



NOAA Data Management Planning

Ruth Duerr
National Snow and Ice Data Center



Session Agenda

1:30 to 1:40

Overview of the current DMP landscape at NOAA

1:40 to 2:00

Data Sharing for NOAA Grants Procedural Directive

Ingrid Guch, NOAA NESDIS, STAR

The Data Sharing Procedural Directive for NOAA Grants was developed and approved this past year. Ingrid will give a brief overview of the new policy.

2:00 to 2:30

Implementation of Data Management Planning Procedural Directive

Jim Sargent, NOAA NMFS, SEFSC

The Data Management Planning procedural directive was developed and approved this past year. This year's focus will be on learning how to implement it and working together to making DMP a meaningful exercise that benefits all aspects of the data lifecycle. Jim will present the basic requirements laid out in the PD, and discuss the nascent workflow concepts for getting plans approved and submitted to the repository. He'll also talk about a collaborative approach to designing the wiki as a resource for those writing plans, with links to existing guidance, tools, examples, as well as a system for tracking issues.

2:30 to 3:30

Workshop on Data Management Planning

Ruth Duerr, NSIDC and the Data Conservancy

Once you have a DMP, it is time to implement it! Ruth will walk through each of the sections of the DMP Template describing what you should think about when writing your plan as well as how to implement the plan during the course of your research. This includes such topics as working with a long-term archive; as well responsibly using data produced by others.

3:30 to 3:45

Data Management Planning at the Office of Ocean Exploration

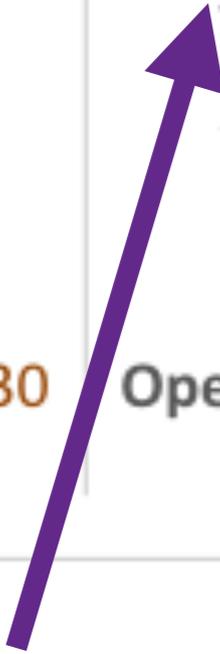
Susan Gottfried, NOAA NESDIS, NODC NCDDC (GDIT contractor)

A brief talk about Data Management Planning at the Office of Ocean Exploration will serve as a jump off point for an open discussion of existing practices and experiences.

3:45 to 4:30

Open discussion: Real-life examples, experiences, Q&A

Jacqueline Mize



EDMC Procedural Directives

Data Management Planning PD

Plan, in advance, how you will preserve, document and distribute your data.

Archive Procedure

What to archive, how to submit to archive.

Data Documentation
How to apply ISO 19115 metadata for discovery, use & understanding.

Data Access & Discovery
Provide on-line services so your data can be found and retrieved.

Data Sharing by NOAA Grantees
State how you will share data, and share within 2 years.

Data Citation
Use unique identifiers to allow data to be referenced and tracked.

In preparation



Outline

- ESIP data management training
- Why do a data management plan?
- Case studies
- The parts of a NOAA DMP
 - Description of data to be managed
 - Points of Contact
 - Data Stewardship
 - Documentation
 - Data Sharing
 - Initial Storage and Protection
 - Long-term Archiving and Preservation



ESIP Data Management Training

- With support from NOAA and the Data Conservancy, the ESIP Federation has been working on developing data management training for scientists and data managers
- ~ 50 draft modules have been contributed to date
- Each module is being peer-reviewed and may be published in the ESIP commons
- Published modules will have voiceovers professionally developed so that the module can be made available in webinar format.
- Draft modules are available at http://wiki.esipfed.org/index.php/Data_Management_Course_Outline
- Quite a few of the slides in my presentation today come from these modules



Outline

- ESIP data management training
- **Why do a data management plan?**
- Case Studies
- The parts of a NOAA DMP
 - Description of data to be managed
 - Points of Contact
 - Data Stewardship
 - Documentation
 - Data Sharing
 - Initial Storage and Protection
 - Long-term Archiving and Preservation



Top three reasons

- To make your research easier, more efficient, and less expensive
- To enhance your reputation
- Because many funding agencies require it



Making Your Research Easier and Cheaper

The 5 P's matter!

(prior planning prevents poor performance)

What? You think it's cheap to recover data off a

- Broken DVD
- A burned up memory stick
- A drowned laptop
- A crashed hard drive





Making Your Research Easier and Cheaper

Accidents do happen!



Don't you think it would be more efficient

- If you didn't have to remember
 - the name of that file?
 - and the directory where you put it?
 - the units those measurements were taken in?
 - which sample site was which?
 - etc.

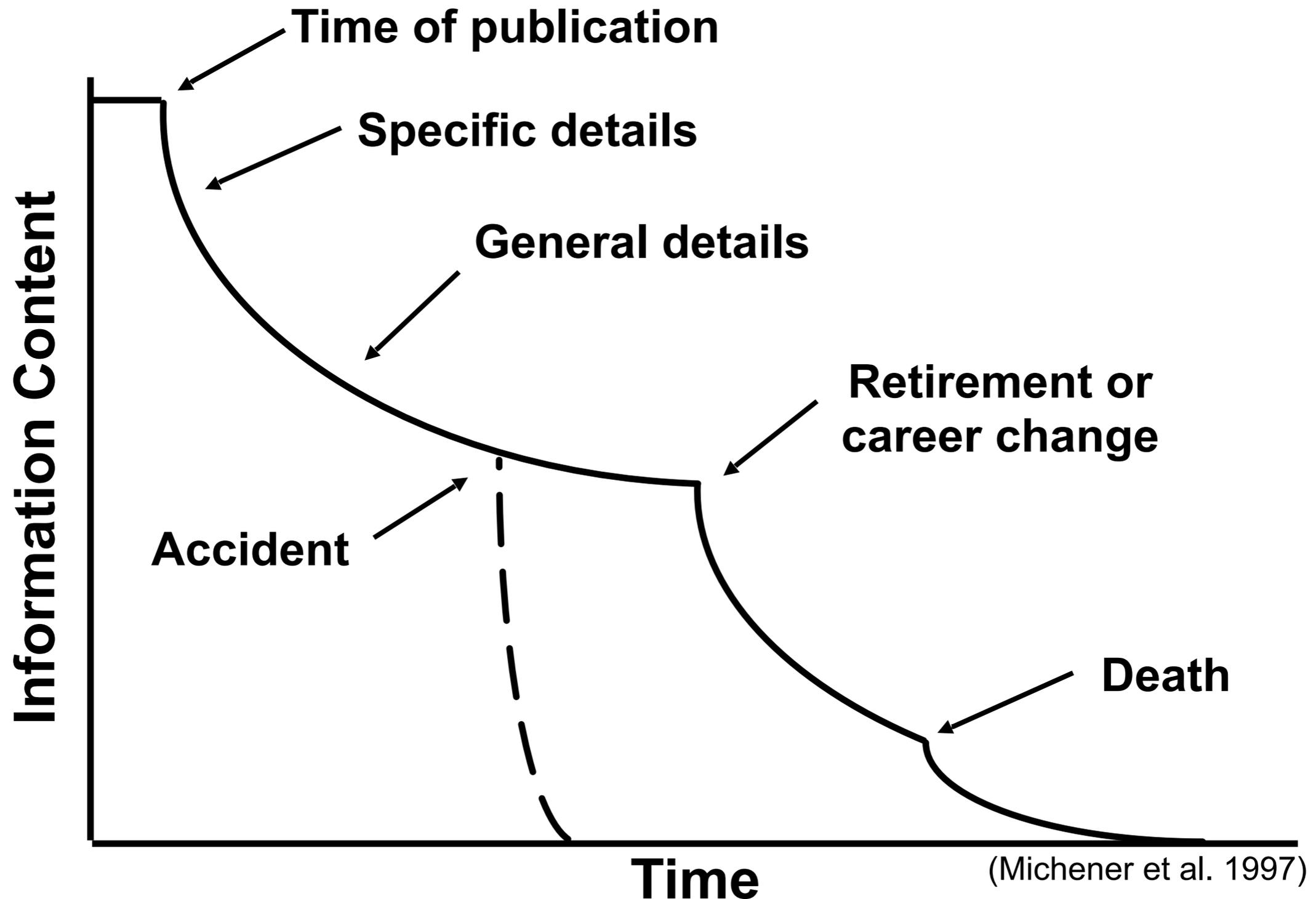


Making Your Research Easier and Cheaper

Write it down!



Poor data practice results in loss of information





Enhancing Your Reputation

- Sloppy, unorganized and undocumented data is hard to share
- What does sloppy, unorganized, and not well documented data say about the rest of your work?
- What if you get hit by a bus?
- Your data may well be your most important legacy!



Scientific Reputation

- Reputation is central to the scientific community
- Researchers build a reputation by producing valuable results, contributing constructively to scientific debates, and being good colleagues
- Peer-recognition influences one's employment opportunities, promotion at work, and ability to win further research funding



Reputation and Data - Why

- Data re-use is growing in importance in almost all scientific fields.
 - Data re-use depends on the availability of trust-worthy 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 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 reputations 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

Many funding agencies require data management plans



- NSF now requires a two page data management plan submitted with every proposal
- NASA Earth Sciences requires a data management plan for its Earth science missions, projects, and grants and cooperative agreements.
- NIH requires that projects with budgets greater than \$500k/yr include a data sharing plan in their proposals.
- NOAA now requires "data managers of all data production projects and systems" to write a DMP



Outline

- ESIP data management training
- Why do a data management plan?
- **Case studies**
- The parts of a NOAA DMP
 - Description of data to be managed
 - Points of Contact
 - Data Stewardship
 - Documentation
 - Data Sharing
 - Initial Storage and Protection
 - Long-term Archiving and Preservation



Environmental Change

- Environmental change is well documented
 - Climate observations
 - Proxy data: ice cores, pollen analysis, tree ring dating
 - Plant and animal species shifts
 - Evaluation of biological specimen collections
 - Etc.
- Preserving those records is essential
 - To understand trends
 - To prepare for changing conditions
 - To recommend courses of action
 - To provide a base for future scientific work



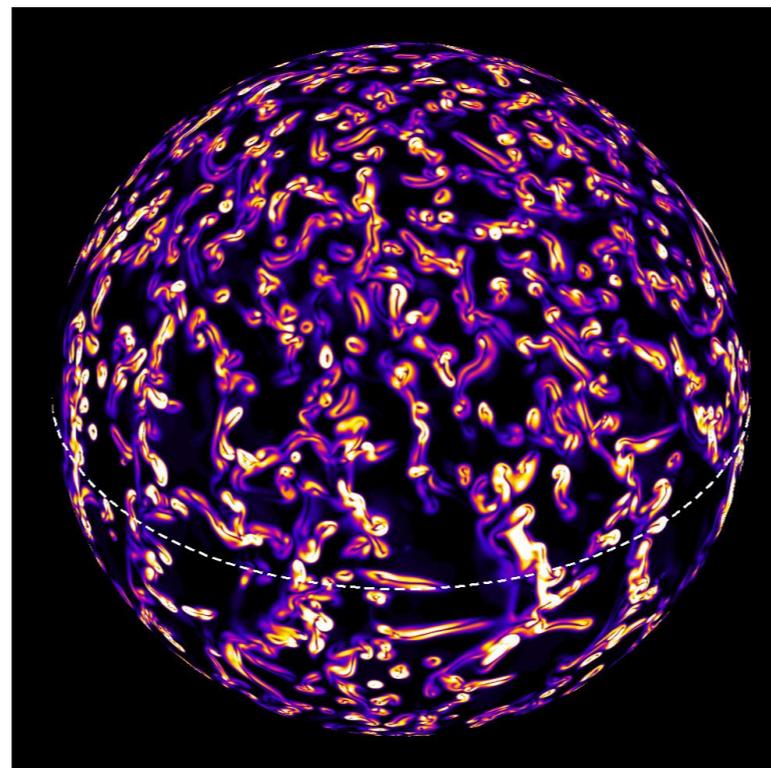


Kinds of Data

Different kinds of data document environmental change in different ways.



Observations



Computational Models



Laboratory Experiments



Observational Data

- Observational data are historical records that cannot be recollected
 - Irreplaceable
 - Very critical to archive
- Documenting data collection methods and equipment is critical to enable:
 - Understanding
 - Evaluation
 - Transparency
 - Reproducibility
 - Trust
 - Secondary use

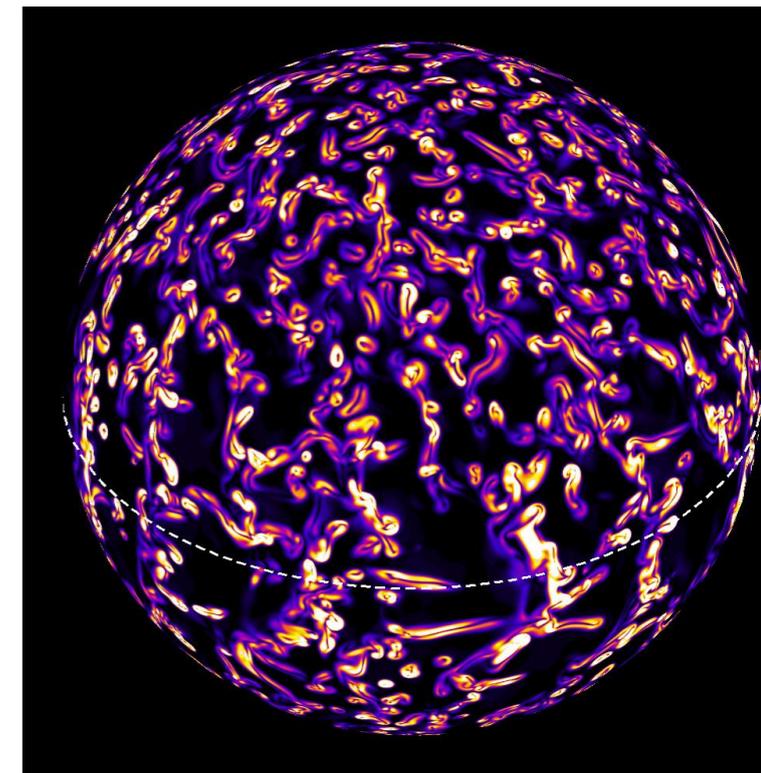




Computational Models and Simulations

- Computational models and simulations produce calculations and predictions of phenomena based on physical theory, algorithms, and observations.
 - Model outputs may not need to be archived, but...
 - Archiving of the model itself and of a robust metadata set (description of the hardware, software, and input data) is essential.
- Ex. The Metaphor project was created to develop a standard way to describe climate models and the data they produce.
 - Metafor worked with climate scientists to create a information model for climate data
 - Tested, developed, and deployed the model for the
 - Metafor: <http://metaforclimate.eu/trac>
 - CMIP5: <http://cmip-pcmdi.llnl.gov/cmip5/>

conceptual





Laboratory Experimental Data

- Experimental data test specific hypotheses in a controlled setting.
 - In principle, experiments can be accurately reproduced and the data need not be stored indefinitely.
 - However, reproducibility can be challenging, even with highly controlled lab settings.
 - Can all of the experimental conditions be precisely reproduced?
 - Nobody could reliably reproduce cold fusion
 - Who will reproduce the CERN Large Hadron Collider?
- Digital “workflow” systems are available for some types of lab work
 - Precise step-by-step description of a scientific procedure
 - Acts as a script for the coordination of research tasks
 - Can enable automated metadata collection and provenance tracking





The parts of a NOAA DMP



General Thoughts on DMPs

- The DMP directive calls for hierarchical organization of DMPs
 - Write it once, if you are duplicating text it should be abstracted up to a higher level document
- The directive allows for alternate templates
- The directive allows for different levels of completion of the DMP content based on:
 - Importance of the data
 - Scope of project
- "Plans are nothing; planning is everything." - Eisenhower, Dwight D.
 - Getting hung up on the form of a template is pointless, the real issue is whether it makes you think about all the questions you need to think about!



Description of data to be managed

- Name of the Dataset or data collection project
- Keywords that could be used to characterize the data, and vocabulary from which those keywords were obtained (e.g., GCMD, CF Conventions, etc.)
- Summary description of the data to be generated
- Anticipated temporal coverage of the data
- Anticipated geographic coverage of the data
- What data types will you be creating or capturing?
- How will you capture or create the data? Where will this plan be stored electronically besides in the NOAA DMP II?
- What volume of data is anticipated to be collected in the Project Time Frame?
- Will the data contain Personally Identifiable Information or any information whose distribution may be restricted by law or national security?



Points of Contact

- NOAA's Data Management Integration Team (DMIT) representative?
- Overall point of contact for the data collection
- Who is responsible for data quality verification?
- Who is responsible for answering data questions?
- Who is responsible for data documentation and metadata activities
- Who is responsible for the data storage and data disaster recovery activities?
- Who is responsible for ensuring adherence to this data management plan, including ensuring that appropriate resources are available to implement the data management plan?

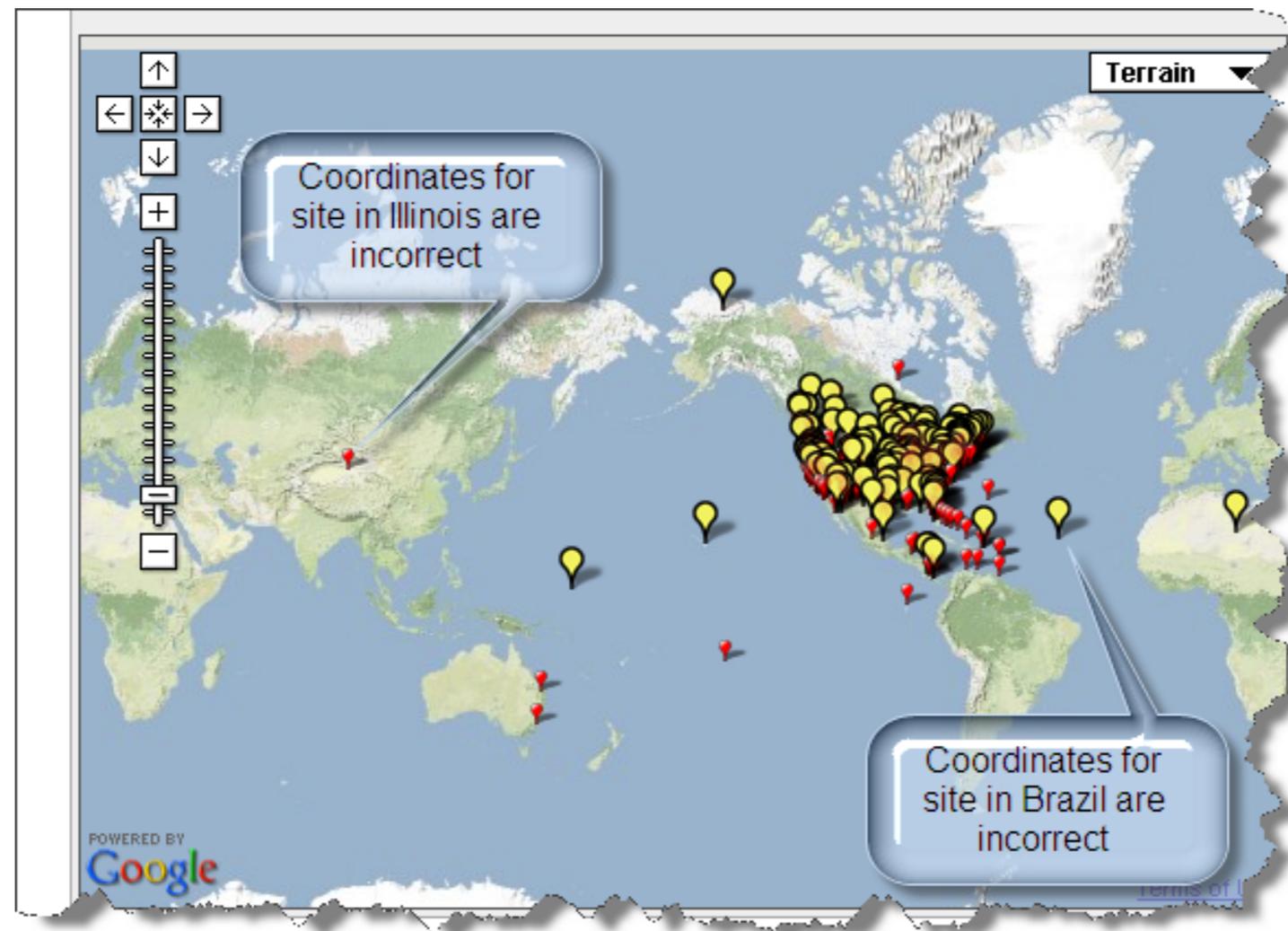


Data Stewardship

- What quality control procedures will be employed?
- What is the overall lifecycle of the data from collection or acquisition to making it available to customer?

Perform basic quality assurance

- *No better QA than to analyze data*
- Perform and review statistical summaries
- Plot data and assess errors



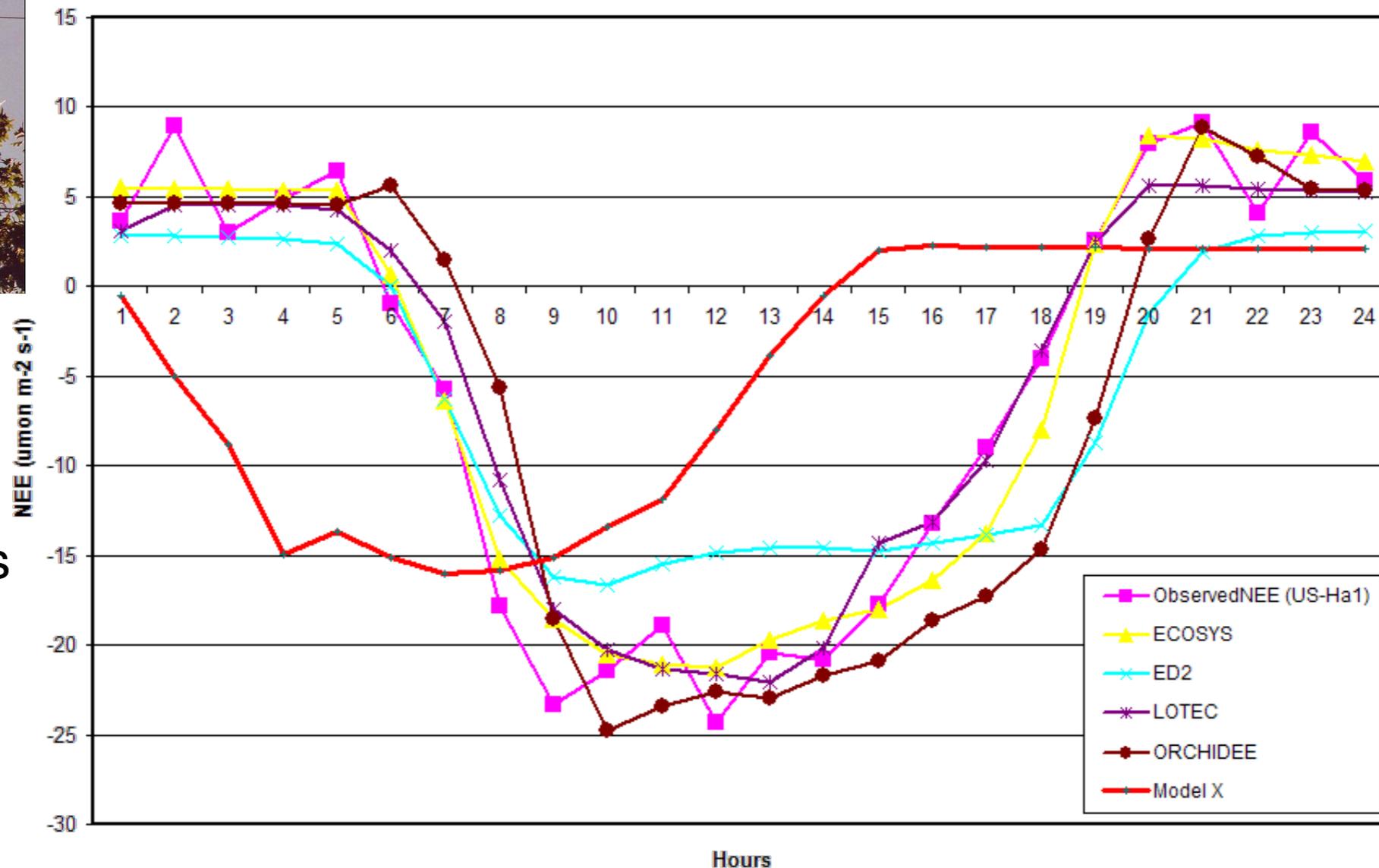
Perform basic quality assurance (con't)

Plot information to examine outliers



Model-Observation Intercomparison

Harvard Forest Flux Tower
Hourly CO2 Flux (2000-06-15)



Model X uses UTC time, all others use Eastern Time

Data from the North American Carbon Program Interim Synthesis (Courtesy of Yaxing Wei, ORNL)

Organize files logically



- Make sure your file system is logical and efficient



Biodiv_H20_heatExp_2005_2008.csv

Biodiv_H20_predatorExp_2001_2003.csv

Biodiv_H20_planktonCount_start2001_active.csv

Biodiv_H20_chla_profiles_2003.csv

...



Assign descriptive file names

- File names should be unique and reflect the file contents
- Bad file names
 - Mydata
 - 2001_data
- A better file name might be
 - bigfoot_agro_2000_gpp.tif
 - BigFoot is the project name
 - Agro is the field site name
 - 2000 is the calendar year
 - GPP represents Gross Primary Productivity data
 - tif is the file type – GeoTIFF
- But only if you document the naming convention!

A STORY TOLD IN FILE NAMES:

Location: C:\user\research\data

Filename	Date Modified	Size	Type
data_2010.05.28_test.dat	3:37 PM 5/28/2010	420 KB	DAT file
data_2010.05.28_re-test.dat	4:29 PM 5/28/2010	421 KB	DAT file
data_2010.05.28_re-re-test.dat	5:43 PM 5/28/2010	420 KB	DAT file
data_2010.05.28_calibrate.dat	7:17 PM 5/28/2010	1,256 KB	DAT file
data_2010.05.28_huh??.dat	7:20 PM 5/28/2010	30 KB	DAT file
data_2010.05.28_WTF.dat	9:58 PM 5/28/2010	30 KB	DAT file
data_2010.05.29_aaarrgh.dat	12:37 AM 5/29/2010	30 KB	DAT file
data_2010.05.29_#*\$@*&!!.dat	2:40 AM 5/29/2010	0 KB	DAT file
data_2010.05.29_crap.dat	3:22 AM 5/29/2010	437 KB	DAT file
data_2010.05.29_notbad.dat	4:16 AM 5/29/2010	670 KB	DAT file
data_2010.05.29_woohoo!!.dat	4:47 AM 5/29/2010	1,349 KB	DAT file
data_2010.05.29_USETHISONE.dat	5:08 AM 5/29/2010	2,894 KB	DAT file
analysis_graphs.xls	7:13 AM 5/29/2010	455 KB	XLS file
ThesisOutline!.doc	7:26 AM 5/29/2010	38 KB	DOC file
Notes_Meeting_with_ProfSmith.txt	11:38 AM 5/29/2010	1,673 KB	TXT file
JUNK...	2:45 PM 5/29/2010		Folder
data_2010.05.30_startingover.dat	8:37 AM 5/30/2010	420 KB	DAT file



Define the contents of your data files

- Content flows from science plan (hypotheses) and is informed from requirements of final archive.
- Keep a set of similar measurements together in one file
 - same investigator,
 - same methods,
 - time basis
 - same instrument
- No hard and fast rules about contents of each files.



Use consistent data organization

- Be consistent in file organization and formatting
- Spreadsheet Example
 - don't change or re-arrange columns
 - Include header rows
 - first row should contain file name, data set title, author, date, and companion file names
 - column headings should describe content of each column, including one row for parameter names and one for parameter units
 - only one sheet per spreadsheet
 - only one table per sheet



Data Formats – Best Practices

- Don't use a proprietary format!
 - These have a short shelf life and will probably become unreadable after a few years
- Don't invent your own format!
 - No one but you will have the tools to read it
- Use open source, well-documented, community-based standard formats where ever possible especially if they are self-describing



Data Documentation

- Which metadata repository will be used to document this data collection?
- In addition to discovery level metadata, what additional metadata or other documentation is necessary to fully describe the data and ensure its long term usefulness? How will that metadata be collected and updated? Is there a requirement to document this data collection in other metadata repositories?
- What standards will be used to represent data and metadata elements in this data collection. Note: The EDMC Data Documentation Procedural Directive calls for the use of ISO 19115 and related standards for data documentation.



Data Documentation Suggestions

- If the data has potential long-term utility for climate change studies consider looking at the Provenance and Context Content Standard that is being developed under the ESIP auspices
 - [http://wiki.esipfed.org/index.php/Provenance and Context Content Standard](http://wiki.esipfed.org/index.php/Provenance_and_Context_Content_Standard)
 - Based on the results of the USGCRP, "Global Change Science Requirements for Long-Term Archiving", Report of the NASA-NOAA Workshop, Sept 28-30, 1998 which can be found at [http://wiki.esipfed.org/images/4/40/USGCRP Long-Term Archiving.pdf](http://wiki.esipfed.org/images/4/40/USGCRP_Long-Term_Archiving.pdf)
 - It has been used to develop a standard for NASA missions
 - Eventual goal is to turn it into an IEEE or ISO standard



Data Sharing

- Will the data be made available to the public? If so, what is the expected date of first availability? Is this a one time data collection, or an ongoing series of measurements? Will there be a Principal Investigator hold or other delay between data collection and publication, and if so for how long?
- If the data are not to be made available to the public, explain why and under what authority distribution may be restricted.
- Will users be subject to any access conditions or restrictions, such as submission of non-disclosure statements, special authorization, or acceptance of a licensing agreement?
- What data access protocols will be used to enable data sharing?
- In what catalogs will these services or data be made registered to enable discovery by users and other Catalogs?



Initial Data Storage and Protection

- Where and how will the data be stored initially (i.e., prior to being sent to a long-term archive facility)?
- How will the data be protected from accidental or malicious modification or deletion? Discuss data back-up, disaster recovery/contingency planning, and off-site storage relevant to the data collection.
- If there will be limitations to data access, how will these data be protected from unauthorized access? How will access permissions be managed? What process is to be followed in the event of unauthorized access?



Backing up your data files

- Create back-up copies often
 - Ideally three copies
 - original, one on-site (external), and one off-site
 - Frequency based on need / risk
 - Higher value data should be backed up more often
 - Sensor data collected at high frequency should be backed up more frequently
- Ensure that all backup copies are identical to the original files
 - Use checksums or file comparisons



Test your backups

- Automatically test backup copies of files frequently to ensure they are viable
 - Media degrade over time
 - Test copies using check sum or file compare
- Be certain that you can recover from a data loss
 - Periodically test your ability to restore information (at least once a year)
 - Simulate an actual loss, by trying to recover solely from the backed up copies



Long-term Archiving and Preservation

- In what NOAA Data Center (NODC, NCDC, NGDC) will the data be archived and preserved? Have you begun discussions with that Data Center regarding your intended submission?
- If you have not identified a NOAA Data Center, what is your long-term strategy for maintaining, curating, and archiving the data?
- How will the costs of long-term data archiving be provided and maintained?
- What transformations or procedures will be necessary to prepare data for preservation or sharing? What related information will be submitted to the archive to enable future use and understanding of the data.
- Identify the Record Schedule applicable to these data and provide the retention time for these data.