Data Exchange and Publishing: Data Sets, Supplemental Materials and Standards

Todd Carpenter
Managing Director, NISO
December 2011

ALCTS Webinar: Datasets for Publication, Standards and Issue
National Information Standards Organization

Non-profit industry trade association accredited by ANSI

110+ Organizational Members
Divided roughly in thirds among publishers, libraries and support service providers

Mission of developing and maintaining technical standards related to information, documentation, discovery and distribution of published materials and media

Volunteer driven organization: 400+ contributors who are spread out across the world
Standards are familiar, even if you don’t notice
Data Publishing
Increasingly, scientific breakthroughs will be powered by advanced computing capabilities that help researchers manipulate and explore massive datasets.
Science data isn’t what it used to be

Datasets for Publication: Standards and Issues
Challenges of Data Publishing/Sharing

Data

- is massive in scale.
- can be constantly changing.
- is difficult to describe/catalog.
- is often difficult to integrate with other data.
- can be complex to analyze (tools required).
- is hard to peer review.
- is challenging to curate.
- is challenging to preserve.
Data – Complexities

Storage
Provenance
Legal Issues
Technical Interoperability
Metadata
Privacy
Identification
Citation
The fact data sets are constantly changing poses the following problems (not exhaustive, BTW)

- How does someone rerun your experiment, when the data set isn’t now what it was when you ran your experiment
- For citation – You have to cite the subset of the data that existed on XXX date when you ran your analysis
- How can you preserve a snapshot of a dataset?
- How do you describe the dataset at that point in time?
- Problems of managing the metadata of each data deposit
- Problems of synchronization across repositories
“…The amount of information created, captured, or replicated exceeded available storage for the first time in 2007. Not all information created and transmitted gets stored, but by 2011, almost half of the digital universe will not have a permanent home.”

Data Complexity – Legal Issues

In US, you can’t copyright facts

Copyright in data collections
  – Form
  – Presentation
  – Privacy

When data crosses international lines, it assumes the legal characteristics of host country

*NODALITIES Magazine #9*: “Open data and the law” By Jordan S. Hatcher, Open Data Commons
Data Complexity – Provenance

• How can you be certain that the data set you are looking at hasn’t been altered or changed?
• Can you trust the data source?
• Has the data been updated/added to/or appended since the analysis was initially run
• Dealing with issues of abstraction
Data Citation
Creating a Culture

Building a Culture of Data Citation

CREATE

1. Australian researcher creates a research dataset and a publication related to the dataset.
2. Dataset is stored in a publicly accessible repository.
3. Researcher uses ANDS services to mint a Digital Object Identifier (DOI) for the dataset.
4. DOI is used in data citation.

REWARD

5. Citation metrics services (e.g., Scopus, Web of Knowledge) accumulate citation references to the dataset and publication.
6. Funding and research groups review publication and dataset citation metrics.
7. Researchers future funding and promotion influenced by dataset citation metrics.

USE

8. Research community use the DOI to access the dataset and carry out related research.
9. Research community generate new publications using the DOI to reference the dataset.

Source: ANDS Data Citation Guide
We all know what a citation looks like? Right?

### Scholarly Electronic Publishing Bibliography

#### 4.1 General Works: Research (Multiple-Types of Electronic Works)


But what about citing a data set?


Source: Citations for SEER Databases


Source: International Polar Year

How to Cite This Data Set

Citation Format: Author (Publication Date), Collection Name, Image Name, Processing Level, Publisher, Publisher Location, Product Coverage Date.

Citation Parameters Example:
- Author: NASA Landsat Program (specify otherwise if processed outside USGS)
- Publication Date: (specify processing date from metadata)
- Collection Name: Landsat (specify ETM+, TM or MSS) scene
- Image Name: (specify scene ID from the name or naming convention)
- Processing Level: (specify as indicated in metadata, for instance: L1G, GeoCover, Orthorectified, Terrain Corrected, or SLC-Off)
- Publisher: USGS (specify otherwise if processed outside USGS)
- Publisher Location: Sioux Falls (specify otherwise if processed outside USGS)
- Product Coverage Date: (specify acquisition date from metadata or naming convention)


Source: Global Land Cover Facility


Source: ICPSR

Income inequality

Source: OECD

Source: The Economist


Source: ICPSR


Source: ICPSR


Source: ICPSR


Source: ICPSR


Source: ICPSR
CODATA Group on Data Citation

- Approved by the CODATA 27th General Assembly in Cape Town in October, 2010

TASKS

- Survey existing literature and existing data citation initiatives.
- Obtain input from stakeholders in library, academic, publishing, and research communities.
- Hold at least one meeting and a workshop to help establish a solid foundation of the state of the art and practices in this area.
- Work with the ISO and major regional and national standards organizations to develop formal data citation standards and good practices.
Issues Requiring Attention

- **Technical**
  - Metadata Standards

- **Scientific**
  - Publication and peer review

- **Institutional**
  - Promotion & Tenure
  - Reward infrastructure

- **Financial**
  - Infrastructure support

- **Sustainability**

- **Identifiers**
  - DOI, PURL, ARC

- **Legal/Intellectual Property Rights**
  - Accessibility and reuse rights

- **Socio-cultural**
  - Culture of citing data
  - Proper attribution
Accomplishments so far

- Compiled a significant bibliography
- Hosted a two-day meeting in Berkeley, CA on these issues in August 2011
- Formulated a survey for repository managers, publishers and funding organizations

**Next steps:**
- Undertake survey
- Report out of Berkeley meeting
- Begin drafting report => Due in October
Supplemental Materials
How pressing is this problem?
Supplemental Materials Project Background

• Began with a short survey conducted by Sasha Schwarzman at AGU in fall 2009
• Distributed results on CrossRef tech listserv
• Lots of discussion during CrossRef ‘09 fall meeting
• Organization of thought leader meeting in DC
• Partnership with NFAIS and draft proposal
• Approved as Joint WG in spring 2010
• WG is nearing completion of draft RP
• Public comment period in 1Q 2012
Supplemental Materials Project

Goal is to focus on the publication structures necessary for communicating supplemental materials as part of the publications process

NOT Addressing every and all data questions
Content that users want & need?

• The core questions:
  – What is critical to understanding?
  – What is supplemental to understanding?
  – What is ancillary to understanding?

The form of the content, i.e., text, video, audio, data, applications, is **NOT** the key to whether something is supplemental
## Content types: Integral, Additional, Related

<table>
<thead>
<tr>
<th>Importance (function)</th>
<th>Integral content</th>
<th>Additional content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Curated by</td>
<td>Pseudo supplemental – essential but treated <em>as if it were</em> supplemental</td>
<td>Truly supplemental – relevant and useful but optional</td>
</tr>
<tr>
<td><strong>Publisher</strong></td>
<td>Multimedia</td>
<td>Multimedia</td>
</tr>
<tr>
<td></td>
<td>Chemical structures</td>
<td>Chemical structures</td>
</tr>
<tr>
<td></td>
<td>Crystallographic structures</td>
<td>Crystallographic structures</td>
</tr>
<tr>
<td></td>
<td>Datasets (smaller)</td>
<td>Datasets (smaller)</td>
</tr>
<tr>
<td></td>
<td>Computer algorithms</td>
<td>Computer algorithms</td>
</tr>
<tr>
<td><strong>Institutional repository or Official data center</strong></td>
<td>Genetic sequences, Protein structures, Crystallographic structures, Datasets (large), etc.</td>
<td>Related content</td>
</tr>
<tr>
<td><strong>Individual</strong></td>
<td>Ephemeral – not appropriate for hosting supplemental materials</td>
<td></td>
</tr>
</tbody>
</table>

*Not integrated into the version of record*
### Some draft business guidelines

<table>
<thead>
<tr>
<th>Section</th>
<th>Integral content</th>
<th>Additional content</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Selecting / Peer reviewing</strong></td>
<td>At the same level as core article</td>
<td>May not be reviewed at the same level</td>
</tr>
<tr>
<td><strong>Copyediting</strong></td>
<td>At the same level as core article. Should be noted if not</td>
<td>May not be edited at the same level. If so, should be noted</td>
</tr>
<tr>
<td><strong>Referencing within article</strong></td>
<td>Cite/link at the same level as table or fig. No ref. list entry, for this content is part of article</td>
<td>Provide in-text citation and link at the appropriate point in text, rather than at the end</td>
</tr>
<tr>
<td><strong>Citing from other pubs</strong></td>
<td>Not to be cited separately. Cite article as a whole</td>
<td>Can be cited separately</td>
</tr>
<tr>
<td><strong>References within sup. mat.</strong></td>
<td>Integrate references into the ref. list of the core article</td>
<td>Keep references separate from the core article ref. list</td>
</tr>
</tbody>
</table>
# Some draft business guidelines

## Preserving
- Preserve at the same level as the core article
- Provide the same metadata markup
- Include in migration plans

## Intellectual property rights
- Treat rights in the same manner as the rights for the core article
- Anyone who has access to online article should also have access to Integral content

### Additional content
- Take preservation into consideration when accepting
- If uncertain about preservation, have author submit to a trusted repository and link to it
- Determination of rights for Additional content may differ and should be transparent to users
Technical Practice - Themes

**Metadata Schema**
- Descriptive Metadata
- Physical Metadata
- Packaging/Manifest

**Identifiers & Linking**

**Grouping Objects**

**Challenges**
- Granularity
- Recurrence
- Relationships
- Heterogeneity
- Hierarchy

**Overall Issues**
- Perceived importance
- Costs
- Business models
What does the future hold?
Too many players, doing too many things?

Source: [XKCD](https://xkcd.com)

December 7, 2011
Datasets for Publication: Standards and Issues
Many potential areas for work in sharing of scientific data including:

Systemic metadata
   What is the form of this information?
   What are its structural components?

Archival issues
   Storage, physical level, but also migration issues

Bibliographic information for discovery, delivery and reuse

Rights issues – Ownership, recognition, sharing, privacy

Data Equivalence – How does one know that this thing and that are equivalent (i.e., contain same data)?
DataCite

- International consortium
  - Establish easier access to scientific research data on the Internet
  - Increase acceptance of research data as legitimate, citable contributions to the scientific record
  - To support data archiving
  19 members from 11 countries
Other Initiatives & Links

- CODATA/ICSTI – Task Group on Data Citation Standards and Practices
- DataCite – datacite.org
- NISO/NFAIS Group on Supplemental Journal Materials
- National Academies Board on Research Data and Information
  - For Attribution: Developing Data Attribution and Citation Practices and Standards in Berkely, CA. August 2011 Report due in January
- Cite Datasets and Link to Publications by Digital Curation Centre
- ScienceCommons - Scholar’s Copyright Project
- International Council for Scientific and Technical Information (ICSTI)
  - Multimedia Search and Retrieval, and Interactive Journal Articles Projects
- Dublin Core Metadata Initiative - Science and Metadata Community
- DataNet projects funded by NSF, launched in 2009
- NISO/OAI ResourceSync project funded by Alfred P. Sloan Foundation
- And on, and on, and on, and on…..
Information Standards Quarterly (ISQ)

Summer 2010 Issue of ISQ
Special issue on Enhanced Journal Articles
Article-level Enhancements in the Humanities and Social Sciences
Hosting Supplemental Materials
Archiving Supplemental Materials
DOIs - Linking and Beyond
Supplemental Materials Survey
Report on NISO/NFAIS project

NOW AVAILABLE
OPEN ACCESS
www.niso.org/publications/isq