Information Management

Corinna Gries
• Data Management Best Practices
• Information Management Considerations for Collaborative Projects
• Data Publication
• Resources
• Questions
Data Management Best Practices

- Data Management Plan
- Data Consistency
- Data Quality
- Data Files
Manage Your Data

Managing your data before you begin your research and throughout the research life cycle is essential to ensure usability, preservation and access. Federal agencies and other funders now require that grant awardees include a data management plan with their grant proposals. We can help you meet these new requirements.

UC3 is ready to consult with UC faculty members and researchers as you develop data management plans. Please contact UC3 directly, or contact your UC campus representative for more information.

Data Planning Checklist:

A. Your data management plan
B. Funding agency requirements
C. Creating your data
   1. What types of data will be produced for your project?
   2. What identifiers will you use for your data?
   3. How will you document your data?
   4. How much data will the project produce?
   5. How often will the data change or be updated, and will versions need to be tracked?
D. Organizing your data
   1. What file formats will be produced for your project and what kinds of data management risks do they present?
   2. How will you organize your files into directories and what naming conventions will you apply to both?
E. Managing your data
   1. Who is responsible for managing and controlling the data?
   2. For what or whom are the data intended?
   3. How long must the data be retained?
   4. How secure are the data? Do you have a procedure for backing up the data?
F. Sharing your data
   1. Does project funding require your data to be shared or publicly accessible?
   2. When and where do you intend to publish or distribute your data?
   3. How do I cite data?
   4. Are there issues with privacy or intellectual property?
G. Tools and expertise
   1. Where can I go for more help?
   2. Can Merritt help me manage my data?
   3. Can EZID help me create and manage long-term identifiers?
   4. Can I use Chromopolis to preserve my data?

Credit to MIT Libraries for permission to use and adapt their pages and to members of the UC3 community.
Please send us any comments about these guidelines.

Last updated December 14, 2012
Document owner: Perry Willett
Data Consistency

- Site Identification
  - Less meaning is better – no dates, no treatment indicator, etc.
  - Latitude Longitude, Geographic Place Names Thesauri
- Parameters
  - EPA Storet codes
- Species Names
  - Authority databases
    - ITIS: Yellow Bass, TSN 167683, Morone mississippiensis Jordan and Eigenmann in Eigenmann, 1887
- Dates and Times
  - Date format, time zone, daylight savings, am pm vs. military time
  - ISO
    - YYYYMMDDTHHMMSSZ (UTC)
    - YYYY-MM-DDTHH:MM:SS ±HH
Data Quality

• In the field (calibration, set up, etc.)
• Basic
  • Columns line up, each row has the same number of columns
  • Each column contains one type of data
  • Number of records is correct
  • File integrity - checksum, bytes
• Automated quality control
  • Basic statistics: range checks, step functions, outliers, long term drift
• Manual quality control
  • Graphing, integrating, analyzing
• Gap filling
  • Modeled, extrapolated from nearby station data
Data Files

• Use descriptive file names
  • ntl_wtemp_1981_2013_qcl1_v1.csv
  • No special characters
  • Version number
  • Quality control level (e.g. 0 = raw, 1 = automated QC, 2 = eyes on, 3 = gap filled)
  • Consistent over time

• Store in a non-proprietary format
  • Table: ASCII comma or tab delimited .csv, .txt
  • Raster: ASCII grid, Geo-tiff, NetCDF
  • Vector: Shapefile, KML
  • Image: jpg?
Information Management for Collaborative Projects

- Information Management for Communication
- Data
  - Secure, back up
  - Protect from changing
  - Data versioning
  - Data quality control levels
- Metadata
- Protocols
  - Document approaches for sampling and data manipulations
  - Versions
- Project Management System
  - Planning documents
  - Meeting notes
  - Code management
  - Progress tracking
Data Publication

• Document the Data (metadata)
  • What, When, Where, How, Who, Why
  • Discover, Assess, Access
  • Metadata format (EML, ISO, FGDC)

• Data Repository
  • DataONE, CUAHSI, OPeNDAP (Networks of repositories)
    • GLEON
    • KNB
    • LTER
    • DRYAD

• Unique Identifier – citable unit
Metadata Tools

- ReadMe text file, word document, template
- DataUp
- Morpho
- DEIMS (Drupal Environmental Information Management System)
- ArgGIS
NTL Metadata Template

- If submitting an Excel spreadsheet, please make sure it does not contain any formulas and comments on cells. If you need comments put them in their own column.
- Is this dataset similar enough to an earlier submitted dataset that metadata can be used from that earlier submission? (list a dataset title)

- Dataset Title
  - (be descriptive, more than 5 words):

- Short name or nickname you use to refer to this dataset:

- Abstract
  - (include what, why, where, when, and how)

- Investigators
  - (list in order as for a paper)

- Other personnel names and roles
  - (field crew, data entry etc.)
  - This document liberally borrows from similar documents at SBC and GCE
NTL Metadata Template

- **Keywords**
  - (list and separate by comma, please check out these resources [http://vocab.lternet.edu](http://vocab.lternet.edu), )

- **Choose one or more major subject areas**
  - primary production
  - population dynamics
  - organic matter and flux
  - inorganic matter and flux
  - disturbance

- **Timeframe**
  - Begin date
  - End date
  - Data collection ongoing/completed

- **Geographic location**
  - Verbal description:
  - North bounding coordinates (decimals)
  - South bounding coordinates (decimals)
  - East bounding coordinates (decimals)
  - West bounding coordinates (decimals)

- **Taxonomic species or groups**

- **Methods**
  - (please be specific, include instrument descriptions, or point to a protocol online)
### NTL Metadata Template

- **Data Table**
- Column name: exactly as it appears in the document. Please avoid special characters, dashes and spaces.
- Description: please be specific, it can be lengthy
- Unit: please avoid special characters and describe units in this pattern: e.g. microSiemenPerCentimeter, microgramPerLiter, absoptionPerMolePerCentimeter
- Code explanation: if you use codes in your column, please explain in way: e.g. LR=Little Rock Lake, A=Sample suspect, J=Nonstandard routine followed
- Data format: please tell us exactly how the date and time is formatted: e.g. mm/dd/yyyy hh:mm:ss
- If a code for ‘no data’ is used, please specify: e.g. -99999

<table>
<thead>
<tr>
<th>Column name</th>
<th>Description</th>
<th>Unit or code explanation or date format</th>
<th>Empty value code</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
DataUp

- Excel plug-in
DataUp

CREATE METADATA

It generates and applies metadata based on a pre-defined schema. Some of the metadata values are automatically populated by the system – the remaining must be supplied by you. Some of the metadata fields are required, designated by red asterisk, and some are optional. Please make sure you provide at least all the required metadata values.
DataUp

![Excel spreadsheet with DataUp interface](image)
Morpho
Morpho

Welcome to Morpho!

Current profile: enviroeducator

Change profile: enviroeducator
Create a new profile...

Network Status: NOT Logged In

If you do not choose to login, you will be able to access only "public" network files as a Guest User.

Login to network using current profile:
Password: login

Work with your data...

Create a new data package...
Open an existing data package...
Search for an existing data package...
Documenting the ‘HOW’

• Sampling approach
• Sample processing
• Event documentation (change of method, observer, natural events) – data flagging
• Data Manipulations
  • Scripting the process
    • R, Matlab, workflow processor
  • Model output documentation
    • Version, parameter settings, assumptions
Resources

- [http://www.dataone.org/](http://www.dataone.org/) DataONE repository
- [http://www.opendap.org/](http://www.opendap.org/) OPeNDAP repository
- [http://his.cuahsi.org/](http://his.cuahsi.org/) CUAHSI Hydrologic Information System
- [http://knb.ecoinformatics.org/index.jsp](http://knb.ecoinformatics.org/index.jsp) KNB data repository and Morpho metadata editor
- [https://dmp.cdlib.org/](https://dmp.cdlib.org/) Data Management Plan Tool
- [http://www.getty.edu/research/tools/vocabularies/tgn/](http://www.getty.edu/research/tools/vocabularies/tgn/) Getty Thesaurus of Geographic Names
Resources

• [http://his.cuahsi.org/ontologyfiles.html](http://his.cuahsi.org/ontologyfiles.html), CUAHSI’s water ontology


Questions