Data Publishing: A form of Scholarly Communication

1665
Data, if any, were part of the printed publication

Now
Vast quantities of digital data (and code) cannot be part of the printed publication

350 years of scientific publishing, with words and data
To make data discoverable, accessible and reusable, we need:

1. **Data Citation**, to reference and find data
2. **Data Repositories**, to host and access data
3. **Information about the data**, to understand and reuse them
Dataverse Software: A Data Publishing framework

... for a wide range of repositories

- Harvard Dataverse
- DANS - Dutch Dataverse
- Odum Institute Dataverse

Public, Generic Repositories
Institutional Repositories
Curated Data Archives Repositories
Dedicated to sharing, archiving and citing research data.
Dataverse 4.0: Enables and Enhances Data Publishing

- A **data citation** compliant with the Data Citation Principles
- Rich **metadata** to describe and find datasets from multiple domains
- Support for **public and restricted** data, open data license and terms of use
- Rigorous **workflows** to publish data, with support for new versions of the data
Data Citation
A Brief History of Citing Data

1906
Chicago Manual of Style: author/creator, title, dates, publisher or distributor

1979
- ASBR ("Data File" type)
- MARC (machine readable catalog)
- Domain Repositories (e.g., GenBank)

1959
First scientific digital repositories (e.g., World Data Center, ICPSR)

1999 - Now
- Growth of Data Repositories (e.g., NESSTAR, Dataverse, Dryad, Figshare, Zenodo)
- DOI services for Data (e.g., DataCite in 2009)

Altman & Crosas, 2013, “The Evolution of Data Citation: From Principles to Implementation” IASSIST Quarterly
Joint Declaration of Data Citation Principles

1. Importance
2. Credit and Attribution
3. Evidence
4. Unique Identification
5. Access
6. Persistence
7. Specificity and Verifiability
8. Interoperability and flexibility

https://www.force11.org/datacitation
Principle 2: Credit and Attribution

Principle 4, 5, 6: Unique Id Access Persistence

Principle 7: Specificity and Verifiability

Principle 8: Interoperability and flexibility:
Repository exports citation metadata in XML, JSON formats

Altman & King, 2007. A Proposed Standard for the Scholarly Citation of Quantitative Data.
metadata
Schematic Diagram of a **Dataverse** in Dataverse 4.0

Schematic Diagram of a **Dataset** in Dataverse 4.0

Container for your data, documentation, and code.
Three Metadata Levels

Generic Metadata
Includes data citation metadata fields (Examples: title, authors, persistent id, description)

Domain Specific Metadata
Examples:
- Social Science Metadata (DDI)
- Life Sciences (ISA-Tab)
- Astronomy (VO)

File Metadata
Examples (automatic):
- For Tabular Files: Column information
- For FITS Files: Header information
### Datasets: Life Sciences Metadata

<table>
<thead>
<tr>
<th><strong>Design Type</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Case Control</td>
<td></td>
</tr>
<tr>
<td>Cross Sectional</td>
<td></td>
</tr>
<tr>
<td>Not Specified</td>
<td></td>
</tr>
<tr>
<td>Parallel Group Design</td>
<td></td>
</tr>
<tr>
<td>Perturbation Design</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Factor Type</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
</tr>
<tr>
<td>Biomarkers</td>
</tr>
<tr>
<td>Developmental Stage</td>
</tr>
<tr>
<td>Cell Surface Markers</td>
</tr>
<tr>
<td>Cell Type/Cell Line</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Measurement Type</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA Methylation Profiling (Bisulfite-Seq)</td>
</tr>
<tr>
<td>DNA Methylation Profiling (MeDIP-Seq)</td>
</tr>
<tr>
<td>Histone Modification (ChIP-Seq)</td>
</tr>
<tr>
<td>Protein-RNA Binding (RIP-Seq)</td>
</tr>
<tr>
<td>Transcription Factor Binding (ChIP-Seq)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Organism</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Danio rerio</td>
</tr>
<tr>
<td>Homo sapiens</td>
</tr>
<tr>
<td>Mus musculus</td>
</tr>
<tr>
<td>Rattus norvegicus</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Cell Type</strong></th>
<th></th>
</tr>
</thead>
</table>
### Datasets: Astronomy and Astrophysics Metadata

<table>
<thead>
<tr>
<th>Type</th>
<th>Facility</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image</td>
<td></td>
</tr>
<tr>
<td>Mosaic</td>
<td></td>
</tr>
<tr>
<td>EventList</td>
<td></td>
</tr>
<tr>
<td>Spectrum</td>
<td></td>
</tr>
<tr>
<td>Cube</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Object</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Spatial Resolution</th>
<th>Spectral Resolution</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Bandpass</th>
<th>Central Wavelength (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Wavelength Range Minimum (m)</th>
<th>Wavelength Range Maximum (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dataset Date Range Start</th>
<th>Dataset Date Range End</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Public vs Restricted
Terms, Licenses and Restrictions

- **Public Dataset**
  - CC0 License
  - Metadata is public
  - Files are public

- **Dataset with Restricted Files**
  - CC0 License
  - Metadata is public
  - Files are restricted
  - Access Terms are defined in dataset

- **Dataset with Terms of Use**
  - Metadata is public
  - Terms of Use are defined in dataset (CC0 can’t apply)
  - Files might be public or restricted
Workflows
Draft, Published and Versions

Upload Data

Draft Dataset

Published Dataset, v1

Draft

Published Dataset, v1.1

Published Dataset, v2

Dataset in review, can be shared with collaborators

Once published, dataset cannot be unpublished (only deaccessioned)

Minor version for small changes to dataset description

Major version for new versions of data files

Data Citation becomes public

Data Citation doesn’t change

Data Citation changes
Multiple Roles for Multiple Workflows

- **Editor**: Upload Data + Edit Metadata
- **Manager**: Upload Data + Edit Metadata + Set File Restrictions + License and Terms
- **Curator**: Upload Data + Edit Metadata + Set File Restrictions + License and Terms + Grant Access + Publish Dataset

+ Custom Roles
Data Processing, Analysis, and Visualizations
Download in Original format or Preservation format (does not depend on software package)

Tabular Data: Converted to Preservation format
Tabular Data: Explore and Analyze with TwoRavens
Geospatial Data: Visualize in WorldMap
Demo acknowledgement: Dwayne Liburd, Sonia Barbosa
Not only Expanding in Features, but also in Size

Federated Dataverse Installations:

874 Dataverses
55,539 Datasets
754,816 Files
1,173,733 Downloads

1. DANS - Netherlands
2. Fudan University
3. OCUL Scholars Portal
4. Abacus
5. University of Alberta Libraries
6. Harvard Dataverse
7. Heidelberg University
8. ODUM Institute
9. UiT The Arctic University of Norway
10. UnBraL Fronteiras

@dataverseorg
1,480 Followers

IQSS
www.iq.harvard.edu
datascience.iq.harvard.edu

Highcharts.com
What’s coming
Many Harvard researchers are subject to open-data policies from the journals publishing their articles or the agencies funding their research. Many others simply want to open up their data to realize the benefits of transparency, collaboration, data citation, research acceleration, and reproducibility. ODAP is a program to help them.

ODAP will offer advice and instruction on how to deposit data files in the Harvard Dataverse. When privacy is an issue, ODAP will offer advice on how to make data files as open as privacy constraints will allow. Since anyone in the world may deposit in Dataverse, ODAP’s online assistance should help researchers everywhere. However, when online assistance isn’t enough, ODAP staffs and volunteers will offer personal assistance to Harvard faculty, students, fellows, and postdocs. We encourage other institutions to offer personal assistance to their own researchers as well, and can work with them on how to do that.

If you’re interested in providing open access to your data, you should also be interested in providing open access to the research articles reporting your analysis and conclusions. If you’re at Harvard, we welcome your research publications, especially your scholarly articles, in our open-access repository, Digital Access to Scholarship at Harvard (DASH). For more details, see the Office for Scholarly Communication. However, the present web site is about opening access to data files, not opening access to texts.

What the Harvard Community is Saying

"Data is one of the most vital resources of the 21st century. Fortunately, with Harvard’s Dataverse, data associated with research can be stored and made accessible freely so others around the world can replicate studies and reuse it for new purposes. The Harvard Library will link publications in DASH, Harvard’s institutional repository, to data in Dataverse, enriching the pool of open access information."

Sarah Thomas
Vice President for the Harvard Library
Roy E. Larsen Librarian for the Faculty of Arts and Sciences
Beyond 4.0

- Integration with other Systems:
  - DASH
  - ORCID
  - Journal Systems (in addition to OJS)
  - Archivematica
  - iRODS

- Support for Sensitive Data:
  - Secure Storage
  - DataTags
  - Analysis with Privacy Preserving Algorithms

- Data Citation with Dataset Provenance
- Expanding APIs!
1st Annual Dataverse Community Meeting
Institute for Quantitative Social Science (IQSS) at Harvard University
Tuesday, June 9, 2015 at 9:00 AM - Thursday, June 11, 2015 at 5:00 PM (EDT)
Cambridge, MA

Registration Information

<table>
<thead>
<tr>
<th>TYPE</th>
<th>END</th>
<th>QUANTITY</th>
</tr>
</thead>
<tbody>
<tr>
<td>RSVP</td>
<td>Jun 9, 2015</td>
<td>Free</td>
</tr>
</tbody>
</table>

**Who's Going**

Connect to see which of your Facebook friends are going to 1st Annual Dataverse Community Meeting.

**Share 1st Annual Dataverse Community Meeting**

Email | Share | Tweet | Like

**Event Details**

You are cordially invited, as active contributors or interested members of the international Dataverse community, to take part in the 1st Annual Dataverse Community Meeting, which includes a Repository API workshop. The meeting will take place June 9-11, 2015 and will be hosted by the Institute for Quantitative Social Science (IQSS) at Harvard University.

**When & Where**

CGIS South, Tsai Auditorium
1730 Cambridge Street
Cambridge, MA 02138

Tuesday, June 9, 2015 at 9:00 AM
Thursday, June 11, 2015 at 5:00 PM (EDT)

Add to my calendar
Thank You

mcrosas@iq.harvard.edu

@mercecrosas

http://datascience.iq.harvard.edu/team