http://www.massobs.org.uk/annual_report.htm

http://infoontech.files.wordpress.com/2011/08/archive-shelving.jpg

Scientific Data Preservation

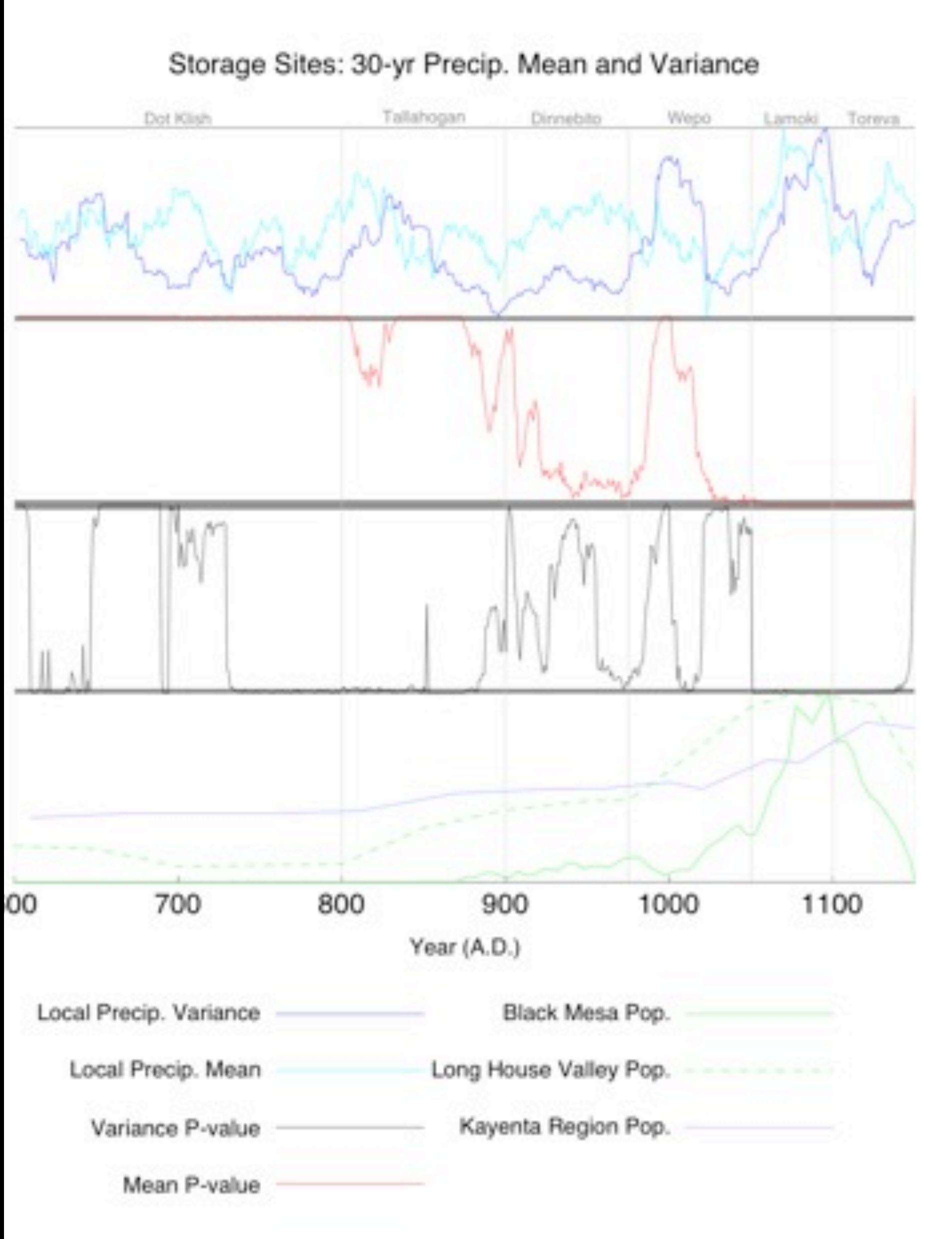
Karl Benedict
EDAC, University Libraries, Dept. of Geography
University of New Mexico

Why Me?

- Earth Data Analysis Center
- University Libraries
- Geography Department
- Federation of Earth Science Information Partners
- Foundation for Earth Science

A Story

- Dissertation work with two large data collections from the late-60s through early 80s
- Over 2200 archaeological sites, paleoclimate and modern meteorological data, publicly available environmental data
- Focus on data integration and modeling - not feasible without well documented data in understandable and usable formats



Roadmap

- Core principles
- Strategies
- Resources

Core Principles

- Data quality/safety
  - During research
  - Following research
- Documentation
  - Discovery
  - Use
  - Understanding
- Sustainability
  - Data and metadata formats/standards

Data Quality and Safety

I learn with great satisfaction that you are about committing to the press the valuable historical and State papers you have been so long collecting. Time and accident are committing daily havoc on the originals deposited in our public offices. The late war has done the work of centuries in this business. The last cannot be recovered, but let us save what remains; not by vaults and locks which fence them from the public eye and use in consigning them to the waste of time, but by such a multiplication of copies, as shall place them beyond the reach of accident.

Letter to Ebenezer Hazard
Philadelphia, 2/18/1791.
Thomas Jefferson

Documentation

ESIP
The Federation of Earth Science Information Partners
Fostering connections to make data matter

Data creators will provide sufficient metadata (defined as all the information necessary for data to be independently understood by users and to ensure proper stewardship of the data) to the data repositories responsible for long-term archival.

- Interagency Data Stewardship Guidelines. http://commons.esipfed.org/node/419

Sustainability

- Library of Congress - http://www.digitalpreservation.gov/formats/intro/format_eval_rel.shtml

Sustainability

Seven sustainability factors

- Library of Congress - http://www.digitalpreservation.gov/formats/intro/format_eval_rel.shtml

Sustainability

Seven sustainability factors

1. Disclosure. Degree to which complete specifications and tools for validating technical integrity exist and are accessible to those creating and sustaining digital content. A spectrum of disclosure levels can be observed for digital formats. What is most significant is not approval by a recognized standards body, but the existence of complete documentation.

- Library of Congress - http://www.digitalpreservation.gov/formats/intro/format_eval_rel.shtml

Sustainability

Seven sustainability factors

1. Disclosure. Degree to which complete specifications and tools for validating technical integrity exist and are accessible to those creating and sustaining digital content. A spectrum of disclosure levels can be observed for digital formats. What is most significant is not approval by a recognized standards body, but the existence of complete documentation.

2. Adoption. Degree to which the format is already used by the primary creators, disseminators, or users of information resources. This includes use as a master format, for delivery to end users, and as a means of interchange between systems.

- Library of Congress - http://www.digitalpreservation.gov/formats/intro/format_eval_rel.shtml

Sustainability

Seven sustainability factors

1. Disclosure. Degree to which complete specifications and tools for validating technical integrity exist and are accessible to those creating and sustaining digital content. A spectrum of disclosure levels can be observed for digital formats. What is most significant is not approval by a recognized standards body, but the existence of complete documentation.

2. Adoption. Degree to which the format is already used by the primary creators, disseminators, or users of information resources. This includes use as a master format, for delivery to end users, and as a means of interchange between systems.

3. Transparency. Degree to which the digital representation is open to direct analysis with basic tools, such as human readability using a text-only editor.

- Library of Congress - http://www.digitalpreservation.gov/formats/intro/format_eval_rel.shtml

Sustainability

Seven sustainability factors

1. Disclosure. Degree to which complete specifications and tools for validating technical integrity exist and are accessible to those creating and sustaining digital content. A spectrum of disclosure levels can be observed for digital formats. What is most significant is not approval by a recognized standards body, but the existence of complete documentation.

2. Adoption. Degree to which the format is already used by the primary creators, disseminators, or users of information resources. This includes use as a master format, for delivery to end users, and as a means of interchange between systems.

3. Transparency. Degree to which the digital representation is open to direct analysis with basic tools, such as human readability using a text-only editor.

4. Self-documentation. Self-documenting digital objects contain basic descriptive, technical, and other administrative metadata.

- Library of Congress - http://www.digitalpreservation.gov/formats/intro/format_eval_rel.shtml

Sustainability

Seven sustainability factors

1. Disclosure. Degree to which complete specifications and tools for validating technical integrity exist and are accessible to those creating and sustaining digital content. A spectrum of disclosure levels can be observed for digital formats. What is most significant is not approval by a recognized standards body, but the existence of complete documentation.

2. Adoption. Degree to which the format is already used by the primary creators, disseminators, or users of information resources. This includes use as a master format, for delivery to end users, and as a means of interchange between systems.

3. Transparency. Degree to which the digital representation is open to direct analysis with basic tools, such as human readability using a text-only editor.

4. Self-documentation. Self-documenting digital objects contain basic descriptive, technical, and other administrative metadata.

5. External Dependencies. Degree to which a particular format depends on particular hardware, operating system, or software for rendering or use and the predicted complexity of dealing with those dependencies in future technical environments.

- Library of Congress - http://www.digitalpreservation.gov/formats/intro/format_eval_rel.shtml

Sustainability

Seven sustainability factors

1. Disclosure. Degree to which complete specifications and tools for validating technical integrity exist and are accessible to those creating and sustaining digital content. A spectrum of disclosure levels can be observed for digital formats. What is most significant is not approval by a recognized standards body, but the existence of complete documentation.

2. Adoption. Degree to which the format is already used by the primary creators, disseminators, or users of information resources. This includes use as a master format, for delivery to end users, and as a means of interchange between systems.

3. Transparency. Degree to which the digital representation is open to direct analysis with basic tools, such as human readability using a text-only editor.

4. Self-documentation. Self-documenting digital objects contain basic descriptive, technical, and other administrative metadata.

5. External Dependencies. Degree to which a particular format depends on particular hardware, operating system, or software for rendering or use and the predicted complexity of dealing with those dependencies in future technical environments.

6. Impact of Patents. Degree to which the ability of archival institutions to sustain content in a format will be inhibited by patents.

- Library of Congress - http://www.digitalpreservation.gov/formats/intro/format_eval_rel.shtml

Sustainability

Seven sustainability factors

1. Disclosure. Degree to which complete specifications and tools for validating technical integrity exist and are accessible to those creating and sustaining digital content. A spectrum of disclosure levels can be observed for digital formats. What is most significant is not approval by a recognized standards body, but the existence of complete documentation.

2. Adoption. Degree to which the format is already used by the primary creators, disseminators, or users of information resources. This includes use as a master format, for delivery to end users, and as a means of interchange between systems.

3. Transparency. Degree to which the digital representation is open to direct analysis with basic tools, such as human readability using a text-only editor.

4. Self-documentation. Self-documenting digital objects contain basic descriptive, technical, and other administrative metadata.

5. External Dependencies. Degree to which a particular format depends on particular hardware, operating system, or software for rendering or use and the predicted complexity of dealing with those dependencies in future technical environments.

6. Impact of Patents. Degree to which the ability of archival institutions to sustain content in a format will be inhibited by patents.

7. Technical Protection Mechanisms. Implementation of mechanisms such as encryption that prevent the preservation of content by a trusted repository.

- Library of Congress - http://www.digitalpreservation.gov/formats/intro/format_eval_rel.shtml

Dataset Considerations

Significant for all datasets is that they be represented in a structure that reveals the characteristics of individual data items and the relationships among them. A dataset format suitable for preservation must retain the syntactical integrity of both the structure *and individual values, so that automated analysis is possible. Also essential for future usability is an understanding of the semantics of the data elements and their relationships within the dataset. The semantics may be described explicitly within the dataset, described explicitly in an ancillary document (preferably itself machine-processable), or implicit through compliance with a community best practice or external specification.*

http://www.digitalpreservation.gov/formats/content/dataset_quality.shtml

Dataset Considerations

Significant for all datasets is that they be represented in a structure that reveals the characteristics of individual data items and the relationships among them. A dataset format suitable for preservation must retain the syntactical integrity of both the structure and individual values, so that automated analysis is possible. Also essential for future usability is an understanding of the semantics of the data elements and their relationships within the dataset. The semantics may be described explicitly within the dataset, described explicitly in an ancillary document (preferably itself machine-processable), or implicit through compliance with a community best practice or external specification.

http://www.digitalpreservation.gov/formats/content/dataset_quality.shtml

Dataset Considerations

Significant for all datasets is that they be represented in a structure that reveals the characteristics of individual data items and the relationships among them. A dataset format suitable for preservation must retain the syntactical integrity of both the structure and individual values, so that automated analysis is possible. Also essential for future usability is an understanding of the semantics of the data elements and their relationships within the dataset. The semantics may be described explicitly within the dataset, described explicitly in an ancillary document (preferably itself machine-processable), or implicit through compliance with a community best practice or external specification.

represented in a structure

relationships

syntactical integrity

structure

individual values

understanding of the semantics

http://www.digitalpreservation.gov/formats/content/dataset_quality.shtml

Strategies

Preservation Resources

- Use existing individual resources
- Leverage existing shared resources
- Acquire new individual or shared resources
- Use commodity/hosted (e.g. cloud) resources

Use Existing Resources

- Pros
  - Little if any initial cost
  - Resources already well understood
- Cons
  - Potential lack of alignment with preservation needs
  - May be insufficient for long-term

http://www.flickr.com/photos/27784972@N07/

Shared Resources

http://www.flickr.com/photos/ben_grey/

- Pros
  - May be able to better use underutilized resource
  - Resource is a known commodity
- Cons
  - Resource decisions are shared
  - Potential contention for resource as it is consumed

Acquire New Resources

- Pros
  - Resources can be specified for need
  - Growth may be built into resource planning
- Cons
  - Requires expenditure of funds
  - Adequate funds may not be available for acquisition
  - New resource may have related costs (i.e. labor, administration)

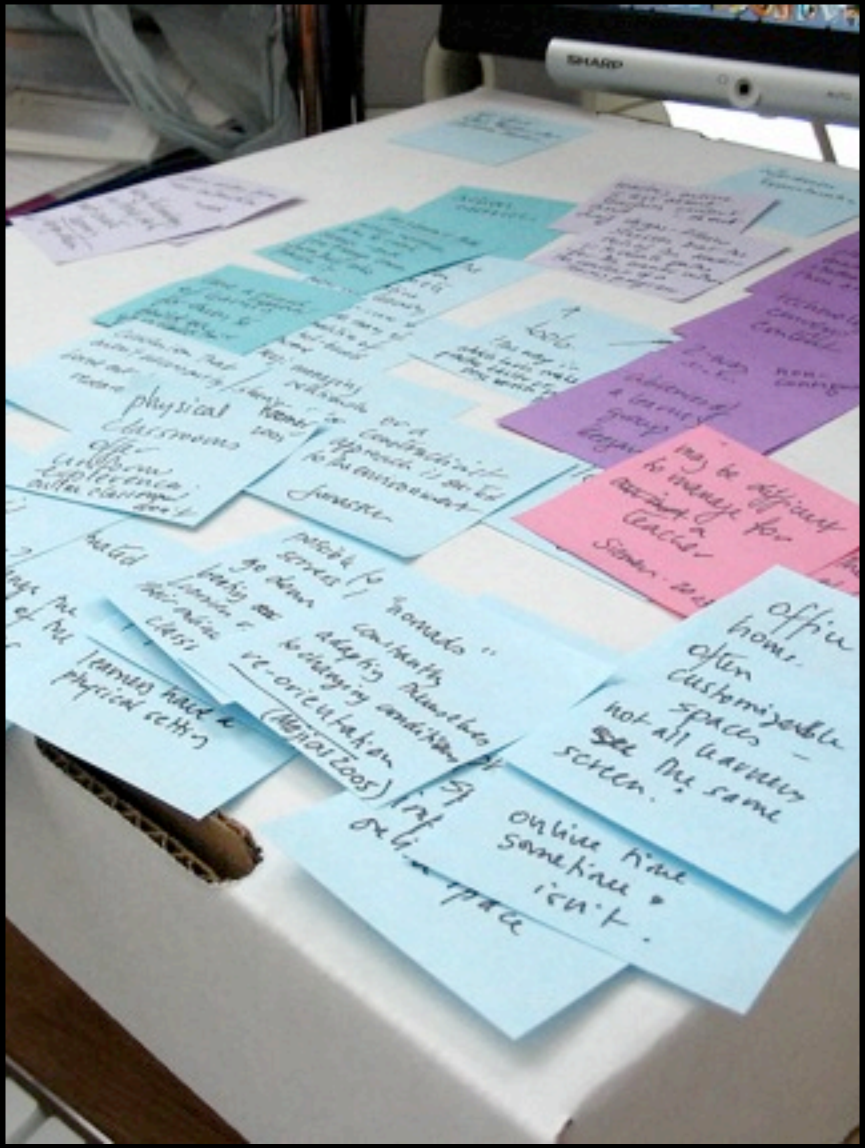
http://www.flickr.com/photos/59937401@N07/

Hosted Resources

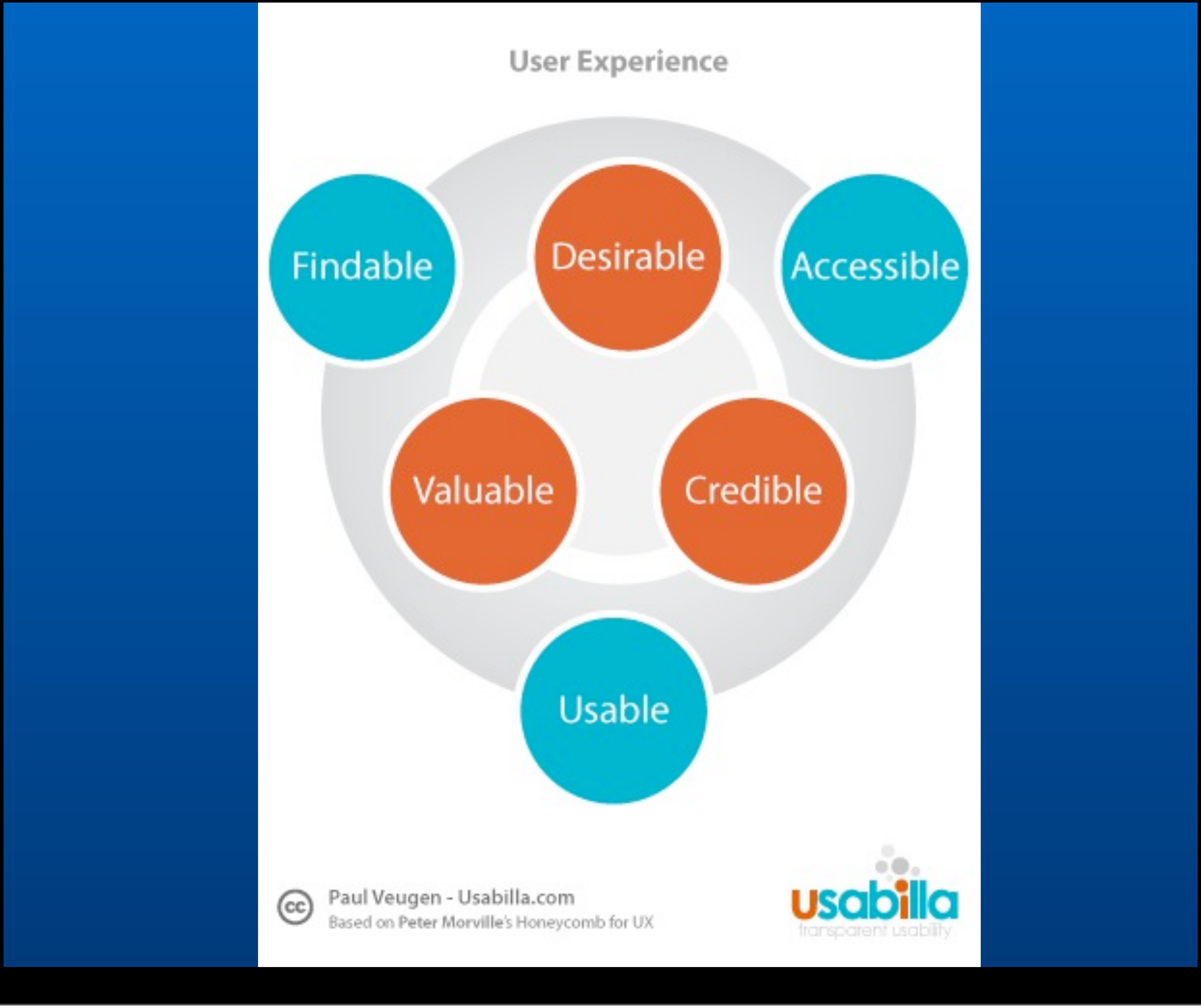
- Pros
  - Cost can scale with demand/need
  - Local administrative and infrastructure costs can be reduced
- Cons
  - Reduced control over data
  - Risk of provider going out of business
  - May not exactly match needs

Documentation

- Start Early
- Elements to Record
  - Who
  - Where
  - When
  - What
  - Why
  - How
- Know target documentation content standard at start to ensure coverage


physical classrooms
may be difficult to manage for teacher
office home. often customizeable spaces - not all learners see the same screen.
online time sametime isn't.
re-orientation

Format Selection

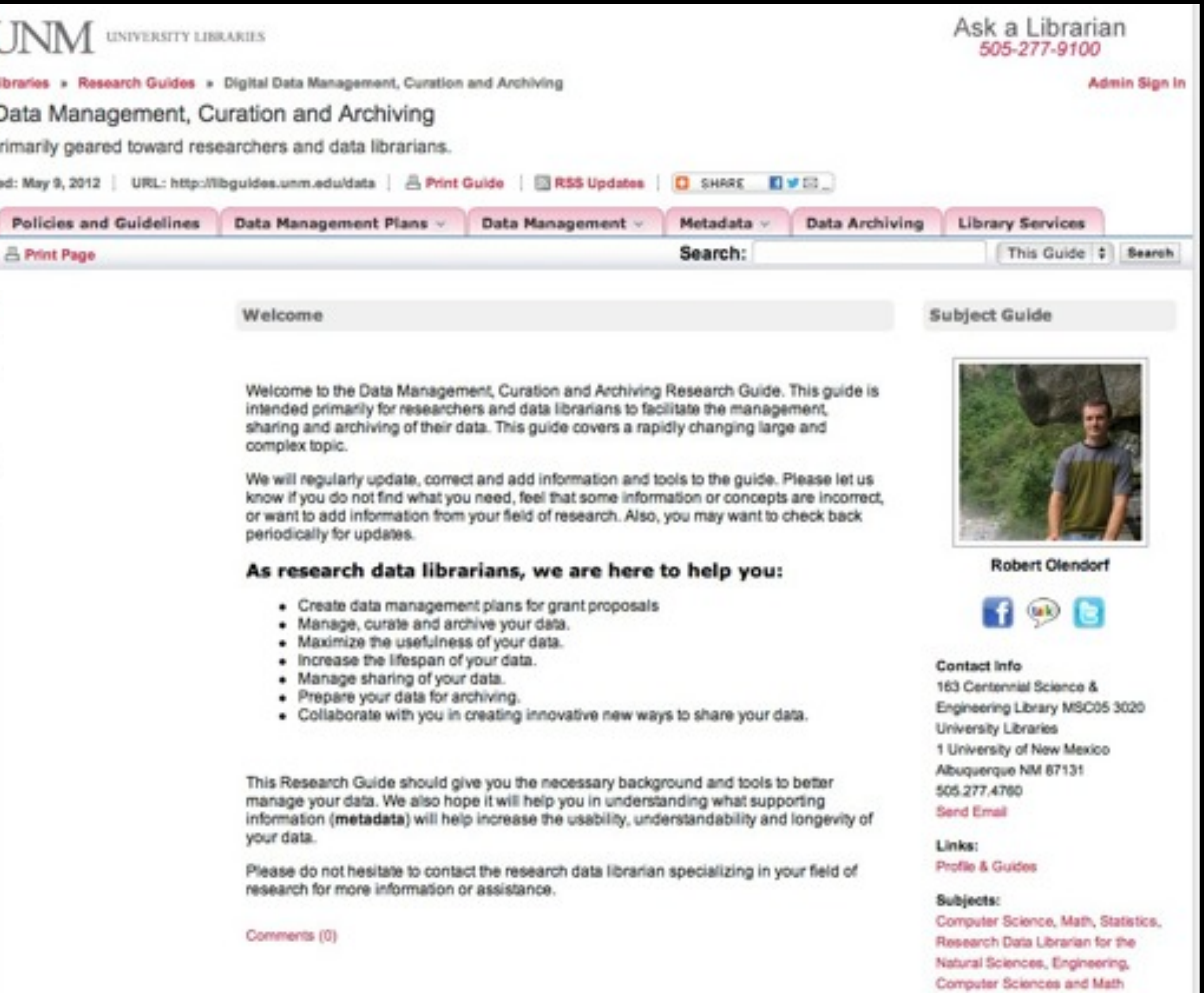


Resources

- UNM LoboVault / UNM Libraries data curators
- UNM Research Storage Consortium
- NM Resource Geographic Information System / NM EPSCoR
- DataONE Member Node(s), OneShare
- Community Repositories
- Further Reading

http://www.flickr.com/photos/kalexanderson/

UNM LoboVault & Data Curators



UNM UNIVERSITY LIBRARIES

Ask a Librarian
505-277-9100

University Libraries » Research Guides » Digital Data Management, Curation and Archiving

Admin Sign In

Digital Data Management, Curation and Archiving

A guide primarily geared toward researchers and data librarians.

Last Updated: May 9, 2012 | URL: http://libguides.unm.edu/data | Print Guide | RSS Updates | SHARE

Welcome

Welcome to the Data Management, Curation and Archiving Research Guide. This guide is intended primarily for researchers and data librarians to facilitate the management, sharing and archiving of their data. This guide covers a rapidly changing large and complex topic.

We will regularly update, correct and add information and tools to the guide. Please let us know if you do not find what you need, feel that some information or concepts are incorrect, or want to add information from your field of research. Also, you may want to check back periodically for updates.

As research data librarians, we are here to help you:

- Create data management plans for grant proposals
- Manage, curate and archive your data.
- Maximize the usefulness of your data.
- Increase the lifespan of your data.
- Manage sharing of your data.
- Prepare your data for archiving.
- Collaborate with you in creating innovative new ways to share your data.

This Research Guide should give you the necessary background and tools to better manage your data. We also hope it will help you in understanding what supporting information (**metadata**) will help increase the usability, understandability and longevity of your data.

Please do not hesitate to contact the research data librarian specializing in your field of research for more information or assistance.

Comments (0)

Subject Guide

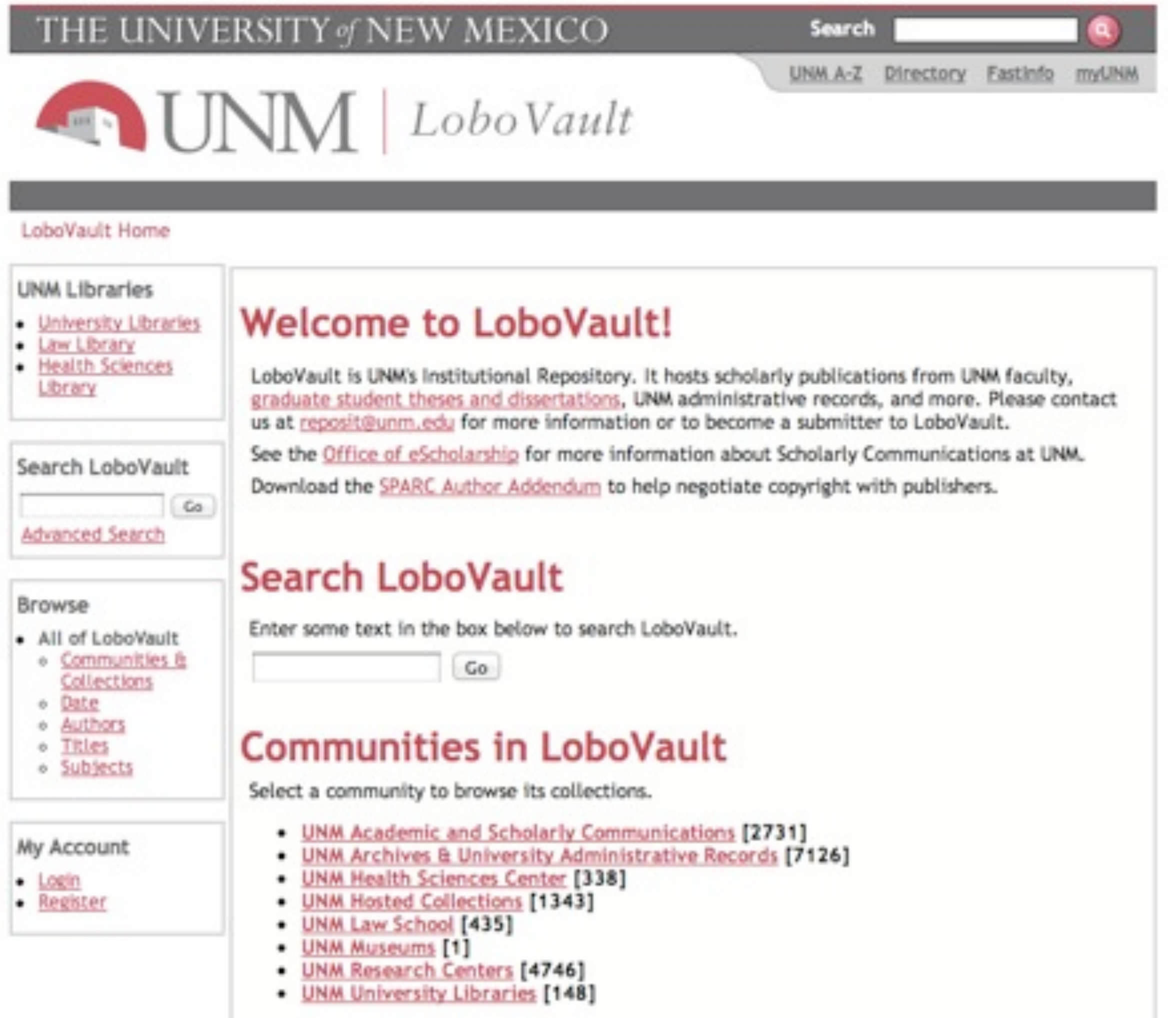
Robert Olendorf

Contact Info
163 Centennial Science &
Engineering Library MSC05 3020
University Libraries
1 University of New Mexico
Albuquerque NM 87131
505.277.4760
Send Email

Links:
Profile & Guides

Subjects:
Computer Science, Math, Statistics, Research Data Librarian for the Natural Sciences, Engineering, Computer Sciences and Math

http://libguides.unm.edu/data



THE UNIVERSITY of NEW MEXICO

UNM | LoboVault

LoboVault Home

UNM Libraries

- University Libraries
- Law Library
- Health Sciences Library

Search LoboVault

Advanced Search

Browse

- All of LoboVault
  - Communities & Collections
  - Date
  - Authors
  - Titles
  - Subjects

My Account

- Login
- Register

Welcome to LoboVault!

LoboVault is UNM's Institutional Repository. It hosts scholarly publications from UNM faculty, graduate student theses and dissertations, UNM administrative records, and more. Please contact us at reposit@unm.edu for more information or to become a submitter to LoboVault.

See the Office of eScholarship for more information about Scholarly Communications at UNM.

Download the SPARC Author Addendum to help negotiate copyright with publishers.

Search LoboVault

Enter some text in the box below to search LoboVault.

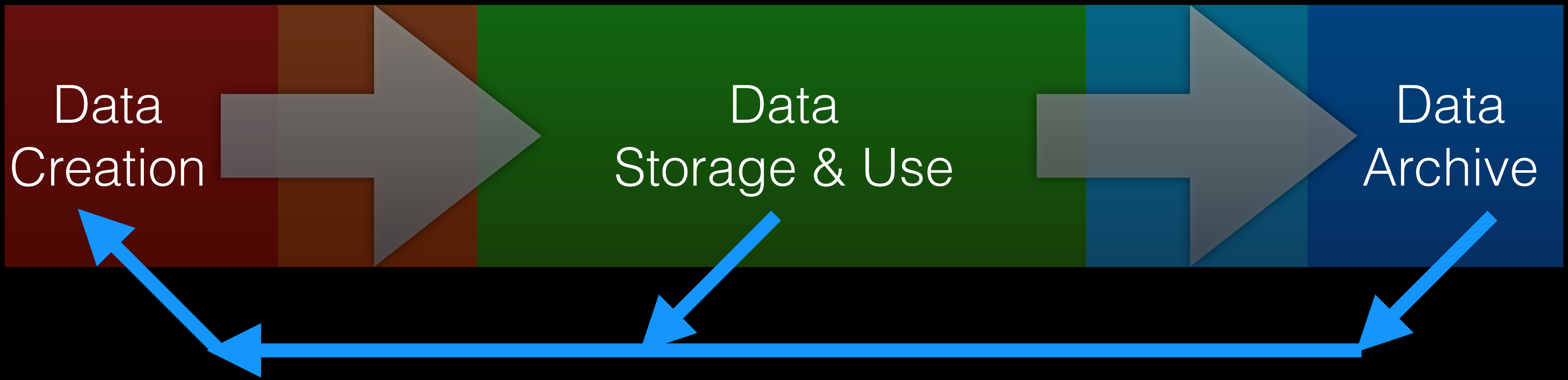
Communities in LoboVault

Select a community to browse its collections.

- UNM Academic and Scholarly Communications [2731]
- UNM Archives & University Administrative Records [7126]
- UNM Health Sciences Center [338]
- UNM Hosted Collections [1343]
- UNM Law School [435]
- UNM Museums [1]
- UNM Research Centers [4746]
- UNM University Libraries [148]

http://repository.unm.edu

UNM Research Storage Consortium



Data Creation

Data Storage & Use

Data Archive

LTER Network Office

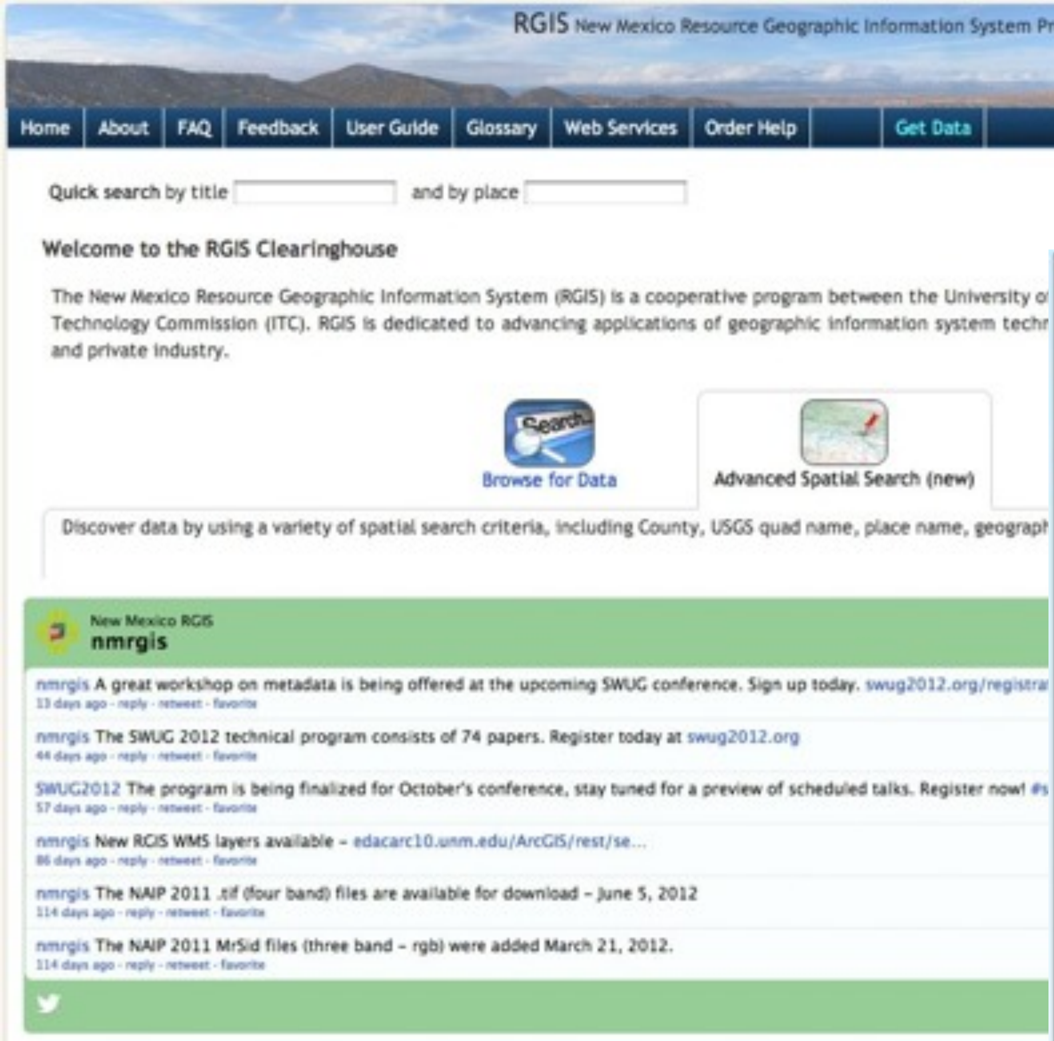
UNM Center for Advanced Research Computing

DataONE

UNM University Libraries

EDAC

NM RGIS & EPSCoR Data Portals



RGIS New Mexico Resource Geographic Information System Program
Home
About
FAQ
Feedback
User Guide
Glossary
Web Services
Order Help
Get Data
Browse for Data
Advanced Search new version
Search
Browse for Data
Advanced Spatial Search (new)

Quick search by title and by place

Welcome to the RGIS Clearinghouse

The New Mexico Resource Geographic Information System (RGIS) is a cooperative program between the University o
Technology Commission (ITC). RGIS is dedicated to advancing applications of geographic information system techn
and private industry.

Discover data by using a variety of spatial search criteria, including County, USGS quad name, place name, geograph

New Mexico RGIS
nmrgis

nmrgis A great workshop on metadata is being offered at the upcoming SWUG conference. Sign up today. swug2012.org/registra
13 days ago · reply · retweet · favorite

nmrgis The SWUG 2012 technical program consists of 74 papers. Register today at swug2012.org
44 days ago · reply · retweet · favorite

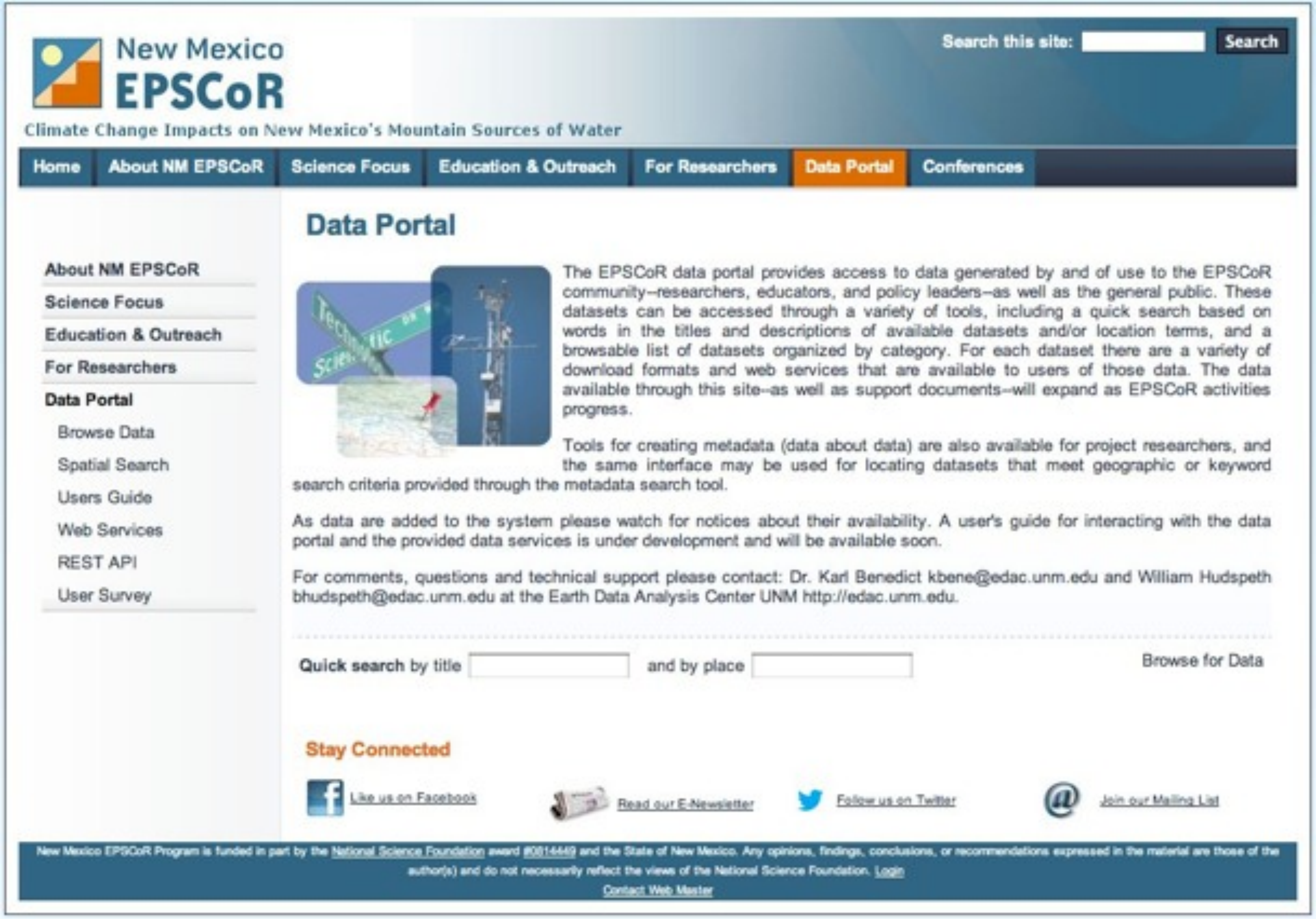
SWUG2012 The program is being finalized for October's conference, stay tuned for a preview of scheduled talks. Register now! #s
57 days ago · reply · retweet · favorite

nmrgis New RGIS WMS layers available – edacarc10.unm.edu/ArcGIS/rest/se...
86 days ago · reply · retweet · favorite

nmrgis The NAIP 2011 .tif (four band) files are available for download – June 5, 2012
114 days ago · reply · retweet · favorite

nmrgis The NAIP 2011 MrSid files (three band – rgb) were added March 21, 2012.
114 days ago · reply · retweet · favorite

http://rgis.unm.edu



Data Portal

The EPSCoR data portal provides access to data generated by and of use to the EPSCoR community–researchers, educators, and policy leaders–as well as the general public. These datasets can be accessed through a variety of tools, including a quick search based on words in the titles and descriptions of available datasets and/or location terms, and a browsable list of datasets organized by category. For each dataset there are a variety of download formats and web services that are available to users of those data. The data available through this site–as well as support documents–will expand as EPSCoR activities progress.

Tools for creating metadata (data about data) are also available for project researchers, and the same interface may be used for locating datasets that meet geographic or keyword search criteria provided through the metadata search tool.

As data are added to the system please watch for notices about their availability. A user's guide for interacting with the data portal and the provided data services is under development and will be available soon.

For comments, questions and technical support please contact: Dr. Karl Benedict kbene@edac.unm.edu and William Hudspeth bhudspeth@edac.unm.edu at the Earth Data Analysis Center UNM http://edac.unm.edu.

Quick search by title and by place Browse for Data

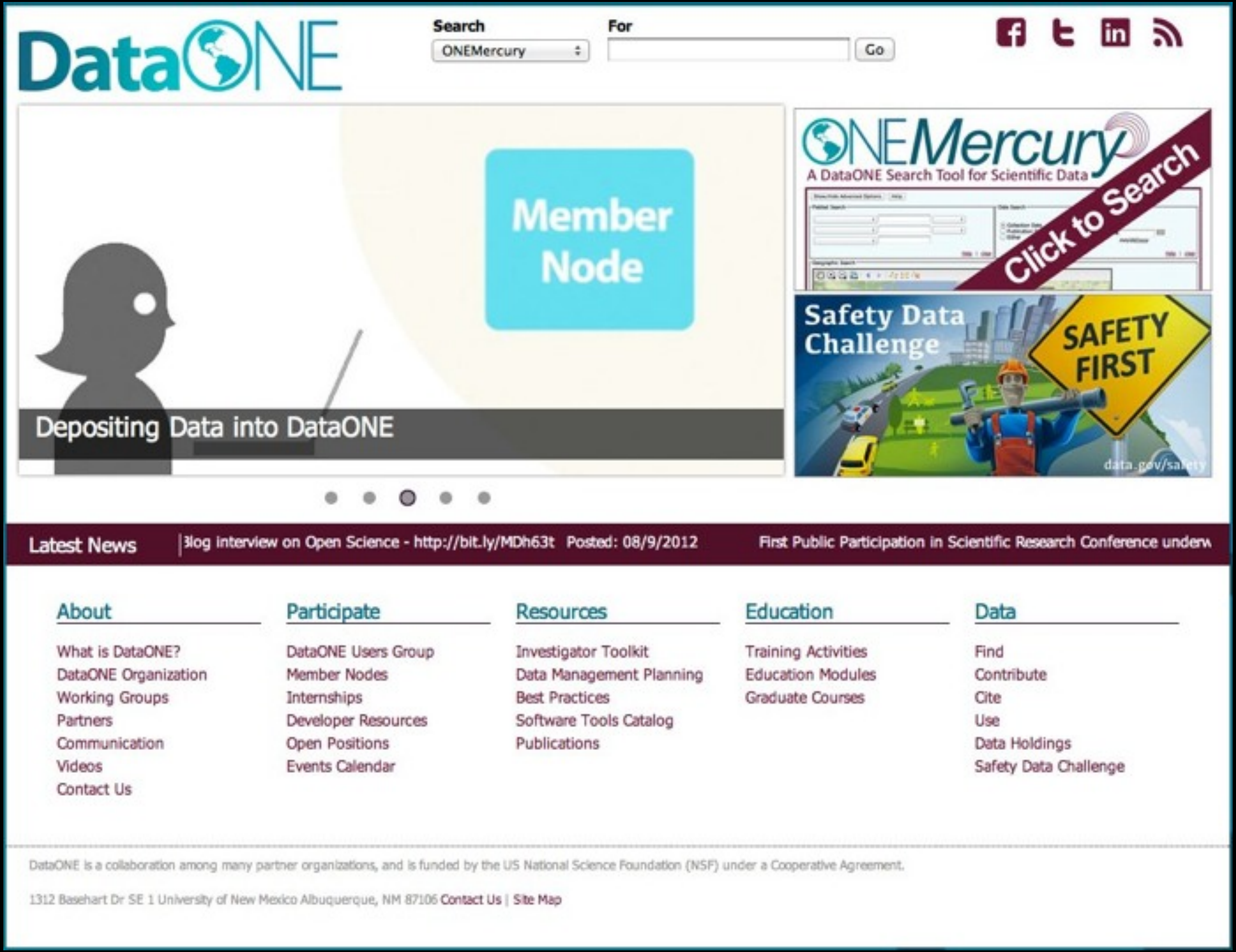
Stay Connected

Like us on Facebook Read our E-Newsletter Follow us on Twitter Join our Mailing List

New Mexico EPSCoR Program is funded in part by the National Science Foundation award #0814449 and the State of New Mexico. Any opinions, findings, conclusions, or recommendations expressed in the material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Login
Contact Web Master

http://nmepscor.org/dataportal

DataONE



Blog interview on Open Science - http://bit.ly/MDh63t Posted: 08/9/2012 First Public Participation in Scientific Research Conference underw

DataONE is a collaboration among many partner organizations, and is funded by the US National Science Foundation (NSF) under a Cooperative Agreement.

1312 Basehart Dr SE 1 University of New Mexico Albuquerque, NM 87106 Contact Us | Site Map

http://www.dataone.org/

Community Repositories

GEO.DATA.GOV

figshare

EMBL-EBI

BioModels Database

DRYAD

GenBank

CUAHSI HIS

Sharing hydrologic data

World Data Center for Cosmic Rays

The Long Term Ecological Research Network

Check out Databib for a list of over 300 repositories:

http://databib.org/index.php

Additional Background Material

- ESIP Federation *Interagency Data Stewardship Guidelines*
  http://commons.esipfed.org/node/419
- ESIP Federation *Data Citation Guidelines for Data Providers and Archives*
  http://commons.esipfed.org/node/308
- Library of Congress *Sustainability of Digital Formats Planning for Library of Congress Collections*
  http://www.digitalpreservation.gov/formats/index.shtml
- UNM Libraries *Digital Data Management, Curation and Archiving* Research Guide
  http://libguides.unm.edu/data
- DataONE *Best Practices database*

  http://www.dataone.org/best-practices

Questions?