provide valuable information with which to refine and improve their LOD catalogues. It is thus possible to identify potential properties that are not properly used to describe the bibliographic information. In the same way, they can measure the extent to which entities are described by means of a sufficient number of properties. For example, a librarian could be interested in assessing if the authors contain at least a name, a date of birth and an identifier matching a specific pattern.

The ShEx schemas provided in this study can be used as a starting point for other institutions willing to assess their LOD. In this way, the schemas could be further refined with additional node constraints as well as the incorporation of new vocabularies to assess further datasets. The adoption and use of this methodology in other contexts is also feasible as is shown in the variety of datasets and vocabularies used to assess the methodology.

With regard to the methodology, this approach is limited to one data quality dimension. In order to improve the methodology, additional data quality dimensions and criteria could be used such as license, completeness and trustworthiness (see Table 2). In addition, the ShEx schemas are based on the most relevant classes in each dataset.

Table 6

Evaluation overview for the four datasets. For each dataset we display the total number of triples, the number of classes and properties assessed, the total number of evaluated items, how many tests passed, failed and did time out (TO). The last column shows the result for the data quality criterion *Consistency of statements with regard to relation constraints*.

Dataset	Triples	Classes	Properties	Tests	Pass	Fail	TO	m_{conRelat}
BNB	151,779,391	5	20	6872	6377	495	0	0,93
BnF	334,457,101	10	20	8982	8164	818	0	0,91
LOV	998,333	8	20	4945	4243	702	79	0,86
NLF	40,000,000	10	20	8977	8161	816	24	0,91

Table 7

Evaluation results aggregated by class for the BNB dataset.

Classes	Pass	Fail	TO
bibo:Book	1000	0	0
foaf:Person	960	40	0
foaf:Organization	626	374	0
skos:Concept	902	6	0
blterms:PublicationEvent	937	63	0
blterms:ProjectPublicationevent	998	2	0
wd:Q5	954	10	0

Table 8

Evaluation results aggregated by class for the BNF dataset.

Classes	Pass	Fail	TO
frbr-rda:Work (part 1)	998	2	0
frbr-rda:Work (part 2)	995	5	0
frbr-rda:Expression	988	12	0
frbr-rda:Manifestation	995	5	0
foaf:Person	944	56	0
foaf:Organization	1000	0	0
skos:Concept	898	102	0
geo:SpatialThings	1000	0	0
wd:Q5	346	636	0

Table 9

Evaluation results aggregated by class for the NLF dataset.

Classes	Pass	Fail	TO
schema:Person	997	3	0
schema:PublicationEvent	1000	0	0
schema:Organization	1000	0	0
schema:CreativeWork	991	2	7
loc:Work	959	28	13
schema:Place	1000	0	0
schema:CreativeWorkSeries	585	415	0
schema:Book	996	4	0
schema:Periodical	632	364	4
schema:Collection	1	0	0

Table 10

Evaluation results aggregated by class for the LOV dataset.

Classes	Pass	Fail	TO
foaf:Person	838	162	0
voaf:Vocabulary	361	367	55
foaf:Organization	164	16	0
lexvo:Language	186	0	0
rev:Review	100	0	0
voaf:DatasetOccurrences	1000	0	0
dcat:Distribution	843	157	0
dcat:CatalogRecord	750	0	0
dcat:Catalog	1	0	0

5. Conclusions

Libraries are using Semantic Web technologies to publish and enrich their catalogues. While LOD repositories can be reused in innovative and creative ways, data quality has become a crucial factor for identifying a dataset for reuse.

Based on previous research, we defined a methodology described in Section 3 to assess the quality of LOD repositories published by libraries that uses ShEx as a main component. The methodology was applied to four use cases, resulting in a collection of ShEx schemas that can be tested online and reused by other institutions as a starting point to evaluate their LOD repositories. Our evaluation showed that ShEx can be useful to assess LOD data published by libraries. In addition, ShEx can be used as documentation since it provides a human-readable representation that helps librarians and researchers to understand the data model.

Future work to be explored includes the improvement of the ShEx definitions and the inclusion of additional use cases. Moreover, the extension of the ShEx validation tool in terms of libraries requirements such as common classes and properties used by libraries will be analysed.

Acknowledgements

This research has been funded by the AETHER-UA (PID2020-112540RB-C43) Project from the Spanish Ministry of Science and Innovation.

Appendix A. List of prefixes

The prefixes in Table 11 are used to abbreviate namespaces throughout this paper.

Appendix. References

[1] Research Libraries UK, A manifesto for the digital shift in research libraries, 2020, [Online; accessed 20-October-2020]. https://www.rluk.ac.uk/digital-shift-manifesto/.

[2] T. Padilla, Responsible Operations: Data Science, Machine Learning, and AI in Libraries, OCLC Research, 2019. https://doi.org/10.25333/xk7z-9g97.

[3] T. Padilla, L. Allen, H. Frost, S. Potvin, E. Russey Roke and S. Varner, Final Report — Always Already Computational: Collections as Data, Zenodo, 2019. doi:10.5281/zenodo.3152935.

[4] 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.

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

[6] J. Godby, S.-Y. Karen, W. Bruce, D. Kalan, D. Karen, F.E. Christine, S. Folsom, X. Li, M. McGee, K. Miller, H. Moody, H. Tomren, and C. Thomas, Creating Library Linked Data with Wikibase: Lessons Learned from Project Passage, OCLC Research, 2019. https://doi.org/10.25333/faq3-ax08.

[7] Association of Research Libraries, ARL Task Force on Wikimedia and Linked Open Data. ARL White Paper on Wikidata: Opportunities and Recommendations, 2019. https://www.arl.org/wp-content/uploads/2019/04/2019.04.18-ARL-white-paper-on-Wikidata.pdf.

[8] I.W.G. on Guidelines for National Bibliographies, Best Practice for National Bibliographic Agencies in a Digital Age, 2019. https://www.ifla.org/ES/node/7858.

[9] M. Frosterus, M. Dadvar, D. Hansson, M. Lappalainen and S. Zapounidou, Linked Open Data: Impressions & Challenges Among Europe's Research Libraries, Zenodo, 2020. doi:10.5281/zenodo.3647844.

[10] G. Bahnemann, M. Carroll, P. Clough, M. Einaudi, C. Ewing, J. Mixter, J. Roy, H. Tomren, B. Washburn and E. Williams, Transforming Metadata into Linked Data to Improve Digital Collection Discoverability: A CONTENTdm Pilot Project., OCLC Research, 2021. doi:https://doi.org/10.25333/fzcv-0851.

[11] K. Smith-Yoshimura, Transitioning to the Next Generation of Metadata, OCLC Research, 2020. doi:https://doi.org/10.25333/rqgd-b343.

[12] 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.

[13] 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.

[14] G. Candela, P. Escobar, R.C. Carrasco and M. Marco-Such, Evaluating the quality of linked open data in digital libraries, *Journal of Information Science* **0**(0) (2020), 0165551520930951. doi:10.1177/0165551520930951.

[15] E. Prud'hommeaux, I. Boneva, J.E.L. Gayo and G. Kellogg, Shape Expressions Language 2.1, 2019. http://shex.io/shex-semantics/index.html.

[16] K. Thornton, H. Solbrig, G.S. Stupp, J.E. Labra 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*, P. Hitzler, M. Fernández, K. Janowicz, A. Zaveri, A.J.G. Gray, V. Lopez, A. Haller and K. Hammar, eds, Springer International Publishing, Cham, 2019, pp. 606–620. ISBN 978-3-030-21348-0.

[17] Tim Berners-Lee, Linked-data design issues. W3C design issue document, 2006. http://www.w3.org/DesignIssues/LinkedData.html.

[18] World Wide Web Consortium, RDF 1.1 Concepts and Abstract Syntax, 2014. https://www.w3.org/TR/rdf11-concepts/.

[19] World Wide Web Consortium, SPARQL 1.1 Query Language, 2013. https://www.w3.org/TR/sparql11-query/.

[20] T.W. Cole, M.-J. Han, W.F. Weathers and E. Joyner, Library Marc Records Into Linked Open Data: Challenges and Opportunities, *Journal of Library Metadata* **13**(2–3) (2013), 163–196. doi:10.1080/19386389.2013.826074.

[21] IFLA, *IFLA Study Group on the FRBR. Functional Requirements for Bibliographic Records*, IFLA Series on Bibliographic Control, München, 1998.

[22] RDA Steering Committee, RDA Toolkit: Resource Description and Access, 2012, [Online; accessed 2-October-2020].

[23] IFLA, IFLA Library Reference Model (LRM), 2017. https://www.ifla.org/files/assets/cataloguing/frbr-lrm/ifla-lrm-august-2017_rev201712.pdf.

[24] B. D'Arcus and F. Giasson, Bibliographic Ontology, 2009. http://bibliontology.com/.

[25] Library of Congress, Bibliographic Framework Initiative. https://www.loc.gov/bibframe/.

[26] Working Group on FRBR/CRM Dialogue, Definition of FRBRoo. A Conceptual Model for Bibliographic Informationin Object-Oriented Formalism, 2015. http://www.cidoc-crm.org/frbroo/.

[27] T. Aalberg and M. Zumer, Looking for Entities in Bibliographic Records, in: *Digital Libraries: Universal and Ubiquitous Access to Information, 11th International Conference on Asian Digital Libraries, ICADL 2008, Bali, Indonesia, December 2-5, 2008. Proceedings*, G. Buchanan, M. Masoodian and S.J. Cunningham, eds, Lecture Notes in Computer Science, Vol. 5362, Springer, 2008, pp. 327–330. doi:10.1007/978-3-540-89533-6_36.

Table 11
Common prefixes used to designate RDF vocabularies.

prefix	URI
bibo	http://purl.org/ontology/bibo/
blterms	http://www.bl.uk/schemas/bibliographic/blterms#
bneonto	http://datos.bne.es/def/
bnf-onto	http://data.bnf.fr/ontology/bnf-onto/
bnfroles	http://data.bnf.fr/vocabulary/roles/
dcat	http://www.w3.org/ns/dcat#
dcmitype	http://purl.org/dc/dcmitype/
dcterms	http://purl.org/dc/terms/
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/
geo	http://www.w3.org/2003/01/geo/wgs84_pos#
lexvo	http://lexvo.org/ontology#
loc	http://id.loc.gov/ontologies/bibframe/
owl	http://www.w3.org/2002/07/owl#
prov	http://www.w3.org/ns/prov#
rdaa	http://rdaregistry.info/Elements/a/
rdac	http://rdaregistry.info/Elements/c/
rdae	http://rdaregistry.info/Elements/e/
rdam	http://rdaregistry.info/Elements/m/
rdaw	http://rdaregistry.info/Elements/w/
rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#
rdfs	http://www.w3.org/2000/01/rdf-schema#
rdau	http://rdaregistry.info/Elements/u/
rev	http://purl.org/stuff/rev#
schema	http://schema.org/
skos	http://www.w3.org/2004/02/skos/core#
vaan	http://purl.org/vocab/vann/
voaf	http://purl.org/vocommons/voaf#
wdt	http://www.wikidata.org/entity/
wd	http://www.wikidata.org/entity/
wemi	http://rdvocab.info/RDARelationshipsWEMI/
xsd	http://www.w3.org/2001/XMLSchema#

[28] D. Vila-Suero, B. Villazón-Terrazas and A. Gómez-Pérez, datos.bne.es: A library linked dataset, *Semantic Web* **4**(3) (2013), 307–313. doi:10.3233/SW-120094.

[29] G. Candela, P. Escobar, R.C. Carrasco and M. Marco-Such, Migration of a library catalogue into RDA linked open data, *Semantic Web* **9**(4) (2018), 481–491. doi:10.3233/SW-170274.

[30] K. Coombs, Federated Queries with SPARQL, 2016. https://www.oclc.org/developer/news/2016/federated-queries-with-sparql.en.html.

[31] 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.

[32] B. Baillieul John, O. Hall Larry, M.F. Moura Jose, S. Hemami Sheila, G. Setti, G. Grenier, B. Forster Michael, F. Zappulla, J. Keaton, D. McCormick and L. Moore Kenneth, The first IEEE workshop on the Future of Research Curation and Research Reproducibility, OpenBU, 2017. https://open.bu.edu/handle/2144/39028.

[33] E. Prud'hommeaux, RDF Validation Service, 2006, [Online; accessed 02-April-2021].

[34] A. Langer, V. Siegert, C. Göpfert and M. Gaedke, SemQuire - Assessing the Data Quality of Linked Open Data Sources Based on DQV, in: *Current Trends in Web Engineering - ICWE 2018 International Workshops, MATWEP, EnWot, KD-WEB, WEOD, TourismKG, Cáceres, Spain, June 5, 2018, Revised Selected Papers*, 2018, pp. 163–175. doi:10.1007/978-3-030-03056-8_14.

[35] G. Sejdiu, A. Rula, J. Lehmann and H. Jabeen, A Scalable Framework for Quality Assessment of RDF Datasets, *CoRR* **abs/2001.11100** (2020). https://arxiv.org/abs/2001.11100.

[36] 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.

[37] World Wide Web Consortium, Shapes Constraint Language (SHACL), 2017, [Online; accessed 02-April-2021].

[38] E. Prud'hommeaux, J.E. Labra Gayo and H. Solbrig, Shape Expressions: An RDF Validation and Transformation Language, in: *Proceedings of the 10th International Conference on Semantic Systems*, SEM '14, Association for Computing Machinery, New York, NY, USA, 2014, pp. 32–40–. ISBN 9781450329279. doi:10.1145/2660517.2660523.

[39] H.R. Solbrig, E. Prud'hommeaux, G. Grieve, L. McKenzie, J.C. Mandel, D.K. Sharma and G. Jiang, Modeling and validating HL7 FHIR profiles using semantic web Shape Expressions (ShEx), *Journal of Biomedical Informatics* **67** (2017), 90–100. doi:https://doi.org/10.1016/j.jbi.2017.02.009. http://www.sciencedirect.com/science/article/pii/S1532046417300345.

[40] 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.

[41] J.C.J. van Dam, J.J. Koehorst, P.J. Schaap and M. Suárez-Diez, The Empusa code generator: bridging the gap between the intended and the actual content of RDF resources, *CoRR* **abs/1812.04386** (2018). http://arxiv.org/abs/1812.04386.

[42] H. García-González and J.E.L. Gayo, XMLSchema2ShEx: Converting XML validation to RDF validation, *Semantic Web* **11**(2) (2020), 235–253. doi:10.3233/SW-180329.

[43] F. Radulovic, N. Mihindukulasooriya, R. García-Castro and A. Gómez-Pérez, A comprehensive quality model for Linked Data, *Semantic Web* **9**(1) (2018), 3–24. doi:10.3233/SW-170267.

[44] 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*, 2014, pp. 747–758. doi:10.1145/2566486.2568002.

[45] V. Charles, J. Stiller, P. Király, W. Bailer and N. Freire, Data Quality Assessment in Europeana: Metrics for Multilinguality, in: *Joint Proceedings of the 1st Workshop on Temporal Dynamics in Digital Libraries (TDDL 2017), the (Meta)-Data Quality Workshop (MDQual 2017) and the Workshop on Modeling Societal Future (Futurity 2017) co-located with 21st International Conference on Theory and Practice of Digital Libraries (TPLD 2017), Thessaloniki, Greece, September 21, 2017*, 2017. http://ceur-ws.org/Vol-2038/paper6.pdf.

[46] P. Király, Validating 126 Million MARC Records, in: *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage*, DATeCH2019, Association for Computing Machinery, New York, NY, USA, 2019, pp. 161–168–. ISBN 9781450371940. doi:10.1145/3322905.3322929.

[47] P. Király, Towards an Extensible Measurement of Metadata Quality, in: *Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage*, DATeCH2017, Association for Computing Machinery, New York, NY, USA, 2017, pp. 111–115–. ISBN 9781450352659. doi:10.1145/3078081.3078109.

[48] M. Daquino, A computational analysis of art historical linked data for assessing authoritativeness of attributions, *Journal of the Association for Information Science and Technology* **71**(7) (2020), 757–769. doi:https://doi.org/10.1002/asi.24301. https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.24301.

[49] J. Baxmeyer, K. Coyle, J. Dyla, M. Han, S. Folsom, P. Schreur and T. Thompson, Linked Data Infrastructure Models: Areas of Focus for PCC Strategies, 2017. https://www.loc.gov/aba/pcc/documents/LinkedDataInfrastructureModels.pdf.

[50] J.E. Labra Gayo, E. Prud'hommeaux, I. Boneva and D. Kontokostas, *Validating RDF Data*, Synthesis Lectures on the Semantic Web: Theory and Technology, Vol. 7, Morgan & Claypool Publishers LLC, 2017, pp. 1–328. doi:10.2200/s00786ed1v01y201707wbe016.

[51] I. Boneva, J. Dusart, D. Fernández Alvarez and J.E.L. Gayo, Shape Designer for ShEx and SHACL Constraints, 2019, Poster. https://hal.archives-ouvertes.fr/hal-02268667.

[52] S.E. Sim, S.M. Easterbrook and R.C. Holt, Using Benchmarking to Advance Research: A Challenge to Software Engineering, in: *Proceedings of the 25th International Conference on Software Engineering, May 3-10, 2003, Portland, Oregon, USA*, 2003, pp. 74–83. doi:10.1109/ICSE.2003.1201189.

[53] S.S. Heckman and L. Williams, On establishing a benchmark for evaluating static analysis alert prioritization and classification techniques, in: *Proceedings of the Second International Symposium on Empirical Software Engineering and Measurement, ESEM 2008, October 9-10, 2008, Kaiserslautern, Germany*, 2008, pp. 41–50. doi:10.1145/1414004.1414013.

[54] B. Spahiu, A. Maurino and R. Meusel, Topic profiling benchmarks in the linked open data cloud: Issues and lessons learned, *Semantic Web* **10**(2) (2019), 329–348. doi:10.3233/SW-180323.

[55] E. Prud'hommeaux, T. Baker, Glenna, J.E.L. Gayo, mrolympia, andrawaag, L. Werkmeister and D. Booth, shex.js - Javascript implementation of Shape Expressions, Zenodo, 2018. doi:10.5281/zenodo.1213693.

[56] P. Vandenbussche, G. Atemezing, M. Poveda-Villalón and B. Vatant, Linked Open Vocabularies (LOV): A gateway to reusable semantic vocabularies on the Web, *Semantic Web* **8**(3) (2017), 437–452. doi:10.3233/SW-160213.