Engage in Culture Change: Fundamentals for Working in a Data Centric World

Viv Hutchison
US Geological Survey
Denver, CO
About the Presenter…

Viv Hutchison

- US Geological Survey
- Core Science Analytics and Synthesis Program
- Data Management Coordinator
- Location: Denver, CO
- vhutchison@usgs.gov
Agenda

- An Intro to DataONE and USGS
- Creating a culture of good data stewards: why data management skills are needed
- Skills and Knowledge for good data management
- What assessments show us
- Education approaches:
  - DataONE
  - USGS
The DataONE Vision and Approach

Providing universal access to data about life on earth and the environment that sustains it, as well as the tools needed by researchers

1. Building community

2. Developing sustainable data discovery and interoperability solutions

3. Enabling science through tools and services
DataONE Principles

1. Data should be part of the permanent scholarly record and requires long-term stewardship.

2. Sharing and reuse maximize the value of data to environmental science.

3. Science is best served by an open and inclusive global community.

4. The data environment is dynamic and requires evidence-based decision-making about practice and governance.
Community Engagement: A Working Group Model

Working Groups

DataONE engages and explores key areas of the scientific data lifecycle using the Working Group model. Working Groups concentrate the intellectual efforts of a team of experts during intense one week meetings resulting in collaborative outcomes to achieve the DataONE goals. Working Groups foci will evolve with the project, however, initial working groups are identified as:

- Community Education and Engagement
- Data Integration and Semantics
- Data Preservation and Metadata
- Distributed Storage
- Federated Security
- Public Participation in Science and Research
- Scientific Exploration, Visualization and Analysis
- Scientific Workflows and Provenance
- Sociocultural Issues
- Sustainability and Governance
- Usability and Assessment
New DataONE Education Resources

DataONE provides CC0 access to a suite of ten training modules that you can download and include in your data management teaching. Modules cover all aspects of the Data Life Cycle from managing data, through collection, description, QAQC, analysis and preservation ...

more
US Geological Survey

- *Science for a Changing World*

- The USGS serves the Nation by providing reliable scientific information to describe and understand the Earth; minimizing loss of life and property from natural disasters; managing water, biological, energy, and mineral resources; and enhancing and protecting our quality of life.
US Geological Survey

Headquarters in Reston, VA

7 Science Mission Areas:
- Climate and Land Use Change
- Core Science Systems
- Ecosystems
- Energy and Minerals
- Environmental Health
- Natural Hazards
- Water

- Science Programs are contained within Mission Areas
- Distributed science centers and research stations with multi-disciplinary foci
US Geological Survey: Community for Data Integration (CDI)

- Community of Practice
  - Lead development and implementation of the USGS data integration strategy
  - Provide recommendations for implementation of data integration guidelines
  - Promote Bureau-wide data management and integration
- Over 250 members
- Monthly webinars
- Multiple Working Groups
- Some funded projects
USGS CDI: Data Management Working Group

- Data Management Working Group
- Purpose:
  - Elevate the practice of data management such that it is seen as a critical activity in the pursuit of USGS science
- Representation from across the Survey
- Monthly Webinars + Focus Groups
Creating a culture of good data stewards: Reasons data management skills are needed from our graduates…
We are Facing a Data Deluge

Data is collected from sensors, sensor networks, remote sensing, observations, and more - this calls for increased attention to data management and stewardship.

Photo courtesy of www.carboafrica.net

Photo courtesy of http://modis.gsfc.nasa.gov

Photo courtesy of http://www.futurlec.com

Photo courtesy of CIMMYT on Flickr

Image collected by Viv Hutchinson

CC image by Tajai on Flickr

CC image by Tajai on Flickr

CC image by Tajai on Flickr
Science is more interdisciplinary, computational, and data intensive
And scientists are still geographically spread out.
Critical that scientists place importance on managing data well

- Science is more collaborative
  - Properly managed data is more easily shared with colleagues, across groups
- Data re-use and sharing concepts gaining attention
Making the Case for Well-Managed Data:

- Creates efficiencies in how science is done
- Improves provenance in the science iteration process
- Supports scientific review and integrity
- Improves reproducibility in science
- Ensures integrity of science data assets
- **Maximizes the effective use and value of data**
Realities of Science

“Please forgive my paranoia about protocols, standards, and data review. I'm in the latter stages of a long career with USGS (30 years, and counting) and have experienced much. Experience is the knowledge you get just after you needed it.

Several times, I've seen colleagues called to court in order to testify about conditions they have observed.

Without a strong tradition of constant review and approval of basic data, they would've been in deep trouble under cross-examination. Instead, they were able to produce field notes, data approval records, and the like, to back up their testimony.

It's one thing to be questioned by a college student who is working on a project for school. It's another entirely to be grilled by an attorney under oath with the media present.”

-Nelson Williams
Eastern Region
USGS Water
Good data management supports Open Science Movement...

...governments and funding agencies are requiring data management
What skills and knowledge are needed in this data-centric, sharing era of science?
Research and Data Life Cycle Integration

Proposal writing
Ideas
Research
Publication

Plan
Collect
Assure

Analyze
Integrate
Discover
Preserve
Describe

ESA 2012: How to Manage Ecological Data
Facets of data management

- Data release policies
- Digital Object Identifiers
- Data deposit
- Data reuse
- Open formats
- Repositories
- Documentation
- Proprietary formats
- Data management plans
- Data catalogs
- Raw data
- File formats
- Quality assurance
- Quality control
- Security
- Reproducibility
- Scientific workflows
- Interoperability
- Sensitive data
- Embargoed data
- Data collection
- Software licenses
- Protocols
- Analysis
- Embargoed data
- Personally identifying information
Needed: Skills and Knowledge in Science Data Management

Focus Training and Workshops on:

- Awareness of the Data Lifecycle
- Best Practices in Data Management:
  - Data Management Planning
  - Data collection and organization
  - Quality control and assurance
  - Metadata
  - Workflows
  - Data Preservation & Sharing

USGS
Best Practices in Data Management: Examples of Good Habits

- Create a data management plan
- Define the contents of your data files
- Use consistent data organization
- Use stable file formats
- Assign descriptive file names
- Preserve information
- Perform basic quality assurance
- Provide documentation
- Protect your data
In addition to training, scientists will need credit for their efforts…

- Data citation
- Use of digital object identifiers
- Credit for publication of data given same weight as publication of scientific papers

...and scientists will need tools
What Can Assessments Show Us?
DataONE: Community Engagement and Education “Assessment Working Group”

- Conducting assessments of scientists, librarians and libraries, and other stakeholders
- Helps to define areas to concentrate effort -- in workshops, training, and community engagement

Samples of Results of DataONE Scientist Surveys

- “More than half of the respondents (56%) reported that they did not use any metadata standard and about 22% of respondents indicated they used their own lab metadata standard.”

- Less than 6% of scientists are making “All” of their data available via some mechanism.

- 2/3rd report that organizational help and support is lacking
What does this tell us?

- Based on these findings, there appear to be opportunities for libraries and librarians...

- Another DataONE survey of libraries and librarians sought to discover answers about librarians role in Research Data Services:
  - Do academic librarians have the background, skills and education to provide RDS?
  - What are librarian attitudes regarding the importance of RDS?
  - What factors contribute to or inhibit engagement of librarians in RDS?
Academic Library Responses

- Total response rate: 302
- Composed of librarians at Academic Research Libraries (223) plus the staff of library directors who took the libraries survey, combined with librarians at UC libraries (79)
- 223 were librarians working in metadata, digital collections, sciences, scholarly communications, health, e-science, e-resources, IRs, data, other.
DataONE survey: Libraries and Librarians

- Level of my knowledge & skills?
- Role in partnering with researcher?
- Are RDS priority?
- Level of participation with data?
- Level of involvement with metadata?
- Is there an agency repository that accepts data?
- Role of librarian discovering data?
- Role of the librarian to help preservation?
- Stewardship role (select & deselect)?
Conclusions from Librarian/Library Survey (unpublished)

- Over two-thirds of respondents have provision of RDS as an occasional or integral part of their job.

- Some have the knowledge, skills, and opportunities to provide RDS.

- They believe RDS are important and consistent with library mission and role.

- Libraries are at an early point in transition to RDS—requiring resetting of priorities, realignment of responsibilities, and opportunities to develop skills.
Education Approaches

DataONE and USGS
DataONE
Data Management
Education Efforts
Workshops at Conferences

- ½ day workshops
- Content of workshops cover:
  - Data Lifecycle
  - Data Management planning
  - Data Collection and Management
  - Quality Assurance/ Quality Control
  - Metadata
  - Data Preservation
- Each workshop has a hands-on component

http://www.dataone.org/training-activities
DataONE Data Management Education Offerings

Walter E. Dean Environmental Information Management Institute

- University of New Mexico, Informatics Training Laboratory
- 3-week course
- MS students and professionals with a BS in biology, geology, ecology, or other environmental sciences, environmental engineering, geography or science librarianship
- Conceptual and practical hands-on training that allows them to effectively design, manage, analyze, visualize, and preserve data and information

http://library.unm.edu/services/instruction/eimi.php
DataONE: Data Management Education Modules

http://www.dataone.org/education-modules

DataONE Community
Engagement and Education
Working Group
DataONE: Data Management Education Modules

Why Data Management
Data Sharing
Data Management Planning
Data Entry and Manipulation
Data Quality Control and Assurance
Data Protection and Backups
Metadata
How to Write Good Quality Metadata
Data Citation
Analysis and Workflows
Testing the Modules

- 2-day workshop 2012
- Solicited graduate student volunteers in various earth science backgrounds to participate
- Pre and mid and post course learning assessments
Feedback from 2-day Module Test

- Variety of real-world examples and stories should be used to illustrate important points;

- Include information and links to tools and resources for data management and sharing;

- Use text sparingly on workshop slides - more on online version;
Feedback, con’t

· Always describe or define jargon;

· Take data management experience levels into account when planning lesson content for each topic;

· Always include information about best practices;

· If modules are to be used in workshop format rather than as stand-alone online lectures, redundant information should be removed.
Obstacles to Best Practices

Several of the recommended best practices for data sharing and data citation were seen as particularly difficult to implement due to lack of:

1) agreed-upon standards;
2) tools or services to help monitor use of data in publications and websites;
3) information about the process and timing of acquiring a DOI or other identifier;
4) formalized systems for coordinating with data and journal publishers;
5) cultural, institutional, and financial support.
Feedback and Lessons Learned

Improving data management and sharing through training: evaluation of training workshop feedback

A two-day Data Management Short Course was held at NCEAS in Santa Barbara on May 23-24, 2012 and led by members of the DataONE Community Education and Engagement (CEE) working group in cooperation with Earth Science Information Partnerships (ESIP). A primary goal of the workshop was to field test the Data Management Education Lessons that have been developed by the CEE working group. A series of online surveys that included a pre-course survey and four surveys during the workshop (one at the end of each morning and each afternoon session) was used to collect feedback. This poster highlights key insights gained through this evaluation process.

Adapt course materials to suit experience level of audience

As part of a pre-course survey, participants provided information about their current data management practices or plans related to data sharing, data management planning, data quality control/quality assurance, metadata creation and maintenance, analysis and workflows, data citation, and backups and protection. Participants already engaging in more advanced data management and sharing practices assigned significantly lower ratings to the quality \( (r=-0.664, df=13, p=0.07) \) and relevance \( (r=-0.681, df=13, p=0.05) \) of the information presented in the course than those less experienced in data management and sharing.

Participants' Evaluation of Information Relevance

- For audiences with significant data management and sharing experience, 1) minimize focus on why researchers should manage/share/backup/cite data (perceived as "preaching to the choir"), and 2) include more advanced information about QA/QC techniques, metadata creation, and data citation.
- Always include information about best practices.
- When best practices were not explicitly identified during a presentation, a number of participants mentioned wanting this type of information.
- If lessons are to be used in a workshop format rather than as stand-alone lectures, redundant information should be removed.

Guidelines for information design

- A variety of concrete or "real-world" examples and stories should be used to illustrate important points.
- When participants reported on the most memorable items from the lessons, they were most likely to mention such examples or stories, and the most common suggestion overall was to include more information of this type.
- Include information about (and links to) tools and resources for data management and sharing.
- Information about and/or links to tools or resources were pointed out as memorable information almost as often as concrete or real-world examples or stories, and were also the next most requested type of information.
- Use text sparingly on presentation slides.
- Several participants suggested that two versions of the information be made available: a set of minimal-essay slides for use during presentation, and a reference document (e.g., slides with more text or a white paper).
- Always describe or define jargon.
- Many participants noted the need for more attention early in the workshop to definitions of basic concepts such as "metadata", "data management", "discovery", "tags", "EML", "quality assurance", and "quality control".

Obstacles to best practices

Participants perceived obstacles to implementing a number of the best practices suggested during the short course, and identified lack of time and lack of training as the most important reasons for these obstacles. Several of the recommended best practices for data sharing and data citation were seen as particularly difficult to implement due to lack of 1) agreed-upon standards, 2) tools or services to help monitor use of data in publications and websites, 3) information about the process and timing of acquiring a DOI or other identifier, 4) formalized systems for coordinating with data and journal publishers, and 5) cultural, institutional, and financial support.

Support provided for DataONE by US National Science Foundation award #0830944 under a Cooperative Agreement. For more information see www.DataONE.org.

Special thanks to Rebecca Koskinel, Amber Sudden, Mark Schiffler and CEE working group members Cliff Duve, Wendy Gram, Stephanie Hampton, Heather Herkel, Carly Strasser, Kristin Vanderkitt and Linda Wayne, who provided feedback on draft versions of the evaluation surveys.
USGS
Data Management Education Efforts
USGS Workshops

- Data Management Training
- Sponsored by USGS + taught by Bureau of Land Management (BLM)
- 2011
- Select USGS scientists and data managers
- Feedback positive to host more workshops and do more educating
Supporting and Enabling USGS Data Management

How do I...

...Get Started?
Why should you manage your data? What's the value of data management? Learn the basics in an Overview of Data Management.

...Create a Data Management Plan?
Learn how to create a data management plan and why it can help you.

...Evaluate my Data Acquisition Options?
Understand your choices before you make them.

...Organize my Data?
Use USGS guidelines to organize your data to benefit yourself and others.

...Create Metadata?
All published datasets require metadata. Learn how to create it and how it helps you.

...Backup my Data?
Better safe than sorry. Avoid disaster with regular backups.

...Preserve my Data?
Find out more about archiving, disposition, persistent Digital Object Identifiers (DOIs), and repositories.

...Dispose of Old Data?
Federal regulations govern the disposition of data. Learn how this affects you.

...Publish my Data?
Find out how to make your data available and visible, and about restrictions on data publishing.


"Challenging scientific questions require the analysis and integration of information and data across scientific disciplines. Data integration within the USGS is a prerequisite for joining international efforts to develop worldwide science collaboration and computing platform[s] that can address future challenges."

Applying best practices to managing scientific data lays the necessary foundation for data integration. The purpose of this Web site is to support and enable USGS scientists in data management.

You can learn about data management techniques and best practices, download sample documents, and find references to authoritative

Data Modeling Zone 2012
Data Modeling Zone 2012 contains four tracks and a combination of half-day and 75-minute sessions from fundamental to advanced levels designed to help you develop an understanding of data modeling concepts, techniques, and approaches; fine tune your relational and dimensional modeling and data vault skills; and take a hands-on PowerDesigner®, ER/Studio® course. Baltimore, Maryland - November 12th-15th 2012

Meetings & Conferences

The National Map (TNM) Users Conference and the fourth annual Community for Data Integration Training and Workshop
Please join us for the second biennial The National Map (TNM) Users Conference and the fourth annual Community for Data Integration Training and Workshop event. The four day event will be hosted by the USGS National Geospatial Program (NGP) and the Community for Data Integration (CDI). Attendees will have the opportunity to collaborate and interact with other users.
USGS Data Management Website

Data Management Planning

A documented sequence of intended actions to identify and secure resources and gather, maintain, secure, and utilize data holdings comprises a Data Management Plan. This also includes the procurement of funding and the identification of technical and staff resources for full lifecycle data management. Once the data needs are determined, a system to store and manipulate the data can then be identified and developed.

Data Management Plans

Planning for Data Management involves answering questions about the data: Do they already exist? How will they be obtained or collected? What is the schedule and budget for data collection? How will the data be checked and certified? What are the likely uses for the data? How will the data be stored, accessed, and protected? A good data management plan provides a strategy for how you will answer all of these questions. Learn more about data management plans.

Organize Files and Data

File organization with a logical, clear structure and labeling system enables not only others to access your data, but makes it easier for you to find your own data as well. These are the public's data, and we have a stewardship responsibility to the data. Thinking about how you will organize your files and data early on in planning may save you from having to reorganize and rename files later on. Learn more about organizing files and data.

Data Standards

Standards make it easier to create, share, and integrate data by making sure that there is a clear understanding of how the data are represented and that the data you receive are in a form that you expected. Learn more about data standards.

What the USGS Geological Survey Manual Says:

The USGS Manual Chapter 1100.1 - Information Product Planning discusses planning for information products, which include data products:

"Policy: Planning for information products begins as early as possible during the evolution of a project. A written planning document must be developed prior to production for each information product. An information product plan will ensure adequate management and budgeting for all elements of the information lifecycle including planning, development, dissemination, documentation, storage, evaluation, and disposition."

Note: Information Product: An information product is the culmination of scientific communication or knowledge such as facts, data, or interpretations in any medium (e.g., print, digital, Web) or form, including textual, numerical, graphical, cartographic, or audiovisual, to be disseminated to a defined audience or customer, scientific or nonscientific, internal or external.

Data & File Formats

Standardization of Data & File formats makes the data easier to work with both for the original researcher and for others who will use the data. Learn more about data formats.

Data Templates

Creating Data Templates for data collection, data storage, and metadata saves time and increases consistency. Utilizing form validation increases data entry reliability. Learn more about data templates.

Data Stewardship: Roles and Responsibilities

Data management is a crucial foundation of your professional work. The data you collect and analyze are a national resource. They are not just your data or my data, they are USGS data - paid for by taxpayers' money - that are being used to make all types of management decisions, many of which have substantial economic and even health and safety consequences. All data are "corporate data" but it can be appropriate for certain data to be maintained and kept at lower levels of the organization. You must evaluate their role in your job and how well you manage your data in your everyday work. Learn more about data stewardship.

Accessibility

Privacy

Terms and Conditions


URL: http://original-www.usgs.gov/datamanagement

Page Contact Information: E-mail Us

Page Last Modified: Thursday, August 01, 2013
USGS Online Modules

- Value of Data Management
- Data Management Planning
- Best Practices for Preparing Science Data to Share
USGS Policies and Workflows

- Policy development:
  - Data Management
  - Foundational Policy
    - Metadata Policy
    - Data Release
    - Software Release
- Workflows for:
  - Data Release Processes

1. FDS is the Federal Information Dissemination System. Refer to URL: http://dissemination.usgs.gov/about.html
2. Under FDS approval, non-interpretable information products are dedicated to the Science Center Director
Thank you!

Comments & Questions are Welcome

Viv Hutchison
US Geological Survey
vhutchison@usgs.gov

“Professor LaVonne had many enemies in the entomological world, detective, but if you examine that data label, you’ll find exactly when and where he was—shall we say—‘collected.’”