Research Data Management and Institutional Repositories

2014 LIS Research Symposium
UNISA
Presentation overview

- Data and Research Data Management (RDM)
- Importance of research data and research data management
- Supporting RDM
  - Repositories in a data ecosystem
  - Institutional Repositories and RDM
    - Requirements for research data repositories
    - Nature of research data
    - The management of research data as a digital objects
- Examples
- Closing remarks
What is research data?

“Research data, unlike other types of information, is collected, observed, or created for purposes of analysis to produce original research results.”

Edinburgh University 2010
Importance of research data & RDM

“Well-managed data in digital form have great potential to be searched, accessed, mined and reused. Data may be examined to validate research results; it could be consulted by researchers of related interest and save time and resources in data re-collecting; data may even be re-purposed to answer questions unrelated to the context in which it was first generated or gathered. The value of data grows significantly as the data form more accessible collections.”

“The awareness of the data deluge phenomenon, the potential impact of data reuse, and the desire to maximize the return on investment of research funding all led to increasing amounts of discussion on research data management.”

Wong 2009
Data management is the process of controlling the information generated during a research project. Managing data is an integral part of the research process. How data is managed depends on the types of data involved, how data is collected and stored, and how it is used - throughout the research lifecycle.
Supporting RDM

Research Data Management

- Data re-use services
- Data curation services
- RDM services
- IT Platforms
- Policies Resources Capacity
Research Data Management Platforms

PRIMARY RESEARCH
Data creators

Proposed writing & project start-up

Data creation

Data analysis

Research in-process (organising, storing, collaboration)

Prepare for deposit of data

Ingest

Preservation

Dissemination

Sharing data

Using data

Managing & preparing data

Preserving data

DATA CURATION
Data curators

SECONDARY DATA USE
Data users

Discovery of data for re-use

Login
Transfer
Storage
Interfaces
Tools
Repositories in a data ecosystem

• Types of repositories
  • Institutional repositories \ Discipline (subject repositories / data centres)
  • Research \ Community \ Reference data collections
  • Metadata repositories \ aggregators

• Implementation models
  • One or multiple custodians
  • Different data pathways
    • Data producer ➔ Analysis (researcher)
    • Data producer ➔ IR
    • Data producer ➔ IR ➔ Discipline repository
    • Data producer ➔ Discipline repository ➔ Metadata repository

National Science Board 2005; Beagrie, Chruszcz et al. 2008; Wilson, Martinez-Uribe et al. 2011
Repositories in a data ecosystem

• The curation continuum
  • Levels of curation (Rusbridge, 2010)
    – High (high levels of expertise, where subject specialists are involved during the ingest phase of data archiving, adding and cleaning descriptive metadata)
    – Low (greater degree of automation; minimal manual intervention)
  • Information continuum

Wilson, Martinez-Uribe et al. 2011; Treloar, Harboe-Ree 2008
Institutional Repositories

“An IR is a set of services and technologies that provide the means to collect, manage, provide access to, disseminate, and preserve digital materials produced at an institution.”

Shreeves, Cragin 2008

Research outputs

Research data?
IRs and research data

Different opinions!
IRs and research data

As noted above, the collection and preservation of grey literature, data sets, and other unpublished material has been called a distraction from self-archiving and open access, touted as the “primary” purpose of IRs by some open access advocates. Current software packages often are inadequate to manage the range of materials collected (particularly data sets) and fail to provide the range of services that repository managers are finding their users want. Preservation—a stated goal of many IRs—is often an afterthought, as repository managers focus on getting material into IRs first (see Rieh et al., in this issue). In many cases,
Questions to answer

- Who are the users of the services and stakeholders?
- What do they need / require?
- What is the nature of the content?
  - Are data sets unique digital objects with unique requirements?
  - Do data sets require a unique set of services?
- How is the content produced and used?
- How will the platform fit into the bigger data ecosystem?
- What is achievable within the context of the organisation?
  - Other platforms and services in the organisation
  - Commitment of organisation (financial and human resources)
  - Readiness of the organisation
Data related requirements

Nature of research data

1. Heterogeneity
2. Contextual dependence
3. Transient nature
4. Digital nature

Standards & protocols

RDM services
Curation services
Re-use services

Data repository services

Heterogeneity

- Research methodologies
- Practices of researchers
- Kinds of data, sizes, formats and composition
- Data value
- Raw, aggregated data

Flexible and highly scalable
Clarity on what the repository cater for
**Contextual dependence**

- Collection composition
  - Linking items (data files, contextual documents, outputs)
  - Data with outputs or outputs with data
  - Organise into project-based collections (a "bag")

**Data set**

**Spatial aspects of unemployment in South Africa 1991-2007: Municipalities - All provinces**

<table>
<thead>
<tr>
<th>All data sets</th>
<th>Data set details</th>
<th>Documentation</th>
<th>Data files</th>
<th>Outputs</th>
<th>Access conditions</th>
<th>Contact</th>
</tr>
</thead>
</table>

**Data set metadata record**

- **Data set ID**: UNEMPL 1991-2007 Municipalities
- **Title**: Spatial aspects of unemployment

http://curation.hsrc.ac.za/Dataset-278.phtml

Weber 2011
Transient nature

- Snapshots vs “live” data
- Versioning of data sets

“...digital preservation systems designed to steward only final, unchanging materials can only fail faced with real-world datasets and data-use practices.”

Salo 2010

- Authority management (Author IDs)
- Persistent identifiers (DOIs)
- Citation standards
Digital object management

- Data submission (deposit)
  - Ethics requirements for human subjects
    - Consent to share
    - Anonymisation
  - Ownership issues
  - Self-archiving
  - Automation

Basic Use Case

- End user logs into repository using SSO
- Starts a submission and must register with Globus if this is their first time
- Is automatically logged into Globus and the submission tool (SSO)
- Chooses a “Collection” and enters required metadata for that collection
- Creates a new endpoint if required
- Selects an endpoint
- Selects files/directories for transfer
- Logs out and is notified of progress via email

Taylor 2013
Digital object management

• Access (consistent with tools and processes of the research community)
  • Discovery services (browsing, searching, OMP-MIH)
  • Metadata
    • Discovery, determine relevance, make data useable, provenance
    • Compliance with recognized standards of the community
  • Dissemination formats
  • Ways to serve and use data
  • Usage terms and conditions
  • Access management
  • Usage statistics
Digital object management

- Preservation
  - Preservation management
  - Registries
  - Retention period, de-accessioning, destruction
  - Strategies support to file formats and their long-term usability
    - Archival formats
    - Format migration
  - Storage and storage management
    - Multiple copies, multi-media, multi site
    - Back-up
    - Disaster recovery
    - Security
Standards and protocols

- Interoperability
  - Metadata (machine readable in appropriate standard)
- Various standards and protocols
  - Open Archives Initiative Object Reuse and Exchange (OAI-ORE) ([http://www.openarchives.org/pmh/](http://www.openarchives.org/pmh/))
  - Simple Web-service Offering Repository Deposit (SWORD) ([http://swordapp.org/](http://swordapp.org/))
Examples

suggest. It is notable that both Cambridge and KCL in our case studies are developing central repositories to work with departmental facilities and discussing federated local data repositories for research data preservation combining services and skills from central and departmental repositories with data distributed and located at different repositories in the institution. A similar discussion and scoping project is also currently underway at the

Architectures

- Flexible Storage and Metadata Architectures
- De-coupling Ingest, Storage and Use

Witt 2008; Taylor 2013; Beagrie, Chruszcz et al. 2008
Closing remarks

• Not just a one size fits all one vendor solution (install and go)
• RDM Services for small data is labour intensive (Various roles and responsibilities)
• Don’t see IRs in isolation
• Clear vision of aims
• Investigate and experiment
• Collaborate

The verdict?
References


MIX, K. and TAYLOR JR, L., VLA Paraprofessional Forum Twentieth Annual Conference.


Thank you

Building the bridge between research, policy and action

Lucia Lötter
llotter@hsrc.ac.za
www.hsrc.ac.za