Towards an Open Infrastructure for Supporting Relationship Discovery Between Scholarly Assets

Chris Munro
chris.munro@manchester.ac.uk
6th December 2016
www.herc.ac.uk & www.farrinstitute.org

Joint work with:
Phil Couch (Manchester), Jon Johnson (UCL)
In August 2012, ten UK funding agencies awarded four Centres of Excellence in e-health informatics research. The four HIRCs aim to optimize the use of health records in research and address the UK’s capacity building requirements to support a sustainable health informatics research base. Manchester University hosts the Northern HIRC called the Health e-Research Centre (HeRC). HeRC involves collaborations at the Universities of Lancaster, Liverpool and York, and NHS partners across North England.
In 2013, the Farr Institute was created to support the HIRCs’ collective work.

- Farr Institute @ CIPHER
- Farr Institute @ HeRC
- Farr Institute @ Scotland
- Farr Institute @ UCL Partners

Together, they bring a total of 21 academic institutions and two MRC units.
What do we have?

• Tsunami of data, generated/leveraged by social/biomedical sciences
• Current model of discovery centred around large investments – e.g. Research Data Portals
• Difficult for *researchers* and *data investors* (those collecting/enabling the collection of research-ready data) to **discover** the relationships between these scholarly assets
• Many (quite different) metadata models in use, compared to e.g. digital manuscript field
What would we like?

• Open and shared specification of the aggregated assets of scholarly work
• Build on existing portals/systems/resources
• Infrastructure to make the aggregations available, with redundancy
• Allow assets to be co-produced across multiple organizations – they may not share source data (e.g. due to governance)
• Infrastructure available to all
Example use cases

1. A PI wants to understand what data from the study has been used in published research and whether the data produced is being under-utilised.

2. A researcher has obtained data from a Research Data Portal and generates some derived variables and wants to share what it is and how it was generated, and provide a citable link for publication.
Aggregation: Research Objects

Enabling reproducible, transparent research.

Scientific hypothesis

- PUBLICATIONS
- DATA
- RESULTS
- WORKFLOWS
- SLIDES
- METADATA
- LOGS

RDF

Linked Open Data
Executable
Discoverable
Reproducible

www.researchobject.org
Components for prototype

• Research Objects-aggregate assets
• OrientDB
  – distributed graph database
  – Inbuilt front-end for browsing/editing graph database
• Scenarios, (example uses)
Find research object aggregating the ALSPAC study DDI-URN from CLOSER portal:
http://discovery.closer.ac.uk/item/uk.cls.ALSPAC/0d8a7220-c61b-4542-967d-a40cb5aca430
<table>
<thead>
<tr>
<th>Field</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>label</td>
<td>ALSPAC derived data research object</td>
</tr>
<tr>
<td>do-description</td>
<td>ALSPAC derived data research object</td>
</tr>
<tr>
<td>do-identifier</td>
<td>&lt;DOI for the Research object&gt;</td>
</tr>
<tr>
<td>cd-creator</td>
<td>chris munro</td>
</tr>
<tr>
<td>type</td>
<td>ore.Aggregation</td>
</tr>
</tbody>
</table>
Infrastructure

- CLOSER platform
  - ROs aggregating CLOSER content
  - Example study:
    - Questions
    - Studies
    - Sweeps
    - Instruments

- e-Lab platform
  - ROs aggregating e-Lab content
  - Example study:
    - Results
    - Script
    - Data

- Distributed graph database
  - RO aggregating study assets in multiple platforms
  - Web interface:
    - Create ROs
    - Browse ROs
    - Search for ROs
  - Search/Create ROs
  - Online journal
    - Publication

- Search/Create ROs
  - Search/Create ROs
  - Web interface
    - Create ROs
    - Browse ROs
    - Search for ROs

- Studies
- Instruments
- Distributed graph database
- ROs aggregating study assets in multiple platforms
- Example study:
  - Results
  - Script
  - Data
Conclusion

- Bottom up approach - builds on existing resources
- Covers starting points other than the final dataset-below the catalogue level
- Not institution based
- Users can aggregate related resources together, doesn't just harvest
- Supports assets that are stored in different platforms/locations e.g. publication on analysis of pooled datasets, instead of a ‘one platform’ approach