Semantic Web 0 (0) 1
IOS Press

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

Gustavo Candela [a,*], Pilar Escobar [a] and María Dolores Sáez [a] and Manuel Marco-Such [a]
[a] *Departmento de Lenguajes y Sistemas Informáticos, Universidad de Alicante, Spain*
E-mail: gcandela@ua.es

Abstract. Cultural heritage institutions are exploring Semantic Web technologies to publish and enrich their catalogs. Several initiatives such as Labs are based on the reuse of the materials published by cultural heritage institutions in innovative and creative ways. In this sense, quality has become a crucial aspect when identifying and reusing a dataset for research. In this article, we propose a methodology to create Shape Expressions definitions to validate LOD datasets published by digital libraries. This methodology is then applied to two use cases based on datasets published by relevant GLAM 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 digital libraries.

Keywords: Linked Open Data, Data Quality, Digital libraries, Cultural Heritage

1. Introduction

Galleries, Libraries, Archives and Museums (GLAM) institutions have traditionally provided access to digital collections. The diversity of materials 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 use such as data science, Machine Learning, and Artificial Intelligence [2, 3]. Recently, the concept of Lab has appeared as a means to explore innovate and creative ways of reusing and exploiting digital collections published by GLAM institutions [4]. GLAM institutions are exploring user engagement as a means to conduct research with 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 it can be read by computers [5]. The Semantic Web is based on a set of technologies that enable the connection of resources known as *Linked Open Data* (LOD).

The application of the LOD concepts to the digital collections provided by GLAM institutions has become very popular in the research community. Many institutions are exploring the adoption of Resource Description Framework (RDF) to describe their content. In addition, the adoption of collaborative edition approaches has been explored using Wikidata and Wikibase [6, 7] to highlight opportunities for research in community-based collections and collective collection. 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. Moreover, LOD has become crucial for search engines in which meaningful results are the key to unlocking a successful and rich user experience [8].

However, the use and exploitation of the Semantic Web technologies requires complex technical skills and professional knowledge in different areas hindering its adoption. Many aspects must be taken into ac-

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

count 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. Similar to other types of structured data, LOD suffers from quality problems such as inaccuracy, inconsistency, and incompleteness which impedes its reuse and exploitation to its full potential.

Some approaches have assessed the quality of LOD using several methods and techniques [9, 10]. Shape Expressions (ShEx) have emerged as a concise, formal, modeling and validation language for RDF structures, addressing the need of the Semantic Web community to ensure data quality for RDF graphs [11, 12]. A preliminary query-based approach assesses the quality of the LOD published by four relevant DLs [13]. However, none of them provides a systematic and reproducible approach to assess the quality of the LOD published by DLs based on ShEx as a main component.

The purpose of this paper is to introduce a systematic and reproducible approach with which to analyze the data-quality of LOD published by libraries. The methodology is then applied to two LOD repositories published by relevant institutions. The results of this study could then be used to reproduce and extend the methodology and the ShEx definitions, as well as to identify candidate datasets for reuse in innovative and creative ways.

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

The paper is organized 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 DLs using ShEx. Section 4 shows the results of the application of the methodology. The paper concludes with an outline of the methodology adopted, general guidelines for the use of 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, and define standard vocabularies. The goal is that information is useful for humans using a browser but also for machines that may automatically process that data. Linked Data was introduced by Tim Berners-Lee [14] as a essential component of the Semantic Web to create relationships among datasets. In this sense, Resource Description Framework (RDF) [15] lies at the heart of Semantic Web providing a standard model for data interchange on the Web and extending the linking structure of the Web by means of URIs. In addition, SPARQL provides a query language for RDF providing a list of instructions [16].

Libraries have traditionally provided the descriptive metadata of bibliographic records using standards such as MARC.[1] While MARC is the most common formats used in libraries to publish bibliographic information, it has limitations for use as RDF, since MARC was not defined for a Web environment [17].

In this sense, several initiatives provide a more expressive and modern framework for bibliographic information based on Semantic Web technologies. Some examples are Functional Requirements for Bibliographic Records (FRBR) family of conceptual models [18] and Resource Description and Access (RDA) specification [19], IFLA Library Reference Model (LRM) [20], Bibliographic Ontology (BIBO) [21], Bibliographic Framework Initiative (BIBFRAME) [22] and CIDOC Conceptual Reference Model (CRM) [23]. However, the translation of the old records into the new format is not an easy task [24], since libraries usually host large catalogs including many types of resources that in many occasions requires a manual revision for an accurate transformation of the data.

Several major libraries (e.g., OCLC, British Library, National Library of France, publishers, and library catalog vendors) have applied LOD to their catalogs 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 concerning works, authors and subjects. The Biblioteca Nacional de España (BNE) has transformed its catalog to RDF and is available at datos.bne.es [25]. The Biblioteca Virtual Miguel de Cervantes (BVMC) catalog has been transformed to RDF based on the RDA vocabulary to describe the resources [26]. The British National Bibliography (BNB) LOD platform provides

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

access to the British National Bibliography published as LOD and made available through SPARQL.[2]

LOD enhance discovery of and access to cultural heritage providing context about resources by linking bibliographic catalog records to external repositories such as Wikidata, GeoNames and Virtual International Authority File (VIAF). In this sense, 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 the combination of information from datasets that are stored in different places and have different SPARQL endpoints [27]. Figure 1 shows an example of federated SPARQL query run in the Wikidata query browser to retrieve the works of William Shakespeare from the BNB LOD platform.

2.2. *Validating LOD*

Data quality is a crucial aspect for data researchers when making the choice of a dataset for reuse [10, 28]. In this sense, several methods and tools have recently appeared in order to assess the quality of datasets built upon Semantic Web technologies. In addition, the research community have highlighted the need for reproducible research by providing articles, as well as, data and code [29].

Stardog Integrity Constraint Validation (ICV) allows to write constraints that are translated to SPARQL in order to assess RDF triples in a repository [26]. Several approaches provide data quality criteria according to which LOD repositories can be analyzed. They have contributed to understand and specify data quality on several dimensions regarding data quality (e.g., accuracy, completeness, licensing) [10, 30, 31]. These efforts are mostly concentrated on the evaluation of repositories focused on general knowledge rather than specific domains such as cultural heritage or literature.

ShEx enables RDF validation through the declaration of constraints on the RDF model [32]. ShEx 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 the definition of node constraints to define the set of allowed values of a node. For instance, the type of a RDF node, literal datatype, XML String and numeric facets and enumeration of value sets. 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. Figure 2 shows an example of ShEx to validate entities of type person described using FOAF.

There are several implementation of ShEx including shex.js for Javascript,[3] Shaclex for Scala[4] and Java ShEx for Java.[5] In particular, shex.js include 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 enables a reproducible environment with which to reproduce the research result. A collection of ShEx schemas has also been started for several vocabularies.[7]

The use and application of ShEx for validating RDF data has gained international interest in the research community. For instance, the description and validation of Fast Healthcare Interoperability Resources (FHIR) to RDF transformations by means of ShEx [33]. In addition, ShEx is used in several Wikidata projects to ensure data-quality by developing quality-control pipelines [34]. ShEx is also used to facilitate the creation of RDF resources that are validated upon creation [35]. Another approach proposes a set of mappings that can be used to convert from XML Schema to Shape Expressions (ShEx) [36].

With regard to DLs, previous works are focused on the identification of LOD repositories published by DLs and their assessment by means of a data-quality criteria [13]. Another approach is based on Europeana and multilinguality describing the measures defined and providing initial interpretations of the results [37]. A new method and results of validation of several catalogs using MARC as metadata format identifies the structural features of the records and most frequent issues [38]. Moreover, an extensible quality assessment framework which supports multiple metadata schemas

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

[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

```
SELECT ?work ?workLabel WHERE {
  wd:Q692 wdt:P5361 ?id
  BIND(uri(concat("http://bnb.data.bl.uk/id/person/", ?id)) as ?bnbID)
  SERVICE <http://bnb.data.bl.uk/sparql> {
    ?work dct:creator ?bnbID.
    ?work dct:title ?workLabel
  }
}
LIMIT 20
```

Fig. 1. A federated SPARQL query run in Wikidata query browser to retrieve the works of William Shakespeare from the BNB LOD platform.

```
<PersonShape> {                    # An <PersonShape> has:
    foaf:givenName  xsd:string+,   # at least one givenName.
    foaf:familyName xsd:string,    # one familyName.
    foaf:phone      IRI*,          # any number of phone numbers.
    foaf:mbox       IRI            # one FOAF mbox.
}

# Person1 matches PersonShape
<http://domainexample/Person1>
    foaf:givenName  "Gustave" ;
    foaf:familyName "Eiffel" ;
    # no phone number needed
    foaf:mbox       <mailto:ge@domainexample.com>
    .
```

Fig. 2. A ShEx Shape to validate a person described using the FOAF ontology. Person1 matches PersonShape including the required properties.

describes the requirements that need to be considered during the design of such a software [39]. However, to the best of our knowledge, none has been carried out an evaluation of the LOD published by DLs using ShEx.

This paper is based on the previously published benchmarking and data-quality criteria applied to the LOD made openly available by four relevant libraries: the Biblioteca Nacional de España (BNE), the Bibliothèque nationale de France (BnF), the British National Bibliography (BNB) and the Biblioteca Virtual Miguel de Cervantes (BVMC) [13] .

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 3 and is based on 4 steps, which are detailed in the following subsections: (i) selection of a repository, (ii) identification of resources, (iii) definition of ShEx rules and (iv) validation. The output of the validation step is a report describing the results of the evaluation.

We have selected ShEx in this methodology since it has become very popular in the research community. In addition, ShEx enables reproducibility allowing researchers to improve the definitions. Moreover, several tools have recently been developed to automate the validation process.

Although the publication of LOD repositories has recently increased, there are some cases in which the URL is no longer available due to several reasons and making difficult its reuse. In this sense, their exploitation and analysis requires specific knowledge about Semantic Web technologies. However, their promotion by means of prototypes and examples of reuse can help easing entry barriers.

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. Selection

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 to address new challenges by researchers.

In this sense, benchmarks provide an experimental basis for evaluating and comparing the performance of computer systems [40, 41] as well as the possibility of replicating existing results [42]. Previous research are based on four LOD repositories published by libraries –BVMC, BnF, BNB and BNE– that serve as benchmark and discusses the methodology employed to evaluate linked data in DLs [13].

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 1 shows an overview of LOD repositories published by libraries and the vocabulary used.

3.2. Identification of resources

The identification of resources is a crucial step in order to analyse the elements and properties to be assessed by means of ShEx.

Publication workflows in libraries are increasingly complicated since maintenance of the metadata is a dynamic and evolving process [43]. In this sense, bibliographic information is stored as metadata using common resources (e.g. author, work, date). Metadata is available in an increasing number of options including FRBR, BIBFRAME, RDA, Dublin Core (DC), schema.org, Europeana Data Model (EDM) and CIDOC-CRM. In addition, the vocabulary used to describe the contents can be complex such as the particular case of FRBR based vocabularies in which entities typed as Work follow a hierarchical organization including several layers.

Table 2 shows an overview of the main entities in LOD vocabularies used by libraries to publish their catalogs.[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 expressive vocabulary to describe the roles with works, expressions and manifestations.

[8]The prefixes used to abbreviate RDF vocabularies can be found in the appendix (Table 4).

3.3. Definition of ShEx rules

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 rules 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– [44].

ShEx has several serialization formats [45]:

- 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 [12], the ShEx-based validation workflow for DLs consists of:

1. writing a schema for the data type in question;
2. transferring that schema into the DL model of items, statements, qualifiers and references;
3. writing a ShEx manifest for the DL-based schema;

A manifest file include 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*. Figure 4 shows an example of manifest file to test entities typed as Person at the LOD repository published by the BNE. ShEx manifests can be hosted on GitHub so it can be used by online services.

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 that enable combination and modification functionalities in order to build a complex ShEx schema [46].

3.4. Validation results

The last step consists on testing entity data from the DL for conformance to the ShEx manifest defined in the previous step.

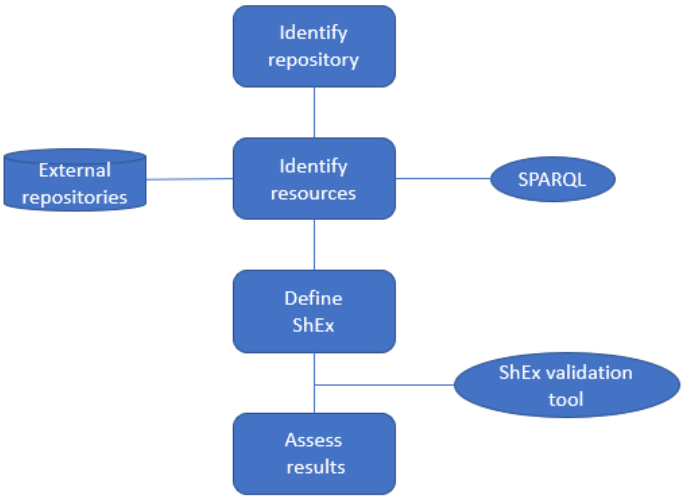

Fig. 3. Methodology for assessing the data-quality of LOD repositories published by digital libraries using ShEx.

Table 1
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	data.cervantesvirtual.com
Bibliothèque nationale de France	FRBR	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 Netherlands	LRM	https://data.bibliotheken.nl

```
{
    "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"
}
```

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

Table 2
Main entities described by LOD vocabularies used by digital libraries to publish bibliographic information.

Vocabulary	Entities
BIBFRAME	Work, Person
BIBO	Event, Agent, Book, Newspaper
CIDOC-CRM	Creation, 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

The ShEx2 Simple Online Validator[9] can be used to test and experiment with ShEx. The prototype provides several examples to show 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 allowing to fix possible issues in the definition of the ShEx rules.

Prototypes and tools like this example enable the reproducibility of the results allowing researchers to replicate, reuse and extend findings, and thereby driving scientific progress. However, There are some aspects to consider when using a LOD dataset published by a DL 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 DLs

According to the previously published benchmarking applied to the LOD made openly available by DLs [13], this section introduces the application of the methodology introduced in Section 3 to two uses cases based on relevant libraries. In the first use case, the BNB Linked Data platform is used by means of its SPARQL endpoint. In the second use case, the validation is performed using the LOD repository published by the BnF. The main difference between the two repositories is the vocabulary used to describe the bibliographic information, in particular the entities and properties.

Both datasets are linked to Wikidata by means of specific properties. In this sense, and in addition to the ShEx definitions created according to the vocabularies used by the DLs, we have created a ShEx schema per DL to assess that the resources linked to Wikidata are typed as human (`wd:Q5`) via the public Wikidata SPARQL endpoint provided by the Wikidata infrastructure. The ShEx definitions have been made grouped by DL in a manifest file (see Figure 3). The project is available on GitHub[10] and has been made citable via Zenodo.[11]

4.1. *The BNB Linked Data Platform*

The BNB Linked Data Platform provides access to the British National Bibliography[12] published as LOD and made available through SPARQL services. 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. The dataset is available under a Creative Commons CC0 licence.[13]

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,[14] 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.[15]

4.1.1. *Validating the dataset*

The identification of resources is performed by means of the SPARQL query in Figure 5 that shows an example of how to retrieve the different classes stored in a RDF repository. However, several classes may be used to type the same resource. For instance, a book

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

[10] https://github.com/hibernator11/ShEx-DLs
[11] https://doi.org/10.5281/zenodo.4022755
[12] https://www.bl.uk/collection-metadata/metadata-services
[13] http://creativecommons.org/publicdomain/zero/1.0/
[14] http://www.bl.uk/schemas/bibliographic/blterms
[15] http://www.bl.uk/bibliographic/pdfs/bldatamodelbook.pdf

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

Fig. 5. A SPARQL query to retrieve the different classes stored in a RDF repository.

can be typed as `bibo:Book`, `schema:Book` and `dc-terms:BibliographicResource`. As a result, Figure 6 shows a summary of the main classes used in the BNB repository.

Once we extracted the main resources described in the repository and identified their type, we extracted the properties per each class using SPARQL queries. For instance, Figure 7 shows a SPARQL query to retrieve the different properties used by the class `bibo:Book`.

A ShEx definition was created per each class to perform the assessment. As an example, the corresponding definition for the `bibo:Book` can be found in Figure 8. All the definitions were included in a manifest file that can be consumed by the ShEx validation tool [47].[16]

In addition, we have created an additional ShEx schema to asses the resources linked to the BNB Linked Data platform from Wikitada by means of the property `wdt:P5361` are typed as human (`wd:Q5`)

4.2. 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.

The dataset integrates several databases including the BnF main catalogue, BnF archives et 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.[17] The dataset can be used via a public SPARQL endpoint or as a dump file.[18]

[16]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

[17]https://data.bnf.fr/en/opendata

[18]http://api.bnf.fr/dumps-de-databnffr

4.2.1. Validating the dataset

An overview of the main classes stored in the LOD repository has been extracted and is shown in Figure 10. A new vocabulary has been defined to describe roles in which resources are linked to the Library of Congress subject headings (LCSH).[19]

Once we extracted the main resources described in the repository and identified their type, we extracted the properties per each class using SPARQL queries. For instance, Figure 11 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.

A ShEx schema was defined per each class to perform the validation as is shown in Figure 12. Similar to the previous use case, all the ShEx schemas were included in a manifest file that is used by the validation tool as an input.[20]

In addition, we have created an additional ShEx schema to assess that the resources linked from Wikidata to data.bnf.fr by means of the property `wdt:P268` are typed as human (`wd:Q5`).

4.3. Results and discussion

In order to assess the datasets, the main resources have been identified and validated using a random sample of 1000 items retrieved per entity and DL from their SPARQL endpoints. In total, 15 ShEx definitions have been created to validate the LOD published by digital libraries. Table 3 shows the description of the classes, ShEx and manifest files used to assess each DL. Figure 9 shows the ShEx validation interface consuming the manifest file and presenting the evaluation results for the BNB Linked Data platform.

When creating the ShEx definitions, preliminary tests were performed to pass the validation and after several iterations we were able to address all the issues. For some classes including a large number of resources, the properties have been 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 404 HTTP error (Request-URI

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

[20]https://rawgit.com/shexSpec/shex.js/wikidata/packages/shex-webapp/doc/shex-simple.html?manifestURL=https://raw.githubusercontent.com/hibernator11/ShEx-DLs/master/bnf.manifest.json

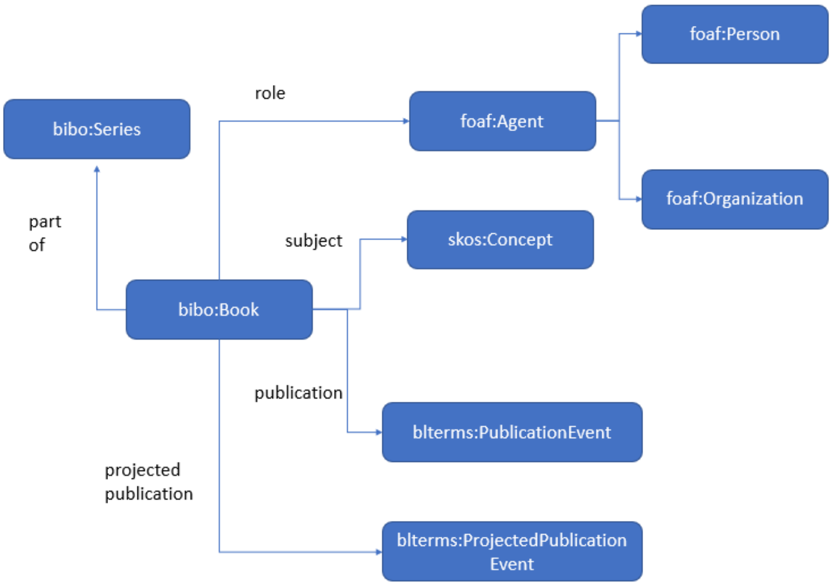

Fig. 6. Main classes retrieved from BNB LOD platform based on BIBO, SKOS and FOAF controlled vocabularies, and how they interact to create meaning.

Table 3

Description of the classes, ShEx definitions and manifest files used to assess each DL and provided in the Github project.

DL	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:Publication	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

Too Large). In this sense, the schemas could be further refined with additional properties and rules to obtain better performance in the results. In some cases, organizations provide a dump file instead of having a public SPARQL endpoint available. For instance, the Library of Congress suggests to download the bulk metadata and use a SPARQL engine to create custom queries such as RDF4J.[21]

Moreover, some resources may not include enough information to be assessed. For instance, the resources typed as `skos:Concept` only includes a `rdfs:label` in the BNB Linked Data platform. In

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

```
PREFIX bibo:
<http://purl.org/ontology/bibo/>

SELECT distinct ?p WHERE {
   ?s a bibo:Book.
   ?s ?p ?o
}
LIMIT 10
```

Fig. 7. A SPARQL query to retrieve the different properties used by the class `bibo:Book`.

some cases the same schema can be used for different classes such as `blterms:PublicationEvent` and `blterms:ProjectedPublicationEvent` since 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 instead of the BIBO vocabulary.

The ShEx Online Validator requires a public SPARQL endpoint that uses HTTPS to test the entities. However, some organizations do not provide this protocol in their services such as the BNE and BVMC.

5. Conclusions

Cultural heritage institutions are exploring Semantic Web technologies to publish and enrich their catalogs. While LOD repositories can be reused in innovative and creative ways, data quality has become a crucial aspect when identifying a dataset for reuse.

Based on previous research, we defined a methodology based on ShEx described in Section 3 for assessing the quality of LOD repositories published by DLs. The methodology have been applied to two use cases.

The expressivity of ShEx enables the identification of inconsistencies and the validation in large RDF datasets. In this sense, ShEx addresses the need of the Semantic Web community to ensure data quality for RDF graphs. ShEx is useful as documentation of the LOD datasets published by DLs since it provides a human-readable representation that helps researchers to understand the data model and evaluate their suitability for reuse in creative and innovative ways promoted by Labs.

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 DLs requirements such as common classes and properties used by DLs will be analysed.

Appendix A. List of prefixes

The prefixes in Table 4 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].

[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*, 2019. ISBN 978-9927-139-07-9.

[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 (ed.), ARL Task Force on Wikimedia and Linked Open Data. ARL White Paper on Wikidata: Opportunities and Recommendations, 2019. https://www.arl.org/storage/documents/publications/2019.04.18-ARL-white-paper-on-Wikidata.pdf.

[8] P. Mika, T. Tudorache, A. Bernstein, C. Welty, C.A. Knoblock, D. Vrandecic, P.T. Groth, N.F. Noy, K. Janowicz and C.A. Goble (eds), The Semantic Web - ISWC 2014 - 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part I, in *Lecture Notes in Computer Science*, Vol. 8796, Springer, 2014. ISBN 978-3-319-11963-2. doi:10.1007/978-3-319-11964-9.

[9] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann and S. Auer, Quality Assessment for Linked Data: A Survey, *Semantic Web Journal* (2015). http://www.semantic-web-journal.net/content/quality-assessment-linked-data-survey.

[10] 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.

[11] 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.

```
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* ;
}
```

Fig. 8. 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.

[12] 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.

[13] 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) (0), 0165551520930951. doi:10.1177/0165551520930951.

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

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

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

[17] 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.

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

[19] RDA Steering Committee, RDA Toolkit: Resource Description and Access, 2012, [Online; accessed 19-November-2018].

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

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

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

[23] CIDOC Documentation Standards Working Group, CIDOC Conceptual Reference Model. http://www.cidoc-crm.org/.

[24] T. Aalberg and M. Zumer, Looking for Entities in Bibliographic Records, in: *Digital Libraries: Universal and Ubiq-*

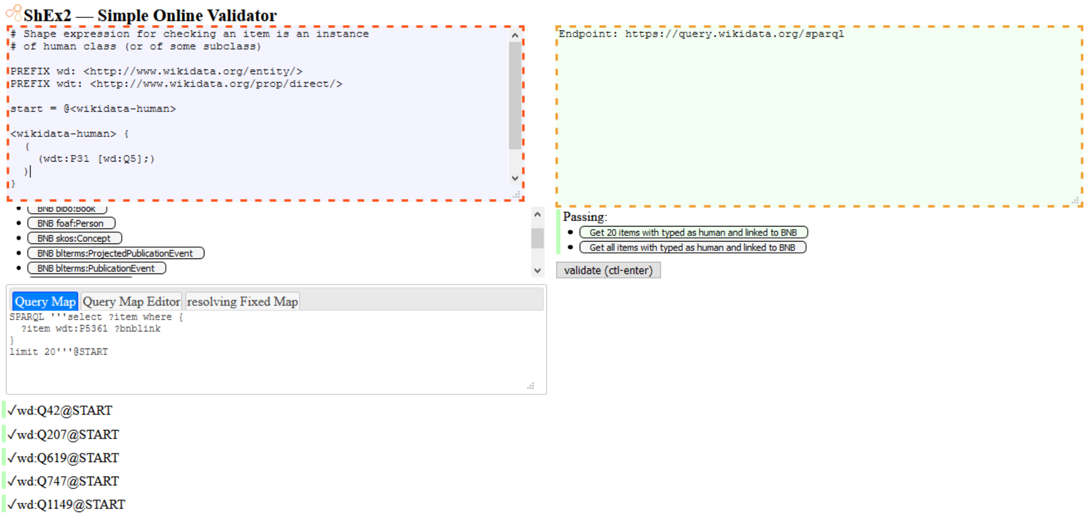

Fig. 9. The ShEx validator interface that uses the manifest file provided for the BNB Linked Data platform to assess each of the ShEx definitions showing the results.

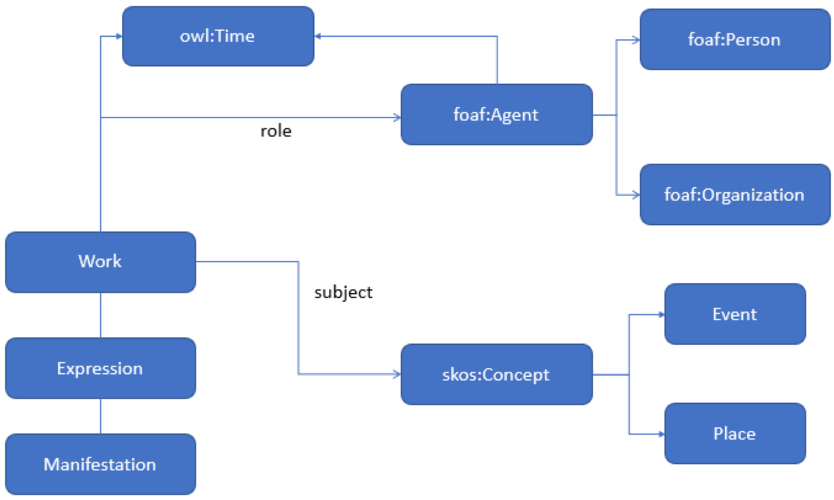

Fig. 10. Overview of the main classes based on FRBR, FOAF and SKOS retrieved from data.bnf.fr and how they interact to create meaning.

uitous 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.

[25] 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.

[26] 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.

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

[28] 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.

[29] 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.

Table 4
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/
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/
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/>
schema	http://schema.org/
skos	http://www.w3.org/2004/02/skos/core#
wdt	http://www.wikidata.org/entity/
wd	http://www.wikidata.org/entity/
wemi	http://rdvocab.info/RDARelationshipsWEMI/
xsd	http://www.w3.org/2001/XMLSchema#

```
PREFIX frbr-rda:
<http://rdvocab.info/uri/
schema/FRBRentitiesRDA/>

SELECT distinct ?p
WHERE {
  ?s a frbr-rda:Work .
  ?s ?p ?o
  FILTER (!regex(?p, "roles")) .
  FILTER (!regex(?p, "relators")) .
} LIMIT 100
```

Fig. 11. A SPARQL query to retrieve the different properties used by the class `frbr-rda:Work`. The `FILTER` instructions exclude the roles and relators properties.

[30] A. Piscopo, Wikidata:Requests for comment/Data quality framework for Wikidata, 2016, [Online; accessed 11-February-2018].

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] J.C.J. van Dam, J.J. Koehorst, P.J. Schaap and M. Suárez-Diez, The Empusa code generator: bridging the gap between

```
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* ;
 }
```

Fig. 12. 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.

the intended and the actual content of RDF resources, *CoRR* **abs/1812.04386** (2018). http://arxiv.org/abs/1812.04386.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] World Wide Web Consortium, Shape Expressions Language 2.1, 2019. https://shex.io/shex-semantics/.

[45] 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.

[46] 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.

[47] 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.

[48] R.Y. Wang and D.M. Strong, Beyond Accuracy: What Data Quality Means to Data Consumers, *J. of Management Information Systems* **12**(4) (1996), 5–33. http://www.jmis-web.org/articles/1002.

[49] G. Candela, P. Escobar, R.C. Carrasco and M. Marco-Such, A linked open data framework to enhance the discoverability and impact of culture heritage, *Journal of Information Science* **0**(0) (0), 0165551518812658. doi:10.1177/0165551518812658.

[50] J. Jett, T.W. Cole, M.K. Han and C. Szylowicz, Linked Open Data (LOD) for Library Special Collections, in: *2017 ACM/IEEE Joint Conference on Digital Libraries, JCDL 2017, Toronto, ON, Canada, June 19-23, 2017*, 2017, pp. 309–310. doi:10.1109/JCDL.2017.7991604.

[51] M.H. Carrasco, S. Luján-Mora, A. Maté and J. Trujillo, Current state of Linked Data in digital libraries, *J. Information Science* **42**(2) (2016), 117–127. doi:10.1177/0165551515594729.

[52] E.T. Mitchell, Library Linked Data: Early Activity and Development, *Library Technology Reports* **52**(1) (2016), 5–13. doi:10.5860/ltr.52n1.