Understanding Information Professionals: A Survey on the Quality of Linked Data Sources for Digital Libraries*

Jeremy Debattista, Lucy McKenna, and Rob Brennan

ADAPT Centre, School of Computer Science and Statistics
Trinity College Dublin, Ireland
debattij@scss.tcd.ie | lucy.mckenna@adaptcentre.ie |
rob.brennan@cs.tcd.ie

Abstract. In this paper we provide an in-depth analysis of a survey related to Information Professionals (IPs) experiences with Linked Data quality. We discuss and highlight shortcomings in linked data sources following a survey related to the quality issues IPs find when using such sources for their daily tasks such as metadata creation.

Keywords: Metadata Quality · Digital Libraries · Linked Data.

1 Introduction

The success of a digital library (DL) is said to be dependent on the quality of the available metadata [8]. In such a broad sense, one could easily answer the question *"what is a good digital library?"*, however, in reality one cannot generalise which digital library is the absolute best for all cases and for everyone. This is due to the fact that defining what constitutes good metadata quality is subjective. Many researchers and librarians themselves tried to define metadata quality, however, their definition is mainly geared towards their institutional needs. These needs are coupled with the Information Professionals' (IP) experience and the role within the library setting, as this would also play an important part of formulating a definition of quality. Furthermore, metadata quality is not just the human's perception that defines quality, but similar to *data quality*, the task-at-hand is a decisive factor for defining quality. The use of Linked Data is gaining momentum within IPs and digital libraries[1]. IPs realised that Linked Data offers many benefits, such as better resource discovery and interoperability [6]. However, Linked Data implementation by IPs has been relatively slow,

* This research has received funding from the Irish Research Council Government of Ireland Postdoctoral Fellowship award (GOIPD/2017/1204) and the ADAPT Centre for Digital Content Technology, funded under the SFI Research Centres Programme (Grant 13/RC/2106) and co-funded by the European Regional Development Fund.

[1] https://www.oclc.org/research/themes/data-science/linkeddata/linked-data-survey.html Date Accessed: 11th July 2018

with issues in relation to the quality of currently available Linked Data resources being a notable challenge [6,9].

We have recently conducted a survey, amongst 185 IPs worldwide, whose purpose is twofold: (a) to have a better understanding of the quality criteria IPs with different experiences and expertise; and (b) to understand better the kind of quality problems these IPs are facing when searching for, or using external data sources. In this short paper we discuss our findings for the latter, that is, discussing the different quality problems these IPs are facing in their day-to-day tasks. The rest of the paper is structured as follows; In Section 2 we discuss the related work. The survey's methodology is described in Section 3. Discussion on the survey findings are discussed in Section 4, whilst in Section 5 we conclude this article with our final remarks and next steps building upon this survey.

2 Related Work

Quality in digital libraries has been discussed in various works throughout the years. However, to the best of our knowledge, there is no large-scale survey that gathers knowledge about metadata quality from IPs. These IPs provide the unique insight based on their varied experience and the institutions they belong to. Identifying and defining *what makes a good digital library* is dependent on a lot of factors, which makes it a statement that cannot be generalised for all. These factors include the stakeholder using the digital library and the task at hand. In order to try and address this statement, Gonçalves et al. [3] defined a formal model for understanding the quality of digital libraries based on the 5S (stream, structural, spatial, scenarios, and societies) theory. Supporting this model, the authors defined 16 dimensions each having a set of measurable metrics. The dimensions *accuracy*, *completeness*, *conformance to standards*, and *consistency* were proposed as candidate quality dimensions affecting the metadata concept. The majority of research work investigating digital library or metadata quality suggested a number of different quality measures, most of them based on their particular use case at hand, but as Park suggested in his work [7], most of the literatures' metric suggestions overlapped.

The broad survey literature on metadata quality in digital libraries led the foundations to our work. The introduction of digital libraries together with the Web of Data brought an upheaval in the way data catalogers generate metadata. More metadata and web resources are being re-used, nonetheless, it does not mean that quality has improved (or otherwise). In this short paper, we discuss the current data quality pitfalls IPs face in their daily tasks.

3 Survey Methodology

The survey analysed for our research formed part of a more extensive survey conducted to explore the attitudes and experiences of Information Professionals (IPs), such as librarians, archivists and metadata cataloguers, with regards to Linked Data. In this paper we refer to these as *digital library consumers*.

Participants in our questionnaire were primarily IPs with experience working in the LAM domain (N = 172). IPs were encouraged to participate regardless of whether they had any prior experience working with the Linked Data. This was done in an attempt to recruit a broad range of participants, rather than just IPs who are highly experienced in Linked Data. Also recruited were researchers and academics with experience in the LAM and/or LD domain (N= 13). This was done in order to gain the perspective of those engaging in current LD and LAM research. The 185 questionnaires that were analysed were classified into two groups: participants who have experience working with Linked Data (N = 54) (group 1), and participants who do not have experience working with Linked Data (N = 131) (group 2). For more information on the survey methodology and participants, we refer the reader to [6].

4 Creating (Linked) Metadata in Digital Libraries - Quality Problems in External Data Sources

Linked Data quality varies from one dataset to another, as Debattista et al. [2] discuss in their recent study of the quality of the 2015 version of the LOD Cloud[2]. In this section we discuss our findings in relation to the various quality issues IPs encounter when consuming these external sources for metadata creation tasks. In order to better understand the kind of quality problems the participants face when consuming external data sources for creating metadata, we asked the following question:

> *Can you give an example of a data quality issue or concern you experience frequently?*

Out of 185 participants, 92 addressed this question. From these 92, answers from 77 could be classified into 14 different quality dimensions. The remaining 15 answers could not be classified as quality problems and are out of scope for this article. Furthermore, some of the respondents mentioned more than one problem in their response and thus in total the number of problems identified is 90. Table 1 aggregates the 14 problems identified by the participants. Overall *semantic accuracy* problems are the most commonly mentioned amongst the 77 participants, whilst lack of *verbosity* was listed as the most commonly cited problem within Group 2 participants.

Semantic Accuracy Problems - The major concern mentioned in both groups was the fact that they have to work with a lot of incorrect data, more specifically dataset not representing the real world library object.The most common pitfall was the presence of incorrect values in data in various fields of catalogue resources. Whilst such issues cannot be pinpointed down to a particular one, there are various metrics that can be deployed in a publishing lifecycle that

[2] `http://lod-cloud.net`

Table 1. Quality pitfalls within external sources.

Problem - Quality Dimension	Group 1	Group 2	Total
Semantic Accuracy	6	9	15
Completeness	7	6	13
Interoperability	6	6	12
Conciseness	2	10	12
Data Formatting / Syntactic Validity	6	4	10
Language Versatility	4	2	6
Availability	3	2	5
Trustworthiness	2	2	4
Interpretability	3	0	3
Licensing	1	2	3
Timeliness	1	2	3
Provenance	0	2	2
Interlinking	0	1	1
Documentation	0	1	1

assess the datasets being produced. For example, one participant mentioned that they often find wrong ISBNs in e-books, as data providers mint their own identifiers rather than using the actual correct one. Another participant highlighted that data extracted using OCR techniques are usually prone to incorrect values. Semantic accuracy can also be a consequence of problems in syntactic validity, but not vice-versa.

Completeness (Data Coverage) Problems - In data quality, a dataset is said to be complete if it is comprehensive enough for the task at hand. This means that even if a dataset is not 100% complete when compared to the real world object, it can still be considered as complete if it meets the consumers' expectations. Participants mentioned that they do not trust that information is correct and complete in crowdsourcing efforts. One of these participants also noted that some content vendors dump their data into shared databases without following any best practices, thus creating noise for data consumers. Another participant noted that completeness of old records is lacking due to them not being updated for compliance with newer standards.

Interoperability Problems - Interoperability is one of the main strengths of the RDF data model, however, in order to ensure maximum interoperability, publishers should try and re-use existing terminology and semantic vocabularies for a particular domain as much as possible. Apart from metadata schemas, in digital libraries we also find a number of controlled vocabularies[3] that can be used when describing a resource. Nonetheless, the responses suggest that there is no consensus on which vocabularies should be used for which purpose.

[3] For example `http://www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset/`

Furthermore, these controlled vocabularies and digital libraries might use different formats, which makes metadata consumption for re-use more challenging. One participant noted that metadata formats (e.g. BIBFRAME[4]) are changing significantly and rapidly from one version to another which might cause interoperability issues between different catalogues that were not updated to the new version. Therefore, ontology maintainers should ensure that appropriate versioning techniques are used, in order to ensure seamless interoperability between the various agents using different versions of a particular dataset.

Conciseness Problems Ambiguity within resources and duplicate copies of the same resource will lead to poor overall quality since it would make it difficult for data consumers to decide which resources one should use for various tasks. The survey shows that in this dimension, ambiguity is a major problem, which could be resolved if a disambiguation process (or authority control in library science) is enforced and unique persistent IDs are used throughout. This could also be linked to an argument one participant raised that local authorities are creating their own resources and that databases such as the Library of Congress should harmonise with the said authorities in order to prevent problems with data duplication within a dataset distribution.

Data Formatting (Syntactic Validity) Problems When dealing with machine-readable formats, syntactic validity is an important aspect in datasets, otherwise such problems might hinder their use as machines would not be able to parse them correctly. These problems are mostly related to the violation of syntactical rules. Common problems mentioned by the participants were incorrect formatting of dates, inconsistencies in names (eg. first name, last name vs last name, first name), and problems caused by OCRd data. Problems in this dimension can directly affect the quality in the semantic accuracy dimension.

Language Versatility Problems Datasets, especially those on the Web, are meant to be used by anyone. A multi-lingual data catalogue is more likely to be re-used by different users/institutes who require the data to be in a specific language. Nonetheless, this does not mean that a dataset should have some resources in one language and some others in another language. One issue raised is regarding the inconsistency of using American English and UK English in terminologies in the authority and subject control data. Another problem is related to the localisation of the machine and the application, where for example one has to use cyrillic alphabet, which is not supported in some international standard authority data (for example Getty Vocabularies `http://www.getty.edu/research/tools/vocabularies/`). In Linked Data, the use of language tags (eg. @en) in string literals is strongly suggested so that data consumers (users, machines) can determine to what extent they can use the data [5].

[4] `https://www.loc.gov/bibframe/`

Nonetheless, processes has to be in place to help encode different transliteration schemes as language tags on their own would not be sufficient.

Availability Problems One of the main Linked Data principles is that resources are decentralised and interlinked together through the Unique Resource Identifiers (URI). Therefore, it is of utmost importance that resources on the web are maintained and are ready to be consumed by machines and humans alike at any time. The most common problems mentioned were, the presence of dead or broken links and the reliability of online services.

Trustworthiness Problems If the data is deemed to be credible, and correct, then the data consumer might consider a data source to be trustworthy. In this survey, some participants voiced different opinion on how they consider a dataset to be trustworthy. For example, one participant noted that "*collaborative effort across multiple industries is required to have trustworthy, unbiased sets of data to work from*". Another participant mentioned that one of the quality criteria he looks at when choosing a fit dataset is whether the work was carried out in his/her institution, implying that the participant trusts (or distrusts) the work done in his/her institution more than others.

Interpretability Problems In Linked Data, interpretability is mostly related to whether a machine is able to process and interpret the data. The concerns mentioned by the participants here are mostly related to the quality of schemas used. An ontology provides formal semantics of a class or a property, therefore, a machine can make sense out of the values that are defined in a dataset. Therefore, having a defunct vocabulary or inconsistencies within the schema itself means that a machine cannot process the data correctly as this data would be without formal meaning. The most pressing issues highlighted by the participants in this regard include (a) Links to published vocabularies go dead; (b) abandoned vocabularies are heading to their death due to the lack of maintainer information; and (c) datasets are using vocabularies with inconsistencies. In Linked Data, these problems have further consequences, for example one would not be able to reason upon data, or it could lead to wrong interlinking in automatic interlinking processes.

Licensing Problems Data, being open or not, should have a license defined in its metadata in order for a data consumer to understand to what extent they can (re-)use the data [2]. If the license is not clearly defined, one might run into intellectual property rights and copyright complications. The three participants highlighting this issue are on a common ground with regard to this topic. On the other hand, when talking about Linked Data datasets **published on the web**, Heath and Bizer [4] state that "it is a common assumption that content and data made publicly available on the Web can be re-used at will. However, the absence of a licensing statement does not grant consumers the automatic right to use that content/data".

Timeliness Problems Freshness and relevance of data sources is also important in metadata creation. Whilst certain values such as book name and author in data catalogues might not be changing frequently, there are some that require changing from time to time, as explained by two participants in the case of outdated authority files[5]. Furthermore, using outdated or broken links as reference pointers is not just an availability problem, but it also a dataset freshness problem, as highlighted by one of the participants. Another common issue mentioned in the survey is that catalogued archives are not being updated in authority files, and thus causing a freshness issue.

Provenance Problems Provenance metadata provides data consumers with the necessary information to understand where the data comes from, who produced it and how. The W3C Data on the Web Best Practices WG [5] highlights the importance of the provision of provenance stating that "published data outlives the lifespan of the data provider projects or organisations". Therefore, it is important that data publishers provide both basic contact information about themselves, but also provenance at a resource or statement level such that these are traceable to the original source. These problems are highly related to trustworthiness, as data consumers might look at provenance information to make decisions on whether to trust a particular dataset or data publisher [2].

Interlinking Problems One of Linked Data principles, having interlinks between resources enable data consumers to discover more (in a follow-your-nose fashion) about a particular entity. For example, data catalogues might not tell us who the spouse of a particular author was, but by linking the author to a data source such as DBpedia, a data consumer might be able to know this information and more. Having interlinks is also a requirement for 5-star Linked Open Data according to Tim Berners-Lee's scheme `https://5stardata.info/en/`.

Documentation Problems Whilst most data resources on the Web should allow for both machine and human consumption, data consumers should be able to understand how to access and use this data. For example, a data source might have a mailing list or even provide information in a human readable format. Nonetheless, when it comes to Linked Data, data publishers can publish such documentation in the dataset metadata using vocabularies such as voID to define regular expressions of typical resource URIs, or even an indication of the vocabularies used in the published dataset. Such documentation makes the dataset more understandable, which in turn could result in more re-use.

5 Final Remarks and Future Direction

When it comes to quality problems within external Linked Data source, IPs point out that most problems are intrinsic in nature, with *semantic accuracy*,

[5] In library science, authority control is the establishment and maintenance of consistent terminology for the identification of concepts across library collections.

completeness, *interoperability* and *conciseness* in the top three places. These survey results are worrying especially for the *semantic accuracy* and *interoperability* dimensions, where Linked Data should excel in. When comparing back to the work in [2], we find that even in the LOD cloud, the average for the usage of undefined classes and properties (related to the *interoperability* dimension) stands around 55%, with a very high standard deviation value. On the other hand, the LOD cloud average for the extensional conciseness metric (related to the *conciseness* dimension) is higher and is around 92%. Therefore, whilst this user study is an indication of the quality gaps within Linked Data sources, it is also an opportunity for the Linked Data publishers to update their publishing mechanisms in order to serve the digital library community better.

This user study is the first step of our quest to support digital libraries and their communities to adopt and improve their services using Linked Data. The next step is to assess the quality of Linked Data sources used in Digital Libraries, making quality metadata publicly available in a quality-based data portal. Furthermore, these quality metadata will be used in an interlinking framework for IPs, where a mechanism suggests different external data sources based on different quality criteria for the task at hand.

References

1. Corcho, O., Poveda-Villalón, M., Gómez-Pérez, A.: Ontology engineering in the era of linked data. Bulletin of the Association for Information Science and Technology **41**(4), 13–17 (5 2015). https://doi.org/10.1002/bult.2015.1720410407
2. Debattista, J., Lange, C., Auer, S., Cortis, D.: Evaluating the quality of the LOD Cloud: An empirical investigation (to appear). Semantic Web (Nov 2017)
3. Gonçalves, M.A., Moreira, B.L., Fox, E.A., Watson, L.T.: "What is a good digital library?" - a quality model for digital libraries. Information Processing & Management **43**(5), 1416 – 1437 (2007)
4. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space. Morgan & Claypool, 1st edn. (2011)
5. Lóscio, B.F., Burle, C., Calegari, N.: Data on the web best practices. W3C recommendation, World Wide Web Consortium (January 2017), `https://www.w3.org/TR/2017/REC-dwbp-20170131/`
6. McKenna, L., Debruyne, C., O'Sullivan, D.: Understanding the position of information professionals with regards to linked data: A survey of libraries, archives and museums. In: Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries (JCDL 2018), Fort Worth, Texas, USA. pp. 7–16 (2018)
7. Park, J.R.: Metadata quality in digital repositories: A survey of the current state of the art. Cataloging & Classification Quarterly **47**(3-4), 213–228 (2009)
8. Tani, A., Candela, L., Castelli, D.: Dealing with metadata quality: The legacy of digital library efforts. Information Processing & Management **49**(6), 1194 – 1205 (2013). https://doi.org/https://doi.org/10.1016/j.ipm.2013.05.003, `http://www.sciencedirect.com/science/article/pii/S0306457313000526`
9. Yoshimura, K.S.: Analysis of an international linked data survey for implementers. D-Lib Magazine **22**(7), 6 (2016)