Data Management Practices for Programming

ESIP Student Fellow – Sophie Hou (hou@illinois.edu)
ESIP Summer Meeting
July 2015

Table of Content

- Introduction: Data Management for Scientists Short Course

- Selected Data Management Topics and Discussion Questions:
  1. How can curating software program/scripts enhance my reputation?
  2. What are the basic curation elements to consider in order to facilitate a deployable open-source software applications/scripts?
  3. How to choose a program/file format and naming convention to enhance interoperability?
  4. What are the metadata content that should be created for the software applications/scripts?
  5. How to provide access to your software applications/scripts for a broader user community?

Introduction

Data Management for Scientists Short Course

- Developed between 2011 and 2013 by ESIP, in cooperation with NOAA and the Data Conservancy.
  - 12 individuals contributed as module authors, and the authors represented 12 different organizations covering federal agencies, academic institutions, information organizations, and data centers,
- Currently, there is a total of 35 modules available.
- The modules can be accessed free of charge through:
  - 1) ESIP Commons (http://commons.esipfed.org/datamanagementshortcourse)
  - 2) ESIPFED Vimeo (http://vimeo.com/album/2142831)
- The Short Course has been presented to these audiences at events with data management focus in order to collect and review their feedback.

Data Management for Scientists Short Course - Continued

Titles for: Responsible Data Use Section

- Citation and Credit
- Copyright and Data
- Data Restrictions

Titles for: Data Management Plans Section

- Why Create a Data Management Plan?
- Elements of a Data Management Plan
    - Identifying the materials to be created
    - Organization and Standards

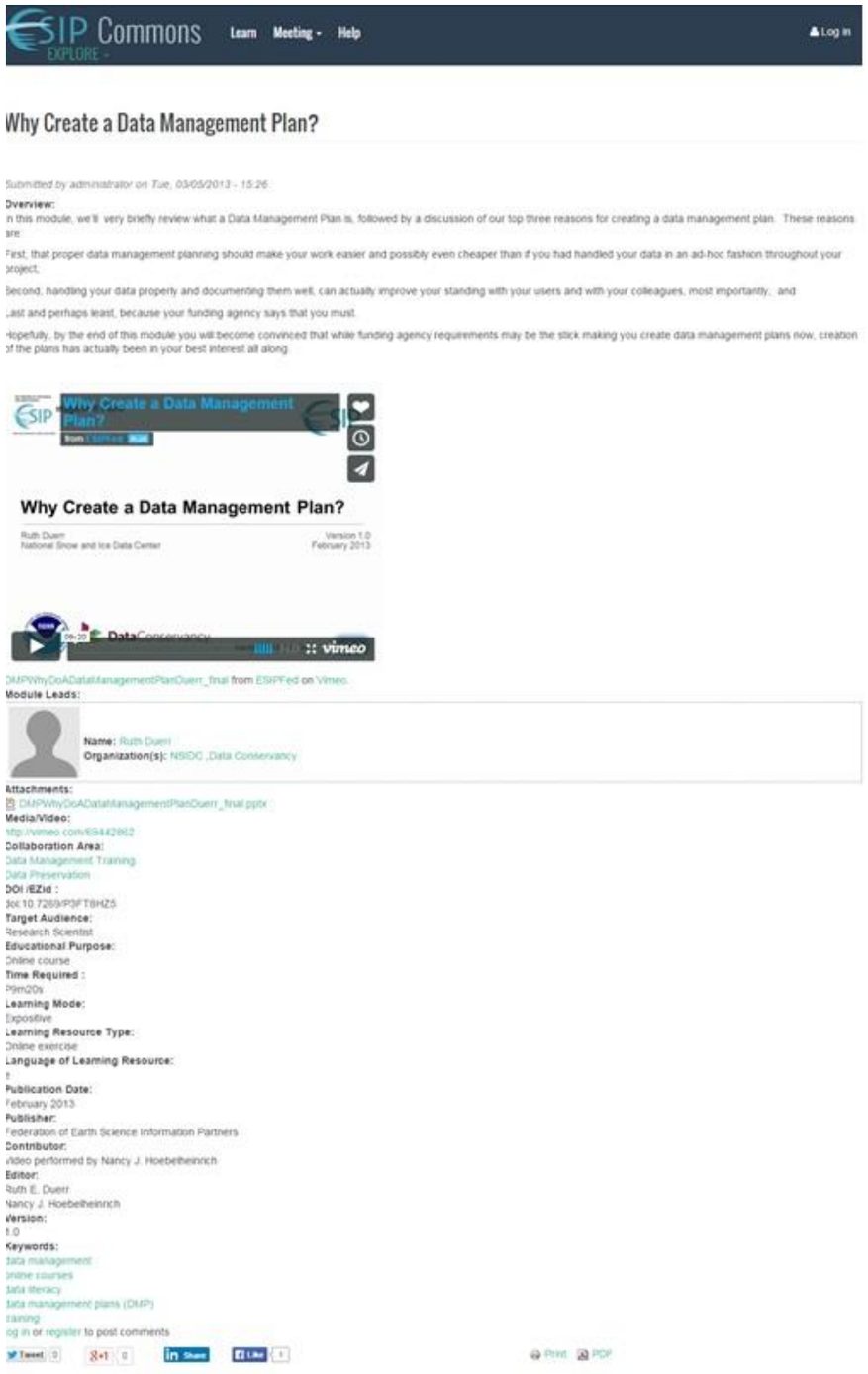
Titles for: Local Data Management Section

- Managing Your Data
    - Assign Descriptive File Names
    - Backing Up Your Data
- Data Formats
    - Choosing and Adopting Community Accepted Standards
    - Using Self-describing Data Formats
- Creating Documentation and Metadata
    - Introduction to Metadata and Metadata Standards
    - Creating a Citation for Your Data
    - Metadata for Discovery
- Working with Your Archive: Broadening Your User Community
- Advertising your data
    - Agency requirements for submitting metadata
    - Using data portals and metadata registries
- Using Data Portals and Metadata Registries: Submitting Metadata to the GCMD
- Providing Access to Your Data
    - Determining your audience
    - Access Mechanisms
    - Tracking Data Usage
    - Handling sensitive data
    - Rights

Titles for: The Case for Data Stewardship Section

- Agency Requirements
    - NASA Data Management Plans
    - NSF Data Management Plans
    - NOAA Administrative Order 212-15: Management of Environmental and Geospatial Data and Information
- Enhancing Your Reputation
- Preserving the Scientific Record
    - Establishing Relationships with Archives
    - Preserving a Record of Environmental Change
    - Case Study 1 – National Snow & Ice Data Center (NSIDC) Glacier Photos
    - Case Study 2 – Arctic Temperature Variability Data

Data Management for Scientists Short Course - Continued



The Case for Data Management: Agency Requirements

from **ESIPFed** PLUS 2 years ago / NOT YET RATED

This is the Federation of Earth Science Information Partners Data Management Short Course, Section: The Case for Data Management, Module: Agency Requirements

This training module is part of the Federation of Earth Science Information Partners (or ESIP Federation's) Data Management Course for Scientists. The subject of this module is "The Case for Data Management: Agency Requirements." This module was authored by Ruth Duerr from the National Snow and Ice Data Center. Besides the ESIP Federation, sponsors of this Data Management Course are the Data Conservancy and the United States National Oceanographic and Atmospheric Administration (NOAA).

Follow + Collections Stats Download vimeo.com/45318853

How can curating software program/scripts enhance my reputation?

Mayernik, M. 2012. “The Case for Data Stewardship: Enhancing Your Reputation.” In Data Management for Scientists Short Course, edited by Ruth Duerr and Nancy J. Hoebelheinrich, Federation of Earth Science Information Partners: ESIP Commons. doi:10.7269/P34M92G1

Scientific Reputation

- Reputation is central to the scientific community
- Researchers build a reputation by:
  - Producing valuable results
  - Contributing constructively to scientific debates
  - Being good colleagues
- Peer-recognition influences:
  - Employment opportunities
  - Promotions
  - Ability to win further research funding

Photo: http://www.flickr.com/photos/sbengineer/3635764985/

Reputation and Data - Why

- Data re-use is growing in importance in almost all scientific fields.
  - Data re-use depends on the availability of trustworthy data sets
  - Trust in data is highly connected to the reputation of the data collectors and data archives
- Having a reputation for collecting and sharing high quality and well documented data makes it more likely that:
  - Other researchers will use your data
  - Other researchers will cite your data
  - Other researchers will share their data with you

Reputation and Data - How

- How to get a reputation for good data management?
  - Make data openly accessible by submitting to open data archives
  - Provide comprehensive metadata
  - Answer questions from data users in a timely manner
- How to ensure that your reputation for data management can grow?
  - Provide proper attribution when you use data collected by someone else
  - Cite data sets in your reference lists
  - Teach proper data management and data attribution to new scientists

Discussion Questions

- Do the same arguments for curating data also apply to researchers who develop software applications and scripts?
  - As evidences of contributions?
  - Allow reusability, reproducibility, and verification?
  - Importance and emphasis of quality?
  - Building collaboration?
  - Availability of documentation/metadata?
- When and how would you currently cite or acknowledge the use of software applications and scripts?
- Does citing software applications and scripts help with building recognition and reputation?

What are the basic curation elements to consider in order to facilitate a deployable open-source software applications/scripts?

Duerr, R. 2013. “Data Management Plans: Elements of a Data Management Plan.” In Data Management for Scientists Short Course, edited by Ruth Duerr and Nancy J. Hoebelheinrich, Federation of Earth Science Information Partners: ESIP Commons. doi:10.7269/P31N7Z22

Elements of a Data Management Plan

Identify:

- Materials to be created
- Organization and standards
- Data access, sharing, and re-use policies
- Backups, archives, and preservation strategy
- Roles and responsibilities

Discussion Questions

- What are some of the “organizations and standards” that might apply to software applications and scripts?
  - Especially for the benefit of
    - Reducing costs
    - Supporting broader sharing
    - Increasing reusability
    - Increasing discoverability
    - Supporting preservation
- What are some of the gaps in community practices?

How to choose a program/file format and naming convention to enhance interoperability?

Cook, R. 2012. “Local Data Management – Managing Your Data: Assign Descriptive File Names.” In Data Management for Scientists Short Course, edited by Ruth Duerr and Nancy J. Hoebelheinrich, Federation of Earth Science Information Partners: ESIP Commons. doi:10.7269/P3F18WNR

Tilmes, C. 2013. “Local Data Management – Data Formats: Choosing and Adopting Community Accepted Standards.” In Data Management for Scientists Short Course, edited by Ruth Duerr and Nancy J. Hoebelheinrich, Federation of Earth Science Information Partners: ESIP Commons. doi:10.7269/P33N21B6

Relevance to Data Management

- Clear, descriptive, and unique file names may be important later when your data file is combined in a directory or FTP site with your own data files or with the data files of other investigators.

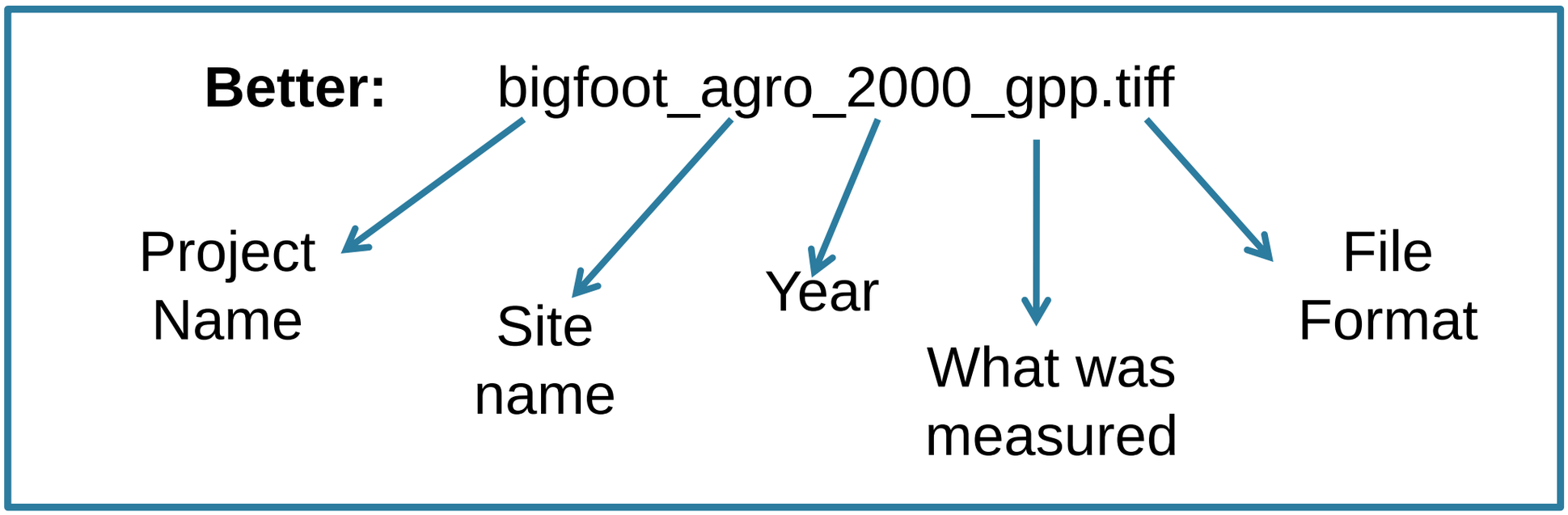
- File names that reflect the contents of the file and uniquely identify the data file enable precise search and discovery.

Assign descriptive file names

- Use descriptive file names
  - Unique
  - Reflect contents
  - ASCII characters only
  - Avoid spaces

Bad: Mydata.xls
2001_data.csv
best version.txt

- Provide an explanation of the convention used to name files



A few guidelines

- Consider your ***archive***:
  - Do they have any recommendations?
- Consider your ***users***:
  - Who wants this data? Why do they want it?
  - What do they want to do with it?
  - Will they be using your data in concert with other data?
- Consider ***heritage***:
  - What worked well for similar data in the past?
  - What could be done better for newly created data?
- Consider ***tools***:
  - Try to use data formats supported by the software you intend to use it with.

Discussion Questions

- What are the factors that influence the selection of software/program format?

- Do the developers and the users necessarily have the same needs and skillset for the software applications/scripts?

- What other information should be included in the naming of a software application/script?

What are the metadata content that should be created for the software applications/scripts?

Olsen, L. and T. Stevens, 2012. “Local Data Management – Creating Documentation and Metadata: Metadata for Discovery.” In Data Management for Scientists Short Course, edited by Ruth Duerr and Nancy J. Hoebelheinrich, Federation of Earth Science Information Partners: ESIP Commons. doi:10.7269/P3JS9NC5

Introduction to Discovery Level Metadata

- A data set description (metadata) that provides information to determine if a particular data set meets the users' needs.
- Typically provides essential information to enable a user to find out if a particular dataset exists, the data's location, and ownership, and how to obtain further information.
- The metadata includes the science discipline of the data, data location, spatial coverage, data provider, data resolution, data quality, etc.
- Discovery level metadata is found in "portals" and metadata registries.
- A controlled keyword vocabulary helps provide a consistent search and discovery of data.

Categories of Discovery Level Metadata

What: Title of Data Set and Keywords Describing the Data Set

Why: Description and Purpose of the Data Set

When: Temporal Coverage of the Data Set

Who: Data Set Creator and Contact

Where: Geographic Extent and Location of Data Set Coverage

How: How the Data Set was Created and How to Access the Data

Discussion Questions

- What are some of the ways that documentation or descriptions of software applications and scripts can be recorded?

- Would a standard metadata format work with the different types software and programming languages?

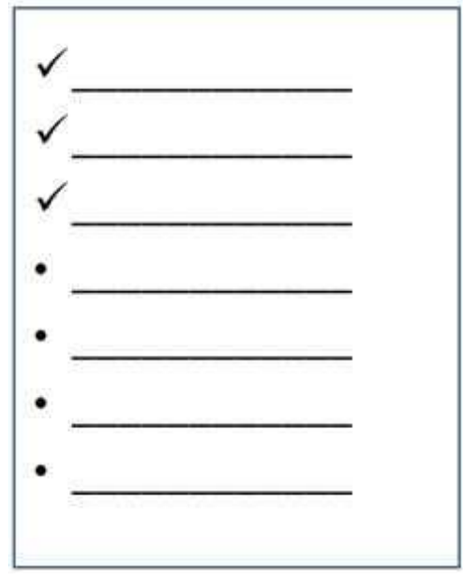
- Who should create the descriptions and who could benefit from the information?

How to provide access to your software applications/scripts for a broader user community?

Downs, R.R. 2013. “Local Data Management – Working with Your Archive: Broadening Your User Community.” In Data Management for Scientists Short Course, edited by Ruth Duerr and Nancy J. Hoebelheinrich, Federation of Earth Science Information Partners: ESIP Commons. doi:10.7269/P3NC5Z41

Develop plan to broaden your user community

- Work with archive to plan steps to broaden community
  - Prioritize activities based on assessment of users, uses, and gaps
- Initiate promotional activities or events
  - Announcements in newsletters, blogs, and relevant listservers
- Identify new opportunities to foster discovery
  - Catalogs, clearinghouses, search engine optimization
- Create new data products
  - Subsets, maps, integrated data, translations, lessons
- Develop new data services
  - Simple analytical tools, advanced tools, visualization capabilities
- Measure success of each planned activity
  - Determine expected outcomes and identify measurement criteria



Methods for broadening your user community

- Increase awareness
  - Ensure that data are being cited when used for publications
  - Improve data discoverability through inclusion in data catalogs
- Promote new and novel uses
  - Describe new uses for your data and feature articles about their use
- Improve capabilities
  - Provide guidance, instruction, and documentation on data use
  - Improve rights and remove any restrictions on data use
- Offer new tools and services
  - Develop easy-to-use tools that enable beginners to use your data
  - Create tools and services that foster data analysis
  - Convert data to formats that facilitate use by common tools
  - Enable integration of your data with other data products and services

Discussion Questions

- Who are some of the individuals or organizations that you have shared your software applications/scripts with?

- Who are some of the other resources outside of the immediate project group that you would like to obtain software applications/scripts from?

- What is the current preferred method for sharing software applications/scripts?