

# Data Management

---

Jon Wheeler [jwheel01@unm.edu](mailto: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

**Context**

# 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

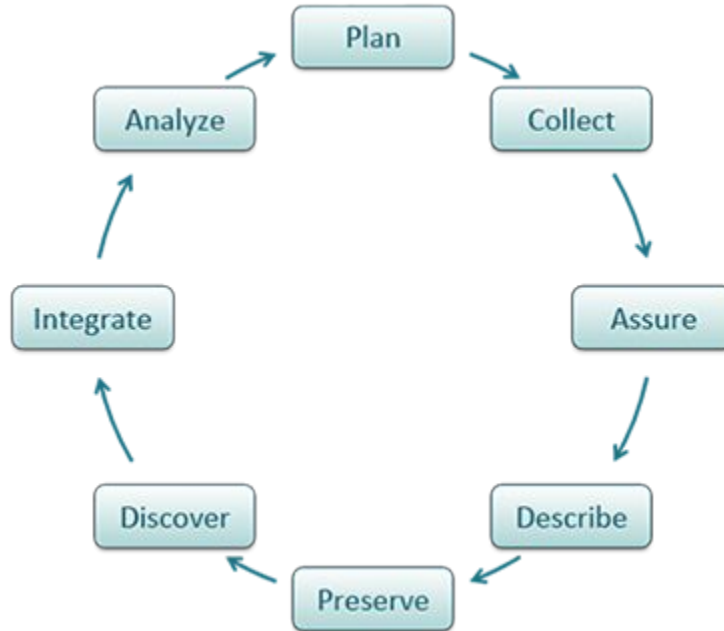
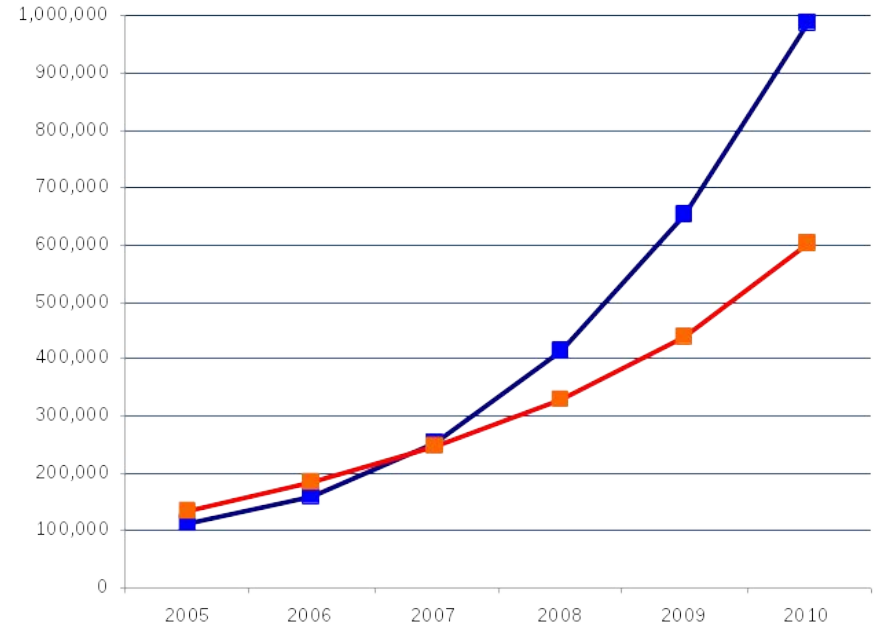


Image credit DataONE

# Challenges Across the Lifecycle

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



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

Information versus Available Storage

# 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?



# **Big Picture Resources**

# 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

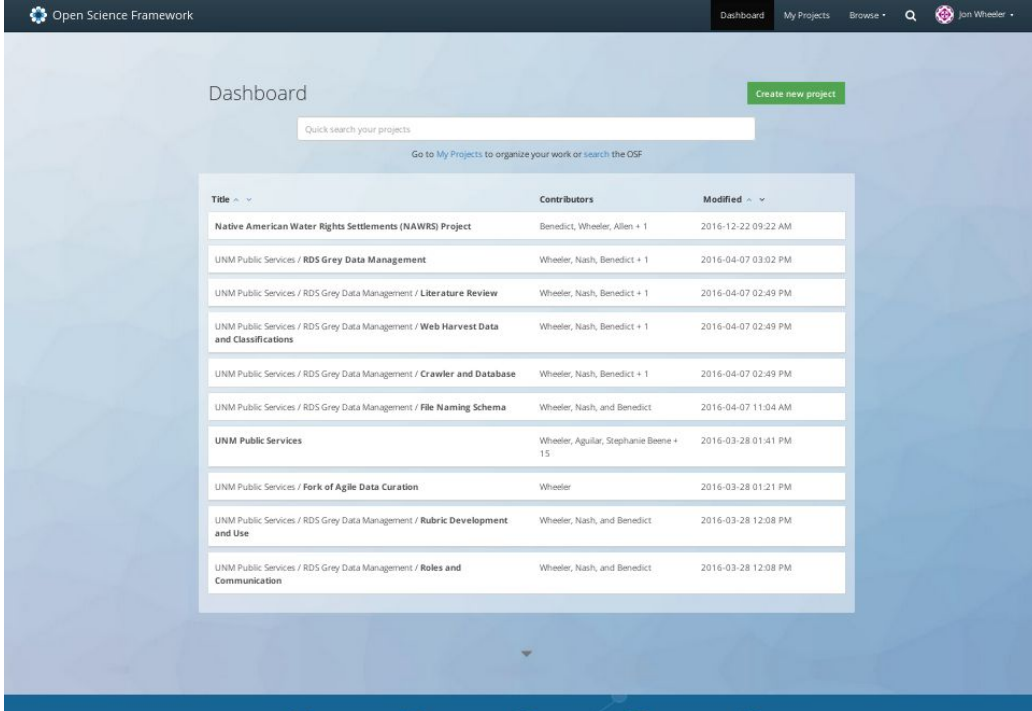
- Roles and responsibilities
- Map needed skills to available staff and identify gaps
- Develop training plans
- Assign responsibilities and monitor
- File management
  - Consistent content
  - Separate data from analysis
  - Keep raw data separate

## 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/>

# OSF Integrations

- Google Drive
- Dropbox
- Github
- AWS
- Figshare
- Dataverse



The screenshot displays the OSF dashboard for user Jon Wheeler. The page features a navigation bar with 'Open Science Framework', 'Dashboard', 'My Projects', and a search icon. A 'Create new project' button is visible in the top right. Below the navigation, there is a search bar and a link to 'Go to My Projects to organize your work or search the OSF'. The main content is a table of projects with columns for Title, Contributors, and Modified.

Title	Contributors	Modified
<b>Native American Water Rights Settlements (NAWRS) Project</b>	Benedict, Wheeler, Allen + 1	2016-12-22 09:22 AM
UNM Public Services / <b>RDS Grey Data Management</b>	Wheeler, Nash, Benedict + 1	2016-04-07 03:02 PM
UNM Public Services / RDS Grey Data Management / <b>Literature Review</b>	Wheeler, Nash, Benedict + 1	2016-04-07 02:49 PM
UNM Public Services / RDS Grey Data Management / <b>Web Harvest Data and Classifications</b>	Wheeler, Nash, Benedict + 1	2016-04-07 02:49 PM
UNM Public Services / RDS Grey Data Management / <b>Crawler and Database</b>	Wheeler, Nash, Benedict + 1	2016-04-07 02:49 PM
UNM Public Services / RDS Grey Data Management / <b>File Naming Schema</b>	Wheeler, Nash, and Benedict	2016-04-07 11:04 AM
<b>UNM Public Services</b>	Wheeler, Aguilar, Stephanie Beene + 15	2016-03-28 01:41 PM
UNM Public Services / <b>Fork of Agile Data Curation</b>	Wheeler	2016-03-28 01:21 PM
UNM Public Services / RDS Grey Data Management / <b>Rubric Development and Use</b>	Wheeler, Nash, and Benedict	2016-03-28 12:08 PM
UNM Public Services / RDS Grey Data Management / <b>Roles and Communication</b>	Wheeler, Nash, and Benedict	2016-03-28 12:08 PM



Private

Make Public



0

## UNM Public Services /

## RDS Grey Data Management

Contributors: [Jon Wheeler](#), [Jacob Nash](#), [Karl Benedict](#), [Jon Wheeler](#)

Date created: 2016-03-18 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

Wiki 

## 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 saf...

[Read More](#)

Files Citation osf.io/m4qre ^

## APA

Wheeler, J., Nash, J., Benedict, K. K., & Wheeler, J. (2016, April 7). RDS Grey Data Management. Retrieved from [osf.io/m4qre](https://osf.io/m4qre)

## MLA

Wheeler, Jon et al. "RDS Grey Data Management." Open Science Framework, 7 Apr. 2016. Web.

## Chicago

Wheeler, Jon, Jacob Nash, Karl K Benedict, and Jon Wheeler. 2016. "RDS Grey Data Management." Open Science Framework. April 7. [osf.io/m4qre](https://osf.io/m4qre).

## Get more citations

## Components

[Add Component](#)[Link Projects](#)

Files



Click on a storage provider or drag and drop to upload

🔍 Filter



Name ^ ▾ Modified ^ ▾

📁 RDS Grey Data Management	
- 🌀 OSF Storage	
- 🗣 Roles and Communication	
- 🌀 OSF Storage	
📄 jw-greydata-roles.md	2016-03-28 05:29 PM
- 🗄 File Naming Schema	
- 🌀 OSF Storage	
- 🗄 Literature Review	
- 🌀 OSF Storage	
- 🖨 Crawler and Database	
- 🌀 OSF Storage	
- 🌐 Google Drive: crawler	
+ 📁 .git	
📄 jw-cd-pyCrawler.py	2016-04-07 06:07 PM

Components

Add Component

Link Projects

🔒 🗣 Roles and Communication ▾

Wheeler, Nash & Benedict  
4 contributions

🔒 🗄 File Naming Schema ▾

Wheeler, Nash & Benedict  
9 contributions

🔒 🗄 Literature Review ▾

Wheeler, Nash, Benedict & 1 more  
7 contributions

🔒 🖨 Crawler and Database ▾

Wheeler, Nash, Benedict & 1 more  
13 contributions

🔒 📄 Web Harvest Data and Classifications ▾

Wheeler, Nash, Benedict & 1 more  
8 contributions

🔒 📝 Rubric Development and Use ▾

Wheeler, Nash & Benedict  
3 contributions

🔒 📺 Media ▾

Wheeler

# 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



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

M. C. Hansen,<sup>1\*</sup> P. V. Potapov,<sup>1</sup> R. Moore,<sup>2</sup> M. Hancher,<sup>2</sup> S. A. Turubanova,<sup>1</sup> A. Tyukavina,<sup>1</sup> D. Thau,<sup>2</sup> S. V. Stehman,<sup>3</sup> S. J. Goetz,<sup>4</sup> T. R. Loveland,<sup>5</sup> A. Kommareddy,<sup>6</sup> A. Egorov,<sup>6</sup> L. Chini,<sup>1</sup> C. O. Justice,<sup>1</sup> J. R. G. Townshend<sup>1</sup>

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 2101 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.

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

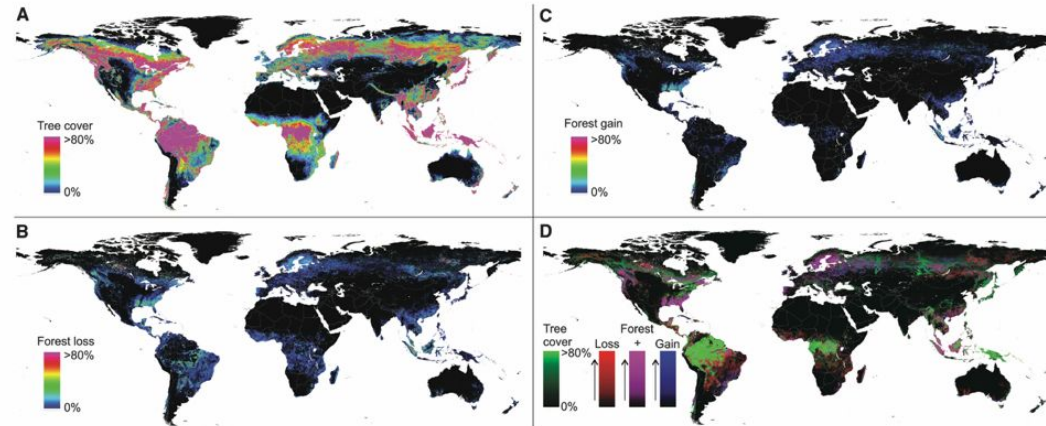
M. C. Hansen et al.

Science 342, 850 (2013);

DOI: 10.1126/science.1244693

<http://science.sciencemag.org/content/342/6160/850>

Slide credit Karl Benedict 2014



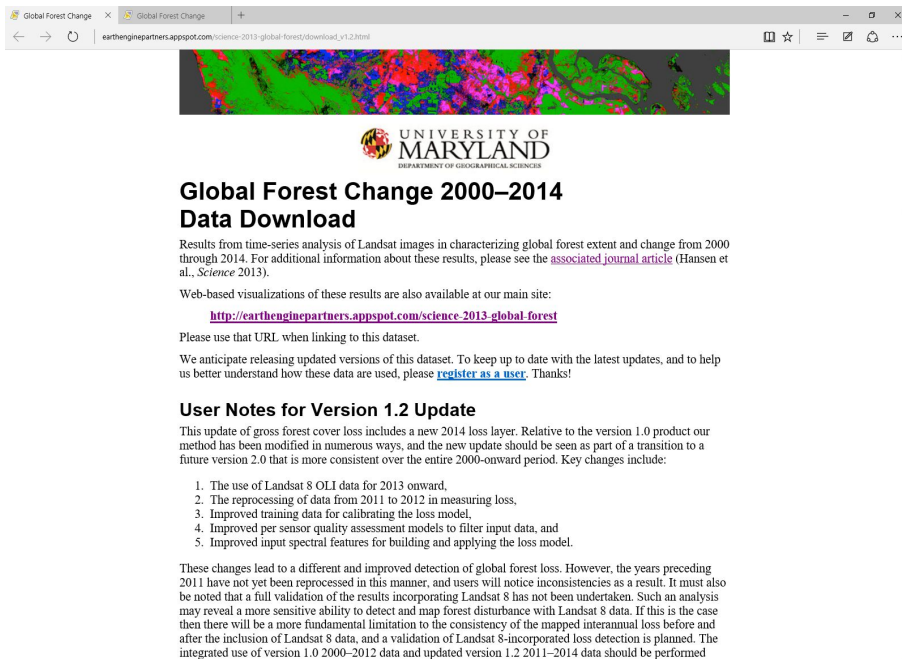
**Fig. 1. (A) Tree cover, (B) forest loss, and (C) forest gain.** A color composite of tree cover in green, forest loss in red, forest gain in blue, and forest loss and gain in magenta is shown in (D), with loss and gain enhanced for improved visualization. All map layers have been resampled for display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

# Questions

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


	1 (impossible)	2	3	4	5 (easy)
Data Discovery					
Access					
Understanding					
Use					

# Scenario 2



Global Forest Change

earthenginepartners.appspot.com/science-2013-global-forest/download\_v1.2.html



## Global Forest Change 2000–2014 Data Download

Results from time-series analysis of Landsat images in characterizing global forest extent and change from 2000 through 2014. For additional information about these results, please see the [associated journal article](#) (Hansen 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 that 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](#). Thanks!

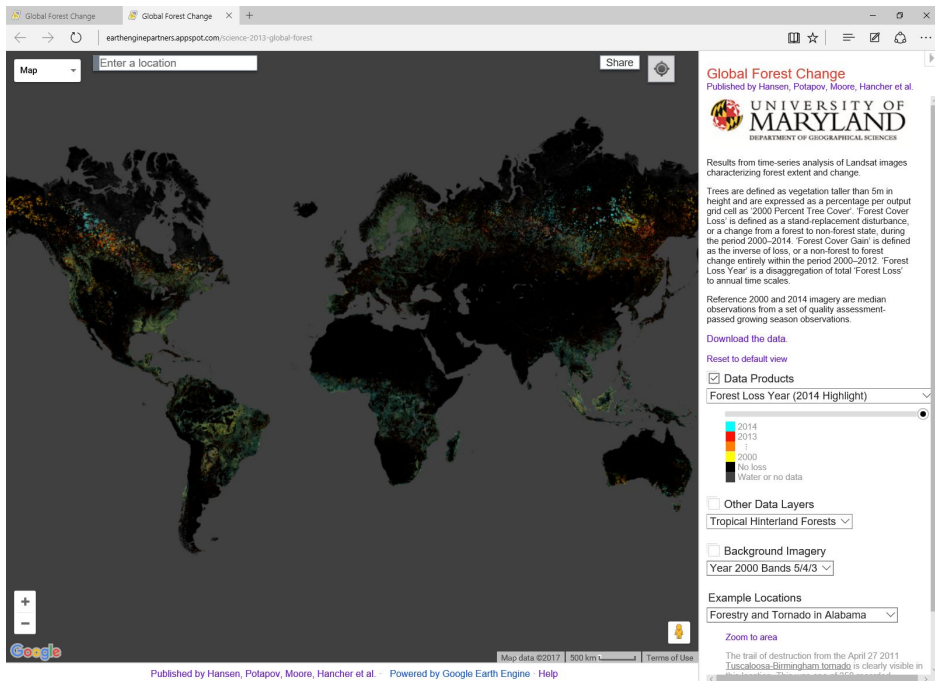
### User Notes for Version 1.2 Update

This update of gross forest cover loss includes a new 2014 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-onward period. Key changes include:

1. The use of Landsat 8 OLI data for 2013 onward,
2. The reprocessing of data from 2011 to 2012 in measuring loss,
3. Improved training data for calibrating the loss model,
4. Improved per sensor quality assessment models to filter input data, and
5. Improved input spectral features for building and applying the loss model.


These changes lead to a different and improved detection of global forest loss. However, the years preceding 2011 have not yet 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 a more fundamental limitation to the consistency of the mapped interannual loss before and after the inclusion of Landsat 8 data, and a validation of Landsat 8-incorporated loss 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/download\\_v1.2.html](http://earthenginepartners.appspot.com/science-2013-global-forest/download_v1.2.html)



Global Forest Change

Published by Hansen, Potapov, Moore, Hancher et al.



UNIVERSITY OF MARYLAND  
DEPARTMENT OF GEOGRAPHICAL SCIENCES

Results from time-series analysis of Landsat images characterizing forest extent and change.

Trees are defined as vegetation taller than 5m in height and are expressed as a percentage per output grid cell as '2000 Percent Tree Cover'. 'Forest Cover Loss' is defined as a stand-replacement disturbance, or a change from a forest to non-forest state, during the period 2000–2014. 'Forest Cover Gain' is defined as the inverse of loss, or a non-forest to forest change entirely within the period 2000–2012. 'Forest Loss Year' is a disaggregation of total 'Forest Loss' to annual time scales.

Reference 2000 and 2014 imagery are median observations from a set of quality assessment-passed growing season observations.

Download the data.

Reset to default view

Data Products

Forest Loss Year (2014 Highlight)

- 2014
- 2013
- 2000
- No loss
- Water or no data

Other Data Layers

Tropical Hinterland Forests

Background Imagery

Year 2000 Bands 5/4/3

Example Locations

Forestry and Tornado in Alabama

Zoom to area

The trail of destruction from the April 27 2011 Tuscaloosa-Birmingham tornado is clearly visible in

Published by Hansen, Potapov, Moore, Hancher et al. Powered by Google Earth Engine Help

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

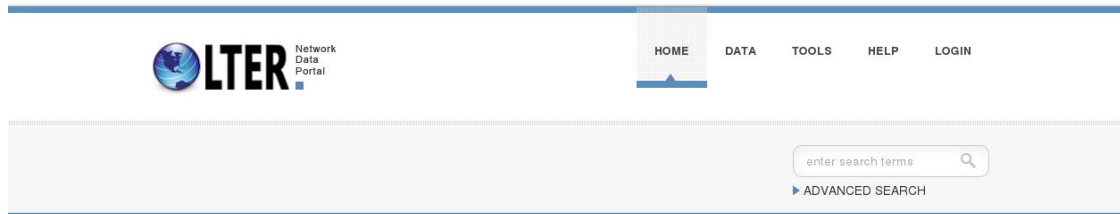
# Questions

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

	1 (impossible)	2	3	4	5 (easy)
Data Discovery					
Access					
Understanding					
Use					

# Metadata Enables

- Discovery
- Use
- Understanding

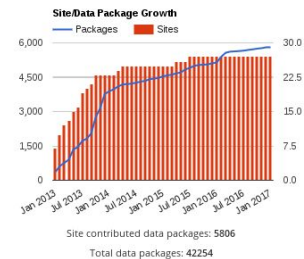


## Welcome to the LTER Network Data Portal

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>
- **Morpho**  
<https://knb.ecoinformatics.org/#tools/morpho>
- You institution's **library!**  
<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

Countries AID systems API Data access Data access restrictions Database access Database licenses Data licenses Data upload Data upload restrictions Enhanced publication Institution responsibility type Institution type Keywords Metadata standards PID systems Provider types Quality management Repository languages Software Syndications Repository types Versioning 

Found 3 result(s)

## DRYAD



Subject(s)

Life Sciences Natural Sciences Social and Behavioural Sciences Biology Zoology Medicine

Microbial Ecology and Applied Microbiology Virology Agriculture, Forestry, Horticulture and Veterinary Medicine

General Genetics Bioinformatics and Theoretical Biology Plant Ecology and Ecosystem Analysis

Evolution, Anthropology Biochemistry and Animal Physiology Geology and Palaeontology

Humanities and Social Sciences Microbiology, Virology and Immunology

Agriculture, Forestry, Horticulture and Veterinary Medicine Basic Biological and Medical Research Plant Sciences

Geosciences (including Geography)

Content type(s)

Standard office documents Scientific and statistical data formats Plain text Structured text

Software applications Source code other

Country

United States United Kingdom International

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

## Ecological Archives



esa's ecological archives

Subject(s)

Life Sciences Biology Agriculture, Forestry, Horticulture and Veterinary Medicine Chemistry

<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

**Focused  
Resources**

# Digital Research Tools

The screenshot shows the homepage of the Digital Research Tools (DIRT) directory. At the top, there is a navigation bar with links for 'About', 'Tools', 'Contribute', and 'Users'. The main header features the DIRT logo, which consists of the letters 'DIRT' in a bold, sans-serif font, with a lightbulb icon integrated into the letter 'I'. Below the logo, the text 'Digital Research Tools' is displayed. A search bar is located in the top right corner, with a 'Search' button. The main content area is divided into several sections: a 'Welcome // ' section with a paragraph describing the directory's purpose; a section titled 'I NEED A DIGITAL RESEARCH TOOL TO...' which lists various tool categories in two columns; a 'LANGUAGES' section with links for 'English' and 'Español'; an 'ABOUT' section with a paragraph and a '(more)' link; and a 'NEWS' section with three news items, each including a date and a '(more)' link.

About Tools Contribute Users

## DIRT

Digital Research Tools

Welcome //

The DIRT Directory is a registry of digital research tools for scholarly use. DIRT makes it easy for digital humanists and others conducting digital research to find and compare resources ranging from content management systems to music OCR, statistical analysis packages to mindmapping software.

**I NEED A DIGITAL RESEARCH TOOL TO...**

Analyze data	Interpret data
Annotate	Model data
Archive data	Analyze networks between my data
Capture information	Organize data
Clean up data	Preserve data
Collaborate	Program
Comment	Publish
Communicate	Record audio/video
Analyze the content of my data	Analyze relationships between pieces of data
Contextualize data	Share
Convert files	Analyze the geographical aspect of my data
Create	Store data
Crowdsource data enrichment/analysis	Analyze the structure of my data

Search

**LANGUAGES**

- English
- Español

**ABOUT**

The DIRT Directory is a registry of digital research tools for scholarly use. (more)

**NEWS**

DIRT plugin available for Commons In A Box (CBOX) Scholarly Network  
27 Mar 2015

DIRT partners with TAPoR to provide "recipes"  
27 Mar 2015

Bring DIRT into your classroom with our "assignment-in-a-box"  
26 Mar 2015

more

<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
- **OpenRefine** <http://openrefine.org/>
- Relational databases
  - **SQLite** <https://sqlite.org/>
- **R** <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/>
- VisTrails  
[https://www.vistrails.org/index.php/Main\\_Page](https://www.vistrails.org/index.php/Main_Page)
- myExperiment <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

Creamer, Andrew, Donna Kafel, Elaine Martin, Regina Raboin. New England Collaborative Data Management Curriculum Module 1: Overview of Research Data Management. Ed. Lamar Soutter Library, University of Massachusetts Medical School. Retrieved Jan 6, 2017. From <http://library.umassmed.edu/necdmc/necdmc>

DataONE Education Module: Data Management. DataONE. Retrieved Jan 6, 2017. From [http://www.dataone.org/sites/all/documents/L01\\_DataManagement.pptx](http://www.dataone.org/sites/all/documents/L01_DataManagement.pptx)

DataONE Education Module: Data Sharing. DataONE. Retrieved Jan 6, 2017. From [http://www.dataone.org/sites/all/documents/L02\\_DataSharing.pptx](http://www.dataone.org/sites/all/documents/L02_DataSharing.pptx)

DataONE Education Module: Data Entry and Manipulation. DataONE. Retrieved Jan 6, 2017. From [http://www.dataone.org/sites/all/documents/L04\\_DataEntryManipulation.pptx](http://www.dataone.org/sites/all/documents/L04_DataEntryManipulation.pptx)

DataONE Education Module: Data Quality Control and Assurance. DataONE. Jan 6, 2017. From [http://www.dataone.org/sites/all/documents/L05\\_DataQualityControlAssurance.pptx](http://www.dataone.org/sites/all/documents/L05_DataQualityControlAssurance.pptx)

DataONE Education Module: Data Protection. DataONE. Retrieved Jan 6, 2017. From [http://www.dataone.org/sites/all/documents/L06\\_DataProtection.pptx](http://www.dataone.org/sites/all/documents/L06_DataProtection.pptx)

DataONE Education Module: Metadata. DataONE. Retrieved Jan 6, 2017. From [http://www.dataone.org/sites/all/documents/L07\\_Metadata.pptx](http://www.dataone.org/sites/all/documents/L07_Metadata.pptx)

Gaudette, Glenn R., and Donna Kafel. 2012. "A Case Study: Data Management in Biomedical Engineering." Journal of eScience Librarianship 1(3): e1027. <http://dx.doi.org/10.7191/jeslib.2012.1027>

Novak Gustainis, Emily R., Darla White, David B. Lowe. New England Collaborative Data Management Curriculum Module 7: Repositories, Archiving, and Preservation. Ed. Lamar Soutter Library, University of Massachusetts Medical School. Retrieved Jan 6, 2017. From <http://library.umassmed.edu/necdmc/necdmc>