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

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

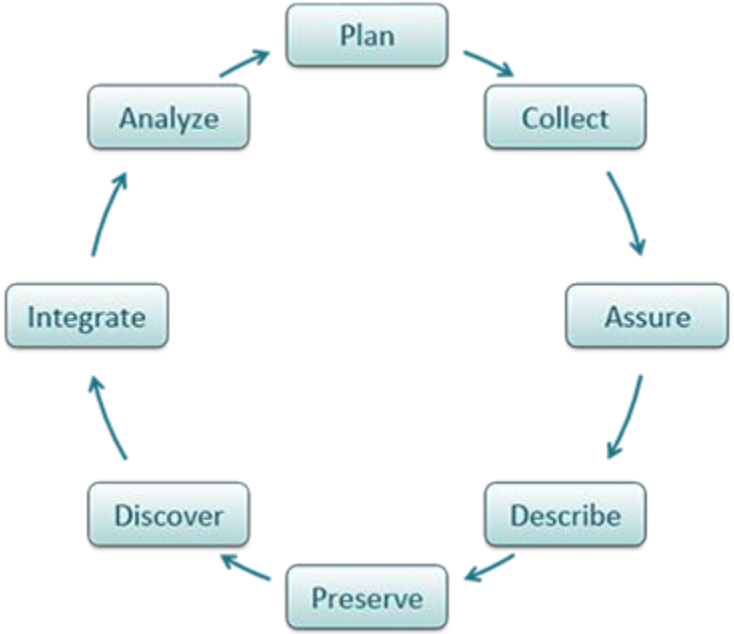
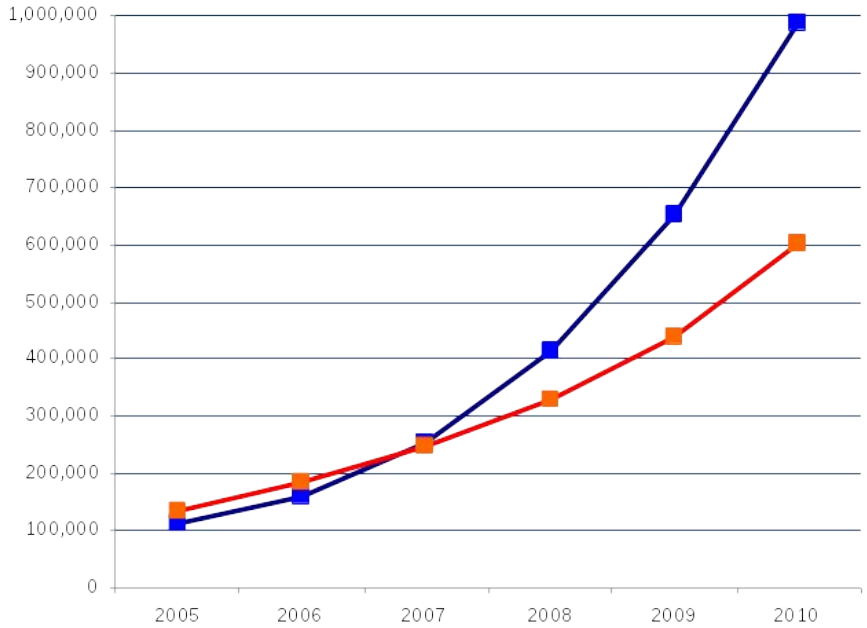


Image credit DataONE

Challenges Across the Lifecycle

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



Information versus Available Storage

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?

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

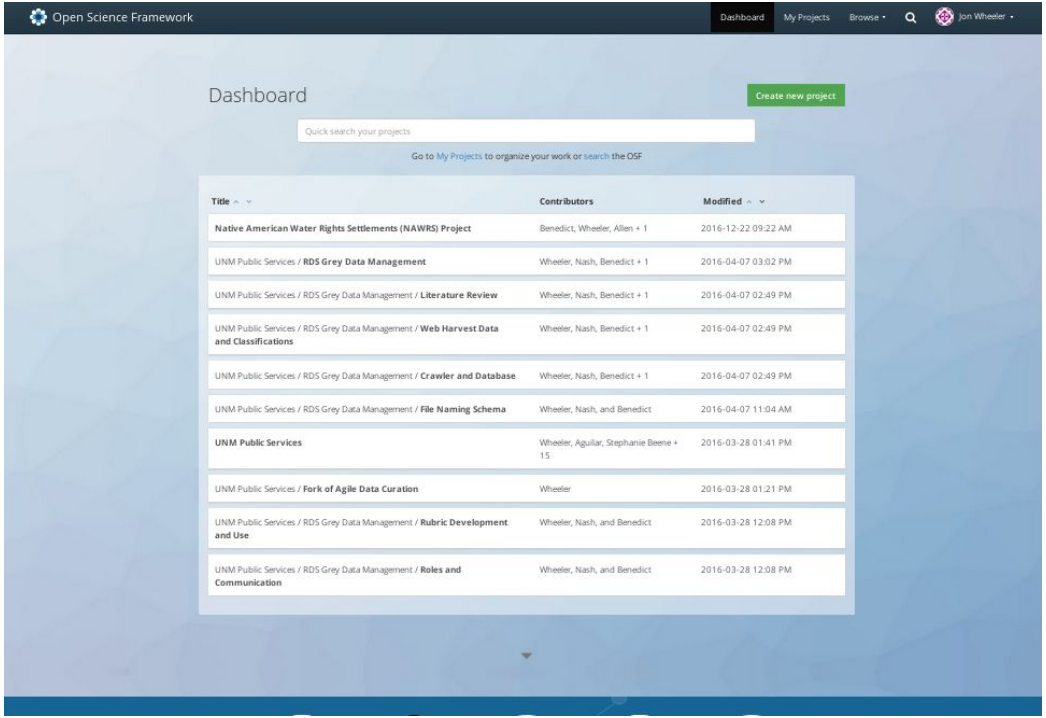
Resources

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

Slide credit DataONE

OSF Integrations

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



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.

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

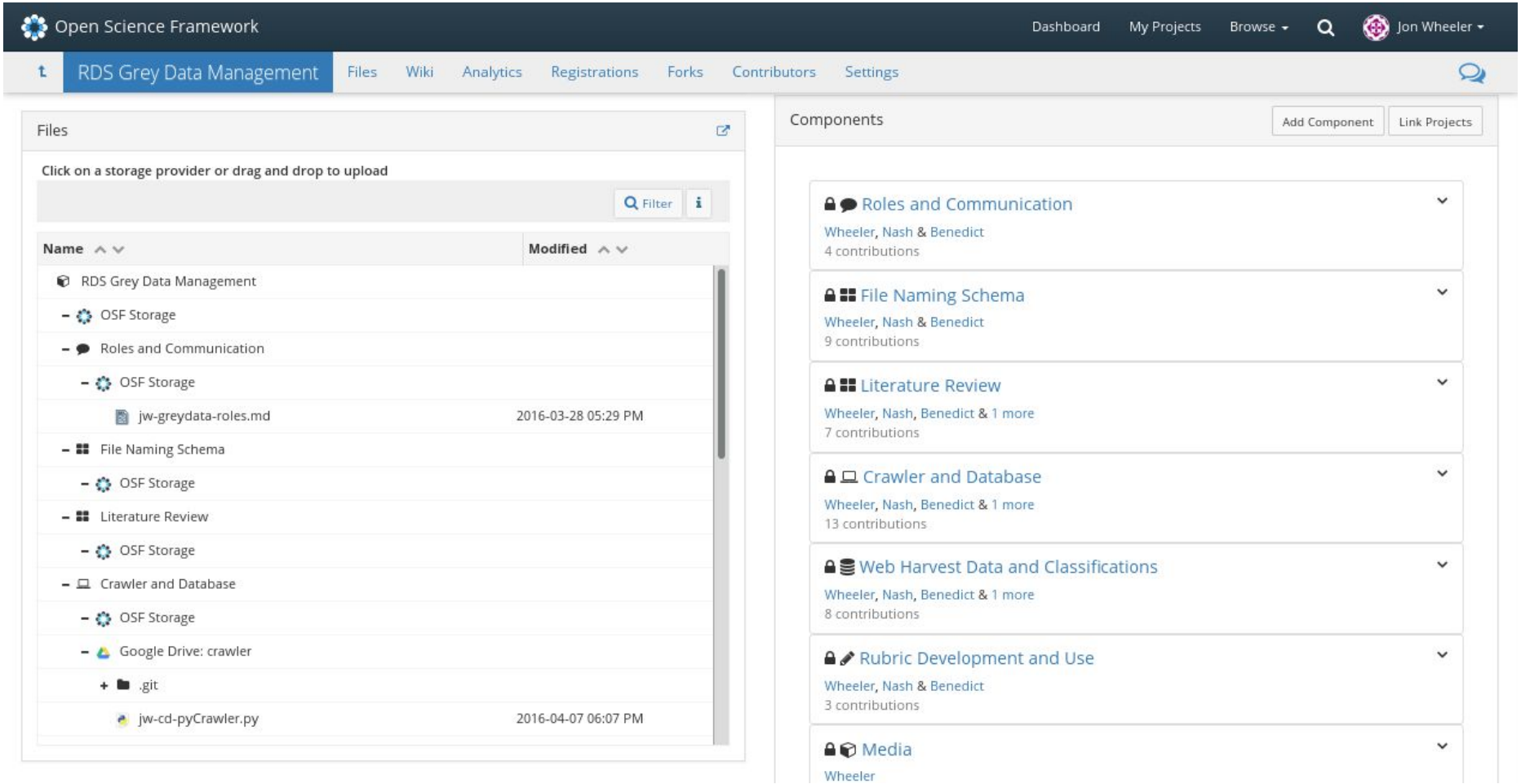
MLA
Wheeler, Jon et al. "RDS Grey Data Management." Open Science Framework, 7 Apr. 2016. Web.

Chicago
Wheeler, Jon, Jacob Nash, Karl K Benedict, and Jon Wheeler. 2016. "RDS Grey Data Management." Open Science Framework. April 7. osf.io/m4qre.

Get more citations

Enter citation style (e.g. "APA")

Components | Add Component | Link Projects


Open Science Framework
Dashboard
My Projects
Browse
Jon Wheeler
RDS Grey Data Management
Files
Wiki
Analytics
Registrations
Forks
Contributors
Settings
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

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...

Slide credit Karl Benedict, 2014

Scenario 1

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

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

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

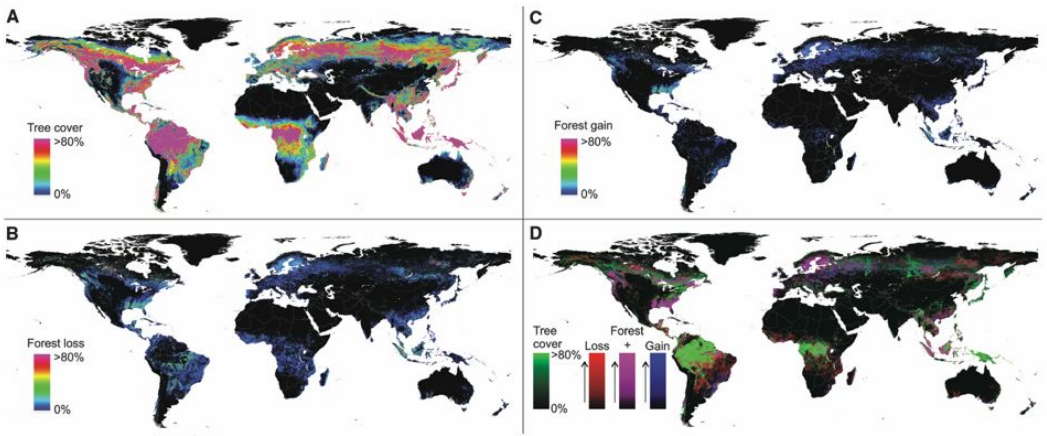


Fig. 1. (A) Tree cover, (B) forest loss, and (C) forest gain. A color composite of tree cover in green, forest loss in red, forest gain in blue, and forest loss and gain in magenta is shown in **(D)**, with loss and gain enhanced for improved visualization. All map layers have been resampled for display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

Slide credit Karl Benedict 2014

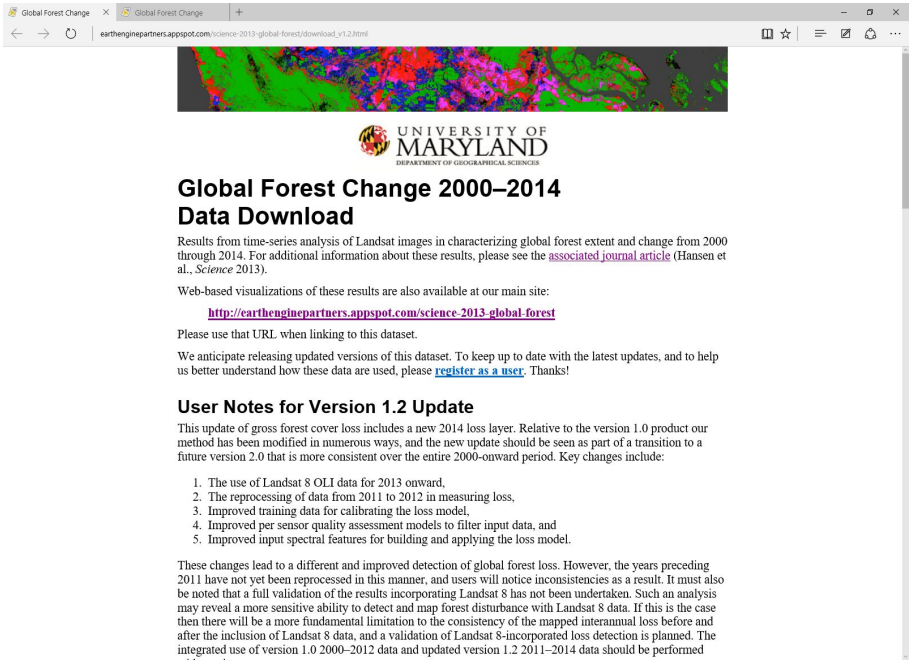
Questions

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

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

Slide credit Karl Benedict 2014

Scenario 2



UNIVERSITY OF MARYLAND
DEPARTMENT OF GEOGRAPHICAL SCIENCES

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!

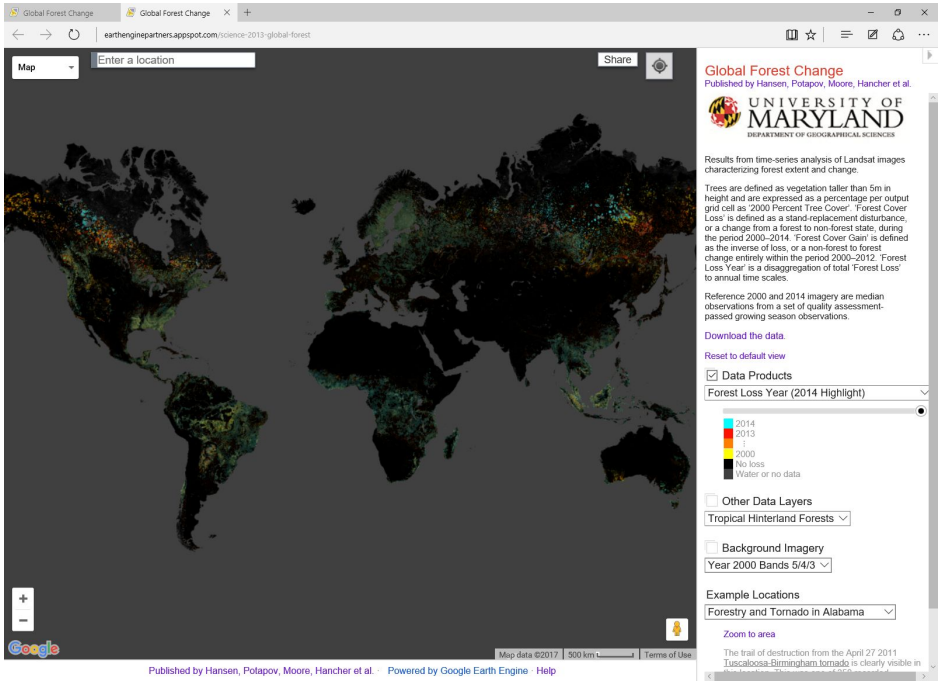
User Notes for Version 1.2 Update

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

Slide credit Karl Benedict 2014

Questions

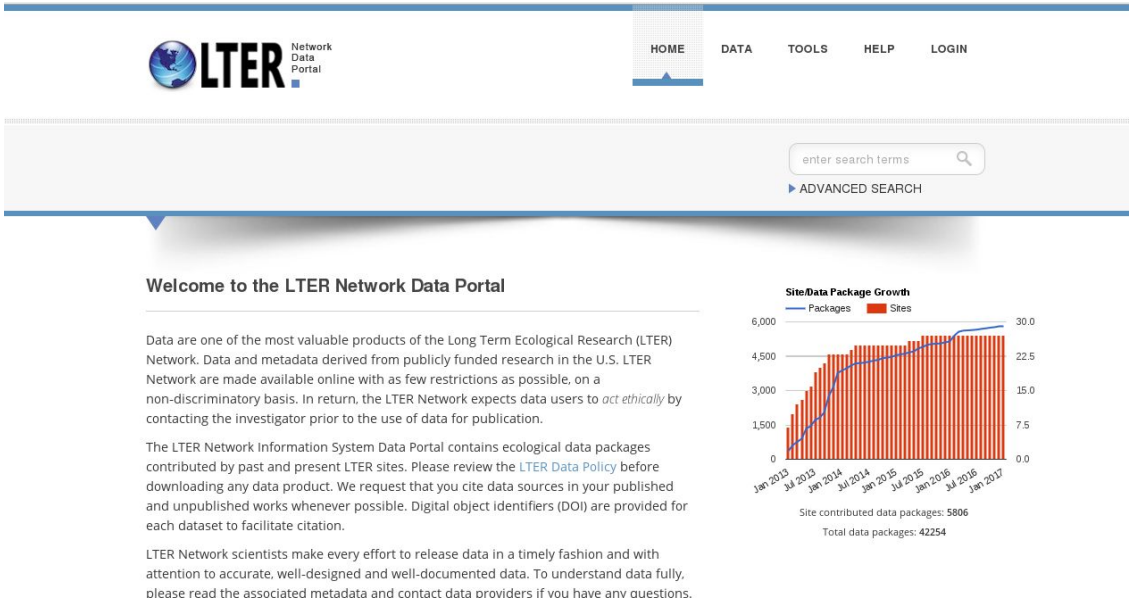
Rate from 1 (impossible) to 5 (easy) the following...

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

Slide credit Karl Benedict 2014

Metadata Enables

- Discovery
- Use
- Understanding



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

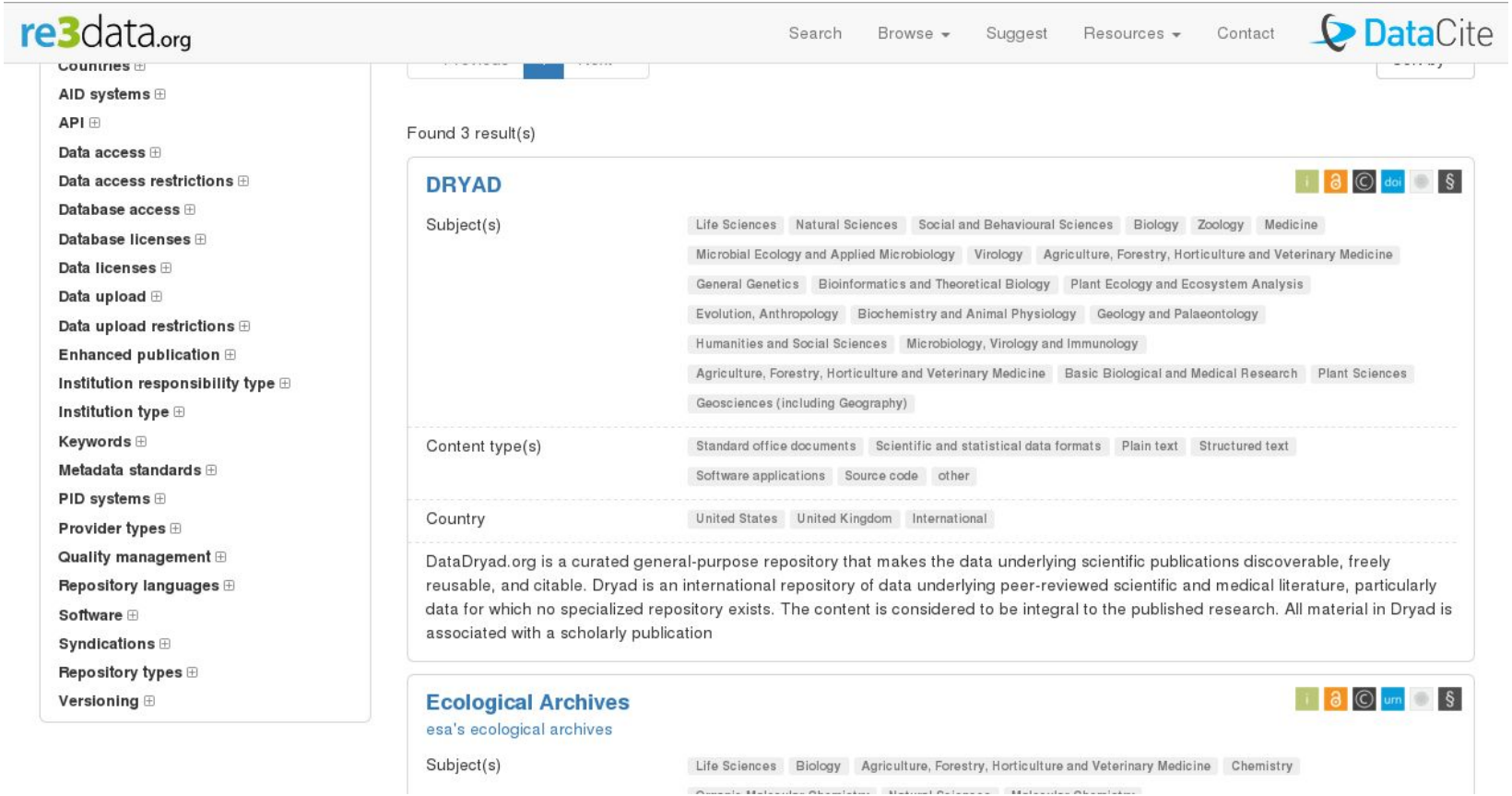
Resources

- **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



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

Slide credit New England Collaborative Data Management Curriculum

Repositories

Domain Specific or General Purpose

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

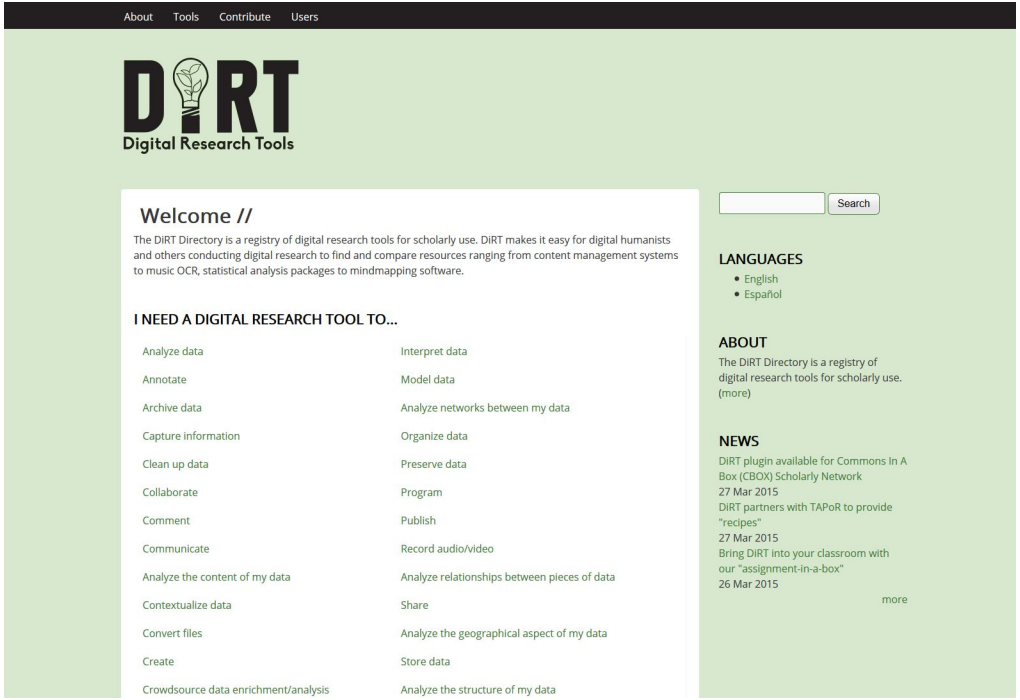
Considerations

- 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
- **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
- 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)

Slide credit DataONE

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

DataONE Education Module: Data Sharing. DataONE. Retrieved Jan 6, 2017. From 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

DataONE Education Module: Data Quality Control and Assurance. DataONE. Jan 6, 2017. From 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

DataONE Education Module: Metadata. DataONE. Retrieved Jan 6, 2017. From 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