An Institutional Approach to Research Data Curation
Collaborators

Brian Caruso
Kathy Chiang
Jon Corson-Rikert
Ann Green
Brian Lowe
Janet McCue

This material is based upon work supported by the National Science Foundation under Grant No. III-0712989.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.
Outline

- Why archive/share data
- Concept: data staging repository
- Is it an institutional repository?
- DataStaR details (approach, partners...)
- Very early lessons/challenges
DataStaR: A Data Staging Repository
DataStaR: A Data Staging Repository

The purpose of DataStaR is to support collaboration and data sharing among researchers during the research process, and to promote publishing or archiving data and high-quality metadata to discipline-specific data centers, and/or to Cornell’s own digital repository.
Common questions / needs:

- I need a place to share (large) data files with colleagues.
- I want to make a data set related to a publication available online.
- Which data should I archive?
- How should data be formatted?
- I want people to ask permission to use my data because...
Promoting the publication of data
Promoting the publication of data

Producer → DataStaR

‘Pre’-submit (pub-ready) → DataStaR

DataStaR → Domain repository

DataStaR → Institutional repository

Repository

Domain repository → DataStaR

Institutional repository → DataStaR

DataStaR → Consumer

Curate

Submit

Disseminate
Promoting the publication of data

DataStaR

Producer

'Pre'-submit (draft)

'Pre'-submit (pub-ready)

DataStaR

Curate

Disseminate

Consumer

Repository

Domain repository

Institutional repository

Submit

Disseminate

Curate

Consumer
Promoting the publication of data

Producer

Preferred start

'Pre'-submit (draft)

'Pre'-submit (pub-ready)

DataStaR

Curate

Disseminate

Consumer

Delayed start

Repository

Domain repository

Institutional repository

Submit

Disseminate

Curate

Consumer
## Is DataStaR an institutional repository?

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Institutional Repository</th>
<th>Domain Repository</th>
<th>DataStaR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constituents</td>
<td>Institution</td>
<td>Discipline</td>
<td>Institution</td>
</tr>
<tr>
<td>Content emphasis</td>
<td>Articles and monographs</td>
<td>Data and/or publications</td>
<td>Data</td>
</tr>
<tr>
<td>Standards: metadata</td>
<td>Generic</td>
<td>Domain-specific</td>
<td>Generic and domain-specific</td>
</tr>
<tr>
<td>Standards: data format</td>
<td>Generic</td>
<td>Domain-specific</td>
<td>Generic → domain specific</td>
</tr>
<tr>
<td>Preservation commitment</td>
<td>May be prepared to manage and migrate SOME formats over time</td>
<td>May be prepared to manage and migrate formats over time</td>
<td>Not a preservation repository, but responsible partner</td>
</tr>
<tr>
<td>Deposit mandate</td>
<td>Voluntary</td>
<td>Voluntary or required</td>
<td>Voluntary</td>
</tr>
<tr>
<td>Deposit process</td>
<td>As automated as possible</td>
<td>May involve significant support</td>
<td>May involve significant support</td>
</tr>
<tr>
<td>Point of engagement in research cycle</td>
<td>Late</td>
<td>May be early, middle, late</td>
<td>May be early, middle, late</td>
</tr>
<tr>
<td>Support for content users</td>
<td>Minimal</td>
<td>May be significant</td>
<td>Minimal</td>
</tr>
<tr>
<td>Tools: data analysis, processing, visualization</td>
<td>None</td>
<td>Domain-specific</td>
<td>None</td>
</tr>
</tbody>
</table>
Infrastructure and services to promote publication of data

What’s in it for researchers:

- Secure sharing
- Easy creation of preliminary metadata
- Tools for creating detailed metadata (various)
- (Help for above!)
- Research group “view”
- Publication assistance

Novel approach to metadata management
Metadata management

- Use semantic web technologies to treat metadata as a web of statements
- Adapt existing software (vitro: http://vitro.mannlib.cornell.edu/)
- Generate OWL representations of metadata schemas

Result for users:

- Consistent interface for creating metadata
- Reuse/recombine information
Metadata management

- Use semantic web technologies to treat metadata as a web of statements
- Adapt existing software (vitro: http://vitro.mannlib.cornell.edu/)
- Generate OWL representations of metadata schemas

Result for users:

- Consistent interface for creating metadata
- Reuse/recombine information
User view

My Research Groups

- Harford Teaching and Research Center
- Agricultural Ecology Program: Upper Susquehanna River Basin

Create New Group  Join a Research Group

My Data Sets

Personal

- Water quality data for Fall Creek (Tompkins County, NY) sampling sites: 1972-1995
- Water quality data for Kashong Creek Watershed (Ontario County and Yates County, NY) sampling sites: 1977-1979

Shared

- Well Logs for wells at the Cornell Department of Animal Science Harford Teaching and Research Center
- Distribution of Cs-137 in stream sediments and stream banks in the upper Susquehanna basin - 2006

Create New Data Set

Recently Updated

Alice Poll edited metadata for Distribution of Cs-137 in stream sediments and stream banks in the upper Susquehanna basin - 2006

David Bouldin created new data set Well Logs for wells at the Cornell Department of Animal Science Harford Teaching and Research Center

Christina Tonitto added as a member of Agricultural Ecology Program: Upper Susquehanna River Basin

Jed Sparks added files to data set Distribution of Cs-137 in stream sediments and stream banks in the upper Susquehanna basin - 2006
User view

Agricultural Ecology Program: Upper Susquehanna River Basin

Recent Activity

- Alice Poll edited metadata for Distribution of Cs-137 in stream sediments and stream banks in the upper Susquehanna basin - 2006
- Tom Butler added as a group member
- Jed Sparks added files to data set Distribution of Cs-137 in stream sediments and stream banks in the upper Susquehanna basin - 2006

Group Members

- David Bouldin: Cornell faculty member
- Tom Butler: visiting scientist
- Chih-Hei Chang: graduate student
- Zach Easton: technician
- Tim Fathay: Cornell faculty member
- Christy Goodale: Cornell faculty member
- Bob Howarth: Cornell faculty member
- Alice Poll: Cornell faculty member
- Jed Sparks: Cornell faculty member
- Christina Tonitto: postdoctoral associate

Data Sets

- Distribution of Cs-137 in stream sediments and stream banks in the upper Susquehanna basin - 2006
- Water quality data for Fall Creek (Tomkins County, NY) sampling sites: 1972-1995
- Water quality data for Kashon Creek Watershed (Ontario County and Yates County, NY) sampling sites: 1977-1979
- Water quality data for southern tributaries to Cayuga Lake (Tomkins County, NY): 1987-1989
- Water quality data for well, stream, and seep samples from the Hartford Teaching and Research Farm (Cortland County, NY): 1974-1994
DataStar

User view

Distribution of Cs-137 in stream sediments and stream banks in the upper Susquehanna basin - 2006

Manage Metadata

- Creator
  - Title

Access

Public
- Title Only
- Metadata
- Files
- Metadata and Files

This Group
- Title Only
- Metadata
- Files
- Metadata and Files

People
- Title Only
- Metadata
- Files
- Metadata and Files

DataStar users who are not members of this group.

Current Files

- File 1: 2.13 mb
- File 2: 876 kb
- File 3: 3.37 mb
- File 4: 9.24 mb
- File 5: 8.3 mb
- File 6: 219 kb

Publish this Data Set
Delete this Data Set
Complete Data Set

DataStar
Home Data Sets Research Groups People Repositories

Hi, David Dashboard Help Logout

Manage Data Set

My Data Sets
- David Boulder's Dashboard
- My Research Groups
- Aquatic Ecosystem Program, Upper Susquehanna River Basin

My Groups

- My Groups
- My Research Groups

My People

- My People
- My Research Groups

Privacy
Terms
Partners

• Upper Susquehanna River Basin Agricultural Ecology Program
• Cornell Biological Field Station
• Cayuga Lake Watershed Network
• Submission mechanism for CUGIR
• Individual researchers

Current status
Challenges / lessons so far:

- Very high level of service → self service
- “Low” barriers may not be
- People are embracing the idea
Thank you

Gail Steinhart  
Research Data & Environmental Sciences Librarian  
Albert R. Mann Library  
Cornell University

gss1@cornell.edu