Engaging Information Professionals in the process of Authoritative Linked Data Interlinking

2018-11-28 at SWIB 2018

Lucy McKenna, Christophe Debruyne & Declan O’Sullivan
ADAPT Centre, Trinity College Dublin, Ireland
Library LD projects

- Increase in uptake & number of libraries implementing LD
- Mostly large institutions/organisations
  - Access to financial & technical resources
- Few implementations use multiple datasets
  - Often single institution initiatives
  - Limited interlinking across datasets
  - Mostly linked to large authorities/controlled vocabularies

LD Survey for Information Professionals

Aims:

1. Explore Information Professionals’ (IPs) knowledge & use of LD
2. Explore the challenges that IPs experience with LD
3. Explore how to overcome these challenges

• Online questionnaire - 50 Questions

• 185 participants
  o Primary Information Professionals from library domain
  o Majority had prior knowledge of the SW (84%) & LD (90%)

McKenna et al. (2018)
Key Findings - Experience with LD

Benefits

- Improved data discoverability & accessibility
- Cross institutional linking & integration – additional context for data interpretation
- Enriched metadata & improved authority control

Challenges

- **Resource Issues:** Dataset/provenance availability & quality, lack of guidelines & use-cases, funding & training, URIs
- **LD Tooling:** Usability issues, unsuitable for needs of LAMs, immature software, technological complexity & learnability
- **Interlinking & Integration:** Ontology & link-type selection, data reconciliation, vocabulary mapping
Key Findings - Potential Solution

• 89% rated LD Tooling specifically designed for IPs as **useful**
  • Reduce technical knowledge gap
  • Encourage increase of LD use in LAMs

• **Requirements**
  • Attuned and adaptable to LAM workflows
  • Hide LD technicalities
  • Aware of common LAM data sources & data quality

• **Importance Measure of Data Quality Criteria**
  • Trustworthiness (66%), Interoperability (51%), Licensing (49%), Completeness (41%), Understandability (40%), Provenance (39%), Timeliness (38%)
1. **Interlinking**
   - Limited interlinking across datasets & institutions
   - Area of particular difficulty in survey & literature
   - Limited guidelines on interlinking library resources

2. **Provenance**
   - Limited guidelines on LD provenance for LAMs
   - Adds to the authority & trustworthiness of LD

3. **LD tooling**
   - Usability issues – mostly designed for technical/LD experts
   - Often not suitable for library workflows or requirements

4. **Library Domain**
   - Majority of survey participants, Data access
Research Question

How can information professionals be facilitated to engage with the process of authoritative linked data interlinking with greater efficacy, ease, and efficiency?

What is Authoritative Interlinking?

- Interlinking – creating a link between two LD resources
- Authoritative - known to be reliable & trustworthy
  - LAMs are an authoritative source of information
  - Provision of provenance data
  - Quality of resources being interlinked

Why Information Professionals?

- Experts in metadata creation, knowledge discovery & authority control
Current Interlinking Frameworks

• RDF Refine, SILK, LIMES, MARiMbA, Catalogue Bridge
  o Majority require a technical knowledge of LD
  o Primarily support owl:sameAs links
  o RDF Refine & MARiMbA
    o Aimed at library domain
    o Access to large-scale datasets e.g. VIAF, LCSH

• Further Requirements
  o Additional link types e.g. dct:relation, schema:isPartOf
  o Interlink with datasets emerging from smaller authoritative institutions
  o Remove need for expert technical/LD knowledge
Research Aims

Develop an **authoritative interlinking framework** specifically designed with the workflows and expertise of IPs in the library domain in mind.

Develop a **provenance model** that expresses the required provenance of interlinks created by IPs.

Design an **interlinking interface** for IPs that guides users through the interlinking process including ontology and link type selection, and provenance data generation.
Methodology

Phase 1
- Concept
- Refine
- User Evaluation

Phase 2
- Prototype
- Mock-Up
- User Evaluation

Phase 3
- Refine
- Deploy
- Use-Case Testing
NAISC – Novel Authoritative Interlinking of Schema & Concepts

NAISC

Identity Resolution Module

Interlinking + Interlink Provenance

Publication Layer

LD Wrapper

Database A
Library, Archive, Museum Metadata

LD Wrapper

Database B
Controlled Vocabulary & Authority Metadata

LD Wrapper

RDF/XML

SparQL Endpoint or

SPARQL

Application Layer

Application Code
NAISC – Resource Selection

Search Internal RDF Dataset using Semantic Faceted Search Tool, SPARQL Endpoint or Web Resource

Collection: 19th Century Social History Pamphlets

Open SemFacet

Primary Resources

<table>
<thead>
<tr>
<th>ID: 3.3</th>
<th>URI:  <a href="http://digital.ucd.ie/data/ivrla:45153">http://digital.ucd.ie/data/ivrla:45153</a></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Description: Letter addressed to His Grace the Archbishop of Cashel, on the subject of a charge purported to have been delivered at Killaloe and Limerick, on the 20th and 22d of June, ult.</td>
</tr>
</tbody>
</table>

Enter & Validate Resource URI

Primary Resource
# NAISC – Resource Selection

## Authorities, Thesauri & Controlled Vocabularies

Use the links below to search an authority for a Secondary Resource URI:

<table>
<thead>
<tr>
<th>Controlled Names</th>
<th>Titles</th>
<th>Subject</th>
<th>Genre</th>
<th>Place</th>
</tr>
</thead>
<tbody>
<tr>
<td>LC/NAF</td>
<td>LC/NAF</td>
<td>AAT</td>
<td>AAT</td>
<td>GeoNames</td>
</tr>
<tr>
<td>ULAN</td>
<td>WorldCat</td>
<td>FAST</td>
<td>TGM</td>
<td>TGN</td>
</tr>
<tr>
<td>ORCID</td>
<td>BNB</td>
<td>LCSH</td>
<td>LCGFT</td>
<td></td>
</tr>
<tr>
<td>VIAF</td>
<td>Europeana</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Search Authoritative External Datasets for Related Resource**

**Plan to provide data quality information for common resources**

## Secondary Resources

**ID:** 3.3.3  **URI:** [http://digital.ucd.ie/data/ivrla:45000](http://digital.ucd.ie/data/ivrla:45000)

**Description:** 19th Century Social History Pamphlets Collection

**Enter & Validate URI for a Related Resource**
Plan to develop a Predicate Recommender that would suggest suitable predicates based on resource & relationship description.

Select Predicate that describes the relationship between the resources.

Primary URI: http://digital.ucd.ie/data/ivrla:45153

Description: Letter addressed to His Grace the Archbishop of Cashel, on the subject of a charge purported to have been delivered at Killaloe and Limerick, on the 20th and 22d of June, ult.

Secondary URI: http://digital.ucd.ie/data/ivrla:45000

Description: 19th Century Social History Pamphlets Collection

Interlink

Select Ontology

OWL

SKOS

Dublin Core

isPartOf

Description:
Letter is part of the pamphlets collection
<table>
<thead>
<tr>
<th>Provenance of Interlinks</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Who created the interlink?</td>
<td>What dataset does the interlink point to?</td>
</tr>
<tr>
<td>How was the interlink created?</td>
<td>Where can the dataset be accessed?</td>
</tr>
<tr>
<td>Why was the interlink created?</td>
<td>What resources are interlinked?</td>
</tr>
<tr>
<td>Where was the interlink created?</td>
<td>What is the relationship between the resources?</td>
</tr>
<tr>
<td>When was the interlink created?</td>
<td>Why was this predicate selected?</td>
</tr>
<tr>
<td>What dataset is the interlink part of?</td>
<td>When was the interlink last modified?</td>
</tr>
<tr>
<td>Who published the dataset?</td>
<td>Who modified the interlink?</td>
</tr>
<tr>
<td>Where can the dataset be accessed?</td>
<td>Why was it modified?</td>
</tr>
<tr>
<td><strong>Provenance of Provenance</strong></td>
<td>[ ]</td>
</tr>
<tr>
<td>When was the provenance data generated?</td>
<td>Who generated the provenance data?</td>
</tr>
</tbody>
</table>
NAISC – Provenance Model

Used 3 graphs:

1. **Interlink Graph**: A Named Graph for a set of interlinks

2. **Provenance Graph**: A prov:Bundle containing a set of provenance descriptions for a set of interlinks

3. **Relationship Graph**: A graph that represents the relationship between an Interlink Graph and a Provenance Graph.

• As a Prov Bundle is an entity we can describe the provenance of the interlink provenance data contained in the bundle.
NAISC Provenance Model

Interlink Graphs

Interlink Graphs

Relationship Graph

<relations>

Provenance Graphs
NAISC – Provenance Ontologies

- Used the Prov Ontology
  - Describe who, where, and when interlinks were created, modified or deleted.
- Extended ontology - NAISCProv to describe what, how, and why interlinks created.
- Added interlink specific sub-classes & properties e.g. naiscProv:Interlink, naiscProv:hasJustification.

- Used Void Ontology for dataset description e.g. void:Dataset, void:sparqlEndpoint, void:dataDump.

- Used Dublin Core & FOAF to further describe entities e.g. dct:title, dct:description, foaf:name, foaf:givenName.
@prefix rr: <http://www.w3.org/ns/r2rml#> .
@prefix rrf: <http://kdeg.scss.tcd.ie/ns/rrf#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .

<#TriplesMap1>
  rr:logicalTable [  
    rr:sqlQuery """"select p.primaryurl, s.secondaryurl, concat(pr.ontologyURI, pr.predicate) predicate from dataset d
    join primaryresource p on d.id = p.dataset_id
    join link l on l.primaryresource_id = p.id
    join secondaryresource s on s.link_id = l.id
    join predicate pr on pr.link_id = l.id
    where d.id = {DATASET_ID} and pr.linkstatus != 'deleted'  """
  ];

  rr:subjectMap [  
    rr:column "primaryurl";
  ];

  rr:predicateObjectMap [  
    rr:predicateMap [  
      rr:column "predicate";
      rr:termType rr:IRI;
    ];

    rr:objectMap [  
      rr:column "secondaryurl";
      rr:termType rr:IRI;
    ];
  ];

].
Future Directions

- Usability Testing of framework, provenance model & interface
- Modify based on feedback
- Addition of Dataset Quality Criteria Scores & Predicate Recommender
- Further Testing
Thank you!

Any Questions?
References

NAISC - Linked Data Application Framework

Application Layer

Application Code

Data Access, Integration & Storage Layer

Web Data Access Module → Vocabulary Mapping Module

SPARQL Endpoint or RDF API

Publishing

Integrated Web Data

Web of Data

Publication Layer

LD Wrapper

Database A: Library, Archive, Museum Metadata

LD Wrapper

Database B: Controlled Vocabulary & Authority Metadata

LD Wrapper

RDF/XML

Other Content

LD Wrapper

HTTP
NAISC – Provenance Ontologies

- Used the **Prov Ontology**
  - Describe who, where, and when interlinks were created, modified or deleted
  - Extended ontology - **NaiscProv** – to describe what, how, and why interlinks created
    - Added interlink specific sub-classes & properties e.g.
      - `naiscProv:Interlink, nasicProv:hasJustification`

- Used **Void Ontology** for dataset description e.g. `void:Dataset, void:sparqlEndpoint, void:dataDump`

- Used **Dublin Core & FOAF** to further describe entities e.g.
  - `dct:title, dct:description, foaf:name, foaf:givenName`
NAISC – Provenance Ontologies
@prefix rr: <http://www.w3.org/ns/r2rml#> .
@prefix rrf: <http://kdeg.scss.tcd.ie/ns/rrf#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .

<#TriplesMap1>
  rr:logicalTable [  
    rr:sqLiteQuery """"select p.primaryurl, s.secondaryurl, concat(pr.ontologyURI, pr.predicate) predicate from dataset d  
    join primaryresource p on d.id = p.dataset_id  
    join link l on l.primaryresource_id = p.id  
    join secondaryresource s on s.link_id = l.id  
    join predicate pr on pr.link_id = l.id  
    where d.id = {DATASET_ID} and pr.linkstatus != 'deleted'  
    """"';
  ];

  rr:subjectMap [  
    rr:column "primaryurl";
  ];

  rr:predicateObjectMap [  
    rr:predicateMap [  
      rr:column "predicate";
      rr:termType rr:IRI;
    ];
  ];

  rr:objectMap [  
    rr:column "secondaryurl";
    rr:termType rr:IRI;
  ];
];
RDF Output

<http://digital.ucd.ie/data/ivrla:45153>
  <http://purl.org/dc/terms/isPartOf>
Naisc – Provenance Model

Justification for interlinking

Reason for interlinking \( \rightarrow \) naisProv:hasJustification

Type

naiscProv:Interlink \( \rightarrow \) rdf:type

prov:startedAtTime

prov:endedAtTime

prov:associatedWith

naiscProv:InterlinkCreationActivity

Justification for interlinking

Uses Reification to describe each statement/interlink

 RDF Statement

rdfs:subject

rdfs:predicate

Predicate

SR_3

dct:isPartOf

SecondaryDataset

Specify the type of activity – Creation, Deletion & Modification

Naisc – Provenance Model

www.adaptcentre.ie
Named Graph \[\text{prov:hasProvenance}\] Prov Bundle