Open Repositories and Interoperability Challenges in UK

Peter Burnhill
Director, EDINA National Data Centre, University of Edinburgh, Scotland UK

DL.org Workshop, The British Academy, London UK, 4 February 2011

Digital Libraries and Open Access: Interoperability Strategies

autobiography as commentary on data facilities

Began as a data manufacturer
- **Scottish Education Data Archive**, late 1970s – mid ’80s
  - Survey statistician: for school leaver, YTS & 16-19 cohort surveys
  - curated as databases: the working capital for research group + ‘guests’

Became a data broker
- **Edinburgh University Data Library**, mid- 1980s & on
  - Providing library and ease of access to data held elsewhere
  - Connected to IASSIST, international group for data librarians/archivists

Learning about interesting spaces [time/place referencing]
- **ESRC Regional Research Laboratory for Scotland** 1986/90
  - Co-director: early days of Geographical Information Systems (GIS)
  - * Inter-Agency Committee on Global Environmental Change Data Task Force

Moved into national data services; learnt more about data curation
- **EDINA national data centre**, mid-1990s to present
  - Director: set-up and continuing development of access/deposit services

- **Digital Curation Centre**, 2004 & 2005
  - Director: set-up & digital curation = ‘data curation’ & ‘digital preservation’

Spot the repository!
Researchers’ viewpoint: a cultural shift?

“You are not finished until you have done the research, published the results, and published the data, receiving formal credit for everything.”

Mark A. Parsons (2006) International Polar Year

Preserve or Perish

“A scholar’s positive contribution is measured by the sum of the original data that he contributes. Hypotheses come and go but data remain.”

in Advice to a Young Investigator (1897) Santiago Ramón y Cajal (Nobel Prize winner, 1906)

Emergence of Digital Library: Information Science

2 traditions/mentalities co-exist in Information Science

1. Document tradition: signifying record-ness
   - non-convergent mentalities working to build the ‘digital library’
     a) modernisation of library services
     b) infrastructure to access complex databases

2. Computational tradition: various uses of formal techniques
   - *non-convergent mentalities working to build the ‘digital library’
     a) modernisation of library services
     b) infrastructure to access complex databases

Emergence of Digital Library: Information Science

“considerable simplification, … helpful to think … of two traditions, or mentalities, even cultures, co-exist in area of Information Science

1. “Approaches based on a concern with documents, with signifying records: archives, bibliography, documentation, librarianship, records management, and the like

2. “approaches based on uses for formal techniques, whether mechanical (such as punch cards and data-processing equipment) or mathematical (as in algorithmic procedures).”

Michael Buckland (UC Berkeley), Presidential Address, American Society for Information Science, JASIS’s 50th (1998)

Semantics of ‘Open Repositories & Interoperability’

• R is for Repository
  - “university-based institutional repository is a set of services … for the management and dissemination of digital materials created by the institution & its community members. … organizational commitment to the stewardship of these digital materials, including long-term preservation where appropriate, as well as organization and access …” (C. Lynch, 2003)
  - Digital repository differs from other digital collections in that:
    * "content is deposited, whether by content creator, owner or third party
    * "architecture manages content as well as metadata
    * "repository offers a minimum set of basic services
    * "must be sustainable & trusted, well-supported & well-managed”


• O is for Open
  - OA (for publications) not the only ‘open’ policy:
    * OER: Open Educational Resources
      - Open means ‘not closed’: making teaching & learning materials visible
    * Open Data
      - Datasets tradition (IASSIST); ‘open/privilege access to databases; open.data.gov
    * Open Source Software
      - OSS has its own way of doing things
**Open Repositories & Interoperability** [in the UK]

*Heretical and Haphazard Thinking About The Brief* ...

- Are Repositories the (only) way to support an Open Agenda? and
- Is Open really what Repositories are for? Or
Is this usage just intended to help us avoid issues of IP and access management?

**And should the focus be on:**

- Interoperability between Repositories?
  or
- Interoperability of Repositories with the wider Internet?

---

**Interoperability Strategies; Interoperability Challenges**

- Whose strategy, and towards what purpose?
  - "within & for the research & education sector"? Or beyond?
  - for the institution, the UK, EU, global anybody; for the researcher?
  - for the machine as user ["Provider/Consumer"]
- Interoperability as technical [& semantic] means to support interworking by persons or systems has challenges
  - policy/technology/infrastructure/management/metadata
  - Internet engineering & semantic web
  * Repositories and Linked Data

**Beyond PUT, KEEP and GET of the singular Repository**

- Connecting Repositories with
  - NOTIFY & EXCHANGE; TRANSFER of objects or metadata only
    * So that the content of the object (or its metadata) may be re-used
- Linking & Identifiers Really Matter!
  - and Registries have a key role as authorities and cross-walks

---

**Maybe can we agree our shared & central task...**

- to ensure ease & continuity of access to (online/digital) scholarly resources
  * for researchers, students and their teachers, now and into the future

**My perspective**

- In a University-based organisation (EDINA) that is a provider of content services & infrastructure services within national (UK) policy framework
  - number of content services based on use of repository software
    * Eprints, DSpace, IntraLibrary
  - number of infrastructure activities
    * OpenDepot.org, OA Repository Junction; OpenURL Router
    - member of SONEX and indirectly of COAR and UK-CORR
- Later focus on on repository-related progress in the UK; where is the value, how this is assessed/expressed?
  - Size of investment in recent times; cost-effectiveness and 'impact'
  - Effort at institutional & inter/national level and the 'shared services' agenda?

---

**Nostalgia for Days of Plenty as we worry about the future**

**JISC as well-funded agent for change**

- JISC Repositories & Preservation Programme - April 2006; March 2009
  "£14m investment in H.E. repository and digital content infrastructure"
- A drive to assist institutions, including JISC RepositoryNet?
  * Repository Support Project; Repository Research Project
  * Intute Repository Search; 'interim repository' | the Depot | OpenDepot.org
- Services/Tools like OpenDOAR; Romeo; SWORD; OARJ
- Check the JISC website
  * http://www.jisc.ac.uk/whatwedo/programmes/inf11/sue2.aspx
  * under the heading of 'key digital repository activities' are 21 funding programmes and 226 funded projects.

& then there were many meetings, including a new 'regular' street event:

- RepoFringe2010: Repository Fringe 2/3 September, Edinburgh
SONEX

- Began with focus on 'deposit opportunities'
  - categorisation of repository types into which authors deposit

- Began with research paper use case:
  multi-author & multi-institutional

- Looking at onward interoperability (SWORD)
  not just technical interoperability but workflow

Day Job: as provider of services/tools & user of software

- EDINA and Edinburgh University Data Library run repositories, with and without JISC
  - Jorum: for learning materials [with Mimas]
  - OpenDepot (the Depot): for research papers
  - ShareGeo: for geo-spatial data
  - DataShare: for research data (institutional, U of Ed)
  - OA Repository Junction as shared service tool
    - using own code and Eprints as an 'escrow' repository during the transfer process.

- All seek to be 'standards-based', reducing need to be mediated

DSpace as open repository software

- Open Architecture: supports flexibility
- Active international development community
- Developer Community provides software add-on and pool of experience
  e.g. creating de-referenceable URIs (RDF/XML) for all in metadata store, to support both registry function and also engage with semantic web

- Implications for Jorum: self-deposit of learning materials
  - Particular requirements, as per Jorum Roadmap, based upon:
    - mediated ingest for multiple items (RSS for metadata; OAI-PMH for objects in IntraLibrary)
    - unmediated deposit (Selenium & SWORD/OARJ) in test-bed development

- cost of ownership, as per JISC OSS Watch:
  "...important that when procuring open source software solutions you also plan to properly resource collaboration work ..."
  - need to embed developments back into DSpace codebase:
    - enables others to use, maintain and develop the new features
    - ensures that your extensions are in main code
    - increases the return on investment by making them available to all
[HCI] User Interface for Unmediated Re-direction/Routing + option for Unmediated Deposit

DataShare as institution’s data repository

Theo Andrew & Ian Stuart (EDINA)

http://oarepojunction.wordpress.com/
task for the Broker is to accept an item for deposit, package it and transfer it

Junction API service-quality, documented at http://oarepojunction.wordpress.com/junction-api/

Junction is a deduction tool via database of repositories takes a deposit object and extracts location information from object to deduce a list of potential targets.

OARJ Project activity: Organisations that hold content for deposit in IRs:

b) As Demonstrator (working with IRs in 7 universities)

1. Set up daily XML export of new records added to UKPMC to OARJ Broker.

1. Earlier work with Nature Pub Gp
   - NPG supply author-submitted manuscripts from journals (with embargo information)
   - OARJ Broker transfer to authors’ institutional repository

Est. Number Of Articles For Transfer During Six Month Period

| Institutional partner | Articles in all NPG journals | Articles in participating NPG journals | Assuming author participation rate of 50%
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cambridge</td>
<td>1429</td>
<td>476</td>
<td>237</td>
</tr>
<tr>
<td>Oxford</td>
<td>962</td>
<td>321</td>
<td>160</td>
</tr>
<tr>
<td>Univ M**</td>
<td>499</td>
<td>166</td>
<td>83</td>
</tr>
<tr>
<td>Univ C**</td>
<td>275</td>
<td>92</td>
<td>46</td>
</tr>
<tr>
<td>Univ Y**</td>
<td>248</td>
<td>83</td>
<td>41</td>
</tr>
<tr>
<td>Edinburgh</td>
<td>194</td>
<td>65</td>
<td>32</td>
</tr>
<tr>
<td>Auckland</td>
<td>53</td>
<td>17</td>
<td>8</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>3660</strong></td>
<td><strong>1220</strong></td>
<td><strong>607</strong></td>
</tr>
</tbody>
</table>

based upon the number of papers published in journals in the NPG portfolio during Jan - June 2010, as recorded in PubMed Central and ISI Web of Knowledge. **Still to be confirmed as a participating institutions
## RSP (Nottingham) Repository Software Survey, Nov. 2010

<table>
<thead>
<tr>
<th>Registered in OpenDOAR</th>
<th>OAI-PMH</th>
<th>SWORD</th>
<th>SWAP</th>
<th>RDF</th>
<th>OAI-ORE</th>
<th>RSS</th>
<th>Atom</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>501 690</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Islandora</td>
<td>DSpace</td>
<td>SU 3.0</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Fedora</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>SRIJSMW</td>
<td></td>
</tr>
<tr>
<td>EQUELLA Ragos</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>SRIJSMW</td>
<td></td>
</tr>
<tr>
<td>16 Hosted</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>SRIJSMW</td>
<td></td>
</tr>
<tr>
<td>Digital Curation Centre</td>
<td>Y</td>
<td></td>
<td></td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>219.50</td>
<td></td>
</tr>
<tr>
<td>Classical Digital Library</td>
<td>Y</td>
<td></td>
<td></td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>219.50</td>
<td></td>
</tr>
<tr>
<td>Y = Yes; O = Optional; (?) = future/pending; ? = I think it’s there</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Re-stating our shared task, to (re-)include data:**

To ensure ease & continuing access to record of scholarship
- research publications and research data

*Consider at least three types of (research) data:*

**A. Supplementary data**
- multimedia files: part of the published article that presents research argument and conclusions
  * more than linear text, limited tabular and graphical display
  * enhances user experience with various multimedia objects

**B. Research dataset(s) upon which conclusions based**
- check analysis of those data to support statements made

**C. Database(s) from which datasets were assembled**
- for reproducibility (exposure to refutation) and new work via alternative analysis and updates to the database(s)

## Have JISC [programme managers] moved on?

“Dealing with institutional processes now, rather than repository technology. Depending on type of content, the projects would fit much more closely in:
- managing research data programme
- research information programme
- open educational resources programme

as they have much more in common with those projects than they do with each other.”

“repositories have found their core business proposition via the REF and making sure Universities list research outputs to obtain research ratings
- have not succeeded in making the business case that IRs should be doing the job of archiving, a core library platform, or the job of an institutional demonstrator/poster space.

Repositories fit in the ‘University Enterprise Stack’ by virtue of being a system that delivers a business solution to a real financial problem.”

## Citation, then linking

**Citation of database(s) (Type C data)**
- for reproducibility (exposure to refutation)
- to prompt new work via alternative analysis and updates to the database(s)
- to credit those who curate the data needed for scholarship

**Citation of the datasets used (Type B data)**
- verification of analysis, that the figures and conclusions accurately reflect those data

Plus hyperlink to the dataset **from** the published article
... and back again **from** the dataset **to** the published article

+ Links to presentations, blogs, websites, funders etc related to the same research activity and same researcher(s) (Type D data?)
Standards to cite data (A long running saga)

There is no universal standard for citing data and computer files, but ...

  - ISO 690: 1987 Bibliographic references - Content, form and structure

Obtaining the citation at source

1. **CIESIN**
   "Most of our datasets and products contain a suggested citation on the Web site as to where the data was obtained"
   "Whenever possible, we urge you to cite the use of data and web resources in the reference section"

2. **How to Cite Statistics Canada Products:**
   "This guide has been developed for authors, editors, researchers, academics, students, librarians and data librarians. It describes, in three steps, how to build your reference when citing Statistics Canada products"
   - http://www.statcan.gc.ca/pub/12-591-x/12-591-x2006001-eng.htm

Get it from those who make the data available: the data publishers

cf Cataloguing in Publication!

Linked Data ... Is this shared understanding?

A note from Tim Berners Lee now in circulation proposes 4 steps:

1. Use URIs as names for things
2. Use http URIs so that people [& computers?] can look up those names
3. When someone looks up a URI, provide useful information using the standards (RDF, SPARQL)
4. Include links to other URIs, so that they can discover more things.

- may become the principles/rules/definition of ‘Linked Data’

To formal publishing into data infrastructure

DISC-UK DataShare Project
(Edinburgh, LSE, Oxford, Southampton)

From informal storage and sharing
DataShare2

[Diagram showing DataShare2 and related concepts]

Time for me to stop

Hoping that I have left some space/place for questions

- Thank you

Acknowledgements
Theo Andrew & Ian Stuart, Pablo de Castro,
Gareth Waller & Robin Rice,
Dave Flanders & Andy McGregor

Ease and Continuity of Access to Data in Difficult Times

End of an era? End of the R word?
1. Embedded in domain-specific processes, but with wider context
   a) Engage, connect and get leverage from Internet Engineering
      * W3C and the commercial/retail world
   b) Linkage and Identifiers really, really matter in m2m world

2. Moving from technology to policy & practice
   some domain-specific, some common to repositories
   a) Collection management: active curation & Linked relationships
      * versions, of data | articles | learning materials
      * Collections, ‘see also’ Curation as value-added linkage between items
   b) First point of public issue (availability); Take-down regimes

3. Institutional stewardship responsibility
   - for content that ‘we need’ for research and education
   - including data and other materials manufactured from within ‘our world’
     * born-digital (and digitised) content

4. What of the (new) shared services imperative?
   - Who does what, at what level/scale?

COAR: Confederation of Open Access Repositories

- 48 members drawn largely from Europe, but including both JISC & CNI, and also EDINA (University of Edinburgh)
- Work Plan for 2010/12, including
  1. Advocacy on behalf of OA and repositories (Rs) [both together?]
  2. Populating (OA) Rs
  3. Best practice documents
  4. Facilitate and ensure data interoperability of (across?) Rs
  5. Interoperability with other systems (such as CRIS systems)
  6. Support national helpdesk
  7. Guidance on how Rs will form essential elements for global e-infrastructure
  8. Promote R manager profession
  9. Provide advice & guidance on suitable R infrastructure technologies
  10. Global (meta)data store
  11. Strategic partner other infrastructure-related initiatives worldwide
Sound & Pictures: access to new data sources

- **20th Century is the first fully audio-visual century**
  - With new forms of research material to use and to master
- **EDINA as platform for downloadable film, video and audio**
  - Licensed for use in learning, teaching and research
  - Wide range of subject coverage, including documentary film
1. **Film & Sound Online**
   - 600 hours of film, digitised for downloading
2. **NewsFilm Online**
   - 3000 hours of material from ITN & Reuters
   - Over 4TBs of clips to download
- **Plus Education Image Gallery of still photography**
- **Visual and Sound Materials Portal**
  - Discovering all sorts of audio-visual material

UK-CORR: UK Council of Research Repositories

individual rather than institutional, UKCORR-discussion@jiscmail.ac.uk

UK has ‘rich heterogeneous repository landscape’ (C.Awre); lurk following comment from Dorothea Salo:
- US mainly about OA full texts; UK mainly about … serving research assessment!
- Is there more to IRs than the REF: lots of bibliographic records & little full text?
- Should IRs only accept full text, not metadata only?
1. in absence of a CRIS, our IR had to do REF (Lancaster & Northampton)
2. was OA but then RAE2008, but should aim to include all (Oo)
3. motive for IR was digital preservation, with different REF system; funder mandate compliance for OA; visibility via OA (Oxford/Bodleian)
4. RAE/REF is opportunity to engage institution-wide (Warwick)
5. Advent of CRIS (which don’t manage outputs well) may be opportunity for IRs to have role, including use of ‘metadata only’ as lever to obtain full text (Hull)
6. RAE/REF has different goals to OA and IRs with low % of full text may undermine OA movement (Nottingham)