Engaging Librarians in the Process of Interlinking RDF Resources

Lucy McKenna(✉)

ADAPT Centre, Trinity College Dublin, Dublin, Ireland
lucy.mckenna@adaptcentre.ie

Abstract. By publishing metadata as RDF and interlinking these resources with other RDF datasets on the Semantic Web, libraries have the potential to expose their collections to a larger audience, increase the use of their materials, and allow for more efficient user searches. Despite these benefits, there are many barriers to libraries fully participating in the Semantic Web. Increasing numbers of libraries are devoting valuable time and resources to publishing RDF datasets, yet little meaningful use is being made of them due to lack of interlinking. The goal of this research is to explore the barriers faced by librarians in participating in the Semantic Web with a particular focus on the process of interlinking. We will also explore how interlinking could be made more engaging for this domain.

Keywords: Engagement · Interlinking · Library · Linked Data · Semantic Web

1 Introduction

The Semantic Web (SW) is a Web of Data where the relationships between data are defined in a common machine readable format [1,2]. These relationships are known as Linked Data (LD), which describes a set of principles for publishing and interlinking data on the web [3].

From the perspective of a library, participating in the SW could greatly influence metadata quality and information discovery. By freeing metadata from library databases and sharing it on the SW, libraries could make their resources more visible, leading to an increase in the use of library data and, as such, an increase in the number of library patrons [4]. Publishing to the SW would also allow libraries to share their metadata with greater ease, thus enhancing metadata accessibility and quality. This could lead to a reduction in the amount of time spent creating metadata, reducing library costs [5]. In addition, the process of interlinking RDF resources with those emerging from other cultural heritage institutions and beyond could allow researchers to be directed to a web of related data based on a single information search [3].

Despite these benefits, relatively few libraries are fully participating in the SW as a result of the many barriers faced by librarians in engaging with SW and

E. Blomqvist et al. (Eds.): ESWC 2017, Part II, LNCS 10250, pp. 216–225, 2017.
DOI: 10.1007/978-3-319-58451-5_16

LD technologies [6,7]. Although increasing numbers of libraries are publishing metadata in RDF [3,8], few have successfully interlinked their data with other RDF resources - a central aspect of the SW.

2 Motivation

The goal of this research is to explore the barriers faced by librarians in participating in the SW, with a particular focus on the process of interlinking. Increasing this group's engagement in the process of interlinking will also be explored. It was decided to focus on this area due to the values provided by increased library participation on the SW. These values are twofold; firstly, as described above, use of LD has the potential to open up library collections and increase metadata quality. Secondly, librarians and library metadata has much to offer the SW in terms of data quality and credibility.

2.1 The Potential Role of Librarians in SW Development

It could be argued that LD generation could be conducted by technical experts or via crowd-sourcing, rather than by librarians. However, librarians have been successfully working in areas of information access and knowledge discovery for centuries and, thus are already ideally placed to play a leading role in this domain.

Librarians are experts at using controlled authorities and vocabularies when creating bibliographic metadata. This allows for consistent identification and linking of similar concepts and entities across records, resulting in more efficient catalogue searches. Libraries have developed many reliable authorities for controlling forms of names, titles and subjects, and countless controlled vocabularies for describing subjects, genres, languages, and locations. A number of these resources are already available as LD, thus, rather than duplicating what has already been created, these resources could be used to consistently identify concepts and entities across the SW [9]. Being familiar with the use of these authorities and vocabularies, librarians already have the expertise to establish this.

Since anyone can publish RDF metadata and interlink datasets on the SW, as the Web of Data grows, there will be an increased need to identify who completed these tasks in order to establish the degree of metadata credibility. As authoritative sources of information, it is believed that LD generated by librarians will be treated with increased credibility over that generated by non-authoritative sources [9–11]. Therefore, it is likely that LD generated by librarians will be used with increased frequency [10,11].

From the above it can bee seen that librarian's have the expertise to evolve the SW into a rich and trustworthy information network [12,13].

2.2 Challenges Faced by Libraries in Participating in the SW

As mentioned, RDF datasets are being published by a growing number of libraries [3,8], yet few are integrating these datasets with those emerging from other organisations [9], possibly because this is one of the most challenging areas of LD implementation [14]. This is also likely due to the fact the tools required to complete such data integration are limited [14] and that little usability testing of these tools has been completed with users, or potential users, of the SW who do not have a technical background [11]. As one of the fundamental prerequisites of the SW is the existence of large amounts of meaningfully interlinked resources [15], it is key that institutions not only publish RDF datasets but also interlink their data with others.

In 2015 the Online Computer Library Center (OCLC) conducted a worldwide survey investigating the use of LD in libraries. Of the 79 responses received, 112 LD projects were reported, the benefits of which included exposing data to a larger audience, enhancing the library's metadata, improving search accuracy, and combining LD datasets [6]. Barriers to using LD included; difficulty establishing links, lack of authority control, difficulty learning how to implement LD, lack of information outlining useful applications of LD in libraries, and difficulties incorporating LD generation into existing workflows [6].

Other reported [7,16] challenges faced by libraries when attempting to participate in the SW include:

- Cataloguing software can be inflexible in adapting to SW requirements.
- Many libraries use MARC21 format for generating bibliographic records, however the MARC data model is inadequate for direct use on the SW, and the processes and technologies used for transforming these records to a more suitable format are time-consuming and challenging.

The above indicates that, although librarians understand the benefits that the SW offers, they have difficulty engaging fully with it as LD technologies are not tailored to the library domain, or to the needs and expertise of librarians.

Librarians are an example of current, and potential, domain expert users of LD who may not have the technical expertise required to work with available LD technologies, but who have the potential to progress the development of the SW if given the opportunity to fully engage with it. Therefore it would be important to focus on exploring how librarians, who may not have a technical background, could engage more in the process of interlinking.

Interviews. Two informal semi-structured interviews were conducted with librarians who work as metadata catalogers, of both physical and digital assets respectively, in a large university library. These interviews were conducted in order to further investigate the challenges faced by librarians in engaging with LD. Both librarians had over a decade of experience working in bibliographic data management and both were familiar with the concepts of the SW and LD.

Common themes emerging from the interviews included:

- The librarians both noted that, although many libraries are exporting their catalogues in RDF, little further use is being made of the data.
- The above led to further discussion regarding the librarians' desire to be able to use published RDF datasets by interlinking them with data in their library catalogues.

- The librarians expressed a desire for a bespoke tool that would allow them to create links with LD datasets.
- The librarians highlighted a need for a tool that would allow them to create RDF records as part of their current cataloguing workflow. Both strongly stated that ideally such an interface would create RDF in the background of the cataloguing process.
- The above discussion led to both librarians expressing a certain level of frustration with current proprietary cataloguing software which does not facilitate the generation of RDF records, RDF ingestion, or interlinking.
- The interviewees highlighted how most current LD technologies do not target librarians or metadata cataloguers.
- The interviewees also expressed that libraries have a lot to offer the SW, both in terms of providing authority control and controlled vocabularies, and in providing "information about people, places and events... that would really bring value to the internet". It was also mentioned that the use of these controls provide a greater capacity for filtering searches.
- The above led to further discussion surrounding the librarians' concerns regarding authority control on the SW. The librarians felt that some current LD resources could be better controlled with increased validity if librarians were involved in the metadata creation process.
- Librarians need more use cases of LD being used effectively to enhance the visibility of resources and improve information searches in the library setting for them to allocate the necessary time and funding to LD creation.
- Difficulties in using MARC to create data in a format that is SW compatible were expressed.
- The interviewees both noted that it would be useful for librarians to have some basic training in coding, RDF and LD technologies so that they could better express their needs and so that they could interact with RDF datasets with greater ease.

From the interviews it was apparent that the librarians felt that libraries need more LD resources and tools targeting their specific needs in order for them to use the SW to its full potential. This was highlighted as a significant gap, with the interviewees feeling that most LD tools are not designed with librarians and their work processes in mind.

3 State of the Art

As prior sections of this paper indicate, librarians are currently unable to engage fully with the process of interlinking as a result of LD technologies being designed primarily for technical experts. The following section discusses some of the existing LD interlinking technologies as well as library projects that used LD interlinking in order to identify how interlinking is currently being achieved in the library domain.

OpenRefine [17] is an open source application that can be used for data cleanup and for transforming data to other formats. RDF Refine [18] is an extension of this tool which adds a GUI for exporting results in RDF and also allows for the reconciliation of collection specific vocabularies against controlled vocabularies expressed in RDF. This provides a way for users to interlink their LD resources to existing LD datasets by generating owl:sameAs links.

Silk [19] is a link discovery framework tool that can be used to generate links between related data items from different LD datasets. Like RDF Refine, the tool supports the generation of owl:sameAs links between resources as well as other types of RDF links. Silk also provides a Silk Workbench GUI for the creation of link specifications. LIMES [20] is another link discovery framework for the Web of Data and, like Silk, it can be used to discover links between LD resources using a GUI.

In 2013 the Library of the University of Nevada, Las Vegas embarked upon a focused LD project [21,22] where a collection of records from their Digital Library was uplifted to RDF and published as LD using OpenRefine. When creating their metadata records the library used controlled vocabularies and authorities from the Library of Congress [23], the FAST [24] subject heading schema, and the Getty Thesaurus of Geographic Names [25], among others. As these vocabularies are available as LD, the LD generated by the library was interconnected with these vocabularies. At the time of publication further interlinking between the library's dataset and other LD resources on the web was not completed, however it was reported that linking to DBpedia [26] and Europeana [27] was planned. In the case of OpenRefine and RDF Refine there are few examples of where the tool provided the user with a means of interlinking to datasets other than controlled vocabularies or large-scale general resources. There also appears to be little research exploring the usability of the tool from a librarian's or other domain expert user's perspective.

Research using Silk and LIMES has involved successfully creating links between large-scale LD resources such as DBpedia and GeoNames [28]. For instance, Swissbib, the meta-catalogue of Swiss University Libraries and the Swiss National Library, is currently being integrated into the SW with the generation of LD from their bibliographic metadata [29]. In this project the libraries' LD datasets have been successfully interlinked with DBpedia and The Virtual International Authority File (VIAF) [30] using both SILK and LIMES. However, as with OpenRefine, there are no apparent examples of the tools being used to successfully interlink with smaller LD resources exported from other libraries or cultural institutions. Additionally, little research with librarian or other domain expert users' of LD appears to have been completed.

Other methods of interlinking can be seen in the British National Biography LD project. In July 2011 the British National Library released the this resource as LD, converting the chosen records to RDF using XSLT [31]. Similar to the University of Nevada's Library project, the British National Library linked to library domain data sets such as VIAF and Library of Congress Subject Headings [32], by matching authorised headings in the records with the corresponding URI in the LD datasets available. In addition to this, the library also linked to other external data sets such as GeoNames and Lexvo [33] via Crosswalk Matching.

Although some interlinking was completed in the projects discussed above, the data sets were only linked with general resources or library domain data sets. Further interlinking could have been completed with smaller LD data sets emerging from other libraries or cultural heritage institutions.

4 Problem Statement and Contributions

As indicated above, librarians are currently not using LD interlinking to its full potential. Therefore the research question being investigated as part of this Early Stage PhD is:

- To what extent can librarians engage with LD interlinking tooling to manage, enhance and curate metadata records?

We hope to contribute to SW and library domain by:

1. Providing a means for librarians to engage in the process of interlinking in a more meaningful way.
2. Increase the potential for LD interlinking between institutions, and widen the possibilities of interlinking with smaller datasets.
3. Increase the potential for librarians' LD interlinking accuracy to equate to that of LD researchers.

The library domain will be used as a concrete environment in which this research question can be explored as librarians are considered experts of metadata management, enhancement and curation. With increasing numbers of libraries producing LD there is scope for them to interlink with LD datasets emerging from other cultural institutions, as well those produced by general resources, such as DBpedia or Library of Congress. Publishing LD datasets has the potential to open up and greatly enhance the metadata of digitised collections generated by libraries and other cultural heritage institutions [5]. Allowing the creators and curators of this data to become more involved in the LD process is likely to positively impact the use of LD and the development of LD tools [11].

It is hoped that this research will benefit librarians by providing them with a more automated way to engage in LD interlinking, and as such data curation, thus reducing the amount of time spent on the cataloguing process which could in turn allow librarians to use their skills and expertise on other areas.

5 Research Methodology and Approach

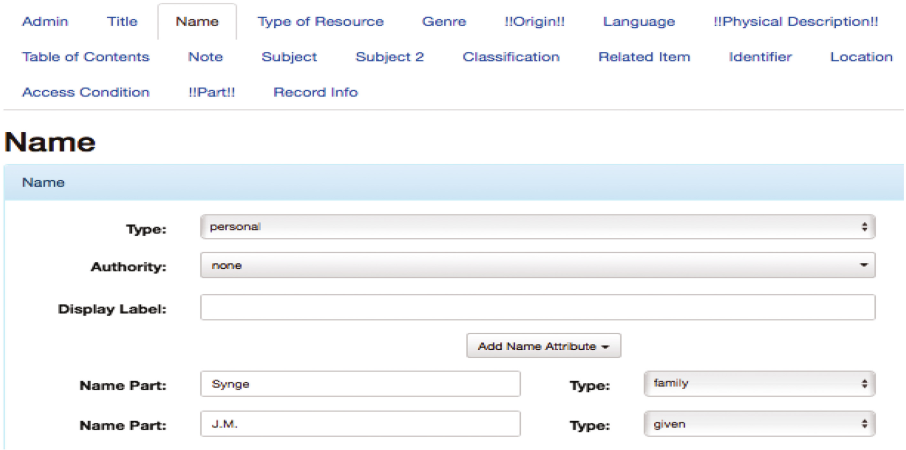
A Design Science approach will be followed for the purpose of this research. Design Science involves creating and evaluating artifacts designed to meet a particular need [34]. The process involves validating proposed artifacts, or products, through an iterative process of design and evaluation in order to design solutions to an identified problem [35].

We plan to initially investigate the current usage trends in respect to LD and LD tooling, and cataloguing tooling, within Ireland and further afield through the means of a state-of-the-art research as well as potential dissemination a survey. Based on these results a number of librarians will be interviewed in order to establish more specific needs and issues in working with LD. These interviews will have a particular focus on LD interlinking tools and the types of interfaces librarians are currently using when creating and managing LD.

A means of increasing librarians' engagement and interaction with LD, particularly at the interlinking stage, will be explored iteratively through a processes of problem identification, solution design, and user testing [36,37]. As we have worked closely with a number of librarians in the past, these librarians and their connections will be contacted to participate in the research.

6 Preliminary Results

A user interface for capturing bibliographic records using the Metadata Object Description Schema (MODS) RDF was developed (see Fig. 1).



Name

Name

Type: personal

Authority: none

Display Label:

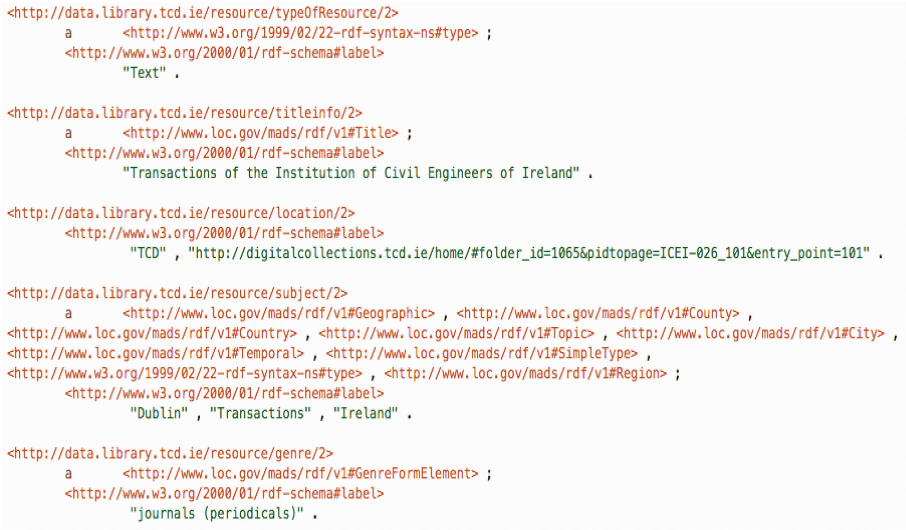
Name Part: Synge Type: family

Name Part: J.M. Type: given

Fig. 1. Screenshot of part of the MODS cataloguing interface.

MODS is an XML schema for a bibliographic element set that can be used to catalogue library materials [38] and MODS-RDF is an OWL/RDF representation of the schema. The interface was developed in collaboration with metadata cataloguers from the Digital Resources and Imaging Services (DRIS) of the Library of Trinity College Dublin (TCD).

The interface was designed to constrain data entry options and to dynamically alter depending on the data entered. This was done to ensure that published MODS records met the requirements of DRIS and the minimum requirements for describing digital cultural-heritage and humanities-based scholarly resources using MODS, as described by the Digital Library Federation Aquifer Initiative [39]. Metadata entered into the interface is stored in a relational database. This data can then be uplifted to MODS-RDF (see Fig. 2) using an R2RML mapping. Generating RDF records would allow for DRIS to publish and link their data on the SW, the benefits of which have been discussed above. The resulting MODS-RDF records can also be queried using SPARQL. SPARQL can be used to integrate data from different resources through the use of federated queries which can be conducted over multiple disparate datasets [40].



```
<http://data.library.tcd.ie/resource/typeOfResource/2>
        a       <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> ;
        <http://www.w3.org/2000/01/rdf-schema#label>
                "Text" .

<http://data.library.tcd.ie/resource/titleinfo/2>
        a       <http://www.loc.gov/mads/rdf/v1#Title> ;
        <http://www.w3.org/2000/01/rdf-schema#label>
                "Transactions of the Institution of Civil Engineers of Ireland" .

<http://data.library.tcd.ie/resource/location/2>
        <http://www.w3.org/2000/01/rdf-schema#label>
                 "TCD" , "http://digitalcollections.tcd.ie/home/#folder_id=1065&pidtopage=ICEI-026_101&entry_point=101" .

<http://data.library.tcd.ie/resource/subject/2>
        a       <http://www.loc.gov/mads/rdf/v1#Geographic> , <http://www.loc.gov/mads/rdf/v1#County> ,
<http://www.loc.gov/mads/rdf/v1#Country> , <http://www.loc.gov/mads/rdf/v1#Topic> , <http://www.loc.gov/mads/rdf/v1#City> ,
<http://www.loc.gov/mads/rdf/v1#Temporal> , <http://www.loc.gov/mads/rdf/v1#SimpleType> ,
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type> , <http://www.loc.gov/mads/rdf/v1#Region> ;
        <http://www.w3.org/2000/01/rdf-schema#label>
                 "Dublin" , "Transactions" , "Ireland" .

<http://data.library.tcd.ie/resource/genre/2>
        a       <http://www.loc.gov/mads/rdf/v1#GenreFormElement> ;
        <http://www.w3.org/2000/01/rdf-schema#label>
                 "journals (periodicals)" .
```

Fig. 2. Sample MODS RDF output from the cataloguing interface.

The cataloguing interface was tested by observing the DRIS metadata cataloguer using the tool to create a bibliographic record for an item in the repository. Results indicated that the metadata cataloguer was satisfied with the progress of the interface and DRIS indicated a strong interest in the ongoing development of the tool. Some interface layout issues and additional requirements were identified. Future iterations of the tool will focus on overcoming these issues.

The aim of developing the cataloging interface was to provide a pathway for DRIS to move towards publishing the bibliographic metadata of their digital collections as RDF. If DRIS begin publishing RDF records, it is hoped that DRIS may be used as an environment in which to explore the usability of LD interlinking tools and interfaces.

7 Evaluation Plan

In keeping with the Design Science approach, evaluation of potential solutions to the problems identified will occur iteratively throughout the research process.

Evaluation will likely involve usability testing which involves recruiting representative users, in this case librarians, to evaluate the degree to which a product meets specific usability criteria [41]. As usability testing seeks holistic information about the product, the setting and the user, it typically requires both quantitative and qualitative data [36].

8 Conclusions

It is hoped that facilitating increased engagement of domain expert users with the process of LD interlinking will aid in the realisation of the full vision of the SW.

Acknowledgments. This study is supported by the Science Foundation Ireland (Grant 13/RC/2106) as part of the ADAPT Centre for Digital Content Platform Research (http://www.adaptcentre.ie/) at Trinity College Dublin.

This research is being completed under the supervision of Prof. Declan O'Sullivan and Dr. Christophe Debruyne, both of the ADAPT Centre, Trinity College Dublin.

References

1. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Sci. Am. **284**, 1–5 (2001)
2. W3C (2015). https://www.w3.org/standards/semanticweb/
3. Hastings, R.: Linked data in libraries: status and future direction. Comput. Libr. **35**, 12–16 (2015)
4. Gonzales, B.M.: Linking libraries to the web: linked data and the future of the bibliographic record. ITAL **33**, 10–22 (2014)
5. Ryan, C., Grant, R., Carragin, E., Collins, S., Decker, S., Lopes, N.: Linked data authority records for Irish place names. IJDL **15**, 73–85 (2015)
6. OCLC: Online Computer Library Center (2017). http://www.oclc.org/research/themes/data-science/linkeddata.html
7. Hallo, M., Lujan Mora, S., Trujillo Mondejar, J.C.: Transforming library catalogs into linked data. In: ICERI (2013)
8. Mitchell, E.T.: Library linked data: early activity and development. Libr. Technol. Rep. **52**, 5–33 (2016)
9. Neish, P.: Linked data: what is it and why should you care? Aust. Libr. J. **64**, 3–10 (2015)
10. Miller, E., Westfall, M.: Linked data and libraries. Serials Libr. **60**, 17–22 (2011)
11. Shvaiko, P., Euzenat, J.: Ontology matching: state of the art and future challenges. IEEE Trans. Knowl. Data Eng. **25**, 158–176 (2013)
12. Greenberg, J.: Advancing the semantic web via library functions. CCQ **43**(3–4), 203–225 (2006)
13. Harper, C.A., Tillett, B.B.: Library of congress controlled vocabularies and their application to the semantic web. CCQ **43**(3–4), 47–68 (2007)
14. Bergman, M.: A decade in the trenches of the Semantic Web (2014). http://www.mkbergman.com/1771/a-decade-in-the-trenches-of-the-semantic-web/
15. Bizer, C., Heath, T., Ayers, D., Raimond, Y.: Interlinking open data on the web. In: 4th European Semantic Web Conference, Austria (2007)

16. Cole, T.W., Han, M.J., Weathers, W.F., Joyner, E.: Library MARC records into linked open data: challenges and opportunities. J. Libr. Metadata **13**(2–3), 163–196 (2013)
17. OpenRefine (2017). http://openrefine.org
18. RDF Refine (2017). http://refine.deri.ie
19. Volz, J., Bizer, C., Gaedke, M., Kobilarov, G.: Silk - a link discovery framework for the web of data. In: LDOW 2009, Spain (2009)
20. Ngomo, A.-C.N., Auer, S.: LIMES - a time-efficient approach for large-scale link discovery on the web of data. In: Proceedings of IJCAI (2011)
21. Lampert, C.K., Southwick, S.B.: Leading to linking: introducing linked data to academic library digital collections. J. Libr. Metadata **13**, 230–253 (2013)
22. Southwick, S.B.: A guide for transforming digital collections metadata into linked data using open source technologies. J. Libr. Metadata **15**, 1–35 (2015)
23. Library of Congress Linked Data Service (2017). http://id.loc.gov
24. OCLC FAST (2016). http://fast.oclc.org/searchfast/
25. Getty Thesaurus of Geographic Names Online (2015). http://www.getty.edu/research/tools/vocabularies/tgn/
26. DBpedia (2017). http://wiki.dbpedia.org
27. Europeana Collections (2017). http://www.europeana.eu/portal/en
28. GeoNames (2017). http://www.geonames.org
29. Bensman, F., Prongu, N., Hellstern, M., Kuntschik, P.: Swissbib goes linked data. In: SWIB (2016)
30. VIAF: The Virtual International Authority File (2016). http://viaf.org
31. Deliot, C.: Publishing the British national bibliography as linked open data. Catalogue Index **174**, 13–18 (2014)
32. Library of Congress Subject Headings (2017). http://id.loc.gov/authorities/subjects.html
33. Lexvo (2016). http://www.lexvo.org
34. Hevner, A.R., March, S.T., Park, S.: Design science in information systems research. MIS Q. **28**, 75–105 (2004)
35. van Aken, J.E.: Management research as a design science: articulating the research products of mode 2 knowledge production in management. Br. J. Manage. **16**, 19–36 (2005)
36. Emanuel, J.: Usability testing in libraries: methods, limitations, and implications. IDLP **29**, 204–217 (2013)
37. Nielsen, J.: Usability 101: Introduction to Usability (2012). http://www.nngroup.com/articles/usability-101-introduction-to-usability/
38. Library of Congress (2017). http://www.loc.gov/standards/mods/
39. Digital Library Federation (2009). https://wiki.dlib.indiana.edu/download/attachments/24288/DLFMODS_ImplementationGuidelines.pdf
40. W3C (2013). https://www.w3.org/TR/sparql11-federated-query/
41. Rubin, J., Chisnell, D.: Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests. Wiley Pub, Indianapolis (2008)