Data Management

Jon Wheeler jwheel01@unm.edu
College of University Libraries and Learning Sciences
University of New Mexico
Objectives

- Understand the impact of data sharing requirements on data management strategies
- Characterize challenges across the data lifecycle
- Identify broadly applicable tools and resources for addressing these challenges
- Identify strategies and resources for efficient data workflow
First Thoughts

In the course of your PhD research, what did you learn about data management that you wished someone had told you before you started?

What do you do differently now? Specifically, which strategies and tools do you use to better manage data?
Data Lifecycle

Plan → Collect → Assure → Describe → Preserve → Discover → Integrate → Analyze → Plan

Image credit DataONE
Challenges Across the Lifecycle

- Organization
- Data entry & processing
- Documentation
- Storage & security

Source: John Gantz, IDC Corporation: The Expanding Digital Universe
Image accessed from DataONE
Data Sharing

Benefits to Self, Science, and Society

- Recognition and reciprocation
- Interdisciplinary and collaborative research
- Research compliance
- Improved data quality, transparency, and trust
- Increased efficiency and innovation
- Better informed decision and policy making

Federal Mandates

- February 2013 OSTP Memo, “Expanding Public Access to the Results of Federally Funded Research”
- May 2013 Executive Order, “Making Open and Machine Readable the New Default for Government Information”
- Public Access Plans
Thinking Ahead

How will the tools and strategies discussed earlier help you to meet Federal requirements for sharing and potentially preserving data? How will they help you meet your own goals?

What additional support and resources are required?
Data Management Plans

- Scale organization and management strategies - improve efficiency
- Support data sharing, publishing, and preservation
- Facilitate re-use

- In January, 2011, the NSF began requiring a (maximum) 2 page data management plan (DMP) to be submitted with all proposals.
- Public access plans of other agencies include DMP requirements.
Components of a DMP

- Data and data formats
- Metadata and documentation
- Policies for access, sharing, and reuse
- Long term storage and management
  - Data protection and privacy
  - Archiving and preservation
- Budget

The DMP Tool is free to use and includes templates for many agency requirements. Customized information for your institution may be available.

https://dmptool.org/
Organization

Strategies

- Roles and responsibilities
- Map needed skills to available staff and identify gaps
- Develop training plans
- Assign responsibilities and monitor
- File management
  - Consistent content
  - Separate data from analysis
  - Keep raw data separate

Resources

- File plan
- File naming formats
- Shell scripts, scripting languages
- Bulk Rename Utility: [http://www.bulkrenameutility.co.uk/Main_Intro.php](http://www.bulkrenameutility.co.uk/Main_Intro.php)
- Open Science Framework [https://osf.io/](https://osf.io/)

Slide credit DataONE
OSF Integrations

- Google Drive
- Dropbox
- Github
- AWS
- Figshare
- Dataverse
UNM Public Services /

RDS Grey Data Management

Contributors: Jon Wheeler, Jacob Nash, Karl Benedict, Jon Wheeler
Date created: 2016-03-16 02:30 PM | Last Updated: 2016-04-07 03:02 PM
Category: Project

Description:
Project space for Public Services collaboration demo and activity.
License: CC By Attribution 4.0 International

Curating Grey Data

Problem Statement
As a type of scholarly communication, data may be taken as a subset of grey literature. Whether or not this is a consensus view (it isn’t!), it is true that data possess many grey characteristics:
- Commonly not subject to peer review
- Primarily used to support internal initiatives
- Document administrative or business processes

Further, this is not really a safe...

Citation

APA

MLA

Chicago

Get more citations
Enter citation style (e.g., "APA")
Metadata

Metadata is: Data ‘reporting’

- **WHO** created the data?
  - Credit researchers and sponsors, document responsibilities and roles

- **WHAT** is the content of the data?
  - What was measured, units, aggregation

- **WHEN** were the data created?
  - Date, time (structured, consistent, standards-based)

- **WHERE** is it geographically?
  - Geographic location (define datum, coordinate system, method)

- **HOW** were the data developed?
  - Instruments, sensors, algorithms, models, software

- **WHY** were the data developed?
  - Purpose for data collection, suggested use, known limitations

- **Access** requirements?
  - Licensing terms, embargo, redistribution, modification

Slide credit DataONE and Karl Benedict 2014
You are starting a new study, and you find a publication that is based on data key to your analysis...
Scenario 1

High-Resolution Global Maps of 21st-Century Forest Cover Change


Quantification of global forest change has been lacking despite the recognized importance of forest ecosystem services. In this study, Earth observation satellite data were used to map global forest loss (2.3 million square kilometers) and gain (0.8 million square kilometers) from 2000 to 2012 at a spatial resolution of 30 meters. The tropics were the only climate domain to exhibit a trend, with forest loss increasing by 2301 square kilometers per year. Brazil’s well-documented reduction in deforestation was offset by increasing forest loss in Indonesia, Malaysia, Paraguay, Bolivia, Zambia, Angola, and elsewhere. Intensive forestry practiced within subtropical forests resulted in the highest rates of forest change globally. Boreal forest loss due largely to fire and forestry was second to that in the tropics in absolute and proportional terms. These results depict a globally consistent and locally relevant record of forest change.

Figure 1. (A) Tree cover, (B) forest loss, and (C) forest gain. A color composite of tree cover in green, forest loss in red, forest gain in blue, and forest loss and gain in magenta is shown in (D), with loss and gain enhanced for improved visualization. All map layers have been resampled for display purposes from the 30-m observation scale to a 0.05° geographic grid.

High-Resolution Global Maps of 21st-Century Forest Cover Change
M. C. Hansen et al.
Science 342, 850 (2013);
DOI: 10.1126/science.1244693

http://science.sciencemag.org/content/342/6160/850
Rate from 1 (impossible) to 5 (easy) the following...

<table>
<thead>
<tr>
<th></th>
<th>1 (impossible)</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5 (easy)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Discovery</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Access</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Understanding</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Use</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Slide credit Karl Benedict 2014
Scenario 2

Global Forest Change 2000–2014

Data Download

Results from time-series analysis of Landsat images characterizing global forest extent and change from 2000 through 2014. For additional information about these results, please see the associated journal article (Fleming et al., Science 2013).

Web-based visualizations of these results are also available at our main site:

http://earthenginepartners.appspot.com/science-2013-global-forest

Please use this URL when linking to this dataset.

We anticipate releasing updated versions of this dataset. To keep up-to-date with the latest updates, and to help us better understand how these data are used, please register as a user. Thank you!

User Notes for Version 1.2 Update

This update of gross forest cover loss includes a new 2015 loss layer. Relative to the version 1.0 product, our method has been modified in numerous ways, and the new update should be seen as part of a transition to a future version 2.0 that is more consistent over the entire 2000-2014 period. Key changes include:

1. The use of Landsat 8 OLI data for 2014 instead.
2. The reprocessing of data from 2011 to 2012 in correcting load loss.
3. Improved masking data for calibrating the load model.
4. Improved forest clearing assessment models to filter agu data, and
5. Improved input spectral features for building and applying the load model.

These changes lead to a different and improved direction of global forest loss. However, the years preceding 2014 have not been reprocessed in this manner, and users will notice inconsistencies as a result. It must also be noted that a full validation of the results incorporating Landsat 8 has not been undertaken. Such an analysis may reveal a more sensitive ability to detect and map forest disturbance with Landsat 8 data. If this is the case then there will be more fundamental limitations to the consistency of the mapped interannual loss before and after the inclusion of Landsat 8 data, and a validation of Landsat 8 incorporated load detection is planned. The integrated use of version 1.0 2000–2012 data and updated version 1.2 2011–2014 data should be performed:


http://earthenginepartners.appspot.com/science-2013-global-forest

Slide credit Karl Benedict 2014
<table>
<thead>
<tr>
<th>Questions</th>
<th>1 (impossible)</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5 (easy)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Discovery</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Access</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Understanding</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Use</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Rate from 1 (impossible) to 5 (easy) the following...
Metadata Enables

- Discovery
- Use
- Understanding

Welcome to the LTER Network Data Portal

Data are one of the most valuable products of the Long Term Ecological Research (LTER) Network. Data and metadata derived from publicly funded research in the U.S. LTER Network are made available online with as few restrictions as possible, on a non-discriminatory basis. In return, the LTER Network expects data users to act ethically by contacting the investigator prior to the use of data for publication.

The LTER Network Information System Data Portal contains ecological data packages contributed by past and present LTER sites. Please review the LTER Data Policy before downloading any data product. We request that you cite data sources in your published and unpublished works whenever possible. Digital object identifiers (DOI) are provided for each dataset to facilitate citation.

LTER Network scientists make every effort to release data in a timely fashion and with attention to accurate, well-designed and well-documented data. To understand data fully, please read the associated metadata and contact data providers if you have any questions.

https://portal.lternet.edu/nis/home.jsp
**Description & Documentation**

**Strategies**

- Incorporate metadata creation early and across all areas of a project
- Budget for metadata creation and consult with experts
- Use a standardized metadata format
  - Dublin Core
  - Darwin Core
  - Ecological Metadata Language (EML)
  - ISO 19115
- Use a controlled vocabulary for keywords

**Resources**

- Collectica for Excel
  [http://www.colectica.com/software/colecticaforexcel](http://www.colectica.com/software/colecticaforexcel)
- Morpho
  [https://knb.ecoinformatics.org/#tools/morpho](https://knb.ecoinformatics.org/#tools/morpho)
- You institution’s library!
  [https://www.openicpsr.org/openicpsr/project/100379/version/V1/view](https://www.openicpsr.org/openicpsr/project/100379/version/V1/view)
Policies for Access, Sharing, & Reuse
Preservation

- Assignment of permanent identifiers (DOI, Handle, etc.)
- Data format conversion and migration
- Metadata enrichment, for example provenance information
DataDryad.org is a curated general-purpose repository that makes the data underlying scientific publications discoverable, freely reusable, and citable. Dryad is an international repository of data underlying peer-reviewed scientific and medical literature, particularly data for which no specialized repository exists. The content is considered to be integral to the published research. All material in Dryad is associated with a scholarly publication.

Eucalyptus Turbinata Seed Morphology
http://www.re3data.org/search?query=dryad
Choosing What to Preserve

- Essential and unique
- Well documented
- Known provenance and ownership
- Supports published research
- Sensitivity and intellectual property
- Completeness
Repositories

## Domain Specific or General Purpose

- Genbank
- LTER Network Data Portal
- FigShare
- Dryad
- Institutional Repositories

## Considerations

- Cost
- Self service or mediated?
- Services
  - DOI or other permanent ID?
  - Format support?
  - Metadata enrichment or validation?
- Integration with other services
- Indexing and reporting
- Licenses

Slide credit New England Collaborative Data Management Curriculum
Focused Resources
Digital Research Tools

http://dirtdirectory.org/
Data Entry & Processing

Strategies

- Use descriptive column and file names
- Use open or non-proprietary data formats
  - Uncompressed text
- Enforce data constraints and validation
- Explicitly encode missing data, and document that encoding
- Use meaningful column headings (short, no spaces)
- Include units
- Provide a data dictionary

Resources

- Spreadsheets
- Relational databases
  - SQLite [https://sqlite.org/](https://sqlite.org/)
- R [https://www.r-project.org/](https://www.r-project.org/)
Analysis & Workflow

Objectives & Strategies

- Facilitate reproducible science
- Increase efficiency and transparency
- Document and preserve
  - Data provenance
  - Inputs & outputs
  - Settings & parameters

Resources

- Kepler [https://kepler-project.org/](https://kepler-project.org/)
- VisTrails [https://www.vistrails.org/index.php/Main_Page](https://www.vistrails.org/index.php/Main_Page)
- myExperiment [http://www.myexperiment.org/home](http://www.myexperiment.org/home)
Storage & Security

Strategies

● Create a detailed backup policy
  ○ Which data?
  ○ How often?
  ○ Where?
● Verify backups
● Use non-proprietary, standard formats
● Backup multiple copies to multiple machines in multiple locations

Resources

● Robocopy (Windows)
● Time Machine (Mac)
● Rsync (Linux)
Backup != Archiving
Thank You
Resources

Benedict, Karl. Data Management Primer. University of New Mexico


DataONE Education Module: Data Quality Control and Assurance. DataONE. Jan 6, 2017. From http://www.dataone.org/sites/all/documents/L05_DataQualityControlAssurance.pptx


