NAISC: An Authoritative Linked Data Interlinking Approach 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

By interlinking internal Linked Data (LD) entities to related LD entities published by authoritative creators and holders of data, libraries have the potential to expose their collections to a larger audience and to allow for richer user searches. While increasing numbers libraries are devoting time to publishing LD, the full potential of these datasets has not been explored due to limited LD interlinking. 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 data interlinking to be a particularly challenging step in the creation of Five Star LD. Consequently, we developed NAISC, an interlinking approach designed specifically for the library domain aimed at facilitating increased IP engagement in the LD interlinking process. Our paper provides an overview of the design and user-evaluation of NAISC. Results indicated that IPs found NAISC easy-to-use and useful for creating LD interlinks.

CCS CONCEPTS

• **General and reference → Evaluation**; **Design**; • **Information systems → Digital libraries and archives**; **Semantic web description languages**; • **Human-centered computing → User studies**; **Usability testing**; *Graphical user interfaces*; *User centered design*;

KEYWORDS

linked data, semantic web, interlinking, library, usability testing

ACM Reference Format:
Lucy McKenna, Christophe Debruyne, and Declan O'Sullivan. 2019. NAISC: An Authoritative Linked Data Interlinking Approach for the Library Domain. In *JCDL'19: The 19th ACM/IEEE Joint Conference on Digital Libraries, June 2–6, 2019, Urbana-Champaign, IL, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/xxxx

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 [7]. 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 [6, 8]. 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 [48]. 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 data to be understood by computers.

LD is classified according to a 5 Star rating scheme and, in order to be considered 5 Star, a LD dataset must contain interlinks to related data [6, 28]. The purpose of LD interlinks are to enhance the knowledge associated with a specific Thing, or entity, such as a person, place, concept or object [43]. These links have the potential to transform the Web into a globally interlinked and searchable database rather than a disparate collection of documents [49], allowing for easier data querying and for the development of novel applications built on top of the Web.

With the Web being one of the first places where people search for information, one domain that would greatly benefit from publishing LD are libraries. By using LD, libraries could improve the discoverability, searchability and interoperability of their data [20], which in turn would increase the use of their resources. 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 [25, 35, 45]. 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 [34]. In response to this, we developed a LD interlinking approach for the library domain called NAISC - the Novel Authoritative Interlinking of Schema and Concepts. Our paper describes the process of developing NAISC, and it is structured as follows; a Background section provides information on LD interlinking and LD provenance. In Related Works we discuss our 2018 LD survey and review LD Interlinking Framework. The Aims of our research are then listed and this is followed by a description of NAISC and its components. Finally we present the Methodology, Findings and Discussion of a user evaluation of the NAISC, as well as our Conclusions and Future Directions.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

JCDL'19, June 2–6, 2019, Urbana-Champaign, IL, USA

© 2019 Association for Computing Machinery.

ACM ISBN xxxx…$15.00
https://doi.org/xxxx

2 BACKGROUND

In the following section LD interlinking and LD provenance are 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 [18]. 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 [39]. 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.

Currently, the majority of interlinks between LD datasets are *identity links* [43]. These are a specific kind of typed link which state that two URIs refer to exactly the same thing i.e. they have the same identity and share the same properties. Identity links, or sameAs statements, are expressed using the owl:sameAs property from the Web Ontology Language[1] (OWL). However, given that the purpose of LD interlinking is to enhance the knowledge associated with an entity [43], and given that LD interlinks are not limited to identity links alone [18], much value could be gained by facilitating LD users to create interlinks that express other relationships. This is particularly relevant given there have been concerns within the LD community that the owl:sameAs property is being used in ways that do not necessarily conform with its definition in OWL [16, 23].

2.1.1 Linked Data Interlinking in the Library Domain. Upon reviewing the data on the Linked Open Data Cloud[2] for some of the leading library LD projects, such as those of the Swedish[3] (LIBRIS), French[4] (BnF), Spanish[5] (BnE), British[6] (BNB) and German[7] (DNB) National Libraries, it was found that the majority of interlinks are to authority files and controlled vocabularies. These authorities include, but are not limited to, the Library of Congress (LoC) LD Service[8], the Virtual International Authority File[9] (VIAF), Gemeinsame Normdatei[10] (GND), the International Standard Name Identifier[11] (ISNI) and GeoNames[12]. Though this is extremely useful, this type of linking predates LD.

The full potential of LD interlinking has yet to be realised within the library domain due to a notable lack of interlinks created for purposes outside of authority control. Interlinking could also be used to provide additional information and context for a given entity. This could be achieved by creating a variety of typed links from internal entities to related entities found in external authoritative resources. Examples in the projects mentioned above include links to encyclopedic data-hubs such as MusicBrainz[13], DBpedia[14] and Wikidata[15]. However, further enrichment could be gained by linking to other institutions such as libraries, archives, museums (LAMs), and Government Bodies.

With one of the fundamental prerequisites of the SW being the existence of large amounts of meaningfully interlinked resources [8], there is a need to explore how IPs can be facilitated to create interlinks beyond those used for authority control.

2.2 Linked Data Provenance

In our paper, LD Provenance refers to the provision of information on the people, institutions, resources, and processes involved in creating a piece of data [37]. This data can be used in order to ascertain whether information is trustworthy and as a means of determining data quality [29, 32]. 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 [38], 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 [24], the Open Provenance Model (OPM) [36], Provenance Authoring and Versioning ontology (PAV) [12], Provenir [44], and the W3C recommended standard, PROV Ontology (PROV-O) [30]. The PROV Data Model, shown in Figure 1, is a Web Oriented provenance standard, developed by the W3C Provenance Working Group [30], for the representation and exchange of provenance information [37]. 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 [30].

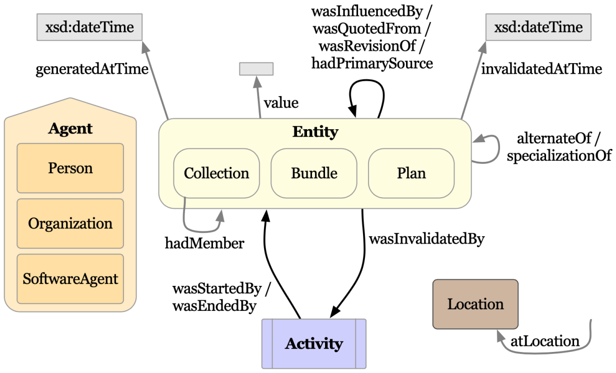


Figure 1: PROV Data Model
Taken from [30]

[1] https://www.w3.org/TR/owl-ref/
[2] https://lod-cloud.net
[3] http://libris.kb.se
[4] http://data.bnf.fr
[5] http://datos.bne.es/inicio.html
[6] http://bnb.data.bl.uk
[7] https://portal.dnb.de
[8] http://id.loc.gov
[9] http://viaf.org/
[10] http://www.dnb.de/EN/lds
[11] http://www.isni.org
[12] https://www.geonames.org/
[13] https://musicbrainz.org
[14] http://wiki.dbpedia.org
[15] https://www.wikidata.org

2.2.1 Provenance of Digital Resources. The Open Archival Information System (OAIS) [13] and Preservation Metadata: Implementation Strategies (PREMIS) [42] 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 [32]. 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 [21]. As such, there is a need for a LD provenance model that captures the data required by the library domain in order to create authoritative interlinks.

3 RELATED WORK

In the following section the results of a survey which explored IPs position with regards to LD are summarised. A brief overview of some existing LD interlinking frameworks is also provided.

3.1 Linked Data Survey

An online questionnaire, consisting of 50 questions, was developed in order to explore IPs position with regards to LD [34]. The survey was completed by 185 IPs, including librarians, archivists and cataloguers, who had experience working the library, archive or museum domain . The majority of participants (56%) came from an Academic Library setting, thus the results of the survey are most applicable to this domain. Additionally, though not a requirement, most participants had some prior knowledge of the SW (84%) and LD (90%). The questionnaire investigated:

(1) IPs' knowledge, views and experience with LD.
(2) IPs' perceived usability of LD tools.
(3) Solutions to the LD challenges experienced by IPs.

The key findings of the survey indicated that IPs considered the primary benefits of LD publication and consumption to include:

(1) Cross institutional linking and integration resulting in additional context for data interpretation and improved cataloguing efficiency.
(2) Improved data discoverability and accessibility.
(3) Enriched metadata and improved authority control.

The main challenges to LD publication and consumption, as experienced by the survey participants, were:

(1) Resource Quality Issues including; LD datatsets and URIs not being maintained, insufficient provenance data, a lack of guidelines and use-cases, and difficulty creating and maintaining URIs. Participants indicated that in order to invest in LD, more useful examples of its application needs to be seen.
(2) LD Tooling Issues including; functional inadequacy for the requirements of the library domain, technological complexity, and difficulty integrating into cataloguing workflows.
(3) Interlinking and Integration Issues including; difficulty selecting appropriate ontologies and link-types, and difficulty with data reconciliation and vocabulary mapping.

Potential solutions to the above challenges were also investigated as part of the survey. Participants had a positive response to the idea of LD tooling designed specifically for IPs, with the vast majority of participants (89%) indicating that they thought such tooling would be useful. The most commonly cited reasons for this being that a bespoke tool could help overcome the technical knowledge gap of IPs, make LD more accessible to IPs, increase the number of LAMs using LD, and create new research opportunities.

Though multiple LD challenges were raised in the survey, we decided to focus our research on the development of a framework, with an accompanying graphical user-interface, that would facilitate increased IP engagement in the process of LD interlinking.

3.2 Linked Data Interlinking Frameworks

The LD interlinking frameworks and tools used by the library projects, mentioned in Section 2.1.1, to create LD interlinks have been summarised in Table 1. MARiMbA[16] was designed specifically for the BnE library LD project, and components of RDF Refine were also designed with the library domain in mind. Both of these tools offer automated identity linking to commonly used datasets in the library domain. MARiMbA does not include a GUI but is controlled via the command line, something which may not appeal to non-technical experts. It can also be seen that all of the tools summarised here only offer automated support for the creation of owl:sameAs links. As mentioned in Section 2.1, there is a need to support the creation of other typed links, not just sameAs statements.

Table 1: LD Interlinking Tools

Tool	RDF Refine[17]	SILK[18]	LIMES[19]	MARiMbA
Data Input	RDF SPARQL	RDF SPARQL CSV	RDF SPARQL CSV	MARC 21 RDF
Link-Types	owl:sameAs	owl:sameAs	owl:sameAs	owl:sameAs
Generation	Automatic, Manual	Manual	Manual	Automatic
Interface	GUI	GUI Web Interface	GUI Web Interface	Cmd Line
Library Datasets	VIAF, LCSH[20] VIVO[21], FAST[22] DBpedia	-	-	VIAF, LIBRIS SUDOC[23], DNB
Domain	Library General Biodiversity	General	General	Library
LD Expertise	Knowledgeable	Knowledgeable	Expert	Knowledgeable

[16] mayor2.dia.fi.upm.es/oeg-upm/index.php/en/technologies/228-marimba/index
[17] http://refine.deri.ie
[18] http://silkframework.org
[19] http://aksw.org/Projects/LIMES.html
[20] http://id.loc.gov/authorities/subjects
[21] https://old.datahub.io/dataset/vivo
[22] http://fast.oclc.org
[23] http://www.sudoc.abes.fr/xslt/

4 AIMS

The Research Question being investigated as part of this study is, "How can information professionals be facilitated to engage with the process of authoritative linked data interlinking with greater efficacy, ease, and efficiency?". With this in mind, and given the conclusions drawn from the literature discussed in Sections 2 and 3, the aims of our study are to:

(1) Propose a LD interlinking framework for the library domain that incorporates the creation LD interlinks with a range of link-types.
(2) Propose a provenance model that expresses the where, when, who, how and why behind the creation of an LD interlink.
(3) Design a graphical user interface (GUI) that provides an instantiation of the proposed interlinking framework and provenance model.
(4) Evaluate the usability and utility of the interlinking framework, provenance model and GUI via user-testing.

Our study contributes to research in the area of LD for libraries by describing a method for IPs to create interlinks between LD resources as well as by developing LD tooling that is accessible to non-technical experts.

5 NAISC

In line with the aims of our study, we developed an interlinking approach specifically for the library domain called NAISC - the Novel Authoritative Interlinking of Schema and Concepts. The NAISC approach encompasses our LD interlinking framework, provenance model and GUI. Figure 2 displays the role of NAISC in the architecture of a LD application.

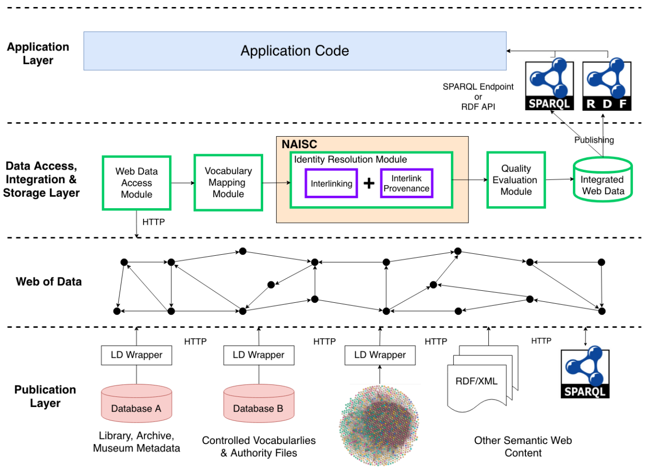


Figure 2: Role of NAISC in a Linked Data Application

5.1 Research Approach

NAISC was developed according to a Design Science (DS) Approach [26] which involves the iterative design and evaluation of an artefact in order to solve an identified problem. This was completed according to the principles of User-Centred Design [46] whereby the user is involved in all stages of development.

5.2 LD Interlinking Framework

The requirements for the LD interlinking framework were distilled from the results of the survey discussed in Section 3.1. These included:

(1) Attuned and adaptable to library workflows.
(2) Designed with the data needs and expertise of IPs in mind.
(3) Option to hide LD technicalities.
(4) Awareness of common data sources and provision of data quality ratings.

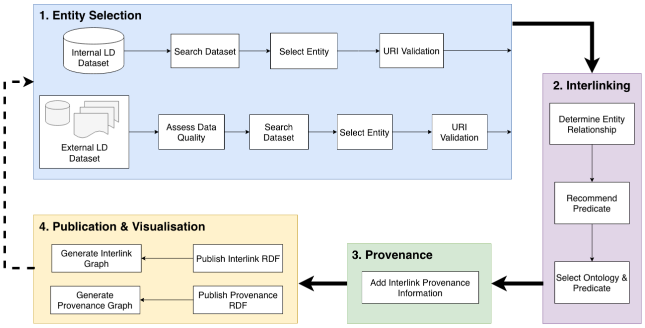


Figure 3: NAISC Interlinking Framework

A cyclical, four-step interlinking framework was subsequently designed, see Figure 3, and the goal of each step is discussed below.

- **Step 1** requires the user to select an internal LD dataset from which a set of URIs will be selected for interlinking. The user is also required to select a set of related URIs from an external LD dataset.
- **Step 2** guides the user through the process of creating a typed link that accurately describes the relationship between an internal and an external URI. Different link-types, or properties, are recommended to the user based on the kind of relationship between the two URIs. This relationship is determined by the user selecting a Relationship Term from a list of six possible options. The link-type properties recommended to the user varies depending on their choice of Relationship Term, for example: Identical - owl:sameAs[24], Similar - ov:similarTo[25], Related - dcterms:relation[26], Represents - sio:represents[27]. These Relationship Terms were taken from research conducted by [22, 23], which discussed the misuse of identity links in LD. The term 'Identical' is used to cover sameAs statements. The term 'Related But Referentially Opaque' refers to instances where two URIs describe the same entity but the properties of the entities are not the same. The term 'Identical But Different Context' describes when two URIs describe the same entity but the URI cannot be re-used in a different context. The term 'Similar' was used to cover instances where two URIs describe different entities but the entities are very similar. The term 'Related' describes instances where two URIs describe different entities that

[24] http://www.w3.org/2002/07/owl#sameAs
[25] http://open.vocab.org/terms/similarTo
[26] http://purl.org/dc/terms/relation
[27] http://semanticscience.org/resource/P138_represents

are related in some capacity. The term ‘Represents’ covers situations where a URI is being used to represent an entity but it is not the entity itself. Finally, the term ‘Other’ was used to describe cases not covered by the other terms.

- **Step 3** involves the generation of provenance data that describes who, where, when, why and how an interlink was created.
- **Step 4** involves the generation of the interlink and provenance RDF. The data is stored in a relational database (RDB) and the RDF triples are generated by uplifting data from the RDB to RDF using an R2RML [15] mapping. This mapping was created using JUMA [1, 2]. Interlinks can be published in Turtle, N-Triples and RDF-XML formats, as well as visualised in graph using GoJS software[28].

5.3 Provenance Model

A set of user requirements for the provenance model were distilled from the results of the LD survey discussed in Section 3.1. These requirements included:

- Allow for different levels of granularity/detail.
- Keep track of modifications to the dataset.
- Link to sources used in the dataset.
- Link to people, organisations, and groups that contributed to the dataset.
- Allow for the explanation/justification of the sources used to create a link.
- Allow for the explanation/justification of the type of link created between resources.

The data provenance requirements of the library domain, which include information on where, when, by whom and how a resource was created, were also considered. [32].

5.3.1 *Ontologies.*

PROV-O was used as the foundation of our interlink provenance model as it is a W3C recommended standard citelebo:w3cProvo. It also provides a model for general provenance descriptions which can then be extended for the needs of domain specific purposes [12]. Existing PROV-O classes, sub-classes and properties were used to describe the who, where and when interlinks were created. We then extended PROV-O, see Figure 4, in order to add interlink specific sub-classes and properties. This extension, called NaiscProv, describes how and why interlinks were created.

The VoID Vocabulary [3] was also used in order to describe the interlinked datasets. Additionally, Dublin Core [9] and FOAF [10] ontologies were used to provide richer descriptions of entities.

5.3.2 *Graph Structure.*

Our Provenance Model, as seen in Figure 5, 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 [11]. Named graphs allow collections of triples to be published as independent units - in this case a set of interlinks associated with a particular dataset or part of a dataset. 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 [19]. In the case of NAISC, an Interlink Graph will contain a set of 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.

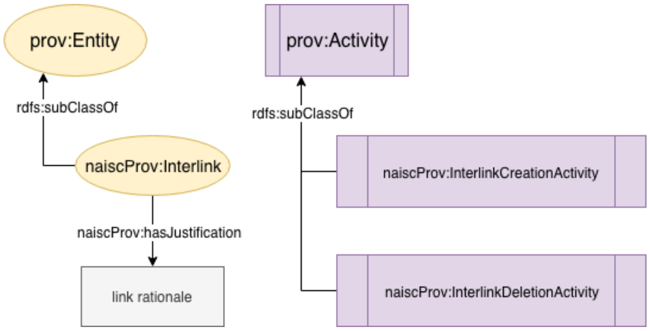

prov:Entity
prov:Activity
rdfs:subClassOf
rdfs:subClassOf
naiscProv:Interlink
naiscProv:InterlinkCreationActivity
naiscProv:hasJustification
link rationale
naiscProv:InterlinkDeletionActivity

Figure 4: NaiscProv PROV-O Extension

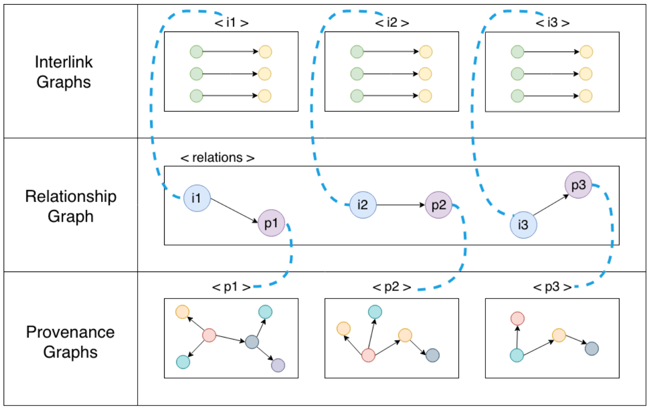


Figure 5: NAISC Provenance Model Graph Structure

(2) Provenance Graph - a prov:Bundle containing the origin data of the statements 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 [30]. 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 using the prov: hasProvenance property.

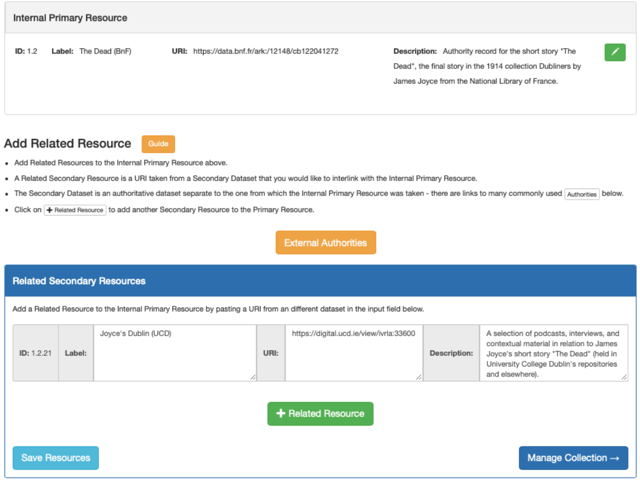
The purpose of these graphs is to allow the user to explore the different sets of interlinks, and also to explore the provenance information for the interlinks. Separating the data in this manner simplifies some of the queries that users could formulate and run over the data whilst still allowing for queries that span across graphs, as facilitated by the relationship layer.

[28] https://gojs.net/latest/index.html

5.4 Graphical User Interface

The purpose of the NAISC GUI[29] is to guide users through the steps proposed in the interlinking framework described in Section 5.2. Using the framework as a guide, we designed a mock-up of the GUI which was subsequently tested by five librarians. As per the DS approach, the results of this evaluation were used to refine the framework and to develop a minimum viable product (MVP) - see Figures 6, 7, 9 and 8 for screenshots. In line with the iterative nature of our research, the usability and utility of the MVP were then evaluated.

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



Internal Primary Resource

ID: 1.2 **Label:** The Dead (BnF) **URI:** https://data.bnf.fr/ark:/12148/cb122041272 **Description:** Authority record for the short story "The Dead", the final story in the 1914 collection Dubliners by James Joyce from the National Library of France.

Add Related Resource

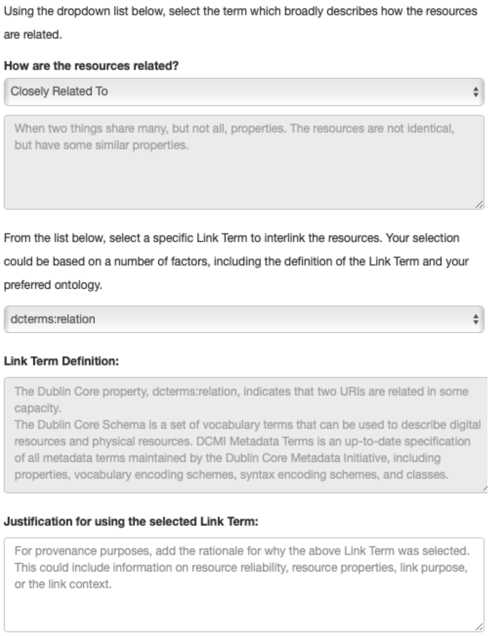
- Add Related Resources to the Internal Primary Resource above.
- A Related Secondary Resource is a URI taken from a Secondary Dataset that you would like to interlink with the Internal Primary Resource.
- The Secondary Dataset is an authoritative dataset separate to the one from which the Internal Primary Resource was taken - there are links to many commonly used Authorities below.
- Click on + Related Resource to add another Secondary Resource to the Primary Resource.

Related Secondary Resources

Add a Related Resource to the Internal Primary Resource by pasting a URI from an different dataset in the input field below.

ID: 1.2.21	**Label:**	Joyce's Dublin (UCD)	**URI:**	https://digital.ucd.ie/view/ivrla:33600	**Description:**	A selection of podcasts, interviews, and contextual material in relation to James Joyce's short story "The Dead" (held in University College Dublin's repositories and elsewhere).

Figure 6: NAISC GUI - Entity Selection



Using the dropdown list below, select the term which broadly describes how the resources are related.

How are the resources related?

Closely Related To

When two things share many, but not all, properties. The resources are not identical, but have some similar properties.

From the list below, select a specific Link Term to interlink the resources. Your selection could be based on a number of factors, including the definition of the Link Term and your preferred ontology.

dcterms:relation

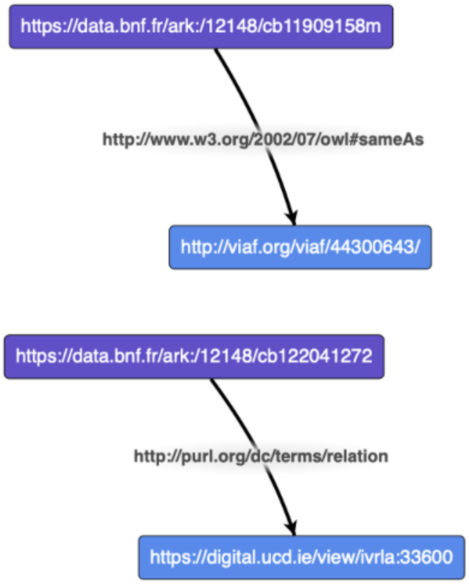
Link Term Definition:

The Dublin Core property, dcterms:relation, indicates that two URIs are related in some capacity.
The Dublin Core Schema is a set of vocabulary terms that can be used to describe digital resources and physical resources. DCMI Metadata Terms is an up-to-date specification of all metadata terms maintained by the Dublin Core Metadata Initiative, including properties, vocabulary encoding schemes, syntax encoding schemes, and classes.

Justification for using the selected Link Term:

For provenance purposes, add the rationale for why the above Link Term was selected. This could include information on resource reliability, resource properties, link purpose, or the link context.

Figure 7: NAISC GUI - Entity Interlinking



https://data.bnf.fr/ark:/12148/cb11909158m
http://www.w3.org/2002/07/owl#sameAs
http://viaf.org/viaf/44300643/
https://data.bnf.fr/ark:/12148/cb122041272
http://purl.org/dc/terms/relation
https://digital.ucd.ie/view/ivrla:33600

Figure 8: NAISC GUI - Interlink Visualisation
Created using GoJS

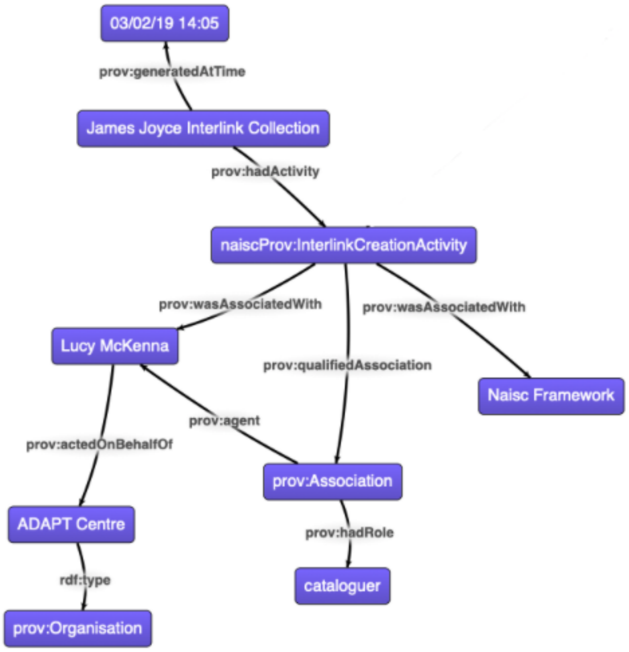


Figure 9: NAISC GUI - Partial Provenance Graph
Created using GoJS

6 USER EVALUATION

Upon completion of a working version of the NAISC GUI, further user testing was undertaken. The methodology, as well as a summary of the key findings of this user-study, are discussed in the following sections.

6.1 Methodology

The user test consisted of four parts - a short pre-test questionnaire, a think-aloud observation, a brief post-test interview and the administration of the Post-Study System Usability Questionnaire (PSSUQ) [31].

6.1.1 Participants. The participants in this study were 15 IPs who had some prior knowledge of LD and the SW. The number of participants that should be recruited for a usability test is a contentious issue, with recommend numbers of participants ranging from 5 [40], 10-12 [27], or more [33], depending on factors such as the complexity of the test, whether the evaluation is formative or summative, and whether quantitative analysis of the results is to be performed. Since there is evidence to suggest that 15 participants can find 90% of usability problems [17], this was the number of participants that we recruited for our study. Non-probabilistic sampling methods were used to recruit participants [14] whereby libraries and information institutions were contacted directly with a request for participants.

6.1.2 Pre-Test Questionnaire. The pre-test questionnaire was used in order to ascertain how participants rated their knowledge of the SW, LD, RDF, URIs and ontologies (Ont). Participants were asked to rate their knowledge on five point scale ranging from 'Not at all Knowledgeable' to 'Extremely Knowledgeable'. The questionnaire also investigated whether participants had ever been directly involved in the implementation of a LD service, and if so, the kinds of LD activities that they gained experience in. The results of the pre-test questionnaire can be found in Section 6.2.1.

6.1.3 Think-Aloud Observation. Think-aloud (TA) observations are a widely used method for the usability testing of software, GUIs, and websites [47]. During a TA, participants are asked to verbalise their thoughts and actions while carrying out a number of scenario-based tasks, thus providing data on the types of difficulties they encounter and highlighting the areas of a system that require further improvement [5, 41]. TAs typically have six to eight tasks [4]. For our study we developed six scenario based tasks which were representative of activities that users would carry out on NAISC. The scenario of our TA was that the participant was a cataloguer for the French National Library (BnF) and that they were required to interlink a set of entities, pertaining to the Irish author James Joyce, from the BnF LD dataset to a set of related entities found in external LD datasets. The BnF LD dataset was used as it offers a rich and varied LD dataset and because participants were likely to be less familiar with its contents over datasets such as LoC and other controlled vocabularies. The six tasks of the TA included:

(1) Creating a new project. This involved setting up a new link-set in which the interlinks pertaining to a particular dataset, or part of a dataset, could be created and contained. In this case the project was created to hold interlinks for the BnF dataset.
(2) Adding an internal URI to the link-set. This involved searching for a specific entity in the BnF dataset, such as a work by James Joyce, and adding its URI to the link-set.
(3) Adding a related URI from an external LD dataset to the link-set. This involved searching an external dataset, such as University College Dublin's Digital Library[30], for a specific resource related to James Joyce.
(4) Selecting an appropriate link-type in order to interlink six internal BnF entities to six external entities. This task required participants to decide on the degree to which the entities in the link-set were related and to subsequently select the link-type which best described this relationship.
(5) Generating the RDF graph and the RDF output, in Turtle, RDF-XML and N-Triples, for the interlinks created.
(6) Generating a sample provenance graph describing how the interlinks were created. This was done for one of the interlinks and used as a means to gather participants' views on our provenance model.

The participants were observed while completing the tasks, their comments were audio-recorded and their work on the GUI was screen-recorded. The results of the TA can be found in Section 6.2.2.

6.1.4 Post-Test Interview. The post-test interview consisted of seven questions which explored the participants' thoughts on the interlinking framework, provenance model and GUI. These questions were:

(1) What is your overall impression of the tool?
(2) What worked well?
(3) What challenges did you encounter?
(4) Are there functions you would like to add or remove from the tool?
(5) What is your impression of the process for selecting link-types in order to link internal and external URIs?
(6) What is your impression of the provenance data stored for the links and interlinking session?
(7) Do you think this tool could be useful for the library domain?

The results of the post-test interview can be found in Section 6.2.3.

6.1.5 PSSUQ. The Post-Study System Usability Questionnaire (PSSUQ) [31] is used for measuring software usability and utility at the end of a user-study. The PSSUQ consists of 19 statements about which the user rates agreement on a seven-point scale from Strongly Agree (1) to Strongly Disagree (7) - thus lower scores indicate fewer usability issues. The results of the PSSUQ can be viewed in four main categories:

(1) System Usefulness (SysUse).
(2) Information Quality (InfoQual).
(3) Interface Quality (InterQual).
(4) Overall Satisfaction.

The PSSUQ was chosen over other questionnaires as it takes both system utility and system usability into account. The results of the PSSUQ can be found in Section 6.2.4.

6.2 Findings

The results for each of the four components of the user-study have been detailed in this section.

6.2.1 Pre-Test Questionnaire. The results of the pre-test questionnaire have been summarised in Table 2. It can be seen that all participants rated themselves as knowledgeable for each of the five concepts, with the majority considering themselves Moderately

[30] https://digital.ucd.ie

Table 2: LD Knowledge Evaluation

Rating / Topic	SW	LD	RDF	URIs	Ont
Not at all Knowledgeable	0	0	0	0	0
Slightly Knowledgeable	2	1	5	4	5
Moderately Knowledgeable	13	14	10	9	8
Very Knowledgeable	0	0	0	2	2
Extremely Knowledgeable	0	0	0	0	0

Knowledgeable. Five of the participants indicated that they had previous experience implementing a LD project. Overall, it can be seen that all participants had some prior awareness and knowledge of LD, but that none rated themselves as Very Knowledgeable in the area, suggesting that no participant was an expert user of LD.

6.2.2 Think-Aloud Evaluation. The results of the TA evaluation have been summarised in Table 3. Here the time (in minutes) that it took each participant to complete each task is documented, as well the degree to which the participant was able to complete the task.

The task which took the longest time for each participant to complete was Task 5, with an average of 18.453 minutes. This was somewhat to be expected given that the task required the participant to complete six separate link-type selection activities, resulting in an average time of 3.075 minutes per activity. In terms of task completion, all but Task 5 had a completeness score of 100%. It is worthy of note, however, that the two participants who were unable to fully complete Task 5 did so due to time constraints, as opposed to the task being too difficult.

In addition to the above, the common challenges which arose during the TA tasks were documented and summarised in Table 4.

6.2.3 Post-Test Interview. The recordings of the interviews were analysed and the key points raised are summarised in Table 6.2.3.

6.2.4 PSSUQ. The combined average scores for each category of the PSSUQ can be seen in Table 6. As mentioned in 6.1.5, the PSSUQ is scored from 1-7, with lower scores indicating fewer usability issues. The average score for each of the components of the PSSUQ was between 2-3. This indicates that participants were generally in agreement with the questionnaire statements and suggests mild usability issues with the GUI.

7 DISCUSSION

In this section, the findings outlined in Section 6.2, will be discussed in relation to each of NAISC's components.

7.1 Interlinking Framework

The results of our study indicate that users found NAISC to be a usable and useful approach for creating LD interlinks. Users found the step-by-step process for selecting an appropriate link-term to create a meaningful interlink between two URIs to be understandable and user-friendly.

It was also noted, however, that the approach is quite time consuming and that automating some of the processes, such as auto-URI ingestion and the addition of a automated predicate recommender, would be useful. However, it was observed that as participants became more familiar with NAISC, they were able to use the GUI with greater ease and complete tasks in less time. This was especially noticeable in Task 5, where participants were able to select appropriate link-types, in order to interlink entities, at a quicker pace over time.

Table 3: Think Aloud Evaluation

Participant	Task 1		Task 2		Task 3		Task 4		Task 5		Task 6	
	Time (mins)	Complete (%)	Time (mins)	Complete (%)	Time (mins)	Complete (%)	Time (mins)	Complete (%)	Time (mins)	Complete (%)	Time (mins)	Complete (%)
1	1.766	100%	3.633	100%	2.433	100%	13.366	100%	2.683	100%	0.066	100%
2	1.383	100%	3.75	100%	3.916	100%	12.133	100%	2.75	100%	0.133	100%
3	0.866	100%	1.833	100%	1.7	100%	8.283	100%	2.016	100%	0.183	100%
4	3.766	100%	4.0	100%	5.566	100%	33.566	100%	2.45	100%	0.366	100%
5	1.5	100%	2.833	100%	3.633	100%	22.25	100%	2.8	100%	0.116	100%
6	1.066	100%	2.483	100%	4.133	100%	14.2	100%	1.633	100%	0.05	100%
7	0.75	100%	4.1	100%	3.816	100%	18.3	100%	0.85	100%	0.2	100%
8	1.633	100%	4.00	100%	3.233	100%	28.033	100%	1.591	100%	0.2	100%
9	0.866	100%	4.1	100%	4.55	100%	23.116	100%	4.13	100%	0.333	100%
10	1.15	100%	3.3	100%	2.933	100%	11.283	100%	1.216	100%	0.45	100%
11	3.9	100%	4.816	100%	4.133	100%	28.75	100%	3.883	100%	0.2	100%
12	2.283	100%	3.966	100%	3.116	100%	18.05	100%	0.083	100%	0.25	100%
13	1.6	100%	6.4	100%	4.1	100%	15.166	66.66%	1.033	100%	0.266	100%
14	0.566	100%	3.35	100%	2.566	100%	14.6	100%	1.05	100%	0.233	100%
15	1.383	100%	6.9	100%	4.583	100%	15.7	66.66%	1.525	100%	0.2	100%
Average	1.631	100%	3.964	100%	3.627	100%	18.453	95.55%	1.983	100%	0.216	100%

Table 4: Think Aloud Task Challenges

Task	Challenges
1	Participants indicated that clearer descriptions of the information required for each field in the collection creation form should be provided.
2	Some participants were unsure which button to click in order to add an internal URI to a collection. Again participants mentioned that the information required for each form field should be defined more precisely.
3	Participants did not always notice the links to the external authorities.
4	Some participants initially found it difficult to identify which two URIs were being interlinked. Participants were not aware that the definition of a Relationship Term would be provided once selected from the dropdown list.
5	Participants suggested adding natural language labels to the graphs in order to improve readability.
6	Again participants suggested adding natural language labels to the graph in order to improve their understanding of the provenance data. They also suggested that having the option to view the provenance data at link-set level, and not just the interlink level, would be useful.

7.2 Provenance Model

The results of the user-study show that participants considered the data captured by the provenance graph to be sufficient for the purpose of curating a set of interlinks. They also indicated that the provision of such data provenance would greatly add to the trustworthiness of the interlinks. However, participants did stress the importance of including natural language labels to the graph so that it can be understood by users who are less familiar with RDF.

7.3 Graphical User Interface

The results of the PSSUQ indicate mild usability issues with the GUI. Navigation issues were noted during activities requiring the participants to add a URI to a link-set. In addition, some participants initially found it difficult to identify which two URIs needed to be interlinked. Future iterations of the tool will use colour coding in order to clearly point to related URIs.

8 CONCLUSIONS AND FUTURE DIRECTIONS

One of the main benefits of LD is the ability to interlink related resources across datasets. However, non-LD experts, such as IPs, are currently unable to engage fully with this process. In response to this we developed the NAISC approach as a means of facilitating increased IP engagement in the LD interlinking process. The results of our study, which evaluated the first iteration of NAISC, demonstrated the successful use of the approach by IPs in order to create LD interlinks via a user-friendly GUI.

Table 5: Interview Findings

QN	Feedback
1	**Overall Impression:** Participants indicated that NAISC was easy and pleasant to use and that, as they became used to the system, the ease and speed of use increased.
2	**Positives:** Participants found that the RDF graphs, of the interlink and provenance data, added to their understanding of the interlinking process.
3	**Challenges:** Participants noted that the process of adding URIs to a link-set was initially confusing due to the labelling of buttons and some of the terminology used.
4	**Additional Functions:** Participants stated that adding a way to view a visualisation of each interlink as it is being created would be a useful function. Participants also mentioned that increased automation for the process of searching for and adding URIs to a link-set would improve their efficiency. The addition of data quality metrics for each of the available external datasets was also suggested.
5	**Link-Type Selection:** The participants indicated that the definitions for each of the Relationship Terms and link-types were useful for deciding on how to express the relation between two URIs. They emphasised that examples should be provided in order to aid in decision making.
6	**Provenance:** Participants stated that the provision of the provenance data was useful and that it would add to the authoritativeness of the interlink data. Participants were also satisfied with the level of detail in provenance data.
7	**Usefulness:** All participants stated that NAISC would be useful for creating interlinks between internal and external LD resources. However, they expressed concerns regarding whether NAISC could be integrated into their current cataloguing systems and workflows.

Table 6: PSSUQ Average Scores

PSSUQ	SysUse	InfoQual	InterQual	Overall
Score	2.45	2.45	2.65	2.07

Future research will involve using the results of this user-study to modify and refine the second iteration of the NAISC approach. The primary focus of the next iteration of NAISC will be to increase the speed and accuracy of the link-type selection process, to provide data quality ratings for external LD datasets, to improve the usability of the GUI and to provide provenance data for all 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] Crotti Junior. A., C. Debruyne, and D. O'Sullivan. 2017. Using a Block Metaphor for Representing R2RML Mappings. In *Proceedings of the Third International Workshop on Visualization and Interaction for Ontologies and Linked Data co-located with the 16th International Semantic Web Conference (ISWC 2017), Vienna, Austria, October 22, 2017.* 1–12. http://ceur-ws.org/Vol-1947/paper01.pdf

[2] Crotti Junior. A., C. Debruyne, and D. O'Sullivan. 2018. Juma Uplift: Using a Block Metaphor for Representing Uplift Mappings. In *2018 IEEE 12th International Conference on Semantic Computing (ICSC)*. 211–218. https://doi.org/10.1109/ICSC.2018.00037

[3] 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/.

[4] Christopher Andrews, Debra Burleson, Kristi Dunks, Kimberly Elmore, Carie S. Lambert, Brett Oppegaard, Elizabeth E. Pohland, Danielle Saad, Jon S. Scharer, Ronda L. Wery, Monica Wesley, and Gregory Zobel. 2012. A New Method in User-Centered Design: Collaborative Prototype Design Process (CPDP). *Journal of Technical Writing and Communication* 42, 2 (2012), 123–142. https://doi.org/10.2190/TW.42.2.c arXiv:https://doi.org/10.2190/TW.42.2.c

[5] Danielle A Becker and Lauren Yannotta. 2013. Modeling a library web site redesign process: Developing a user-centered web site through usability testing. *Information Technology and Libraries* 32, 1 (2013), 6–22.

[6] T. Berners-Lee. 2006. Linked Data. Retrieved November 2018 from https://www.w3.org/DesignIssues/LinkedData.

[7] T. Berners-Lee, J. Hendler, and O. Lassila. 2001. The Semantic Web. *Scientific American* 284, 5 (2001), 1–5.

[8] 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.

[9] 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/.

[10] D. Brickley and L. Miller. 2014. FOAF Vocabulary Specification 0.99. Retrieved November 2018 from http://xmlns.com/foaf/spec/.

[11] 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.

[12] 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.

[13] Panel 2] Consultative Committee for Space Data Systems [CCSDS. 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.

[14] J. Daniel. 2011. *Sampling Essentials: Practical Guidelines for making Sampling Choices* (1st ed.). SAGE Publications Ltd., CA, Chapter Choosing Between Non-Probability Sampling and Probability Sampling, 66–80.

[15] S. Das, S. Sundara, and R. Cyganiak. 2012. R2RML: RDB to RDF Mapping Language. W3C Recommendation 27 September 2012. Retrieved January 2019 from https://www.w3.org/TR/r2rml/.

[16] L. Ding, J. Shinavier, T. Finin, and D.L. McGuinness. 2010. owl: sameAs and Linked Data: An empirical study. In *Proceedings of the Second Web Science Conference.*

[17] L. Faulkner. 2003. Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. *Behavior Research Methods, Instruments, & Computers* 35, 3 (01 Aug 2003), 379–383. https://doi.org/10.3758/BF03195514

[18] Alfio Ferrara, Andriy Nikolov, and François Scharffe. 2011. Data linking for the semantic web. *International Journal on Semantic Web and Information Systems (IJSWIS)* 7, 3 (2011), 46–76.

[19] 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.*

[20] 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.

[21] 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.

[22] H. Halpin, P.J. Hayes, J.P. McCusker, D.L. McGuinness, and H.S. Thompson. 2010. When owl:sameAs IsnâĂŹt the Same: An Analysis of Identity in Linked Data.. In *Proceedings of the 9th International Semantic Web Conference.*

[23] H. Halpin, I Herman, and P.J. Hayes. 2010. When owl:sameAs IsnâĂŹt the Same: An Analysis of Identity Links on the Semantic Web. Retrieved January 2019 from http://www.w3.org/2009/12/rdf-ws/papers/ws21.

[24] 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.*

[25] R. Hastings. 2015. Linked Data in Libraries: Status and Future Direction. *Computers in Libraries* 35, 9 (2015), 12–16.

[26] A. Hevner and S. Chatterjee. 2010. *Design research in Information Systems: Theory and Practice.* Springer Publishing, New York.

[27] W. Hwang and G. Salvendy. 2010. Number of People Required for Usability Evaluation: The 10&Plusmn;2 Rule. *Commun. ACM* 53, 5 (May 2010), 130–133. https://doi.org/10.1145/1735223.1735255

[28] J.G. Kim and M. Hausenblas. 2015. 5 Star Open Data. Retrieved January 2019 from https://5stardata.info/en/.

[29] 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.

[30] 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/.

[31] R.J. Lewis. 2002. Psychometric Evaluation of the PSSUQ Using Data from Five Years of Usability Studies. *International Journal of HumanâĂŞComputer Interaction* 14, 3-4 (2002), 463–488.

[32] 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.*

[33] R. Macefield. 2009. How to Specify the Participant Group Size for Usability Studies: A Practitioner's Guide. *J. Usability Studies* 5, 1 (Nov. 2009), 34–45. http://dl.acm.org/citation.cfm?id=2835425.2835429

[34] 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.

[35] E.T. Mitchell. 2016. Library Linked Data: Early Activity and Development. *Library Technology Reports* 52, 1 (2016), 5–33.

[36] 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.

[37] 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.

[38] P. Neish. 2015. Linked data: what is it and why should you care? *The Australian Library Journal* 64, 1 (2015), 3–10.

[39] 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

[40] J. Nielsen. 2000. Why You Only Need to Test with 5 Users. Retrieved January 2019 from https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/.

[41] J. Nielsen. 2012. Thinking Aloud: The No. 1 Usability Tool. Retrieved January 2019 from https://www.nngroup.com/articles/thinking-aloud-the-1-usability-tool/.

[42] Library of Congress. 2018. PREMIS: Preservation Metadata Maintenance Activity. Retrieved January 2019 from http://www.loc.gov/standards/premis/.

[43] L. Papaleo, N. Pernelle, F. Saïs, and C. Dumont. 2014. Logical Detection of Invalid SameAs Statements in RDF Data. In *Knowledge Engineering and Knowledge Management*, K. Janowicz, S. Schlobach, P. Lambrix, and E. Hyvönen (Eds.). Springer International Publishing, Cham, 373–384.

[44] 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.*

[45] K. Smith-Yoshimura. 2018. Analysis of 2018 international linked data survey for implementers. *Code4Lib* 42 (2018).

[46] D. Travis. 2011. ISO 13407 is dead. Long live ISO 9241-210! Retrieved from January 2019 https://www.userfocus.co.uk/articles/iso-13407-is-dead.html.

[47] M. van den Haak, M. De Jong, and P.J. Schellens. 2003. Retrospective vs. concurrent think-aloud protocols: Testing the usability of an online library catalogue. *Behaviour & Information Technology* 22, 5 (2003), 339–351. https://doi.org/10.1080/0044929031000 arXiv:https://doi.org/10.1080/0044929031000

[48] W3C. 2014. RDF: Resource Description Framework. https://www.w3.org/RDF/.

[49] W3C. 2015. Semantic Web. https://www.w3.org/standards/semanticweb/.