NAISC-L: AN AUTHORITATIVE LINKED DATA INTERLINKING APPROACH FOR THE LIBRARY DOMAIN

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 (LD) [1]. 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) [2]. We also discuss trialling NAISC-L at the Irish Traditional Music Archive.

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 [3]. 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. 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 (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 rating scheme(https://5stardata.info/en/) 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 (https://lod-cloud.net) for some of the leading library LD projects, such as those of the Swedish (LIBRIS), French (BnF), Spanish (BnE), British (BNB) and German (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).

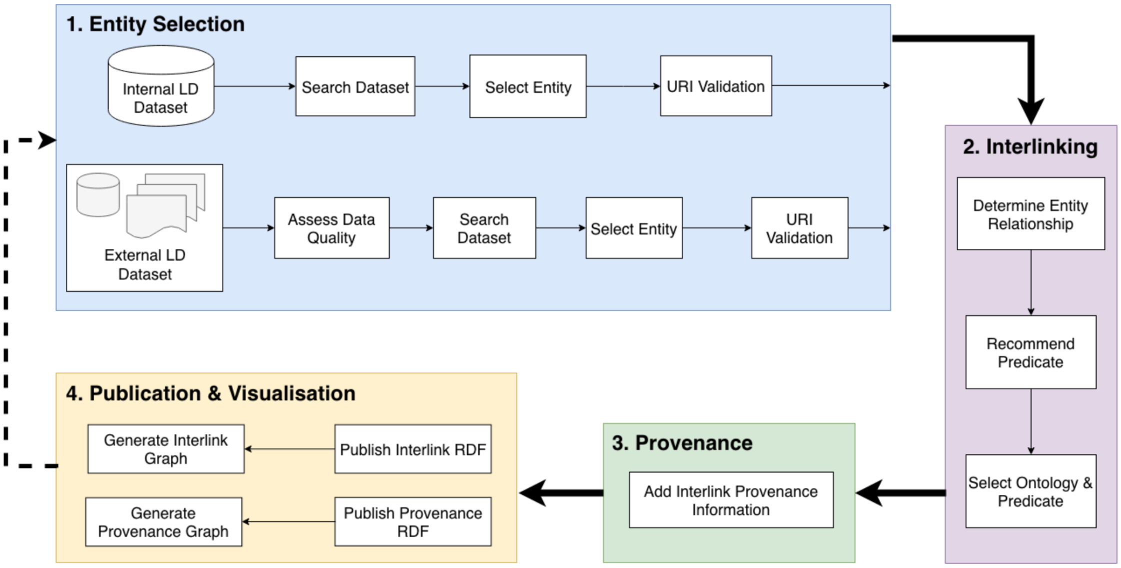


Figure 1 NAISC-L Interlinking Framework (CC BY-SA)

- 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 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 interlinked specific sub-classes and properties. This extension, called NaiscProv (see Figure 2), is used to describe how and why interlinks were created.

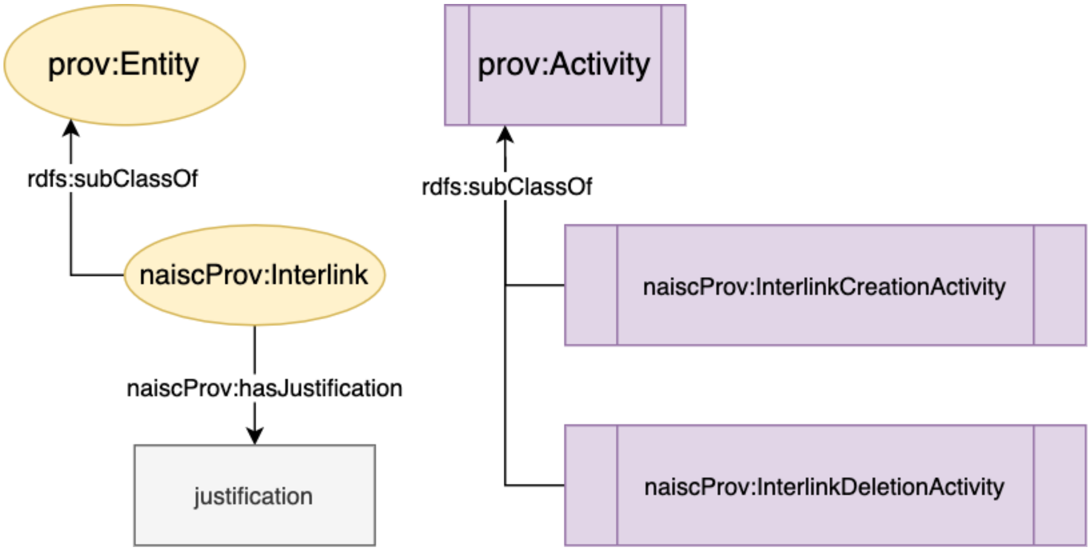
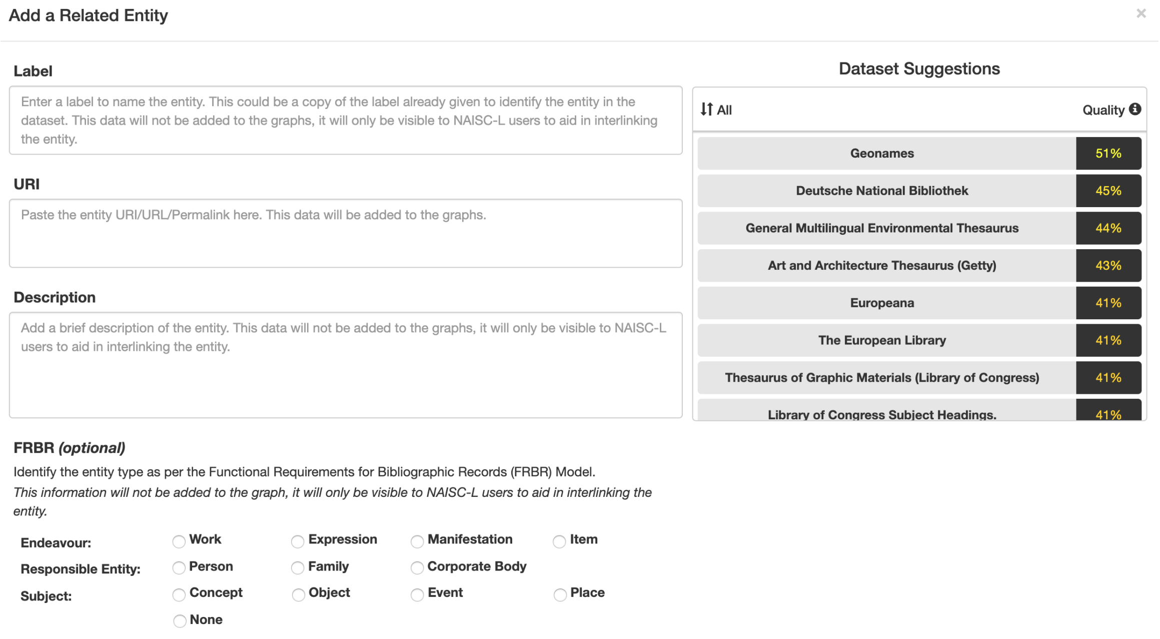

prov:Entity
prov:Activity
rdfs:subClassOf
rdfs:subClassOf
naiscProv:Interlink
naiscProv:InterlinkCreationActivity
naiscProv:hasJustification
justification
naiscProv:InterlinkDeletionActivity

Figure 2: NasicProv PROV-O Extension (CC BY-SA)

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 (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.



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.

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%

FRBR *(optional)*

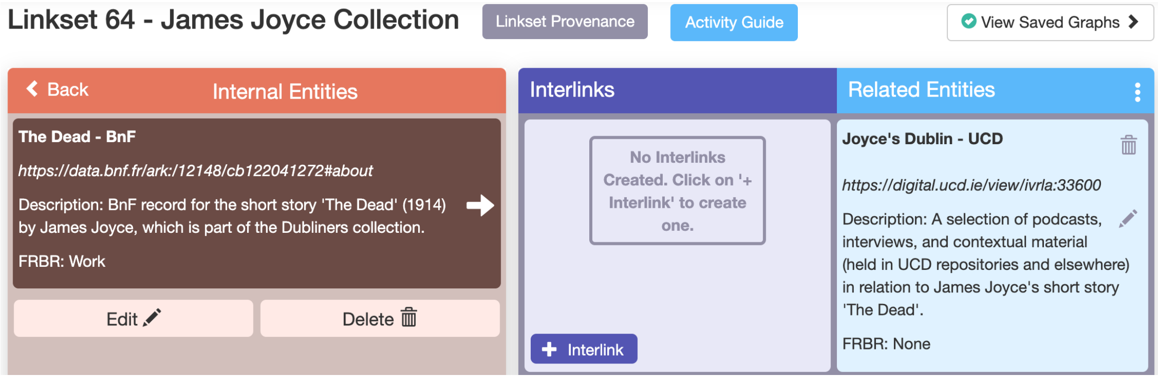
Identify the entity type as per the Functional Requirements for Bibliographic Records (FRBR) Model.

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			

Figure 3: Entity Selection (CC BY-SA)

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 64 - 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.

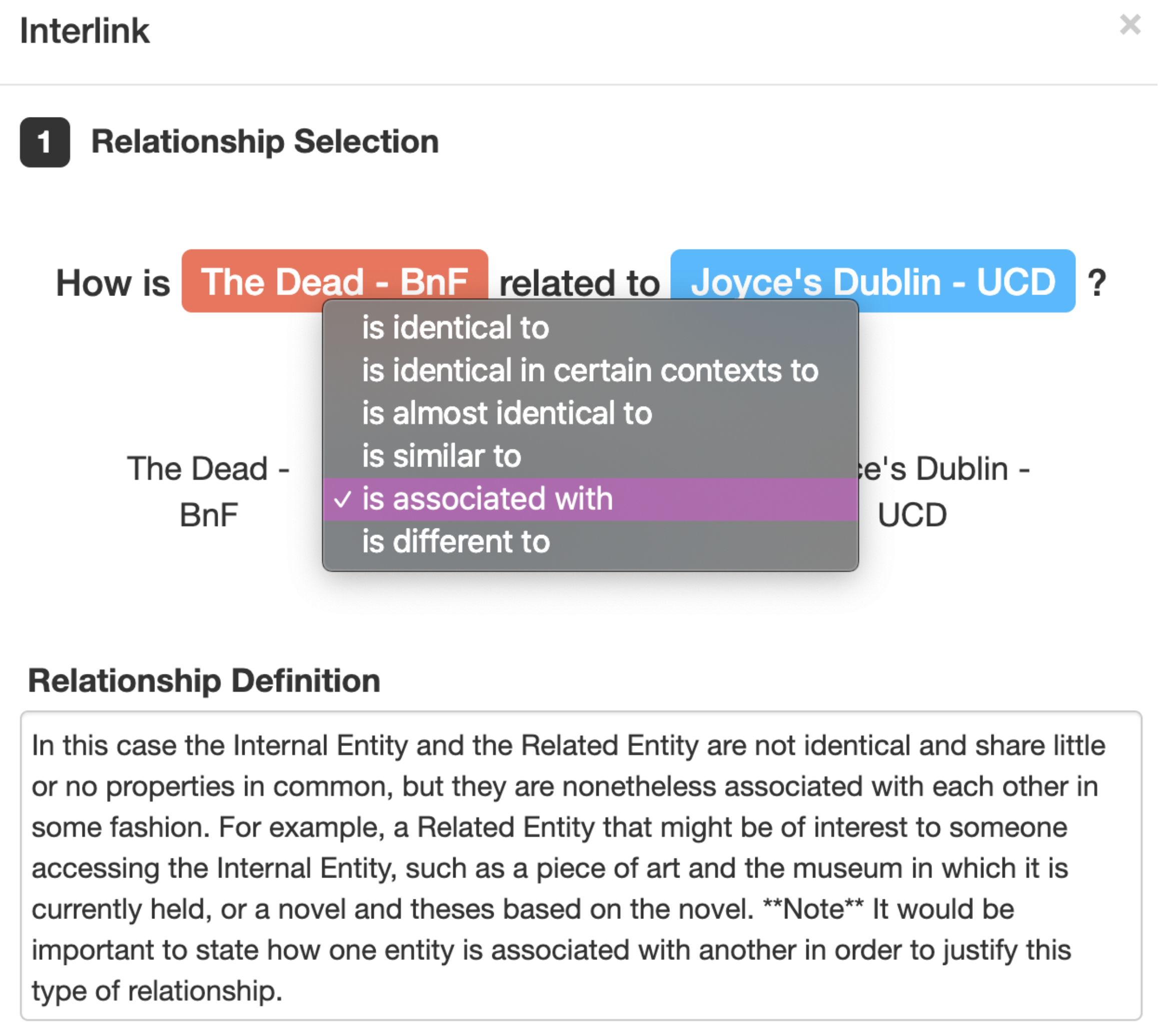
Joyce's Dublin - UCD

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 (CC BY-SA)



Interlink

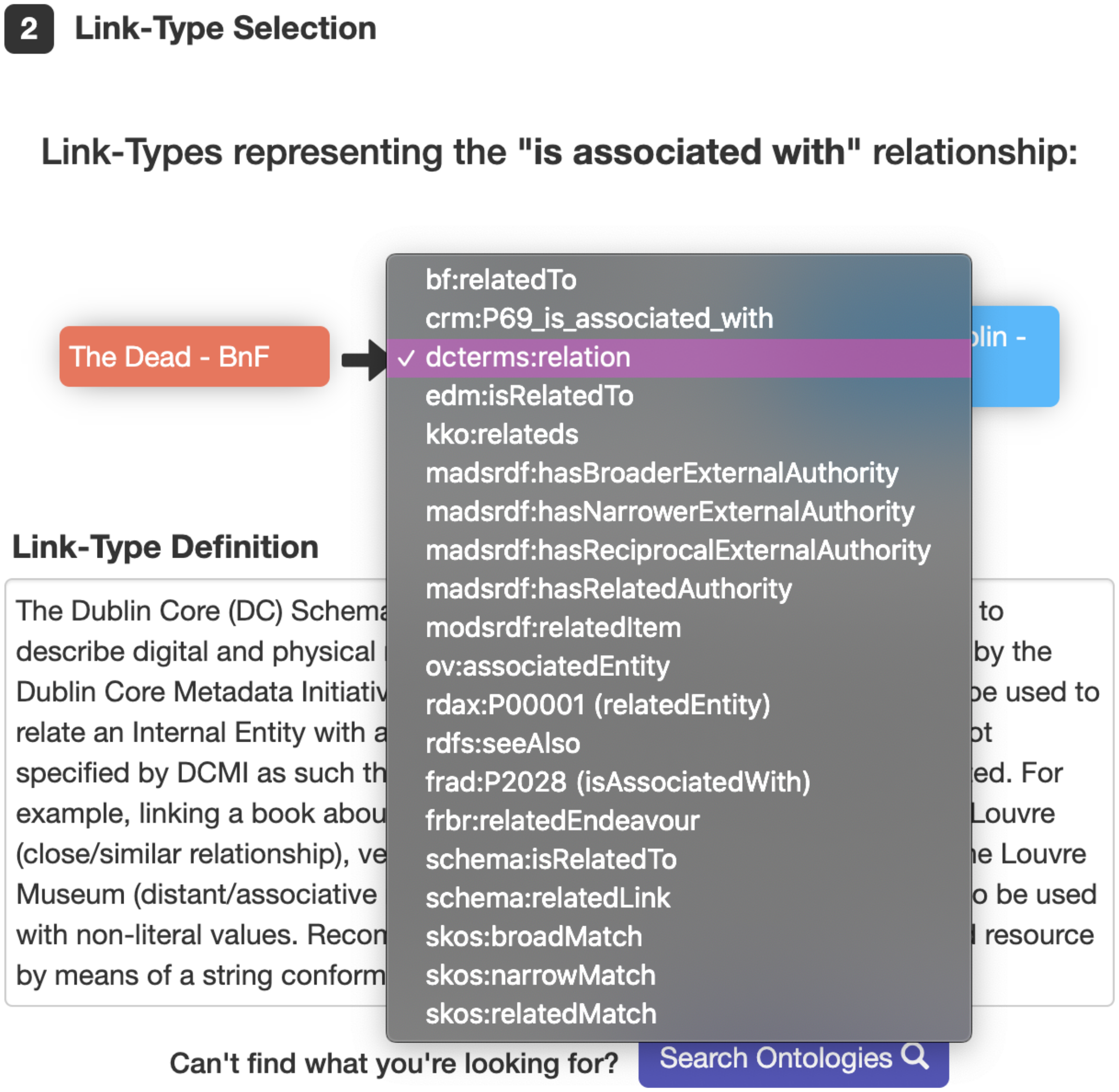
1 Relationship Selection

How is The Dead - BnF related 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.

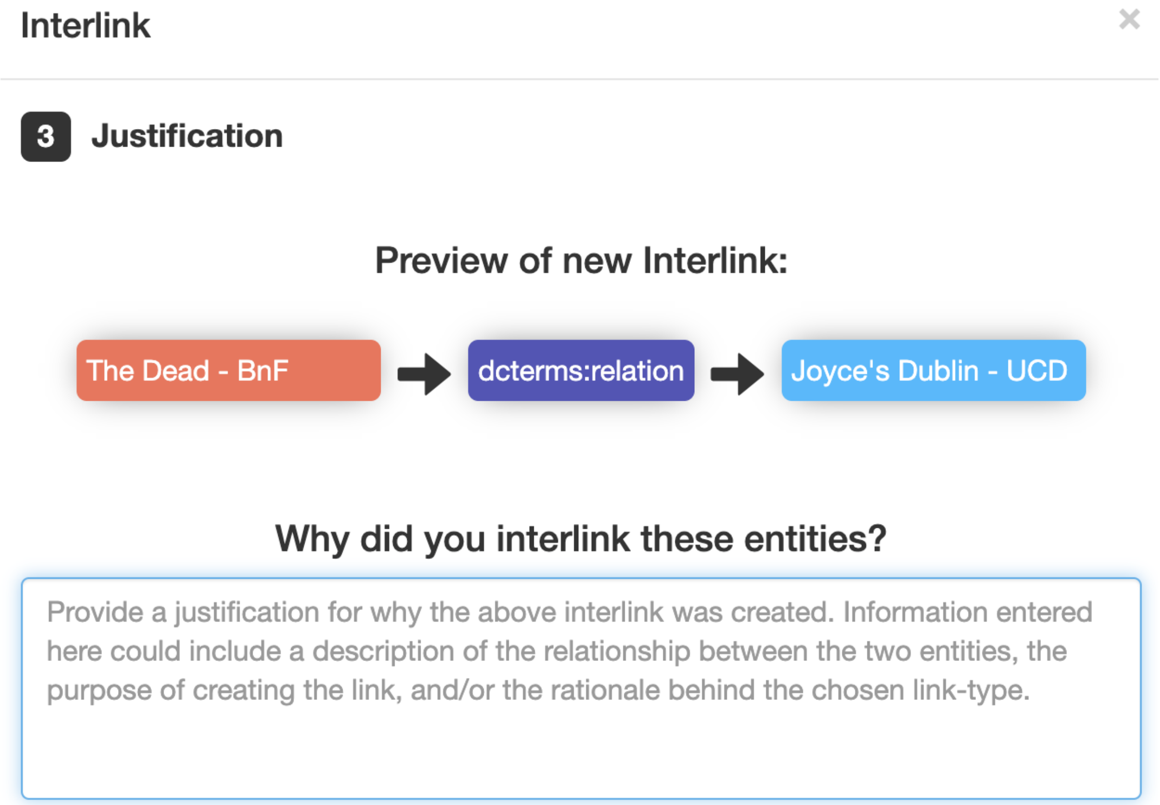
Figure 5: Relationship Type Selection (CC BY-SA)


2 Link-Type Selection
Link-Types representing the "is associated with" relationship:
The Dead - BnF
blin -
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
Can't find what you're looking for?
Search Ontologies

The Dublin Core (DC) Schema … to
describe digital and physical … by the
Dublin Core Metadata Initiativ… be used to
relate an Internal Entity with a … ot
specified by DCMI as such th… ed. For
example, linking a book abou… Louvre
(close/similar relationship), ve… e Louvre
Museum (distant/associative … o be used
with non-literal values. Recon… resource
by means of a string conform

Figure 6: Property Selection (CC BY-SA)

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:

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 (CC BY-SA)

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.

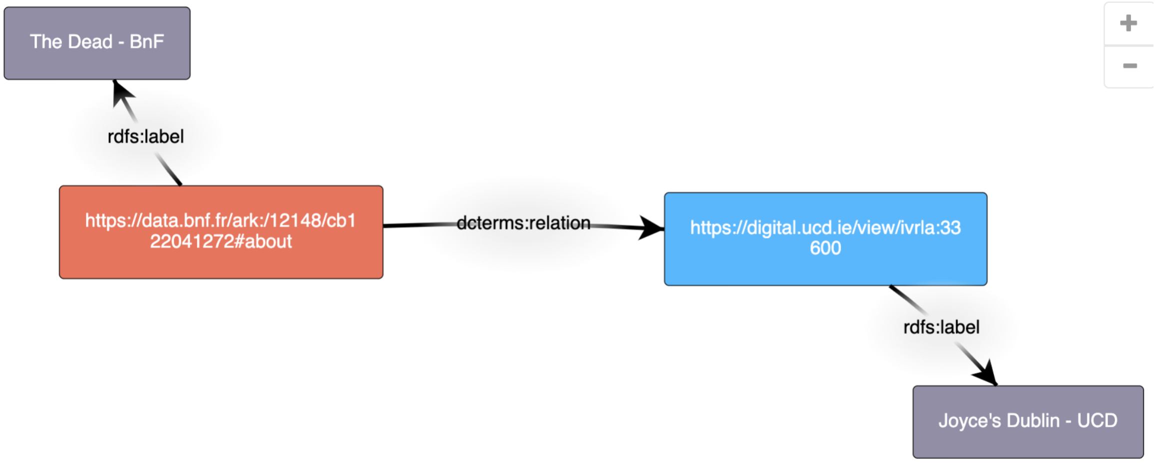

The Dead - BnF
rdfs:label
https://data.bnf.fr/ark:/12148/cb122041272#about
dcterms:relation
https://digital.ucd.ie/view/ivrla:33600
rdfs:label
Joyce's Dublin - UCD
+
−

Figure 8: Interlink Graph Visualisation (created using GoJS) (CC BY-SA)

Using RDF Turtle syntax, Figure 9 below demonstrates the provenance of the interlink (*) 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)

Figure 9: Provenance Data RDF Graph

Trialling NAISC-L at the Irish Traditional Music Archive

The Irish Traditional Music Archive (ITMA) holds a vast collection of materials relating to Irish traditional music, songs and dance. ITMA was recently involved in the LITMUS (Linked Irish Traditional Music) project which focused on the development of the first LD framework tailored to the needs of Irish traditional song, instrumental music, and dance. The project included the development of the LITMUS ontology to represent contemporary and historical Irish traditional music practice, documentation and performance, as well as the LD pilot project. This project involved using 20 years of TG4 Gradam Ceoil (Irish music awards) performance data in order to create a LD dataset which demonstrated the use of the LITMUS ontology and vocabularies.

Over one working week, three Information Professionals (IPs) at ITMA used NAISC-L for a short period each day in order to create a set of 30 interlinks. These interlinks created connections from some of the people/groups mentioned in TG4 Gradam Ceoil LD dataset to related entities in the Virtual International Authority File (VIAF), an OCLC-hosted name authority service. The aim of these interlinks was to provide authoritative information on specific individuals or groups, as well as to create links to other libraries which contributed to the VIAF record for these entities. Such links could be used to guide ITMA patrons to related resources held in other institutions and, in turn, direct VIAF users to data held at ITMA. The IPs who trialled NAISC-L found the framework and interlinking process to be engaging, functional and useful for their needs. The IPs were able to use NAISC-L as part of their cataloguing workflow and, even though the IPs had little to no prior experience with LD, all were able to successfully create interlinks.

Future Directions

During the NAISC-L trial at ITMA, some minor suggestions to improve the GUI were made by the IPs. Using this feedback, the GUI will be updated and, once complete, NAISC-L will be made available as an open access interlinking tool for libraries.

A demo of NAISC-L can be found @ https://www.scss.tcd.ie/~mckennl3/naisc/

References

[1] 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).

[2]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).

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