What Should a Data Repository Do?

MSU ScholComm Summit
Friday, July 22, 2016
Introductions

Bryan Brown
- Web Developer
- ~2 years @ FSU
- IU grad
- @bryjkbrown

Devin Soper
- ScholComm Librarian
- ~1 year @ FSU
- UBC grad
- @devinsophe
1. What is a Data Repository?
2. Exemplar Repositories
3. Core Functionality
4. Public Access Requirements
5. Blue Sky: DR 2.0
6. Future of DigiNole
What is a Data Repository?

Is there a shared, authoritative definition?

- Data sharing
- Data publishing
- Open data

Image credit: https://www.jisc.ac.uk/sites/default/files/open-ideas_0.jpg
Data Sharing & Publishing

Ten repository-specific guidelines:

1. Explicit mission of digital archiving
2. Due diligence to ensure legal compliance
3. Documented processes for managing data storage
4. Long-term digital preservation plan
5. Explicit workflows across the data life cycle
6. Responsibility for access and availability of digital objects
7. Ensuring discovery and persistent identification
8. Ensuring integrity of digital objects & metadata
9. Ensuring the authenticity of digital objects & metadata
10. Technical infrastructure compliant with archival standards (OAIS)

http://www.datasealofapproval.org/en/information/guidelines/
Data Repository Directories

re3data.org

Registry of Research Data Repositories

Browse by subject

Graphical  Text

click to zoom into subjects or to select a bottommost subject in the hierarchy as filter for the re3data search page
ctrl + click on a top subject to select it as filter

http://www.re3data.org/
Data Repository Directories

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

Filter
Subjects
Content Types
Countries
AID systems
API
Certificates
Data access
Data access restrictions
Database access
Database access restrictions
Database licenses
Data licenses
Data upload
Data upload restrictions
Enhanced publication
Institution responsibility type
Institution type

Search...

1 2 3 4 5 6 7 ... 66 Next →

Found 1633 result(s)

ICSU World Data System
International Council for Science World Data System

Subject(s)
- Humanities and Social Sciences
- Life Sciences
- Natural Sciences
- Engineering Sciences

Content type(s)
- Standard office documents
- Images
- Scientific and statistical data formats
- Raw data
- Plain text
- Archived data
- Structured text

Country
- Japan
- International

The Prototype Data Portal allows to retrieve Data from World Data System (WDS) members. WDS ensures the long-term stewardship and provision of quality-assessed data and data services to the international science community and other stakeholders.

http://www.re3data.org/
● Most data repositories are discipline-specific, but a growing minority self-identify as IRs

● Many institutions have launched IRs that focus on data specifically, but traditional IRs serve the same purpose

● Growing need for “generalist” data repositories as data sharing requirements impact research in different fields

● Exciting opportunity for institutions to develop IRs that accommodate data sharing and publication

Exemplar Data Repositories
Focus on social/behavioral sciences

Great faceting

Features codebooks

Support for multiple datasets

Must be a member to use
Exemplar Data Repos

The Dataverse Project

- Discipline agnostic
- Open Source
  - Institutional installations
- Journal integrations
- Interesting previews for certain types of data
  - Harvard has a connected R server (neat!)

Dataverse Category
- Researcher (484)
- Research Project (450)
- Organization or Institution (155)
- Journal (93)
- Teaching Course (14)

Dataverses (1,765)
Datasets (61,997)
Files (332,315)
Exemplar Data Repos

- General purpose
- All open
- Digital preservation
  - Fixity checking
  - Format migration
- Journal integration
- No previews
- Built on DSpace
Open Science Framework

- Discipline agnostic
- Free, even for orgs
- Like GitHub, but for research data
- Focus is on working aspects of data
  - Could be integrated with other data repos?
  - AND vs OR
What Should a DR Do?
Format Agnostic File Storage

- Manages data without having to know what it is
- Support for FILES, not file TYPES
- Preview what you can, skip the rest
- Support for large files (STORAGE!) & complex dataset structures
Functionality

Appropriate Use of Metadata

- Metadata drives display
- Leverage discipline-specific metadata
- Optimal search interfaces
  - Intuitive results display
  - Relevant facets
- Clearly state terms of use
- Preferred (& dynamic!) citations
Digital Preservation

- Commitment to long term hosting of data = digital preservation
- Scheduled checks for file fixity
- Migration to modern file formats
  - Obviously not possible for all file formats, but do what you can
  - When done right, this can yield multiple download format options
Public Access Requirements
What functionality is required to ensure compliance with federal public access mandates?
## Overview of OSTP Responses

<table>
<thead>
<tr>
<th>Funder</th>
<th>Policy Coverage</th>
<th>Policy Stipulations</th>
<th>Support Provided</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Published Outputs</td>
<td>Data</td>
<td>Time Limits</td>
</tr>
<tr>
<td>AHRQ</td>
<td>● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ●</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
http://sparcopen.org/our-work/research-data-sharing-policy-initiative/funder-policies/

https://dmptool.org/
Requirements

1. Archiving and public access
   a. Format-agnostic file storage
   b. Support for large files and compound objects

2. Discovery & persistent identification

3. Long-term preservation
   a. Up to 10 years
   b. Format migration?
Blue Sky
Preservation

- Clear policies on data formatting & availability
- Clear curation services and support levels

Discovery

- Data-specific search criteria
  - validation, reuse, geolocation
- Tagging & rating systems
- Browsing & visualization
Access

- Granular access controls
- Browser-based viewers
- Web-based transfer protocols
- Versioning & time-stamping

Metadata

- Micro-attribution
- Data citation
- Citation relationships
- Granular licensing
Repository 2.0

- Altmetrics, links to citing papers, etc.
- Collaborative workspaces to capture data in context
- Social networking functionality
  - Profiles, comments, likes, notifications
- Rating, tagging, recommendations
- Integrations with other systems

The Road Ahead at FSU
The Future of DigiNole

- Reassess metadata possibilities for data sets (tweak search & display)
- Increase storage capacity
- Fixity checks w/ PREMIS
- Integration with OSF
The Future of DigiNole

- Reassess metadata possibilities for data sets (and tweak search, too)
- Increase storage capacity
- Fixity checks with PREMIS
- Integration with OSF
- Outreach!
Questions?
Thank you!