The Practice of Metadata

The how’s and why’s of metadata at USGS
Presenter

- Viv Hutchison
- USGS Core Science Systems / Core Science Analytics and Synthesis (CSAS) Program
- Denver, CO
- Data Management Program Coordinator for CSAS; Science Data Management Team for CSAS

- Background: MLS from University of Maryland – College Park, 2002
Overview

- USGS science and organization
- Challenges in data management in USGS
- Importance of metadata
- Broad steps to manage data in USGS
- A focus on metadata in USGS
US Geological Survey

- Earth Science
  - Natural Hazards – earthquake, volcano, etc
  - Water
  - Biology
  - Geology

- Characteristics of USGS:
  - Large, distributed science agency
  - Science centers located in every state - sometimes multiple centers
  - Small labs in many locations
Challenges in Data Management: USGS

- Scientists are focused on science and publishing.
- Scientists are given credit for publishing, not “data management”.
- Some scientists view their publically funded research as “their data”.
- Multiple science disciplines throughout the agency – “data silos” - No single repository for accessing data
- Repetition of data documentation throughout agency – project financial database, Pubs Warehouse, metadata creation, etc,
- Interesting misunderstandings about “publishing processes in journals” and “data publishing processes”.
What is being done to help “elevate” data management in USGS?

1) Reorganization of USGS from “disciplines” to “Mission Areas” – promote interdisciplinary science activities
   - Powell Center: Funds USGS-led Working Groups to solve science questions using high performance computing capabilities

2) Publications Warehouse and ScienceBase
   - Pubs Warehouse required for USGS publications – accompanying data and metadata more prominent; managed by the USGS Library
   - ScienceBase – data discovery system leading way towards more global view of USGS data
What is being done to “elevate” data management in USGS?

3) Community for Data Integration

- Organized to advance science progress through shared use of data and information, tools and techniques
- Volunteer community; monthly meetings
- Funded Projects
- Outside Partnerships
- Working Groups –
  - Tech Stack, Data Semantics, Citizen Science, Data Management – Data Policy sub-team; Data Best Practices sub-team
CDI:
Research Data Lifecycle for USGS

- PLAN
- ACQUIRE
- PROCESS
- ANALYZE
- PRESERVE
- PUBLISH / SHARE

- Describe (Metadata, Documentation)
- Manage Quality
- Backup & Secure
Data Management Policies: a new chapter on metadata...

- 502.2 - Fundamental Science Practices: Planning and Conducting Data Collection and Research
CDI: Data Management Website

Metadata and Documentation

Metadata describes information about a dataset, such that a dataset can be understood, re-used, and integrated with other datasets. Information described in a metadata record includes where the data were collected, who is responsible for the dataset, why the dataset was created, and how the data is organized. Metadata generally follows a standard format, making it easier to compare datasets and to transfer files electronically.

Key Points
- Metadata creates longevity for data
- Use metadata to understand and re-use data
- Avoid data duplication
- Reduce workload
- Share reliable information
- Metadata transcends people and time
- Data is not complete without a metadata record
- Identify new partnership opportunities when metadata is shared
- Ensure organizational investment in data
- Use Federal metadata standards
- Keep all documentation associated with your data

Download Key Points Checklist

Metadata is crucial for any potential use or reuse of data; no one can reasonably re-use or interpret data without accompanying metadata that explains how the data set was created, why, where it is geographically located, and details about the structure and meaning of the data.

There are many uses for metadata, even beyond the simple discovery of datasets. Metadata can be used for understanding data, analysis and synthesis, maintaining longevity of a dataset for an organization, tracking the progress of a research project, and demonstrating the return on investment for research at an institution.

Federal agencies are mandated by Executive Order 12906 to use the Federal Geographic Data Committee Content Standard for Digital Geospatial Metadata. A transition to the ISO Standard is currently occurring, the adoption of which is endorsed by the FGDC. Metadata is not just for geospatial data. Consider a biological data spreadsheet containing species occurrence information. It can be documented using a metadata standard to answer questions about the data, which may include, but not be limited to, location data.

Best Practices

[Download checklist for ALL Best Practices]

Writing Metadata

[Download checklist for THJS Best Practices]

- Get organized: Writing good metadata begins with being organized.
  - Organization seems like a logical step, but many choose to start writing a metadata record before having all the necessary information together. It makes a difference to stop, get organized, and make a plan. Begin by gathering your information together, especially if multiple people have the information that you need.
- Use information already developed. Information needed for high-quality metadata records is often already written.
What is being done to “elevate” data management in USGS?

4) Data Rescue Program
   · Limited annual funding dedicated to preserving “orphan” datasets

5) Ad-hoc Teams:
   · Data Release at USGS
     · Use cases: release of old data held at Science Centers with limited documentation; new trend for publications requesting data to accompany the journal article
   · Data Preservation Team
     · Looking at how data can better be preserved in USGS as a part of the research data lifecycle.
Metadata
What is Metadata?

- “Structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage any other resource”
  - National Information Standards Organization

- Answers who, what, when, and why about a dataset. - ISO 19115 standard
Which data products measure the quantities I need?

Does any of this data have configuration problems?

This data is valuable, but will I find it again?

Can I trust these measurements? How were they taken?

How can I track the configuration of my experiment?

Questions metadata can help solve.
What does a metadata record look like?

North American Breeding Bird Survey (BBS)

Identification Information:

Citation:
    Originator: Patuxent Wildlife Research Center, Biological Resources Division, U.S. Geological Survey (USGS)
    Publication Date: 1997
    Title: North American Breeding Bird Survey (BBS)
    Publication Information:
        Publication Place: Laurel, MD
        Publisher: Patuxent Wildlife Research Center, Biological Resources Division, U.S. Geological Survey (USGS)
    Other Citation Details:

Description:

Abstract:
The North American Breeding Bird Survey (BBS), which is coordinated by the Biological Resources Division and Canadian Wildlife Service, is a primary source of population trend and distribution information for most species of North American birds. The BBS was initiated during 1966 by Chan Robbins and his associates at the Patuxent Wildlife Research Center to monitor the populations of all breeding bird species across the continental U.S., Canada, and Alaska. Approximately 2200 skilled observers participate in the survey each year. The BBS has accumulated 30 years of data on the abundance, distribution, and trends for more than 400 species of birds. These data are widely used by researchers, various federal and state agencies, non-governmental organizations, and the general public. Analyses of BBS data by PWRC statisticians have been instrumental in the development of innovative approaches for analyzing trends of wildlife populations.

Purpose:
In the 1960's, chlorinated hydrocarbon pesticides and similar poisons were widely used for spraying not only killed insects but also killed birds, raising serious concerns over its effect. Unfortunately, no long-term regional or continental population data were available for ornithologists to demonstrate declines in bird populations. The Bird Breeding Survey has provided information on bird population trends. Robbins et al. (1986) provided the first continent-wide analysis of the relative abundance of species that are well sampled by the BBS. In addition, the temporal patterns in trends. Populations of permanent resident and short-distance migrants
Importance of Metadata...
Era of Big Data

- Fourth Paradigm: scientific breakthroughs will increasingly be powered by advanced computing capabilities that help researchers manipulate and explore massive datasets.
- Metadata must be preserved when scientific data is generated – Jim Gray
- Further the time/space distance between data producer and re-use, the more detailed metadata that’s required.
Data Sharing: Critical Issue as Science Questions Grow Larger

What will Baltimore look like in 2025 under a plan for sustainability?

The challenge is to integrate into high-resolution:

- thermal satellite imagery of the greater Baltimore area,
- surface observations of meteorological and air quality variables,
- traffic density and emissions data,
- trends in sea level,
- projected infrastructure renovation,
- demographic trends,
- tax base projections, and
- overall economic outlook.

Robust metadata is a key to major data integration.
“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, USGS
The climate scientists at the centre of a media storm over leaked emails were yesterday cleared of accusations that they fudged their results and silenced critics, but a review found they had failed to be open enough about their work.
A new image processing technique reveals something not before seen in this Hubble Space Telescope image taken 11 years ago: A faint planet (arrows), the outermost of three discovered with ground-based telescopes last year around the young star HR 8799.D. Lafrenière et al., Astrophysical Journal Letters

“The first thing it tells you is how valuable maintaining long-term archives can be. Here is a major discovery that’s been lurking in the data for about 10 years!” comments Matt Mountain, director of the Space Telescope Science Institute in Baltimore, which operates Hubble.
Informatics Challenges:

Majority of Earth Science data is undocumented
  - Lacks information on structure and content of data
  - May be impossible to understand data without contacting the original researchers, which is problematic over the long-term

Data are massively dispersed across data centers
  -- Difficulties in accessing critical data

Documentation conventions widely vary
  - Requires large time investment to understand each data set

Data loss
  - Huge investments in research unavailable to future researchers and managers due to lack of data management practices
Information Entropy

- Time of data development
- Specific details about problems with individual items or specific dates are lost relatively rapidly
- General details about data set are lost through time
- Accident or technology change may make data unusable
- Retirement or career change makes access to “mental storage” difficult or unlikely
- Death of developer results in loss of remaining info

(From Michener et al 1997)
What is the value of metadata to organizations?

- Metadata helps ensure investment in data:
  - Documentation of data processing steps, quality control, definitions, data uses, and restrictions
  - Ability to use data after initial intended purpose

- Transcends people and time:
  - Offers data permanence
  - Creates institutional memory

- Advertises research
  - Creates possible new partnerships and collaborations thru data sharing
Metadata at USGS...
Metadata Policy for Federal Agencies

The Executive Order 12906:

- Signed in 1994 by then U.S. President Clinton
- Defines the responsibilities of the Federal Geographic Data Committee (FGDC)
- Outlines three major uses of metadata:
  1. (1) to maintain an organization's *internal investment* in geospatial data
  2. (2) to provide information to data *clearinghouses* and catalogs
  3. (3) to provide information needed to process and interpret data transferred from another organization.
- Requires creation of metadata for data sets from 1995 forward
<table>
<thead>
<tr>
<th>Concern</th>
<th>Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Workload required to capture accurate robust metadata (“It’s too hard”)</td>
<td>Incorporate metadata creation into data development process – distribute the effort; utilize tools with auto capture</td>
</tr>
<tr>
<td>Time and resources to create, manage, and maintain metadata (“It takes too much time”)</td>
<td>Include in grant budget and workflow, research schedule</td>
</tr>
<tr>
<td>Readability / usability of metadata (“I take notes in a text file on my data processes”)</td>
<td>Use a standardized metadata format</td>
</tr>
<tr>
<td>Discipline specific information and ontologies (“My science discipline is special”)</td>
<td>Use ‘profiles’ in standards that require specific information and use specific values</td>
</tr>
</tbody>
</table>
Implementing Metadata: Varied Approaches

Individuals or Teams throughout USGS:

- Team Leader / Project Manager
- GIS Specialist
- Field Personnel
- Database Manager
- Science Staff
- Data Analysis Lead
Implementation of Metadata Policies: USGS CSAS Metadata Program Example

- Record creation assistance
- Training
- Metadata creation tools
- Quality Control
- Clearinghouse

These services are free for USGS and its partner organizations
CSAS Metadata Program: Key Players

- **Viv Hutchison**
  - US Geological Survey
  - Denver, CO
  - Data Management Program Coordinator
  - vhutchison@usgs.gov

- **Giri Palanisamy**
  - Oak Ridge National Lab
  - Oak Ridge, TN
  - Clearinghouse technical manager

- **Colin Talbert**
  - US Geological Survey
  - Fort Collins, CO
  - Metadata quality control

- **Laurel Cepero**
  - NASA
  - Greenbelt, MD
  - Metadata record creation
Welcome

**CSAS Metadata Program Vision:**

The CSAS Metadata Program seeks to achieve a workflow environment that assists USGS scientists with the implementation of critical data management activities.

**CSAS Metadata Program Mission:**

The mission of the CSAS Metadata Program is to provide USGS and its partners with data documentation services that make critical data management activities readily accessible and easy to use. Services provided include: metadata creation and quality control assistance, metadata authoring tools, educational and training workshops, and a searchable metadata clearinghouse.

**FGDC Metadata**

Metadata is a description of the content, quality, lineage, contact, condition, and other characteristics of data. The description of the data is organized in a standardized format using a common set of terms. Metadata is literally "data about data". Metadata records are similar in concept to library catalog records: details about a book such as title, author, and publisher are recorded in a standard way to ease the search for information. Metadata ultimately makes information about data sets more easily accessible to scientists and researchers.

Metadata is a valuable tool. Since metadata records preserve the usefulness of data over time by detailing methods for data collection and data set creation, Metadata greatly minimizes duplication of effort in the collection of expensive data, and fosters, sharing of digital data resources. Metadata supports local data asset management such as local inventory and data catalogs, and external user communities such as Clearinghouses and websites. It provides adequate guidance for end-use application of data such as detailed lineage and context. Metadata makes it possible for data users to search, retrieve, and evaluate data set information by providing standardized descriptions of geospatial and non-geospatial data.

Core Science Analytics and Synthesis (CSAS) Metadata Program offers an extensive array of metadata resources. Explore the links and find resources for your organization to use.
Tools

- https://mercury.ornl.gov/OME

Metadata Collection Form

Your Information:

* Full Name (First Name followed by Last Name): *

* Organization Name: *

Position Name:

* Telephone Number: *

* Email: *

* Address:
  * Street: *

* City: * Country: *

* State/Provence: * Zip Code: *

Describe your data. To create a citation for your data, the following information is needed:

* Title of the Data set (Include "where", "what", and "when" in the title of your data set): *
  e.g. Aquifer Systems and Recharge Potential in Louisiana from LDEQ source data, geographic NAD83, LOSCO(1999)

* Dataset Author/Originator (First Name followed by the Last Name):
  The person(s) who developed the dataset.

Organization Name:
  The organization that developed the dataset

More Author/Originator: +

Is this data set published? ☐ Yes ☐ No

* Publication Date (Format: YYYYMMDD): *

Publication Place:
http://mercury.ornl.gov/clearinghouse/
## Results

### Related Topics
- Algae
- Freshwater plants
- Macrophytes
- Periphyton
- Seagrasses

### Presentation Category
- Maps and Data (3053)
- Tools and Software (343)
- Unknown (343)
- Publications (108)
- Tabular (4)

### Decade
- 2000 to 2009 (2321)
- 1990 to 1999 (647)
- Unknown (270)
- 2010 to 2019 (247)
- 1980 to 1989 (237)
- 1970 to 1979 (231)

### Originator
- Tilman, David (473)
- Jornada Basin LTER (449)
- Hueneke, Laura (354)
- PISCO (333)
- Menge, Bruce (326)

### Viewing Documents 1 - 10 out of 4015

<table>
<thead>
<tr>
<th>Sort By:</th>
<th>Relevance</th>
<th>Publication Date</th>
<th>Provider</th>
</tr>
</thead>
</table>

### PLANTS (ÄJtte)
**Data provider:** GLOBAL BIODIVERSITY INFORMATION FACILITY NETWORK

Database of specimens from the plant collection of Äjtte museum, Jokkmokk, Sweden...

[Get data](#) [View full metadata](#)

### SOUTHWEST NON-NATIVE INVASIVE PLANT DATABASE (SWEMP07)
**Data provider:** METADATA CLEARENCHQUE PRINCIPAL NCDE

The Southwest Exotic Plant Mapping Program (SWEMP) is a collaborative effort between the United States Geological Survey and federal, tribal, state, county and NGO partners in the southwest. This project is an ongoing effort to compile and distribute regional data on the occurrence of non-native inv.

[Get data](#) [View full metadata](#)

### BOTANICAL MUSEUM, DENMARK. DATABASE OF REGISTRATIONS OF RED LISTED PLANTS
**Data provider:** GLOBAL BIODIVERSITY INFORMATION FACILITY NETWORK

Database of registrations of redlisted plants of Denmark, registered as specimens at the museum or in the "TBU"- archives. The Danish Biodiversity Information Facility..

[Get data](#) [View full metadata](#)
| Title: | Ibaraki Nature Museum, Dr. Masatomo Suzuki collection: Vascular Plants (1) |
| Investigator(s): | Ibaraki Nature Museum, Dr. Masatomo Suzuki collection: Vascular Plants (1) (<a>more</a>) |
| Abstract: | Vascular Plants collected by Dr. Masatomo Suzuki in Japan |
| Metadata Reference Information: |  |
| Metadata Date: | N/A |
| Metadata Review Date: | N/A |
| Metadata Future Review Date: | N/A |
| Metadata Contact: |  |
| Organization: | Seichi Kokufozuka |
| Email: |  |
| Phone: | N/A |
| Address: |  |
| City: | N/A |
| State: | N/A |
| Postal: | N/A |
| Phone: | N/A |
| Fax: | Email: |
| Site Information: |  |
| Site: | Eastern Asia |
| North: | 44.3533 |
| West: | 124.157 |
| South: | 24.3473 |
| East: | 145.584 |
| Keyword(s): |  |
| Theme Keyword Thesaurus: | N/A |
| Theme Keyword: | N/A |
| Status: | Complete |
| Access Restriction: | Allow read to users: public |
| Data Center Contact: | City: N/A |
| State: | N/A |
| Postal: | N/A |
| Phone: | Email: N/A |
| Resource Description: | eml.gbif.package.1813 |
| Use Constraints: | N/A |
Southwest Non-native Invasive Plant Database (SWEMP07)

Metadata:

- Identification Information
- Data Quality Information
- Spatial Data Organization Information
- Spatial Reference Information
- Entity and Attribute Information
- Distribution Information
- Metadata Reference Information

Identification Information:

Citation:

Originator:
Kathryn Thomas and Patricia Guertin. U.S. Geological Survey, Southwest Biological Science Center (USGS-SBSC)

Publication Date: 20070508

Title: Southwest Non-native Invasive Plant Database (SWEMP07)

Geospatial Data Presentation Form: tabular digital data

Online Linkage:

Description:

Abstract:
The Southwest Exotic Plant Mapping Program (SWEMP) is a collaborative effort between the United States Geological Survey and federal, tribal, state, county and NGO partners in the southwest. This project is an ongoing effort to compile and distribute regional data on the occurrence of non-native invasive plants in the southwestern United States. The database represents the known sites (represented by a point location, i.e. site) of non-native invasive plant infestations within Arizona and New Mexico, and adjacent portions of California, Colorado, Nevada and Utah. These data, collected from 1911 to 2006, represent the field observations of various state, federal, tribal and county agencies, along with some specimen data from Herbaria. The SWEMP database is published at least once a year and consists of a compilation of all data submitted up to the date of publication.

Purpose:
This dataset was created to provide a regional perspective on non-native invasive plant distributions. It can be used to assist land managers, as well as the public, to review the locations and extent of reported infestations. These data can ultimately help guide management strategies and policies for the control of non-native invasive plant species. All plant species in the database are non-native as defined by the USDA PLANTS database 2007; the extent to which they are invasive has not been determined.
USGS CSAS Clearinghouse Dashboard

Welcome to the NBII Clearinghouse Metadata Provider Dashboard

The NBII appreciates your submissions to the Clearinghouse. We are striving to make useful, robust records available to the scientific community. Metadata records are important in that they provide scientists the ability to share detailed information about their datasets. It is important that we, in our data sharing efforts, strive to maintain high quality metadata - both for data discovery and data use purposes. As an aggregator of metadata records, the NBII harvests metadata from a large number of providers. This dashboard has been created to help you, as an organization providing metadata records to NBII, to check on the status and quality of these records. The dashboard allows you to check on your metadata submissions in a variety of ways.

Thank you for your commitment to providing metadata records to the NBII Clearinghouse! We hope you find the Provider Dashboard a useful experience. Check back often, as new content will be added.

<table>
<thead>
<tr>
<th>Title</th>
<th>Link</th>
<th>Date Broken</th>
</tr>
</thead>
<tbody>
<tr>
<td>1989 Land Cover/Land Use on the Upper Mississippi River System</td>
<td>Missing link</td>
<td>Tue Jan 04 17:32:48 EST</td>
</tr>
<tr>
<td>1990 land use for part of Muddy Creek basin, Rhode River, MD</td>
<td>Missing link</td>
<td>Tue Jan 04 17:29:08 EST</td>
</tr>
<tr>
<td>1995 International Crown Fire Experiment in Northern Canada</td>
<td>Missing link</td>
<td>Tue Jan 04 16:13:04 EST</td>
</tr>
</tbody>
</table>

Total broken links: 1909

Overall Clearinghouse Searches

<table>
<thead>
<tr>
<th>Month</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan</td>
<td>767</td>
</tr>
<tr>
<td>Feb</td>
<td>430</td>
</tr>
<tr>
<td>Mar</td>
<td>406</td>
</tr>
</tbody>
</table>

Most Frequent Keywords Occurrences in all records

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>line</td>
<td>767</td>
</tr>
<tr>
<td>dlg</td>
<td>430</td>
</tr>
<tr>
<td>graph</td>
<td>406</td>
</tr>
</tbody>
</table>
Updates in the Standards World

How standards proliferate:
(see: A/C chargers, character encodings, instant messaging, etc.)

Situation: There are 14 competing standards.

14?! Ridiculous! We need to develop one universal standard that covers everyone's use cases. Yeah!

Soon:

Situation: There are 15 competing standards.
Transitions: FGDC to ISO standard

The international community, through the International Organization for Standardization (ISO) designed a standard for geospatial metadata.

ISO 19115 enables describing:
- Geospatial data sets
- Non-geospatial resources (example: tabular data)
- Services: portals, web mapping
Challenges in Metadata Standard Transitions

- Training – small number of experts on the new standard...how will the science community be trained?
- Tool maturity – new standard = new tools that need to work well with new standard
- Cultural – getting people to use the new standard
Making Metadata Transitions in Large Organizations

- FGDC held a “Metadata Summit” in October, 2011 – 52 participants from 25 different organizations
- ISO Training and Breakout sessions
- Made recommendations to FGDC in 3 areas:
  - Policy and Guidance
  - Tools and Applications
  - Education and Communication
Next Steps

- Agency Working Groups for Metadata to coordinate implementation of the transition
- Train-the-Trainer workshops to re-build the ability to educate
- Contribution to community tools
- Outreach

USGS
Some Parting Summary Thoughts

- Data without metadata has diminished value.
- Small projects grow large. Large projects need metadata to succeed. Don’t forget to plan for metadata when you start a project. It is an additional scaling consideration.
- There are lots of resources and approaches for metadata creation and implementation.
- Transitions in the standards are major undertakings, and take time.
Thank you for your time…Questions?

Contact:
Viv Hutchison (vhutchison@usgs.gov)
USGS – Core Science Analytics and Synthesis (CSAS)