Metadata

Part of EUDAT’s training on the Fundamentals of Data Infrastructures
What we’ll cover

1. What is metadata?
2. What is a metadata standard?
3. The value of metadata
4. What makes good metadata
5. The utility of metadata: data distribution, management & project management
6. How to write good metadata
7. Where do you put metadata?
8. The EUDAT Problem and what we are doing about it
1. WHAT IS METADATA?
What is metadata?

- Commonly interpreted as meaning ‘data about data’

- But according to wikipedia...
  - “The term metadata is ambiguous, as it is used for two fundamentally different concepts…”
  - … Which is where the confusion starts
The Myriad Types of Metadata

• Here begins the religious war!
  – Structural/Control Metadata and Guide Metadata
    • Bretheron&Singley - 1994
  – Technical, Business and Process
    • Ralph Kimball
  – Descriptive, Structural and Administrative
    • National Information Standards Organisation

• And that’s just the start!
  – Hierarchical vs. Flat
  – Keyword/value vs. RDF vs. XML
  – Language
But there is some Sanity!

• Standards!
  – Dublin Core designed to provide minimal set to allow discovery
  – Endorsed by IETF, ISO and NISO
  – ISO/IEC 11179-1 (2004) and subsequent standards

• But there are a lot of them 😞
  – DDI, CDWA, Darwin Core, CSDGM, EML, ISO19115...
Structural, Administrative and Descriptive Metadata

**Structural metadata**

Information about structure of the data
- Organisation of Information
- Detailed description of structure
  - Sections, ‘endian-ness’, magic numbers, magic strings
- ‘Data about the data container’
- Can be more complex that the data itself
  - Trying reading the full MP4 specification!

**Administrative metadata**

Information about file type
- Just because I name a file .pdf does not mean acroread can interpret it!
- Rights management
- Preservation metadata (Data provenance)

**Descriptive metadata**

Classical ‘Data about data’
- Describes content of data
- Often, but not necessarily, searchable
- Dublin Core represents a good minimal subset (**But is it too generic for science?**)
<table>
<thead>
<tr>
<th><strong>Author(s)</strong></th>
<th>Boullosa, Carmen.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Title(s)</strong></td>
<td>They're cows, we're pigs / by Carmen Boullosa</td>
</tr>
<tr>
<td><strong>Place</strong></td>
<td>New York : Grove Press, 1997.</td>
</tr>
<tr>
<td><strong>Physical Descr</strong></td>
<td>viii, 180 p ; 22 cm.</td>
</tr>
<tr>
<td><strong>Subject(s)</strong></td>
<td>Pirates Caribbean Area Fiction.</td>
</tr>
<tr>
<td><strong>Format</strong></td>
<td>Fiction</td>
</tr>
</tbody>
</table>
2. WHAT IS A METADATA STANDARD?
What is a metadata standard?

• A Standard provides a structure to describe data with:
  – Common terms to allow consistency between records
  – Common definitions for easier interpretation
  – Common language for ease of communication
  – Common structure to quickly locate information

• In search and retrieval, standards provide:
  – Documentation structure in a reliable and predictable format for computer interpretation
  – A uniform summary description of the dataset
The metadata schema

- Metadata schema are sets of metadata elements designed for a specific purpose, such as describing a particular type of information resource. The resource that is being identified can be of any kind, but it is typically a dataset.
- Generally specify names of elements and their semantics. Option ally, they may specify content rules for how content must be formulated (for example, how to identify the main title), representation rules for content (for example, capitalization rules), and allowable content values (for example, terms must be used from a specified controlled vocabulary).
  - Metadata schema can be hierarchical in nature where relationships exist between metadata elements and elements are nested so that parent-child relationships exist between the elements. An example is the IEEE LOM schema where metadata elements may belong to a parent metadata element.
  - Metadata schema can also be one-dimensional, or linear, where each element is completely discrete from other elements and classified according to one dimension only. An example is Dublin Core schema which is one dimensional.
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 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 PWRRC statisticians have been instrumental in the development of innovative approaches for analyzing trends of wildlife populations.

Purpose:
The pesticides not only killed insects but also killed birds, raising serious concerns over its effects on bird population trends. Unfortunately, no long-term regional or continental population data were available for most bird species, making it difficult for birders to demonstrate declines in bird populations. The Bird Breeding Survey has proven to be a valuable source of information on bird population trends. Robbins et al. (1986) provided the first continental relative abundance maps for various songbirds based on BBS data. When viewed at continental or regional scales, these maps provide a reasonably good indication of the relative abundance of species that are well sampled by the BBS. In addition, the BBS is a good source of information on temporal patterns in trends. Populations of permanent resident and short-distance migrant (birds wintering primarily in the...
Multiple Metadata Standards Exist: Examples

- **Dublin Core Element Set**
  - Emphasis on web resources, publications
- **FGDC Content Standard for Digital Geospatial Metadata (CSDGM)**
  - Emphasis on geospatial data
  - **Biological Data Profile (BDP) of the CSDGM**
  - Profile to the CSDGM emphasis on biological data (and geospatial)
- **ISO 19115/19139 Geographic information: Metadata**
  - Emphasis on geospatial data and services
Multiple Metadata Standards Exist: Examples

• **Ecological Metadata Language (EML)**
  – Focus on ecological data

• **Darwin Core**
  – Emphasis on museum specimens
  – [http://rs.tdwg.org/dwc/index.htm](http://rs.tdwg.org/dwc/index.htm)

• **Geography Markup Language (GML)**
  – Emphasis on geographic features (roads, highways, bridges)
# Comparing Metadata Standards

<table>
<thead>
<tr>
<th>EML</th>
<th>FGDC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>Title</td>
</tr>
<tr>
<td>Abstract</td>
<td>Abstract</td>
</tr>
<tr>
<td>Entity Description</td>
<td>Entity Type Definition</td>
</tr>
<tr>
<td>Intellectual Rights</td>
<td>Use Constraints</td>
</tr>
</tbody>
</table>
Choosing a Metadata Standard

• Many standards collect similar information
• Factors to consider:
  – Your data type:
    • Are you working mainly with GIS data? Raster/vector or point data? Do you have biological or shoreline information in your dataset?
      – Consider the FGDC Content Standard for Digital Geospatial Metadata with one of its profiles: the Biological Data Profile or the Shoreline Data Profile.
    • Are you working with data retrieved from instruments such as monitoring stations or satellites? Are you using geospatial data services such as applications for web-mapping applications or data modeling?
      – If so, then consider using the ISO 19115-2 standard
    • Are you mainly working with ecological data?
      – Consider Ecological Metadata Language (EML)
Choosing a Metadata Standard

• More Factors to consider:
  – Your organization’s policies: do they state which standard to use?
  – What resources are available to create metadata?

Examples of Tools:
  • **FGDC CSDGM:**
    » Metavist (Forest Service) [http://ncrs.fs.fed.us/pubs/viewpub.asp?key=2737](http://ncrs.fs.fed.us/pubs/viewpub.asp?key=2737)
  
  • **EML:**
    » Morpho [http://knb.ecoinformatics.org/morphoportal.jsp](http://knb.ecoinformatics.org/morphoportal.jsp)
  
  • **ISO:** [http://www.fgdc.gov/metadata/iso-metadata-editor-review](http://www.fgdc.gov/metadata/iso-metadata-editor-review)
    » XML Spy or Oxegyn
    » CatMD

  – Other factors: Availability of human support; instructional materials; use of controlled vocabularies; output formats
3. THE VALUE OF METADATA
The value of metadata

Data developers

Data users

Metadata helps...

Organizations
Concerns About Creating Metadata

Even if the value of data documentation is recognized, concerns remain as to the effort required to create metadata that effectively describe the data.
## Concerns About Creating Metadata

<table>
<thead>
<tr>
<th>Concern</th>
<th>Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>workload required to capture accurate robust metadata</td>
<td>incorporate metadata creation into data development process – distribute the effort</td>
</tr>
<tr>
<td>time and resources to create, manage, and maintain metadata</td>
<td>include in grant budget and schedule</td>
</tr>
<tr>
<td>readability / usability of metadata</td>
<td>use a standardized metadata format</td>
</tr>
<tr>
<td>discipline specific information and ontologies</td>
<td>‘profile’ standard to require specific information and use specific values</td>
</tr>
</tbody>
</table>
What is the Value to Data Developers?

• Metadata allows data developers to:
  – Avoid data duplication
  – Share reliable information
  – Publicize efforts – promote the work of a scientist and his/her contributions to a field of study
What is the Value to Data Users?

- Metadata gives a user the ability to:
  - Search, retrieve, and evaluate data set information from both inside and outside an organization
  - Find data: Determine what data exists for a geographic location and/or topic
  - Determine applicability: Decide if a data set meets a particular need
  - Discover how to acquire the dataset you identified; process and use the dataset
What is the Value to Organizations?

• Metadata helps ensure an organization’s 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 an organization’s research:
  – Creates possible new partnerships and collaborations through data sharing
The phenomenon of “information entropy” associated with research

- Specific details about problems with individual items or specific dates are lost relatively rapidly
- General details about datasets are lost through time
- Accident or technology change may make data unusable
- Retirement or career change makes access to “mental storage” difficult or unlikely
- Loss of data developer leads to loss of remaining information

(From Michener et al 1997)
Information Entropy

Sound information management, including metadata development, can arrest the loss of dataset detail.
4. WHAT MAKES GOOD METADATA
Why use metadata?

• (Good) Metadata allows:
  – Discovery (i.e. searchable)
  – Interpretation
  – Limitations
  – Ownership and restrictions
  – Interoperability
What makes good metadata

• Use of standards
• Controlled Vocabulary
  – Unambiguous keywords
• Simple, Complete and Consistent
• Appropriate Description
• State Limitations
• Avoid special characters (,!@<~ etc.)
• DOI’s to the data
The good and the bad

**Good**
- Metres/second
- 2012-09-10T15:00:01+01:00
- Longitudinal wind speed
- PDF 1.7
- 2008 US Population Statistics
- Barcelona, Venezuela
- Not for ‘Navigation’

**Bad**
- Furlongs/fortnight
- 10th Sept. 2012 15:00:01
- U
- PDF
- Population Statistics
- Barcelona
- ‘For Information Only’
Review your Metadata!

- Could someone use an automatic search to locate the data?
- Can they assess its usefulness?
- Can a novice understand it?
- Is it specific enough?
- Is there enough to use/re-use the data?
- Is anything missing (e.g. Projection)?
- Is it unambiguous?
5. THE UTILITY OF METADATA: DATA DISTRIBUTION, MANAGEMENT & PROJECT MANAGEMENT
A Closer Look: The Utility of Metadata

• Metadata can support:
  – data distribution
  – data management
  – project management

• If it is:
  – considered a component of the data
  – created during data development
  – populated with rich content
Data Distribution via Metadata

data discovery

metadata publication

data portals
The descriptive content of the metadata file can be used to identify, assess, and access available data resources.
Distribution: Metadata Publication

- A metadata collection can be published to the internet via:
  - website catalog
  - web accessible folder (WAF)
  - Z39.50 metadata clearinghouse
  - metadata service
  - geospatial data portal
  - Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)
Distribution: Geospatial Data Portals

- **Examples of metadata search portals:**
  - **Data.gov**
    - Federal e-gov geospatial data portal
      - [http://www.geo.data.gov](http://www.geo.data.gov)
  - **Metacat**
    - Repository for data and metadata
      - [http://knb.ecoinformatics.org/index.jsp](http://knb.ecoinformatics.org/index.jsp)
  - **US Geological Survey**
    - USGS Core Science Metadata Clearinghouse:
      - [http://mercury.ornl.gov/clearinghouse](http://mercury.ornl.gov/clearinghouse)
  - **ArcGIS Online**
    - ESRI sponsored national geospatial data portal
      - [http://www.geographynetwork.com](http://www.geographynetwork.com)
Data Management via Metadata

- Maintenance & Update
- Data Accountability
- Discovery & Re-use
- Data Liability
Management: Maintenance and Update

- Metadata records can be used to track data provenance accuracy

- Data Maintenance:
  - Are the data current?
    - Do we have data older than ten years?
    - Was before some political or geophysical event that resulted in significant change?
  - Are the data valid?
    - Prior to most current source data
    - Prior to most current methodologies

- Data Update:
  - Contact information
  - Distribution policies, availability, pricing, URLs
  - New derivations of the dataset
Management: Data Discovery & Reuse

If you create metadata, other people can discover your data.
If you create metadata, you can find your own data.

• Find your data by:
  – themes / attributes
  – geographic location
  – time ranges
  – analytical methods used
  – sources and contributors
  – data quality
Management: Data Accountability

• Metadata allows you to **repeat scientific process** if:
  – methodologies are defined
  – variables are defined
  – analytical parameters are defined

• Metadata allows you to **defend your scientific process**:
  – demonstrate process
  – increasingly GIS-savvy public requires metadata for consumer information

• Metadata is an exercise in data accountability. It requires you to assess:
  – What do you know about the dataset?
  – What **don’t** you know about the dataset?
  – What **should** you know about the dataset?
Management: Data Liability

- Metadata is a declaration of:
  - **Purpose**
    - the originator’s intended application of the data
  - **Use Constraints**
    - inappropriate applications of the data
  - **Completeness**
    - features or geographies excluded from the data
  - **Distribution Liability**
    - explicit liability of the data producer and assumed liability of the consumer
Project Management via Metadata

- Project Planning
- Project Monitoring
- Project Coordination
- Contract Deliverables
Project Management: Project Planning

• Metadata records can serve as a project design document:
  – descriptions & intent of project
  – geographic and temporal extent of project
  – source data of project
  – attribute requirements of project

• Benefits:
  – expectations are clearly outlined
  – metadata is integrated into the process
  – provides a medium to record progress

• Use metadata to monitor:
  – data development status
  – QA/QC assessments
  – needed changes in approach

*Monitoring requires that the metadata be actively maintained and reviewed!*
Project Management: Project Coordination

• Metadata can be a means to improve communications among project participants using common:
  – descriptions & parameters
  – keywords, vocabularies, thesauri
  – contact information
  – attributes
  – distribution information

• If reviewed regularly by all participants, metadata created early and updated during the project improves opportunity for coordinating:
  – source data
  – analytical methods
  – new information
Project Management: Contract Deliverables

• As a key component of the data, metadata should be part of any data deliverable

• For quality metadata from a deliverable, the record should provide:

  – Citation information
  – Data quality information
  – Accurate geospatial information
  – Clearly defined entities and attributes
  – Distribution information
6. HOW TO WRITE GOOD METADATA
Steps to Create Quality Metadata

- Organize your information
  - Did you write a project abstract to obtain funding for your proposal? Re-use it in your metadata!
  - Did you use a lab notebook or other notes during the data development process that define measurements and other parameters?
  - Do you have the contact information for colleagues you worked with?
  - What about citations for other data sources you used in your project?
- Write your metadata using a metadata tool
- Review for accuracy and completeness
- Have someone else read your record
- Revise the record, based on comments from your reviewer
- Review once more before you publish
Steps to Create Quality Metadata

- Write your metadata using a metadata tool
- Review for accuracy and completeness
- Have someone else read your record
- Revise the record, based on comments from your reviewer
- Review once more before you publish
Steps to Create Quality Metadata

- Review for accuracy and completeness
- Have someone else read your record
- Revise the record, based on comments from your reviewer
- Review once more before you publish
Tips for Writing Quality Metadata

- Do not use jargon
- Define technical terms and acronyms:
  - CA, LA, GPS, GIS: what do these mean?
- Clearly state data limitations
  - E.g., data set omissions, completeness of data
  - Express considerations for appropriate re-use of the data
- Use “none” or “unknown” meaningfully
  - None usually means that you knew about data and nothing existed (e.g., a “0” cubic feet per second discharge value)
  - Unknown means that you don’t know whether that data existed or not (e.g., a null value)
Tips for Writing Quality Metadata

Titles, Titles, Titles…

• Titles are critical in helping readers find your data
  – While individuals are searching for the most appropriate data sets, they are most likely going to use the title as the first criteria to determine if a dataset meets their needs.
  – Treat the title as the opportunity to sell your dataset.

• A complete title includes: What, Where, When, Who, and Scale

• An informative title includes: topic, timeliness of the data, specific information about place and geography
Tips for Writing Quality Metadata

- A Clear Choice: Which title is better?
- *Rivers*
  OR

Greater Yellowstone *(where)* Rivers *(what)* from 1:126,700 *(scale)* U.S. Forest Service *(who)* Visitor Maps *(when)*
Tips for Writing Quality Metadata

• Be specific and quantify when you can! The goal of a metadata record is to give the user enough information to know if they can use the data without contacting the dataset owner.

  • Vague: We checked our work and it looks complete.
  • Specific: We checked our work using a random sample of 5 monitoring sites reviewed by 2 different people. We determined our work to be 95% complete based on these visual inspections.
Tips for Writing Quality Metadata

- Select keywords wisely
- Use descriptive and clear writing
- Fully qualify geographic locations
- Use thesauri for keywords whenever possible
- Example: USGS Biocomplexity Thesaurus (over 9,500 terms)
Tips for Writing Quality Metadata

• Remember: a computer will read your metadata
• Do not use symbols that could be misinterpreted:
  Examples: ! @ # % { } | / \ < > ~
• Don’t use tabs, indents, or line feeds/carriage returns
• When copying and pasting from other sources, use a text editor (e.g., Notepad) to eliminate hidden characters
7. WHERE DO YOU PUT METADATA?
Where to put it?

- No easy answer to this!
- Searchable metadata
  - Accessible via web page or automated search
  - But what you define as searchable is your own business
- Some formats are ‘self describing’
  - Some metadata (of all types) is embedded in the file
  - netCDF, HDF, GRIB
- Some need reference to external documentation
  - MP4, PDF, ERS-SAR data format
- Some data describing collection and some for data objects?
8. THE EUDAT PROBLEM AND WHAT WE ARE DOING ABOUT IT
The EUDAT problem

• Summed up in one word: **Federation!**
• Several Projects already working on this: ANDS, DataONE, Europeana, PanData
• BUT...
  – Most federation systems are based on similar communities
  – EUDAT needs to deal with a wider base of sciences
  – Plus ‘Simple Store’ Use Case
The interdisciplinary problem

• Same term can be used by different disciplines:
  – Species for a chemists and a zoologists
  – Andromeda for an astronomer and a historian
    • Some domain knowledge necessary

• Meanings depend on measurement
  – Sea Surface Temperature can be from surface down to 20m

Need to know how to interpret information!!
EUDAT Plans

- Gather existing schemas (where they exist)
- Define a ‘core’ set of metadata
- Harvest Existing Metadata using OAI-PMH
- Find ways to obtain metadata for communities not using OAI-PMH
- Allow faceted browsing using SOLR/LUCENE
- Provide tools to allow users to comment on metadata
  - To improve quality
The Concept!

Data & Metadata

Simple Store

Replication Store

Metadata Index

Community Store

Community Store

Community Store

Community Store

Portal

Metadata

Metadata
The Implementation

EUDAT non xml comm

EUDAT xml communities

External xml communities

www

MD ingestion service

OAI-PMH server

OAI Harvester

EUDAT Primary MD Store (EPS)

Ontology for natural mapping entities

EUDAT Mapped MD Store (EMS)

Lucene SOLR

EUDAT non xml comm

OAI-PMH server

OAI Harvester
The Research

• Looking at different architectures
  – Single/Distributed MDS
  – Hierarchical MDS
  – Metadata caching
  – Optimising harvester
  – Alternatives to OAI-PMH
Summary

• Metadata is more complex than it seems
  – Not just ‘data about data’
  – Communities need to spend time understanding their metadata
  – Cross disciplinary access mean you/we need to think harder!

• EUDAT is working on a model and architecture
  – But it’s not an easy problem
  – It may take a few iterations to ‘get it right’
Acknowledgements

These slides were produced by TrustIT as part of the EUDAT project (www.eudat.eu). The slides are based on content written by Shaun de Witt (STFC) and the DataOne project.

You are welcome to re-use these slides under the terms of CC BY 4.0 (http://creativecommons.org/licenses/by/4.0/)