eLucidate Article

Title: NAISC-L: An Authoritative Linked Data Interlinking Approach for the Library Domain

Authors: Lucy McKenna, Christophe Debruyne & Declan O'Sullivan

In 2017 we distributed a questionnaire to Information Professionals (IPs) in libraries, archives and museums (LAMs) in order to explore the benefits and challenges that they experienced when using Linked Data[i] (LD). Of the 185 responses, over 60% indicated that LAMs face multiple barriers to using LD particularly in the areas of LD interlinking, tooling, integration, and resource quality. A more in-depth exploration of the interlinking issue highlighted that the processes of ontology and link type selection (determining and describing the relationship between two entities) were areas of particular difficulty. Participants also mentioned that LD tools are often technologically complex and unsuitable for the needs of LAMs. With regards to data integration, participants indicated that mapping between different vocabularies used across datasets poses a significant challenge. Participants also expressed concerns regarding the quality and the reliability of many currently published LD resources.

In response to the results of the survey, we developed a LD interlinking framework and accompanying tool specifically for the library domain. This framework and tool are summarised below, however, a more in-depth description of the framework can be found in McKenna, Debruyne and O'Sullivan (2019)[ii].

The Semantic Web and Linked Data

The Web contains a vast amount of information presented in the form of documents linked together via hyperlinks. In order to find specific resources on the Web, search engines are used to rank webpages based on relevancy via keyword searches. While this is done to great effect, unlike humans, computers have very little understanding of the meaning of data on these webpages nor do they understand how they relate to each other.

The Semantic Web (SW) is an extension of the current Web in which individual units of information/data are given a well-defined meaning, and where the relationships between data are defined in a common machine-readable format[iii]. These units of data are known as entities and an entity could be a person, place, organisation, object, concept or Thing. Linked Data (LD) involves creating unique identifiers for these entities and then linking them together by meaningfully describing how they are related[iv]. Entities can be linked to endless amounts of other related resources, creating a Web of Data.

A LD dataset is structured information encoded using the Resource Description Framework[v] (RDF), the recommended model for representing and exchanging LD on the Web. RDF statements take the form of subject-predicate-object triples, which can be organised in graphs. Subjects and objects typically represent an entity such as a person, place or Thing, and predicate properties describe the relationship between the two. RDF requires that Unique Resource Identifiers (URIs), such as URLs and permalinks, are used to identify subjects and predicates. An object can also be identified by a URI or by a literal (i.e. plain text). These URIs allow for the data to be understood by computers.

Linked Data Interlinking

LD is classified according to a 5 Star[vi] rating scheme and, in order to be considered 5 Star, a LD dataset must contain external interlinks to related data. LD interlinking describes the task creating a relationship between an entity in one LD dataset to an entity in another LD dataset. Interlinks can be used as a way of representing that both entities describe the same Thing or as a way of indicating that they are similar or related to one another in some capacity. Such links have the potential to transform the Web into a globally interlinked and searchable database allowing for richer data querying and for the development of novel applications built on top of the Web.

Upon reviewing the data on the Linked Open Data Cloud[vii] for some of the leading library LD projects, such as those of the Swedish[viii] (LIBRIS), French[ix] (BnF), Spanish[x] (BnE), British[xi] (BNB) and German[xii] (DNB) National Libraries, it was found that the majority of interlinks are to authority files and controlled vocabularies. Although these types of interlinks are extremely useful, there is a notable lack of interlinks created for purposes outside of authority control. For instance, interlinking could also be used to enrich data by linking to external resources that provide additional information and context for a particular entity.

Our Research

The focus of our research was to develop an interlinking framework that would encourage the creation of different kinds of LD interlinks and that was designed with the needs of the library domain in mind. In order to remove some of the challenges experienced by librarians when working with LD, we also developed an accompanying graphical user-interface which was designed to be used by metadata experts rather than technical LD experts.

NAISC-L

NAISC-L stands for Novel Authoritative Interlinking for Semantic Web Cataloguing in Libraries. The word NAISC (pronounced noshk) is also the Irish word for links. The NAISC-L approach encompasses a LD interlinking framework, a provenance model and a graphical user-interface.

The NAISC-L interlinking framework is a cyclical, four-step method to creating an interlink (as outlined below in Figure 1).

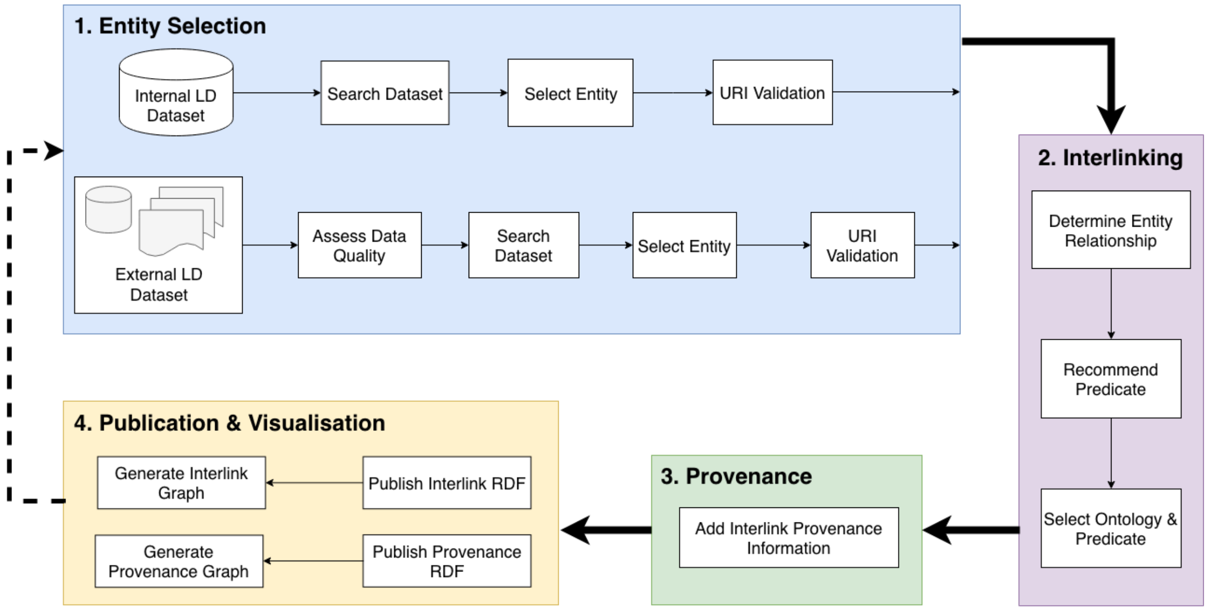


Figure 1 NAISC-L Interlinking Framework

- **Step 1** first requires the user to select entities, from an internal dataset, which they would like to create interlinks from. The user is then required to search for and select entities in external datasets which they would like to create interlinks to.
- **Step 2** guides the user through the process of selecting a property/predicate that accurately describes the relationship between an internal and external entity, thus creating an interlink. This process first requires the user to determine the type of relationship between the two entities using a natural language term e.g. ‘is identical to’, ‘is similar to’, ‘is associated with’. Following this, the user is then presented with a list of properties/predicates which represent the selected relationship type. Using the provided property definitions and examples, the user is then guided to select the property most suitable for interlinking the entities.
- **Step 3** involves the generation of provenance data, using the NAISC-L provenance model, that describes who, where, when, why and how an interlink was created.
- **Step 4** involves the generation of the interlink and provenance RDF data.

The NAISC-L provenance model uses PROV-O[xiii] as its foundation as it is the W3C recommended standard for describing provenance data and because it can be easily extended for domain specific purposes. We used PROV-O to describe who, where and when an interlink was created. We then extended PROV-O to include interlink specific sub-classes and properties. This extension, called NaiscProv (see Figure 2), is used to describe how and why interlinks were created.

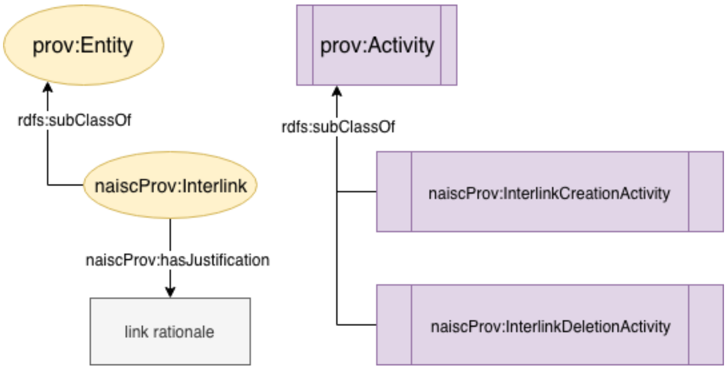
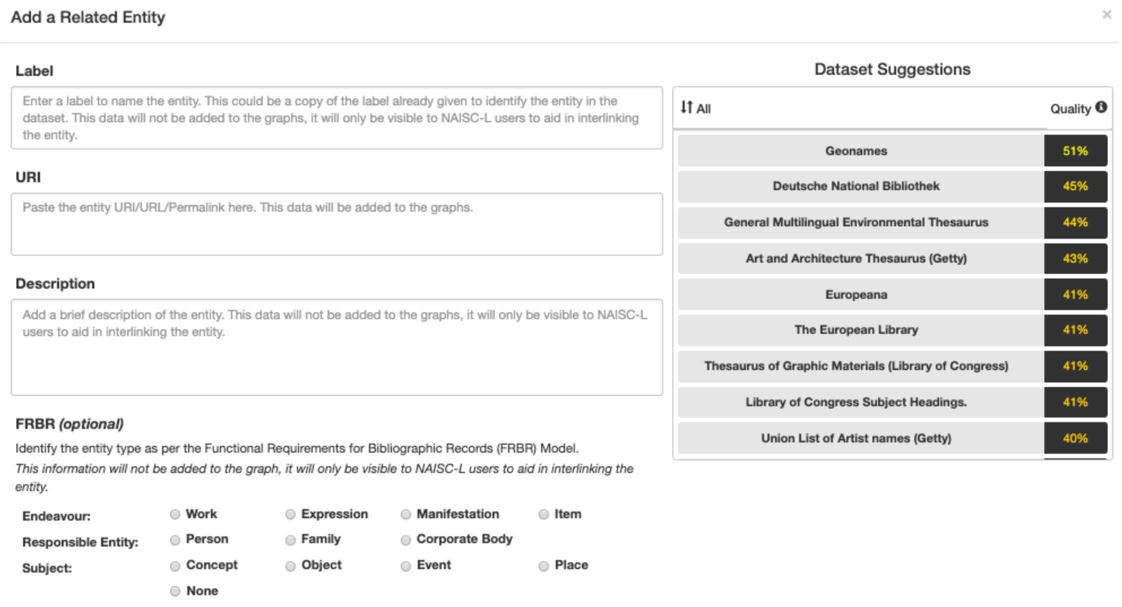

prov:Entity
prov:Activity
rdfs:subClassOf
rdfs:subClassOf
naiscProv:Interlink
naiscProv:InterlinkCreationActivity
naiscProv:hasJustification
link rationale
naiscProv:InterlinkDeletionActivity

Figure 2: NasicProv PROV-O Extension

The above framework and provenance model are accessible to the user via the NAISC-L graphical user-interface (GUI). The purpose of the GUI is to guide users through each of the steps outlined in the framework. An iterative user-centred design approach was followed in the creation of the GUI meaning that Information Professionals were involved in a series of cyclical tool design and testing phases.

Step 1 of the framework is represented on the GUI similarly to the image in Figure 3 below. Here the user can enter a label, URI and a description of a particular entity. The user also has the option of describing the entity as per the Functional Requirements for Bibliographic Records[xiv] (FRBR) model in order to aid in the interlinking process. In the case of selecting a Related Entity, the user is presented with a list of LD datasets in which they can search for a Related Entity. Each of LD datasets were given a quality score based on three quality metrics – Trustworthiness, Interoperability and Licensing. The datasets included in NAISC-L were selected based on the results of the 2017 survey, discussed above, from which a list of commonly used LD datasets was derived. As part of the same survey, participants were asked to select the evaluation criteria they apply when using/searching for external data sources, the results of which informed the quality metrics chosen for dataset analysis. The aim of providing this data quality score was to assist users in selecting high quality and authoritative resources to interlink with.



Add a Related Entity

×

Label

Enter a label to name the entity. This could be a copy of the label already given to identify the entity in the dataset. This data will not be added to the graphs, it will only be visible to NAISC-L users to aid in interlinking the entity.

URI

Paste the entity URI/URL/Permalink here. This data will be added to the graphs.

Description

Add a brief description of the entity. This data will not be added to the graphs, it will only be visible to NAISC-L users to aid in interlinking the entity.

FRBR *(optional)*

Identify the entity type as per the Functional Requirements for Bibliographic Records (FRBR) Model.

This information will not be added to the graph, it will only be visible to NAISC-L users to aid in interlinking the entity.

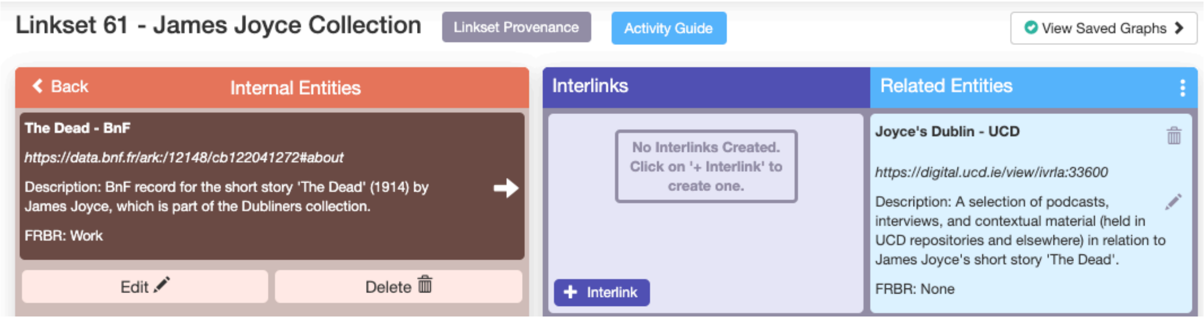
Endeavour:	○ Work	○ Expression	○ Manifestation	○ Item
Responsible Entity:	○ Person	○ Family	○ Corporate Body	
Subject:	○ Concept	○ Object	○ Event	○ Place
	○ None			

Dataset Suggestions

⇅ All	Quality
Geonames	51%
Deutsche National Bibliothek	45%
General Multilingual Environmental Thesaurus	44%
Art and Architecture Thesaurus (Getty)	43%
Europeana	41%
The European Library	41%
Thesaurus of Graphic Materials (Library of Congress)	41%
Library of Congress Subject Headings.	41%
Union List of Artist names (Getty)	40%

Figure 3: Entity Selection

Part 2 of the framework is represented in a step-by-step process which guides the user in selecting the type of relationship type between a pair of entities (see Figure 4 and 5), followed by selecting a property which represents this relationship (see Figure 6).



Linkset 61 - James Joyce Collection

The Dead - BnF

https://data.bnf.fr/ark:/12148/cb122041272#about

Description: BnF record for the short story 'The Dead' (1914) by James Joyce, which is part of the Dubliners collection.

FRBR: Work

No Interlinks Created. Click on '+ Interlink' to create one.

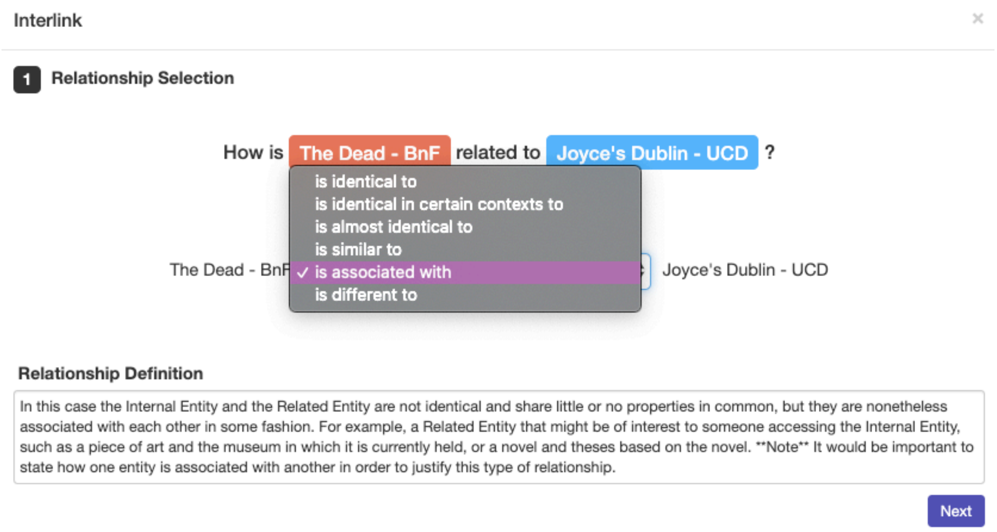
Joyce's Dublin - UCD

https://digital.ucd.ie/view/ivrla:33600

Description: A selection of podcasts, interviews, and contextual material (held in UCD repositories and elsewhere) in relation to James Joyce's short story 'The Dead'.

FRBR: None

Figure 4: Related Entities



Interlink

1 Relationship Selection

How is The Dead - BnF related to Joyce's Dublin - UCD ?

The Dead - BnF

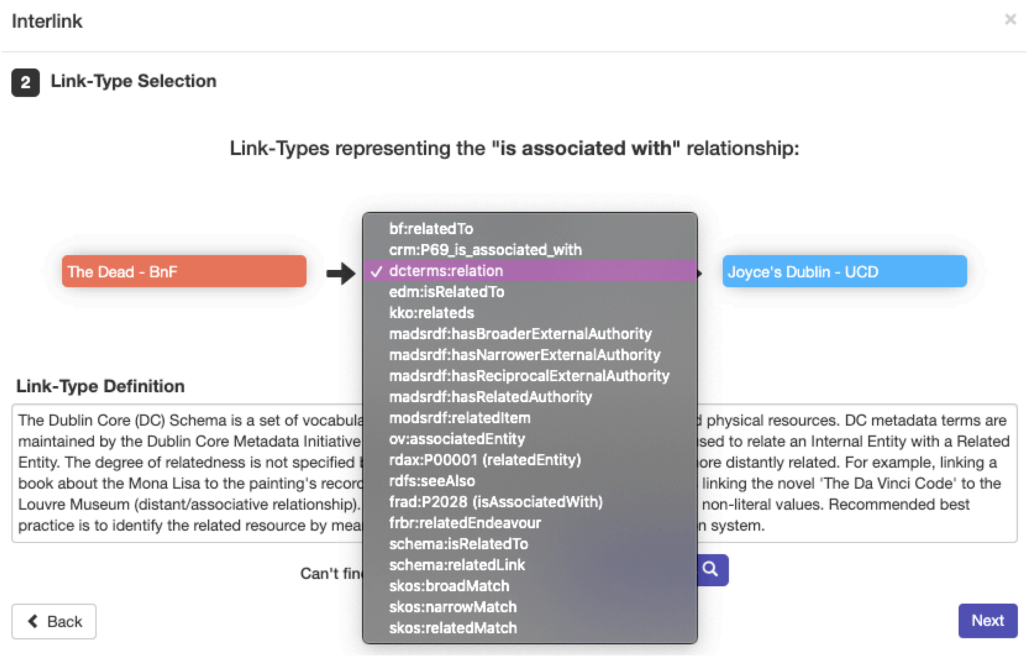
- is identical to
- is identical in certain contexts to
- is almost identical to
- is similar to
- ✓ is associated with
- is different to

Joyce's Dublin - UCD

Relationship Definition

In this case the Internal Entity and the Related Entity are not identical and share little or no properties in common, but they are nonetheless associated with each other in some fashion. For example, a Related Entity that might be of interest to someone accessing the Internal Entity, such as a piece of art and the museum in which it is currently held, or a novel and theses based on the novel. **Note** It would be important to state how one entity is associated with another in order to justify this type of relationship.

Figure 5: Relationship Type Selection



Interlink

2 Link-Type Selection

Link-Types representing the "**is associated with**" relationship:

The Dead - BnF → ✓ dcterms:relation → Joyce's Dublin - UCD

bf:relatedTo
crm:P69_is_associated_with
✓ dcterms:relation
edm:isRelatedTo
kko:relateds
madsrdf:hasBroaderExternalAuthority
madsrdf:hasNarrowerExternalAuthority
madsrdf:hasReciprocalExternalAuthority
madsrdf:hasRelatedAuthority
modsrdf:relatedItem
ov:associatedEntity
rdax:P00001 (relatedEntity)
rdfs:seeAlso
frad:P2028 (isAssociatedWith)
frbr:relatedEndeavour
schema:isRelatedTo
schema:relatedLink
skos:broadMatch
skos:narrowMatch
skos:relatedMatch

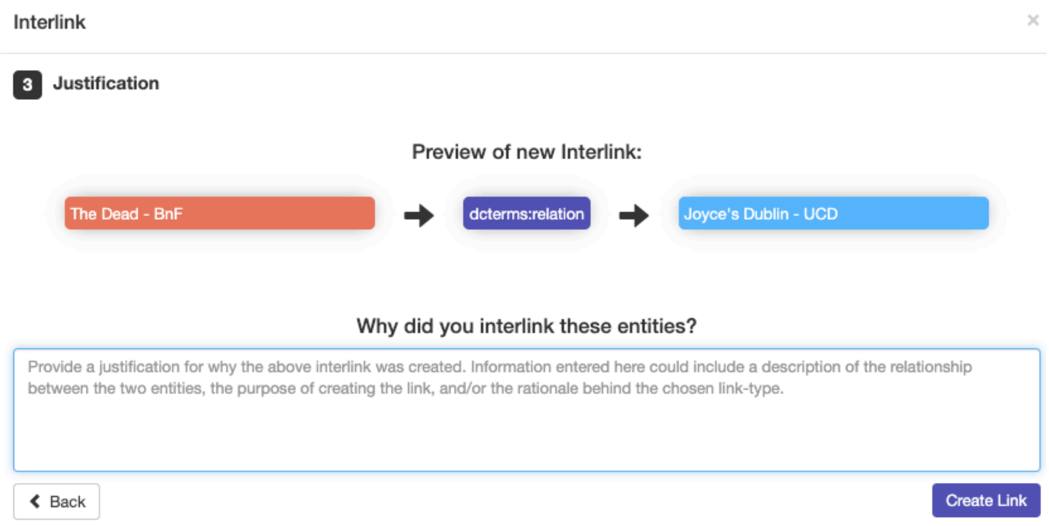
Link-Type Definition

The Dublin Core (DC) Schema is a set of vocabula … d physical resources. DC metadata terms are
maintained by the Dublin Core Metadata Initiative … sed to relate an Internal Entity with a Related
Entity. The degree of relatedness is not specified … ore distantly related. For example, linking a
book about the Mona Lisa to the painting's record … linking the novel 'The Da Vinci Code' to the
Louvre Museum (distant/associative relationship). … non-literal values. Recommended best
practice is to identify the related resource by mea … n system.

Can't fin…

Figure 6: Property Selection

Step 3 of the framework is completed automatically by the tool (e.g. date, time, user), except, when creating an interlink the user is required to enter a justification for the link in order to provide the 'why' portion of the provenance model (see Figure 7).



Interlink

3 Justification

Preview of new Interlink:

The Dead - BnF → dcterms:relation → Joyce's Dublin - UCD

Why did you interlink these entities?

Provide a justification for why the above interlink was created. Information entered here could include a description of the relationship between the two entities, the purpose of creating the link, and/or the rationale behind the chosen link-type.

Figure 7: Justifying an Interlink

Step 4 of the framework is again completed automatically by the tool. The user is presented with an RDF graph and visualisation of the interlinks generated and their corresponding provenance data. Figure 8 below demonstrates a visualisation of a single interlink - in this case a link between the entity for James Joyce's short story 'The Dead' held in the Bibliothèque national de France (BnF) to an entity for a collection of items related to the story held in the library of University College Dublin.

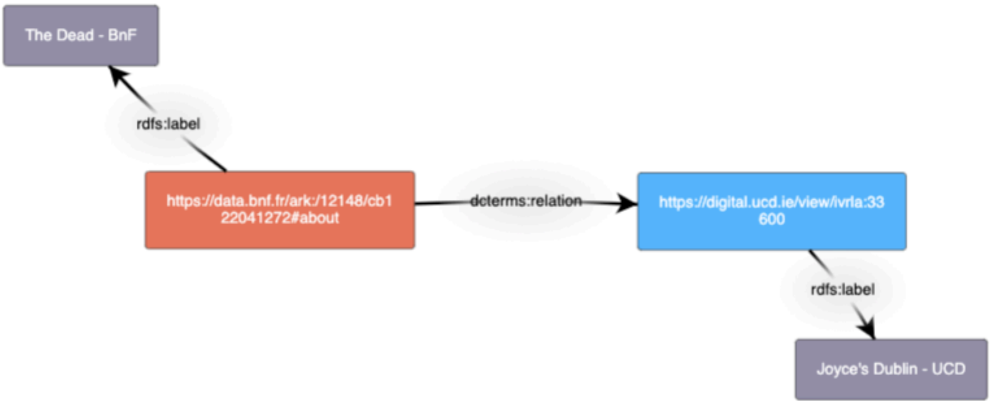
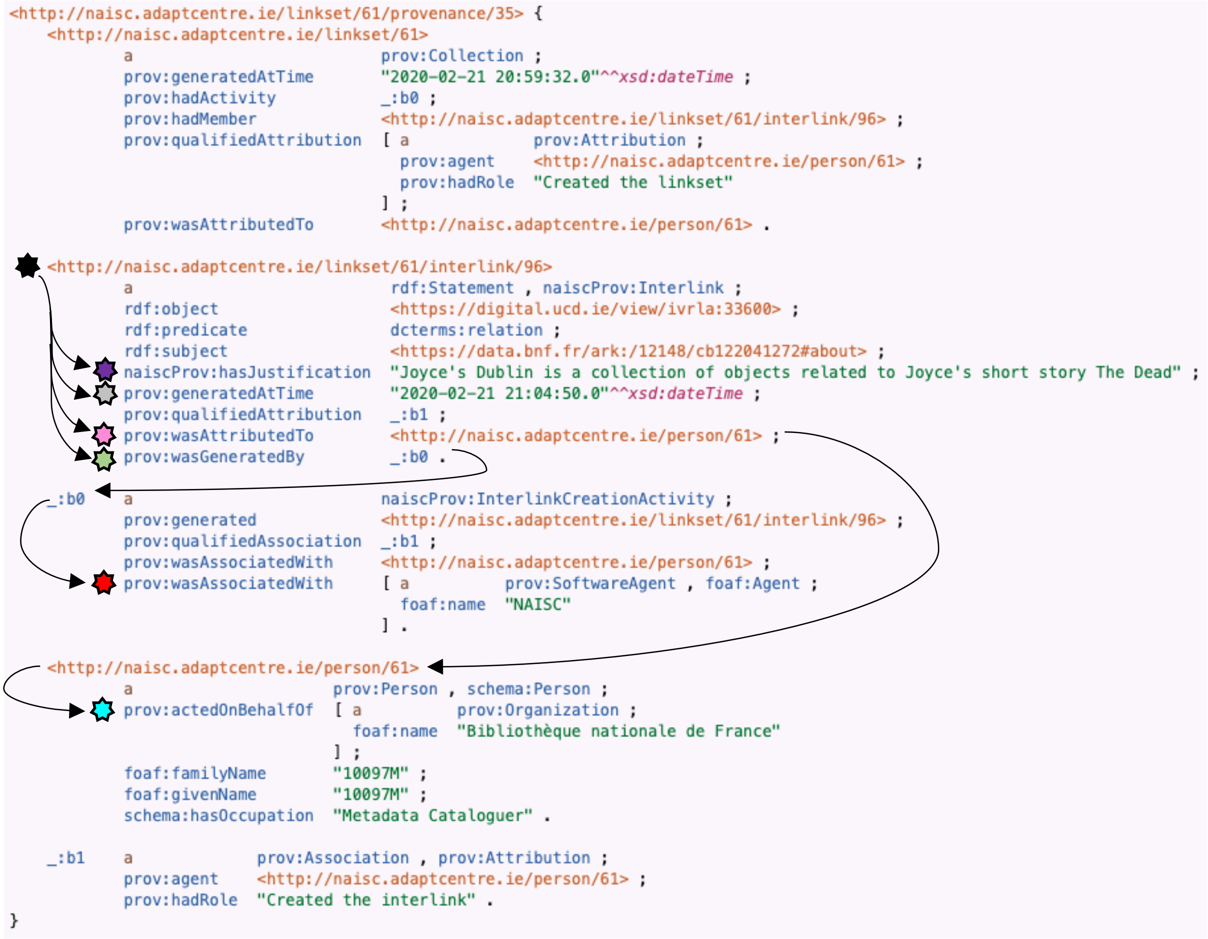

The Dead - BnF
rdfs:label
https://data.bnf.fr/ark:/12148/cb122041272#about
dcterms:relation
https://digital.ucd.ie/view/ivrla:33600
rdfs:label
Joyce's Dublin - UCD

Figure 8: Interlink Graph Visualisation (created using GoJS[xv])

Using RDF Turtle syntax, Figure 9 below demonstrates the provenance of the interlink (*<http://naisc.adaptcentre.ie/linkset/61/interlink/96>) displayed in Figure 8. Note that the URIs in Figure 8 for the subject (orange box), predicate and object (blue box), correspond to the rdf:subject, rdf:predicate and rdf:object in Figure 9. Other provenance information for the creation of the interlink in Figure 9 includes:

- who (*prov:wasAttributedTo),
- what (*prov:wasAssociatedWith),
- where (*prov:actedOnBehalftOf),
- when (*prov:generatedAtTime),
- how (*prov:wasGeneratedBy)
- and why (*naiscProv:hasJustification)



```
<http://naisc.adaptcentre.ie/linkset/61/provenance/35> {
    <http://naisc.adaptcentre.ie/linkset/61>
            a                           prov:Collection ;
            prov:generatedAtTime        "2020-02-21 20:59:32.0"^^xsd:dateTime ;
            prov:hadActivity            _:b0 ;
            prov:hadMember              <http://naisc.adaptcentre.ie/linkset/61/interlink/96> ;
            prov:qualifiedAttribution   [ a             prov:Attribution ;
                                          prov:agent    <http://naisc.adaptcentre.ie/person/61> ;
                                          prov:hadRole  "Created the linkset"
                                        ] ;
            prov:wasAttributedTo        <http://naisc.adaptcentre.ie/person/61> .

    <http://naisc.adaptcentre.ie/linkset/61/interlink/96>
            a                            rdf:Statement , naiscProv:Interlink ;
            rdf:object                   <https://digital.ucd.ie/view/ivrla:33600> ;
            rdf:predicate                dcterms:relation ;
            rdf:subject                  <https://data.bnf.fr/ark:/12148/cb122041272#about> ;
            naiscProv:hasJustification  "Joyce's Dublin is a collection of objects related to Joyce's short story The Dead" ;
            prov:generatedAtTime         "2020-02-21 21:04:50.0"^^xsd:dateTime ;
            prov:qualifiedAttribution    _:b1 ;
            prov:wasAttributedTo         <http://naisc.adaptcentre.ie/person/61> ;
            prov:wasGeneratedBy          _:b0 .

    _:b0    a                           naiscProv:InterlinkCreationActivity ;
            prov:generated              <http://naisc.adaptcentre.ie/linkset/61/interlink/96> ;
            prov:qualifiedAssociation   _:b1 ;
            prov:wasAssociatedWith      <http://naisc.adaptcentre.ie/person/61> ;
            prov:wasAssociatedWith      [ a            prov:SoftwareAgent , foaf:Agent ;
                                          foaf:name    "NAISC"
                                        ] .

    <http://naisc.adaptcentre.ie/person/61>
            a                      prov:Person , schema:Person ;
            prov:actedOnBehalfOf   [ a          prov:Organization ;
                                     foaf:name  "Bibliothèque nationale de France"
                                   ] ;
            foaf:familyName        "10097M" ;
            foaf:givenName         "10097M" ;
            schema:hasOccupation   "Metadata Cataloguer" .

    _:b1    a             prov:Association , prov:Attribution ;
            prov:agent    <http://naisc.adaptcentre.ie/person/61> ;
            prov:hadRole  "Created the interlink" .
}
```

Figure 9: Provenance Data RDF Graph

Future Directions

NAISC-L is currently undergoing a final user-evaluation phase. We invite Information Professionals with an interest in Linked Data to complete this questionnaire which will give you the opportunity to complete a set of interlinking tasks on NAISC-L and to provide us with feedback on your experience.

We are also looking for Information Professionals who would like to trial NAISC-L to create interlinks from their organisation's LD dataset. If you would be interested in trialling NAISC-L, please feel free to contact lucy.mckenna@adaptcentre.ie.

More information on NAISC-L can be found @ https://www.scss.tcd.ie/~mckennl3/naisc/

Questionnaire link @ https://scsstcd.qualtrics.com/jfe/form/SV_cJ9VBQ2BuNbvbcF

Lucy McKenna is in the final year of her PhD in the ADAPT Centre, Trinity College Dublin. Funded by Science Foundation Ireland, ADAPT is a multi-institutional dynamic research centre focused on developing next generation digital technologies. Lucy's research is in the area of Linked Data for libraries, archives and museums. Lucy obtained a Masters in Library and Information Studies from University College Dublin in 2015.

[i] McKenna, L., Debruyne, C., & O'Sullivan, D. (2018). Understanding the Position of Information Professionals with regards to Linked Data: A survey of Libraries, Archives and Museums. In *2018 ACM/IEEE on Joint Conference on Digital Libraries (JCDL)*.

[ii] McKenna, L., Debruyne, C., & O'Sullivan, D. (2019). NAISC: An Authoritative Linked Data Interlinking Approach for the Library Domain. In 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL).

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

[iv] https://www.w3.org/standards/semanticweb/data.html

[v] https://www.w3.org/RDF/

[vi] https://5stardata.info/en/

[vii] https://lod-cloud.net

[viii] http://libris.kb.se

[ix] http://data.bnf.fr

[x] http://datos.bne.es/inicio.html

[xi] http://bnb.data.bl.uk

[xii] https://portal.dnb.de

[xiii] https://www.w3.org/TR/prov-o/

[xiv] IFLA Study Group on the Functional Requirements of Bibliographic Records. 1998. "Functional Requirements of Bibliographic Records: final report." München: K. G. Saur. Available online at http://www.ifla.org/VII/s13/frbr/frbr.pdf.

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