Paving the way to open and interoperable research data service workflows

Progress from 3 perspectives

Angus Whyte, Digital Curation Centre
Rory Macneil, Research Space
Stuart Lewis, University of Edinburgh

Repository Fringe, Edinburgh, 2nd August 2016
Paving the way to open and interoperable research data service workflows

Progress from 3 perspectives

Angus Whyte, Digital Curation Centre
Research data service models, new DCC guidance and some draft ‘design principles’ for institutions integrating research data service workflows

Rory Macneil, Research Space
Integrating the RSpace ELN with University of Edinburgh’s DataShare and Harvard’s Dataverse repositories

Stuart Lewis, University of Edinburgh
DataVault –Jisc Research Data Spring prototype for packaging data to be archived
Guidance from Digital Curation Centre

Aims to effectively support organisations with-
• Providing effective research data services
• Promoting reusability of their research data
• and reproducibility of their research
Models that reflect and help shape reality

Key element of DCC guidance e.g. Curation Lifecycle Model (2008)

DCC research data service model (2016)

RDM policy & strategy ➔ Business plans & sustainability

Data Management Planning

Active data management

Appraisal & risk assessment

Preservation

Access & publishing

Discovery

Training  Advisory services

Some working definitions

**Research data service** – “a means of delivering value to the producers and users of digital objects by facilitating outcomes they want to achieve without the ownership of specific costs or risks” (*derived from ITIL definition of a service*)

Sub-types -

**Active data management** “services used to create or transform digital objects for the purposes of research”

**Preservation**: “services offering to ensure digital objects meet a defined level of FAIRness - findability, accessibility, interoperability, and reusability - for a designated community and period of time”

**Publication**: “services offering to enhance digital objects FAIRness by reviewing their quality on specified criteria, or connecting them to additional metadata”

**Guidance**: “services offering practical guidance on choosing or using the above services”
Draft design principles for integration of research data service workflows

1. **Active data management** services should use open standards to express and expose the objects and metadata they offer to downstream services, including their access and reuse terms.
Draft design principles for integration of research data service workflows

2. **Preservation** and **publication** services should publish policies stating what digital object types they accept, for what communities, and on what terms and conditions.
Draft design principles for integration of research data service workflows

3. **Preservation** and **publication** services should make openly available sufficient metadata to enable reuse of their outputs, including all terms and conditions for third-party access and reuse.
Draft design principles for integration of research data service workflows

4. **Active data management, preservation** and **publication** services should make sufficient detail of their workflows available to support the reproducibility of research that produced the digital objects they act upon
Draft design principles for integration of research data service workflows

5. **Guidance** services should support users of other services to make an informed choice of downstream service capabilities, informed by best practices for reuse and reproducibility.
Draft design principles for integration of research data service workflows

6. Guidelines 1-5 should be implemented using machine-actionable content
Real services are made up of many more parts. Tools and services at researchers' disposal are increasing.
One size does not fit all

‘lifecycles’ are often non-linear

Integration use cases do not follow neat sequences
What about Functional Models? OAIS foundation for Trustworthy Digital Repository standards
RDM platforms can be grafted on, to relate OAIS functions to context info sources.
What best practice models apply ‘upstream’?

Q. How can institutions ensure researchers have informed choice of trustworthy services before they need a repository

• “Whole lifecycle” service models are one solution
• But how should they be governed?
  – Commercial services?
  – Commons?
  – Hybrid?
“On Wednesday 1 June, Elsevier acquired Hivebench to help further streamline the workflow of researchers – putting research data management at their fingertips. The added value of the integration lies in linking Hivebench with Elsevier’s existing Research Data Management portfolio for products and services. The research data that researchers have stored in the Hivebench notebook are linked to the Mendeley Data repository, which will be linked to Pure. This ....adds instant value to the datasets because they become far more suitable for reuse.”
Commons approach
e.g. Principles for Open Scholarly Infrastructures

Principles for Open Scholarly Infrastructures
23 FEBRUARY 2015 14 COMMENTS


*infrastructure* |ˈɪnfrestrʌktʃər| (noun) – the basic physical and organizational structures and facilities (e.g. buildings, roads, power supplies) needed for the operation of a society or enterprise. – *New Oxford American Dictionary*

Everything we have gained by opening content and data will be under threat if we allow the enclosure of scholarly infrastructures. *We propose a set of principles by which Open Infrastructures to support the research community could be run and sustained.* – Geoffrey Bilder, Jennifer Lin, Cameron Neylon
Commons approach
e.g. Open Science Framework

Workflows in effect governed by a ‘mixed economy’ of platforms

So what principles should apply?

Background - RDA Working Group on Data Publishing Workflows

Reviewed 25 examples of repository data publishing workflows, including some integrating with ‘downstream’ services e.g. data journals for peer review


Drafted a reference model and made best practice recommendations-

1. Start small, building modular, open source and shareable components
2. Follow standards that facilitate interoperability and permit extensions
3. Facilitate data citation, e.g. through use of digital object PIDs, data/article linkages, researcher PIDs
4. Document roles, workflows and services
Background - RDA Working Group on Data Publishing Workflows

• Follow up call (Dec 15) for examples of repositories connecting with upstream research workflows e.g. to gather metadata earlier

• Aiming to identify whether recommendations apply, and how the intention to publish data is changing research practices.

• Collected 12 cases - mix of concrete examples, prototypes (e.g. Dendro), conceptual models (e.g. Science 2.0 repositories)

• Report in preparation
Review of upstream workflow examples

Are the services components underpinning research workflows loosely coupled?

– Modular design of workflow components? yes, plenty evidence

– Standard vocabularies and protocols to describe components some evidence

– Significant investment in building trust-based relationships among participants some evidence

– Standardized ways of specifying capabilities and performance requirements limited (e.g. WDS-DSA Catalogue of Requirements)

So do we need more ...

• Definitions – to describe services?
• Design principles – to articulate best practice?
• Capability models - to articulate mutual expectations of service owners?
• Case studies of repository workflow integration?
• Understanding of how tools are actually being integrated, and effect on data publication practices

E.g. DCC How-to describe research data service workflows (forthcoming)
Draft design principles for integration of research data service workflows

1. **Active data management** services should use open standards to express and expose the objects and metadata they offer to downstream services, including their access and reuse terms.

2. **Preservation** and **publication** services should publish policies stating what digital object types they accept, for what communities, and on what terms and conditions.

3. **Preservation** and **publication** services should make openly available sufficient metadata to enable reuse of their outputs, including all terms and conditions for third-party access and reuse.

4. **Active data management, preservation** and **publication** services should make sufficient detail of their workflows available to support the reproducibility of research that produced the digital objects they act upon.

5. **Guidance** services should support users of other services to make an informed choice of downstream service capabilities, informed by best practices for reuse and reproducibility.

6. Guidelines 1-5 should be implemented using machine-actionable content.

Thanks for comments to Suenje Dalmeier Tiessen and Amy Nurnberger, co-chairs of the RDA Working Group on Data Publishing Workflows.
Maintaining trust across the research cycle

Q. How do these principles apply in real cases?

Case Study 1 – Rory Macneill, Research Space
Integrating the RSpace ELN with University of Edinburgh’s DataShare and Harvard’s Dataverse repositories
Current Repository Paradigm

Export data and metadata

Archive or Repository
The reality is (a lot) more complex

<table>
<thead>
<tr>
<th>Research units</th>
<th>Vehicles</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data</td>
<td>Repositories</td>
<td>Capture and structure data</td>
</tr>
<tr>
<td>Files</td>
<td>Archives</td>
<td>Generate metadata</td>
</tr>
<tr>
<td>Links</td>
<td>ELNs</td>
<td>Export data and metadata</td>
</tr>
<tr>
<td></td>
<td>File systems</td>
<td>Establish links to file(s) and databases</td>
</tr>
<tr>
<td></td>
<td>Databases</td>
<td>Track file locations</td>
</tr>
</tbody>
</table>

**Issues**

- Data versus file links
- Forms of data export
- Data from intermediary vehicles
- File location and integrity of links
- Post deposit access - permissions
- Post deposit access - capabilities
- Pre-deposit impact on post-deposit capabilities

Repositories need the ability to easily ingest diverse data types and formats, and links to files, in a structured manner, directly and from other tools.
Towards a new paradigm
The Dataverse – Starfish – RSpace project @Harvard Medical School

Data, files and research

Capture and structure data
Generate metadata

Track file locations

Make data and files available for public access and query
Capture and structure data
Generate metadata
Track file locations

Export data and metadata
In various formats
Using open standards
- PDF
- Word
- HTML
- XML

Access hyperlinks
Track file locations

On premises or Cloud file system
Draft design principles for integration of research data service workflows

1. **Active data management** services should use open standards to express and expose the objects and metadata they offer to downstream services, including their access and reuse terms.

2. **Preservation and publication** services should publish policies stating what digital object types they accept, for what communities, and on what terms and conditions.

3. **Preservation and publication** services should make openly available sufficient metadata to enable reuse of their outputs, including all terms and conditions for third-party access and reuse.

4. **Active data management, preservation and publication** services should make sufficient detail of their workflows available to support the reproducibility of research that produced the digital objects they act upon.

5. **Guidance** services should support users of other services to make an informed choice of downstream service capabilities, informed by best practices for reuse and reproducibility.

6. Guidelines 1-5 should be implemented using **machine-actionable content**.
Case study 2

Stuart Lewis, University of Edinburgh
DataVault – Jisc Research Data Spring prototype for packaging data to be archived
Stuart Lewis
Deputy Director, Library & University Collections
The University of Edinburgh
@JiscDataVault
Research Data Management Services
Data Stewardship

- **DataVault**
  - Long term archival storage
  - First envisaged a few years ago...
What is the DataVault - Analogies

https://www.flickr.com/photos/brookward/8457736952
What is the DataVault - Analogies
Where does it sit?
Where does it sit?
Where does *could* it sit?
Phase 1 (3 months)
Phase 2 (4 months)
Phase 3 (6 months)
Data Vault
Keep your research data safe.

Welcome to the Data Vault!

Please use the 'Create new Vault' button below to get started...

Create new Vault
Welcome to the Data Vault!

As this may be your first time here, please read the help section before continuing.

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Define a dataset in CRIS</td>
<td>You must record a dataset in CRIS to record details of your vault.</td>
</tr>
<tr>
<td>2. Define your storage options</td>
<td>We need to know a little about where you intend to archive data from. Please define your storage options.</td>
</tr>
<tr>
<td>3. Create your first vault</td>
<td>You are now ready to create your first vault!</td>
</tr>
</tbody>
</table>
Vault Name

Brain image scans from project DIFF-5

Description

BBSRC 5th DIFF project. Scans from MRI experiments, stored in RAS format.

Relates to

Please choose a dataset record

- Whole blood mRNA expression profiling of host molecular networks in neonatal sepsis
- Temporal analysis of RNA turnover in Interferon Gamma treated bone marrow-derived macrophages

Group

Medicine and Veterinary Medicine
Vault Name
Brain image scans from project DIFF-5

Description
BBSRC 5th DIFF project. Scans from MRI experiments, stored in RAS format.

Relates to
Whole blood mRNA expression profiling of host molecular networks in neonatal sepsis

Retention Policy
BBSRC

Group
Medicine and Veterinary Medicine

Create new Vault  Cancel
My research project

Images of plants and organisms with associated metadata.

Owner: user1
Retention Policy: EPSRC
Group: Science and Engineering
Created: Jun 14, 2016 8:30:02 AM
Review date: Jun 14, 2026 8:30:02 AM (Status: OK)
Size: 0 bytes

Deposit data
Deposit Note: First set of data

Deposit file or directory:

- Default filesystem (local)
  - .DS_Store
  - active
  - bioimages
    - images.csv
    - license.xml
    - names.csv
    - organisms.csv
    - samples.bin
    - sparql
    - xquery
    - sample data

Deposit size: 1 GB

[Deposit data] [Cancel]
First set of data

- Calculating size
- Transferring

Transferred 700 MB of 1 GB (69 MB/sec)

Note: First set of data
Status: In progress
Size: 0 bytes
Timestamp: Jun 14, 2016 8:38:56 AM
# First set of data

<table>
<thead>
<tr>
<th>Note</th>
<th>Status</th>
<th>Size</th>
<th>Timestamp</th>
</tr>
</thead>
<tbody>
<tr>
<td>First set of data</td>
<td>complete</td>
<td>1 GB</td>
<td>Jun 14, 2016 8:38:56 AM</td>
</tr>
</tbody>
</table>

[Retrieve data]
### First set of data

<table>
<thead>
<tr>
<th>File</th>
<th>Type</th>
<th>Checksum</th>
</tr>
</thead>
<tbody>
<tr>
<td>bioimages/license.xml</td>
<td>application/xml</td>
<td>7bfaf745438d1bd27200f67d7f40705d md5</td>
</tr>
<tr>
<td>bioimages/xquery/index-files.xq</td>
<td>application/xquery</td>
<td>efc6c7648410eafc70d838f6428585e2 md5</td>
</tr>
<tr>
<td>bioimages/xquery/image-rdf-all-hack.xq</td>
<td>application/xquery</td>
<td>a14223f7541240120352f422cc0b7f4b md5</td>
</tr>
<tr>
<td>bioimages/sparql/counting-query.sparql</td>
<td>text/plain</td>
<td>fee33e1b43017b928f638d5674085027 md5</td>
</tr>
<tr>
<td>bioimages/images.csv</td>
<td>text/csv</td>
<td>6e3accf46f1fba3a3b7415bb331294b5 md5</td>
</tr>
<tr>
<td>bioimages/sparql/.DS_Store</td>
<td>application/octet-stream</td>
<td>fa4ade8b137c27ae09fe503ed8c32cf7 md5</td>
</tr>
<tr>
<td>bioimages/xquery/.DS_Store</td>
<td>application/octet-stream</td>
<td>41d509a2c82675c794f77087ed54a3f5 md5</td>
</tr>
<tr>
<td>bioimages/organisms.csv</td>
<td>text/csv</td>
<td>349384c29dc3ded0a96b6c95d6ca149a md5</td>
</tr>
<tr>
<td>bioimages/xquery/organism-html.xq</td>
<td>application/xquery</td>
<td>b7f7f162d04660276557077c6cd2ee9 md5</td>
</tr>
<tr>
<td>bioimages/sparql/inverse-reasoning.sparql</td>
<td>text/plain</td>
<td>027f404d125aba81df236c19116fc5c2 md5</td>
</tr>
</tbody>
</table>
Describe the reason for this retrieve request (who and why) and choose a working directory to retrieve data from the archive

**Retrieve Note:**

Restoring my data

**Target directory:**

- Default filesystem (local)
  - active
  - bioimages
  - sample data

[Retrieve data] [Cancel]
First set of data

- Computing free space
- Retrieving from archive
- Validating data
- Transferring files
- Data retrieve complete

Transferred 840 MB of 1 GB (141 MB/sec)

<table>
<thead>
<tr>
<th>Note</th>
<th>Status</th>
<th>Size</th>
<th>Timestamp</th>
</tr>
</thead>
<tbody>
<tr>
<td>First set of data</td>
<td>Complete</td>
<td>1 GB</td>
<td>Jun 14, 2016 8:38:56 AM</td>
</tr>
</tbody>
</table>

Retrieve data
Metadata already captured from CRIS, plus files from the Vault

First set of data

- Computing free space
- Retrieving from archive
- Validating data
- Transferring files
- Data retrieve complete

Transferred 840 MB of 1 GB (141 MB/sec)

<table>
<thead>
<tr>
<th>Note</th>
<th>Status</th>
<th>Size</th>
<th>Timestamp</th>
</tr>
</thead>
<tbody>
<tr>
<td>First set of data</td>
<td>Complete</td>
<td>1 GB</td>
<td>Jun 14, 2016 8:38:56 AM</td>
</tr>
</tbody>
</table>

Make available online

Retrieve data
What is a Data Vault?
Draft design principles for integration of research data service workflows

1. **Active data management** services should use open standards to express and expose the objects and metadata they offer to downstream services, including their access and reuse terms.

2. **Preservation** and **publication** services should publish policies stating what digital object types they accept, for what communities, and on what terms and conditions.

3. **Preservation** and **publication** services should make openly available sufficient metadata to enable reuse of their outputs, including all terms and conditions for third-party access and reuse.

4. **Active data management, preservation** and **publication** services should make sufficient detail of their workflows available to support the reproducibility of research that produced the digital objects they act upon.

5. **Guidance** services should support users of other services to make an informed choice of downstream service capabilities, informed by best practices for reuse and reproducibility.

6. Guidelines 1-5 should be implemented using **machine-actionable content**.