Modelling the Provenance of Linked Data Interlinks for the Library Domain

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

Christophe Debruyne
ADAPT Centre,
Trinity College Dublin,
Ireland
christophe.debruyne@adaptcentre.ie

Declan O'Sullivan
ADAPT Centre,
Trinity College Dublin,
Ireland
declan.osullivan@cs.tcd.ie

ABSTRACT

As the Web of Data grows, so does the need to establish the quality and trustworthiness of its contents. Increasing numbers of libraries are publishing their metadata as Linked Data (LD). As these institutions are considered authoritative sources of information, it is likely that library LD will be treated with increased credibility over data published by other sources. However, in order to establish this trust, the provenance of library LD must be provided.

In 2018 we conducted a survey which explored the position of Information Professionals (IPs), such as librarians, archivists and cataloguers, with regards to LD. Results indicated that IPs find the process of LD interlinking to be a particularly challenging. In order to publish authoritative interlinks, provenance data for the description and justification of the links is required. As such, the goal of this research is to provide a provenance model for the LD interlinking process that meets the requirements of library metadata standards. Many current LD technologies are not accessible to non-technical experts or attuned to the needs of the library domain. By designing a model specifically for libraries, with input from IPs, we aim to facilitate this domain in the process of creating interlink provenance data.

CCS CONCEPTS

• **General and reference** → **Design**; • **Information systems** → **Data provenance**; **Digital libraries and archives**; **Semantic web description languages**; • **Human-centered computing** → *User centered design.*

KEYWORDS

linked data, semantic web, interlinking, provenance, library

ACM Reference Format:
Lucy McKenna, Christophe Debruyne, and Declan O'Sullivan. 2019. Modelling the Provenance of Linked Data Interlinks for the Library Domain. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3308560.3316518

WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316518

1 INTRODUCTION

The Semantic Web (SW) is an extension of the current Web where data is given well defined meaning and where the relationships between data, and not just documents, are defined in a common machine-readable format - creating a Web of Data [3]. Linked Data (LD) describes a set of best practices for publishing and interlinking this data on the SW, as per the principles defined by the W3C [2, 5]. These principles include the use of HTTP Uniform Resource Identifiers (URIs) as names for entities, such as works, people, places, and events, and also for retrieving data using the existing HTTP stack. A LD dataset is structured information encoded using the Resource Description Framework (RDF), the recommended model for representing and exchanging LD on the Web [30]. RDF statements take the form of subject-predicate-object triples, which can be organised in graphs. RDF requires that URIs are used to identify subjects and predicates - allowing for the resulting data to be understood by computers.

The SW and LD have the potential to transform the Web into a globally interlinked and searchable database rather than a disparate collection of documents [31]. This would allow for easier data querying and processing, and for the development of novel applications built on top of this Web of Data. With the Web being one of the first places where people search for information, one domain that is set to benefit from publishing to the SW are libraries. By using LD, libraries could improve the discoverability, searchability and interoperability of their data [13], which in turn would increase the use of their resources. However, since any individual can publish to the SW, it is crucial that libraries not only publish their metadata, but also the provenance of this metadata to the SW. Provision of provenance information allows potential users to establish the origin and trustworthiness of the data it describes. Given that libraries are considered authoritative sources of information, data from this domain is likely to be treated with increased credibility[25]. As such, if given access to descriptive provenance data, Web users are likely to engage with library LD datasets with increased confidence and frequency.

Though the number of libraries publishing to the SW is growing, uptake is still relatively slow due to the range of challenges faced by these institutions when using LD, including a lack guidelines, financial constraints, data quality concerns, URI maintenance issues, and software complexity [17, 22, 29]. A 2018 survey explored the position of 185 Information Professionals' (IPs) with regards to LD and results highlighted LD interlinking as a task that IPs find to be particularly challenging [21]. In response to this, we developed a LD interlinking approach for the library domain called NAISC - the Novel Authoritative Interlinking of Schema and Concepts. The

aim of NAISC is to improve LD interlinking accessibility to non-technical experts. This being achieved the the iterative design and user-testing of a graphical user-interface (GUI) which guides users through a step-by-step process of interlink generation[1]. NAISC specifically targets IPs by providing access to datasets and ontologies commonly used in the library domain. NAISC also reduces the need for expert LD knowledge by suggesting suitable link-types based links created by the user using natural-language relationship terms.

With one of the fundamental prerequisites of the SW being the existence of large amounts of meaningfully interlinked resources [5], it is not only important that libraries are facilitated to interlink their data to a range of authoritative sources, but that the trustworthiness of such interlinks is established through the provision of provenance data. Thus, as part of NAISC, we developed a provenance model for the LD interlinking process that meets the unique requirements of the library domain.

Our paper describes the NAISC provenance model, and it is structured as follows: a Background section provides information on LD Interlinking and LD Provenance; the Aims and Provenance Requirements for our model are then discussed; this is followed by a description of our Provenance Model with a Demonstration of potential uses. Lastly, the Future Directions and Conclusion of our research are discussed.

2 BACKGROUND

In the following section LD Interlinking and LD Provenance are defined and discussed in the context of our research within the library domain.

2.1 Linked Data Interlinking

Data linking describes the task of determining whether a URI, used to identify an entity, can be linked to another URI as a way of representing that they both describe the same Thing or as a way of indicating that they are related in some capacity [10]. LD interlinks are known as *typed links*, so called because the linking property, or predicate, describes the type of relationship between the subject URI and the object URI [26]. The property used to describe the relationship between two URIs is known as a link-type. In the context of our research, LD interlinking specifically refers to the process of creating an interlink between two URIs from different data sources.

2.2 Linked Data Provenance

Provenance data provides information on the people, institutions, resources, and processes involved in creating a piece of data [24]. This data can be used in order to ascertain whether information is trustworthy and as a means of determining data quality [18, 20]. Since any individual or group can publish to the SW, it is crucial that libraries publish the provenance of their interlinks as this would allow researchers to establish the origin of the data. Given that libraries are considered authoritative sources of information [25], it is possible that interlinks from this domain will be deemed trustworthy and thus used more frequently. In the context of our research, interlinks with rich data provenance are considered authoritative LD interlinks.

There are a number of provenance models that have been developed for use with LD including the Provenance Vocabulary [16], the Open Provenance Model (OPM) [23], Provenance Authoring and Versioning ontology (PAV) [9], Provenir [28], and the W3C recommended standard, PROV Ontology (PROV-O) [19]. The PROV Data Model, shown in Figure 1, is a Web Oriented provenance standard, developed by the W3C Provenance Working Group [19], for the representation and exchange of provenance information [24]. The model can be used to describe the Entities (physical, digital or conceptual object), Agents (person, organisation, software) and Activities involved the process of creating a specific Entity [19].

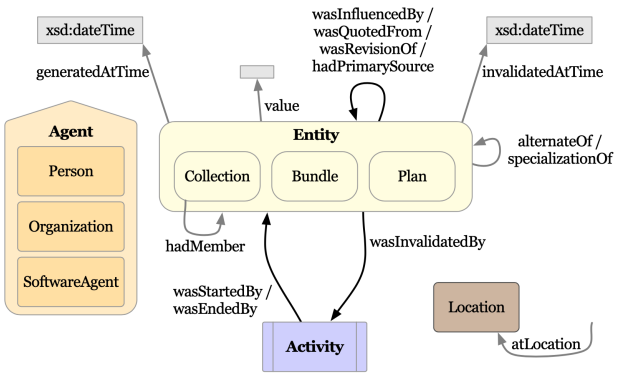

Figure 1: PROV Data Model

2.2.1 Provenance of Digital Resources. The Open Archival Information System (OAIS) [11] and Preservation Metadata: Implementation Strategies (PREMIS) [27] are widely accepted standards for digital preservation. Both OAIS and PREMIS require the provision of provenance information when archiving digital resources so as to maintain their long-term use and preservation. In the library domain, data provenance requires the inclusion information on where, when, by whom and how a resource was created [20]. Given that data provenance is likely to play an important role in establishing the trustworthiness of LD, it seems appropriate that these provenance standards should also be applied to the creation of interlinks. However, LD software typically only provides provenance information on resource ownership, as well as time-stamps for resource creation or modification [14]. As such, there is a need for a LD provenance model that captures the rich data required by the library domain in order to create authoritative interlinks.

3 NAISC PROVENANCE MODEL

As mentioned, we are currently developing an interlinking approach specifically for the library domain called NAISC. A component of the interlinking approach is the NAISC Provenance Model which provides authoritative origin data for the interlinks generated as part of the larger NAISC approach. The design of the provenance model is discussed below.

[1]See https://www.scss.tcd.ie/~mckennl3/naisc for a video demonstration of NAISC.

3.1 Provenance Requirements

A set of user requirements for the provenance model were distilled from the results of our international survey of 185 IPs [21]. The majority of participants (56%) came from an Academic Library perspective, thus the results of the survey are most applicable to this domain. Additionally most participants had some prior knowledge of the SW (84%) and LD (90%). The provenance requirements included:

- Allow for different levels of granularity e.g. view provenance for a set of links and for an individual link.
- Keep track of revisions made to interlinks.
- Link to the dataset/source of a subject or object entity.
- Link to the creator of the interlinks and the provenance data.
- Allow for the justification of linking a pair of entities.
- Allow for the justification of the link-type.

Further requirements for the provenance model were established from a series of ontological competency questions [4, 15], see Table 1. These questions were inspired by common requirements for data provenance on the SW [14].

Table 1: Interlink Provenance Competency Questions

Who created the link?	How can the dataset be accessed?
How was the link created?	Who published the dataset?
Why was the link created?	When was the link modified?
Where was the link created?	Who modified the link?
When was the link created?	How was the link modified?
What resources are linked?	Why was the link modified?
Why was the link created?	Who created the link provenance?
What datasets are linked?	When was the provenance created?

3.2 PROV-O Extension

PROV-O was used as the foundation of our interlink provenance model as it is a W3C recommended standard [19]. It also provides a model for general provenance descriptions which can then be extended for the needs of domain specific purposes [9]. Existing PROV-O classes, sub-classes and properties were used to describe the who (prov:Person), where (prov:Organisation) and when (prov:generatedAtTime) interlinks were created. We then extended PROV-O, see Figure 2, in order to add interlink specific sub-classes and properties. This extension, called NaiscProv, describes how (naiscProv:InterlinkCreationActivity) and why (naiscProv:hasJustification) interlinks were created.

Dublin Core (dcterms) [6] and FOAF [7] ontologies were used to provide richer descriptions of subject and object entities. The VoID Vocabulary [1] was also used in order to describe the datasets, or sources, of these entities.

3.3 Graph Structure

Our Provenance Model, as seen in Figure 3, incorporates three graphs:

(1) Interlink Graph - a named graph containing a set of interlinks. A named graph is a sub-graph that contains a set of triples and that has been assigned a unique name in the form of a URI [8]. Named graphs allow collections of triples to be published as independent units. Named graphs are often used in the process of provenance data generation as they allow for the assertion of statements relating to a specific set of triples in a dataset [12]. In the case of NAISC, an Interlink Graph will contain a set of finalised interlinks associated with a particular dataset, or part of a dataset, that are ready to be published to the SW. Over time, as interlinks are added, modified or deleted from the dataset, new Interlink Graphs will be created. Having these multiple Interlink Graphs allows for the provenance of individual interlinks to be maintained over time in a more simplified manner.

(2) Provenance Graph - a prov:Bundle containing the origin data of the interlinks in an Interlink Graph. In the PROV Data Model, a Bundle is a named set of provenance descriptions that can be used to describe the creation and modification of an entity or group of entities [19]. As a Bundle is itself an entity, the provenance of the Provenance Graph can also be captured. Every Interlink Graph will have a corresponding Provenance Graph which captures who, where, when, why and how the interlinks contained in the graph were created.

(3) Relationship Graph - represents the relationship between an Interlink Graph and a Provenance Graph (prov:hasProvenance).

The purpose of these graphs is to allow the user to explore the different sets of interlinks, and also to explore the provenance information for the interlinks. The Interlink Graph allows the user

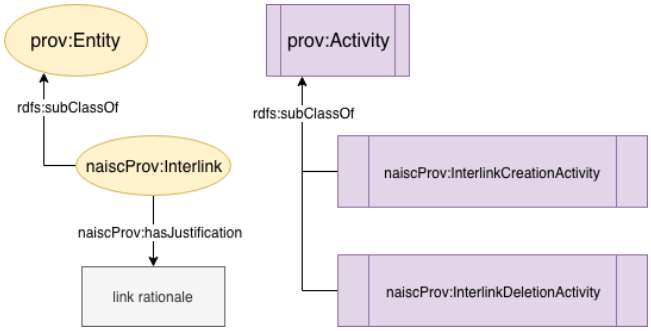

Figure 2: NaiscProv PROV-O Extension

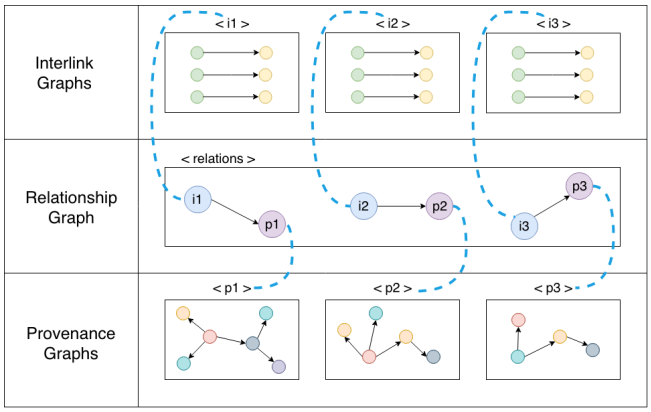

Figure 3: NAISC Provenance Graph Structure

to view a particular set of LD interlinks created using the NAISC Approach. Should the user wish to review the provenance of this set of interlinks, the Relationship Graph can be used to direct the user to the associated Provenance Graph associated with the Interlink Graph. The Provenance Graph provides origin information for each of the interlinks as well as for the interlink creation, revision and deletion processes.

Separating the data in this manner simplifies some of the queries that users could formulate and run over the data whilst still allowing for queries that span across graphs, as facilitated by the relationship layer.

4 DEMONSTRATION

In the following section a simple provenance graph for the creation, revision and deletion of an interlink shall be demonstrated using the NAISC Provenance Model.

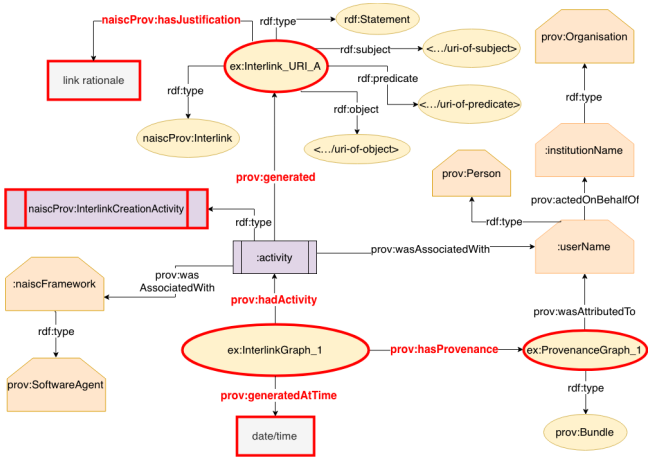

Figure 4: Interlink Creation Provenance Graph

4.1 Interlink Creation

Figure 4 demonstrates a snippet of a Provenance Graph describing the process of creating an interlink. Here an Interlink Graph (ex:InterlinkGraph_1), containing a set of interlinks, is stated to have had a creation activity (naiscProv:InterlinkCreationActivity) which generated a specific interlink (ex:Interlink_URI_A). The interlink is given a Unique Resource Identifier (URI) and its subject, predicate and object are described using RDF Reification. RDF Reification allows for each interlink to be given its own URI, and also allows for use of the naiscProv:hasJustification property. This property provides an opportunity for the inclusion of rationale, or 'why' provenance, data. Information that could be captured here includes justifying why the subject and predicate entities were interlinked, as well as the choice of predicate.

Other important provenance information included in the Provenance Graph is the generation date/time of the Interlink Graph (prov:generatedAtTime), as well as the Agents responsible for creating the link such as NAISC (prov:SoftwareAgent), the cataloguer (prov:Person), and the cataloguer's institution (prov:actedOnBehalfOf). Also included is the provenance of the Provenance Graph itself

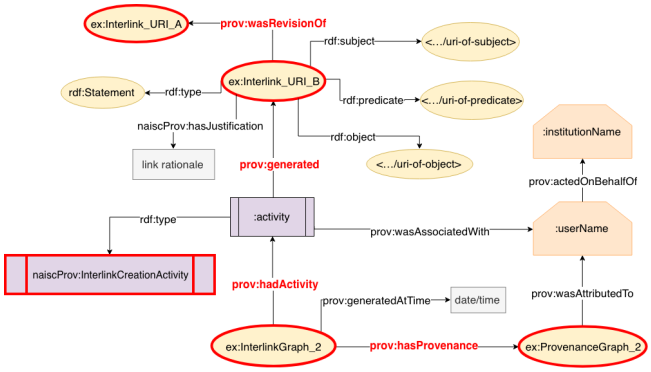

Figure 5: Interlink Revision Provenance Graph

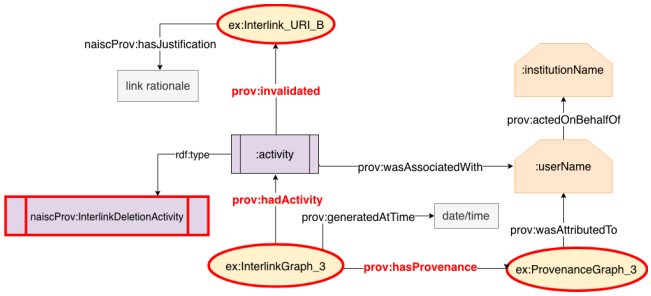

Figure 6: Interlink Deletion Provenance Graph

(prov:hasProvenance). Other information that could be added to the graph includes the source of the subject and/or object entity (dcterms:isPartOf, void:Dataset).

4.2 Interlink Revision

Figure 5 demonstrates a Provenance Graph describing the revision of an interlink using the NAISC Provenance Model. Here a new Interlink Graph (ex:InterlinkGraph_2) is stated to have had a creation activity which generated a new Interlink (ex:Interlink_URI_B). The property prov:wasRevisionOf is used to describe how the new interlink (ex:Interlink_URI_B) is a modified version of (ex:Interlink_URI_A). The rationale for modifying the interlink can be provided through the use of the naiscProv:hasJustification property.

4.3 Interlink Deletion

Figure 6 demonstrates the deletion of an interlink using the NAISC Provenance Model. In this instance a new Interlink Graph (ex:InterlinkGraph_3) is stated to have had a deletion activity (naiscProv:InterlinkDeletionActivity). This activity resulted in the invalidation, or deletion, (prov:inalidated) of ex:Interlink_URI_B. The reason for deleting the interlink can be provided using the naiscProv:hasJustification property. As seen previously, the relationship between the new Interlink Graph (ex:Interlin-kGraph_3), and its predecessor (ex:InterlinkGraph_2) is conveyed using prov:wasRevisionOf.

5 CONCLUSION

One of the main benefits of the SW is the ability to interlink related entities across datasets. However, such interlinks can only be meaningfully used if their origin and creation processes are exposed to users. This data enables the assessment of the context in which the interlink was created as well as its quality and validity. As part of our research we developed an interlink provenance model specifically for the library domain. This provenance model meets the unique requirements of IPs, as discussed in Section 3.1, particularly that of the provision of 'why' provenance information which is not catered for in other provenance models. This provenance data will add to the trustworthiness of library LD which in turn may increase the use of interlinks published as part of these datasets. By incorporating our Provenance Model into the NAISC framework, IPs will be able to easily create richer provenance data for the interlinks they generate in a using NAISC's GUI.

6 FUTURE DIRECTIONS

As mentioned, the NAISC Provenance Model forms part of the NAISC Approach to LD Interlinking. An accompanying graphical user interface has been developed as a means of guiding users through the steps of LD interlink creation and provenance generation, as proposed by our approach. The NAISC Provenance Model will be evaluated as part of the user-testing of the NAISC Approach. This will involve assessing the efficacy of the model in capturing the provenance of interlinks as well as the perceived influence of the provenance data on the trustworthiness of the interlinks.

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.

REFERENCES

[1] K. Alexander, R. Cyganiak, M. Hausenblas, and J. Zhao. 2011. Describing Linked Datasets with the VoID Vocabulary. Retrieved November 2018 from https://www.w3.org/TR/void/.
[2] T. Berners-Lee. 2006. Linked Data. Retrieved March 2019 from https://www.w3.org/DesignIssues/LinkedData.
[3] T. Berners-Lee, J. Hendler, and O. Lassila. 2001. The Semantic Web. *Scientific American* 284, 5 (2001), 1–5.
[4] C. Bezerra, F. Freitas, and F. Santana. 2013. Evaluating ontologies with competency questions. In *In Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, Vol. 3. 284–285.
[5] C. Bizer, T. Heath, and T. Berners-Lee. 2009. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems* 5, 3 (2009), 1–22.
[6] Dublin Core Metadata Initiative Usage Board. 2014. Dublin Core Metadata Initiative Metadata Terms. Retrieved November 2018 from http://dublincore.org/documents/2012/06/14/dcmi-terms/.
[7] D. Brickley and L. Miller. 2014. FOAF Vocabulary Specification 0.99. Retrieved November 2018 from http://xmlns.com/foaf/spec/.
[8] J. J. Carroll, C. Bizer, P. Hayes, and P. Stickler. 2005. Named graphs. *Web Semantics: Science, Services and Agents on the World Wide Web* 3, 4 (2005), 247–267.
[9] P. Ciccarese, S. Soiland-Reyes, K. Belhajjame, A. J. Gray, C. Goble, and T. Clark. 2013. PAV ontology: provenance, authoring and versioning. *Journal of Biomedical Semantics* 4, 1 (2013), 37.
[10] A. Ferrara, A. Nikolov, and F.çois Scharffe. 2011. Data linking for the semantic web. *International Journal on Semantic Web and Information Systems (IJSWIS)* 7, 3 (2011), 46–76.
[11] Consultative Committee for Space Data Systems. 2002. CCSDS 650.0-B-1: Reference Model for an Open Archival Information System (OAIS). Blue Book. Issue 1, 1-1. http://ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents/pdf/CCSDS-650.0-B-1.pdf.
[12] T. Gibson, K. Schuchardt, and E. Stephan. 2009. Application of named graphs towards custom provenance views. In *Paper presented at the First workshop on on Theory and practice of provenance, San Francisco, CA.*
[13] B.M. Gonzales. 2014. Linking Libraries to the Web: Linked Data and the Future of the Bibliographic Record. *Information Technology and Libraries* 33, 4 (2014), 10–22.
[14] P. Groth, Y. Gil, J. Cheney, and S. Miles. 2012. Requirements for provenance on the web. International Journal of Digital Curation. *nternational Journal of Digital Curation* 7, 1 (2012), 39–56.
[15] M. Gruninger and M.S. Fox. 1995. Methodology for the Design and Evaluation of Ontologies. In *In Proceedings of the IJCAI Workshop on Basic Ontological Issues in Knowledge Sharing.*
[16] O. Hartig and J. Zhao. 2010. Publishing and Consuming Provenance Metadata on the Web of Linked Data.. In *Paper presented at the IPAW 2010, Berlin, Heidelberg.*
[17] R. Hastings. 2015. Linked Data in Libraries: Status and Future Direction. *Computers in Libraries* 35, 9 (2015), 12–16.
[18] S. Kumar, M. Ujjal, and B. Utpal. 2013. Exposing MARC 21 Format for Bibliographic Data As Linked Data With Provenance. *Journal of Library Metadata* 13, 2-3 (2013), 212–229.
[19] T. Lebo, S. Sahoo, D. McGuinness, K. Belhajjame, and D. et al Corsar. 2013. Prov-o: The prov ontology. W3C Recommendation. World Wide Web Consortium. Retrieved November 2018 from https://www.w3.org/TR/prov-o/.
[20] C. Li and S. Sugimoto. 2014. Provenance Description of Metadata using PROV with PREMIS for Long-term Use of Metadata. In *Proceedings of the IInternational Conference on Dublin Core and Metadata Applications, Sao Paulo, Brazil.*
[21] L. McKenna, C. Debruyne, and D. O'Sullivan. 2018. 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.* 7–16.
[22] E.T. Mitchell. 2016. Library Linked Data: Early Activity and Development. *Library Technology Reports* 52, 1 (2016), 5–33.
[23] L. Moreau, B. Clifford, J. Freire, J. Futrelle, and Y. et al Gil. 2011. The Open Provenance Model core specification (v1.1). *Future Generation Computer Systems* 27, 6 (2011), 743–756.
[24] L. Moreau, P. Groth, J. Cheney, T. Lebo, and S. Miles. 2015. The rationale of PROV. *Web Semantics: Science, Services and Agents on the World Wide Web* 35 (2015), 235–257.
[25] P. Neish. 2015. Linked data: what is it and why should you care? *The Australian Library Journal* 64, 1 (2015), 3–10.
[26] Georg Neubauer. 2017. Visualization of typed links in Linked Data. *Mitteilungen der Vereinigung ÃŰsterreichischer Bibliothekarinnen und Bibliothekare* 70 (09 2017), 179. https://doi.org/10.31263/voebm.v70i2.1748
[27] Library of Congress. 2018. PREMIS: Preservation Metadata Maintenance Activity. Retrieved January 2019 from http://www.loc.gov/standards/premis/.
[28] S. S. Sahoo and A. P. Sheth. 2009. Provenir ontology: Towards a Framework for eScience Provenance Management. In *Microsoft eScience Workshop, Pittsburgh, PA Oct 15-17.*
[29] K. Smith-Yoshimura. 2018. Analysis of 2018 international linked data survey for implementers. *Code4Lib* 42 (2018).
[30] W3C. 2014. RDF: Resource Description Framework. https://www.w3.org/RDF/.
[31] W3C. 2015. Semantic Web. https://www.w3.org/standards/semanticweb/.