Best Practices for Preparing Ecological Data to Share

Bob Cook

Environmental Sciences Division

Oak Ridge National Laboratory

Presenter

- Bob Cook
  - Biogeochemist
  - Chief Scientist, NASA's ORNL Distributed Active Archive Center for Biogeochemical Dynamics
  - Associate Editor, *Biogeochemistry*
  - Oak Ridge National Laboratory, Oak Ridge, TN
  - cookrb@ornl.gov
  - Phone: +1 865 574-7319

ORNL, Oak Ridge, TN

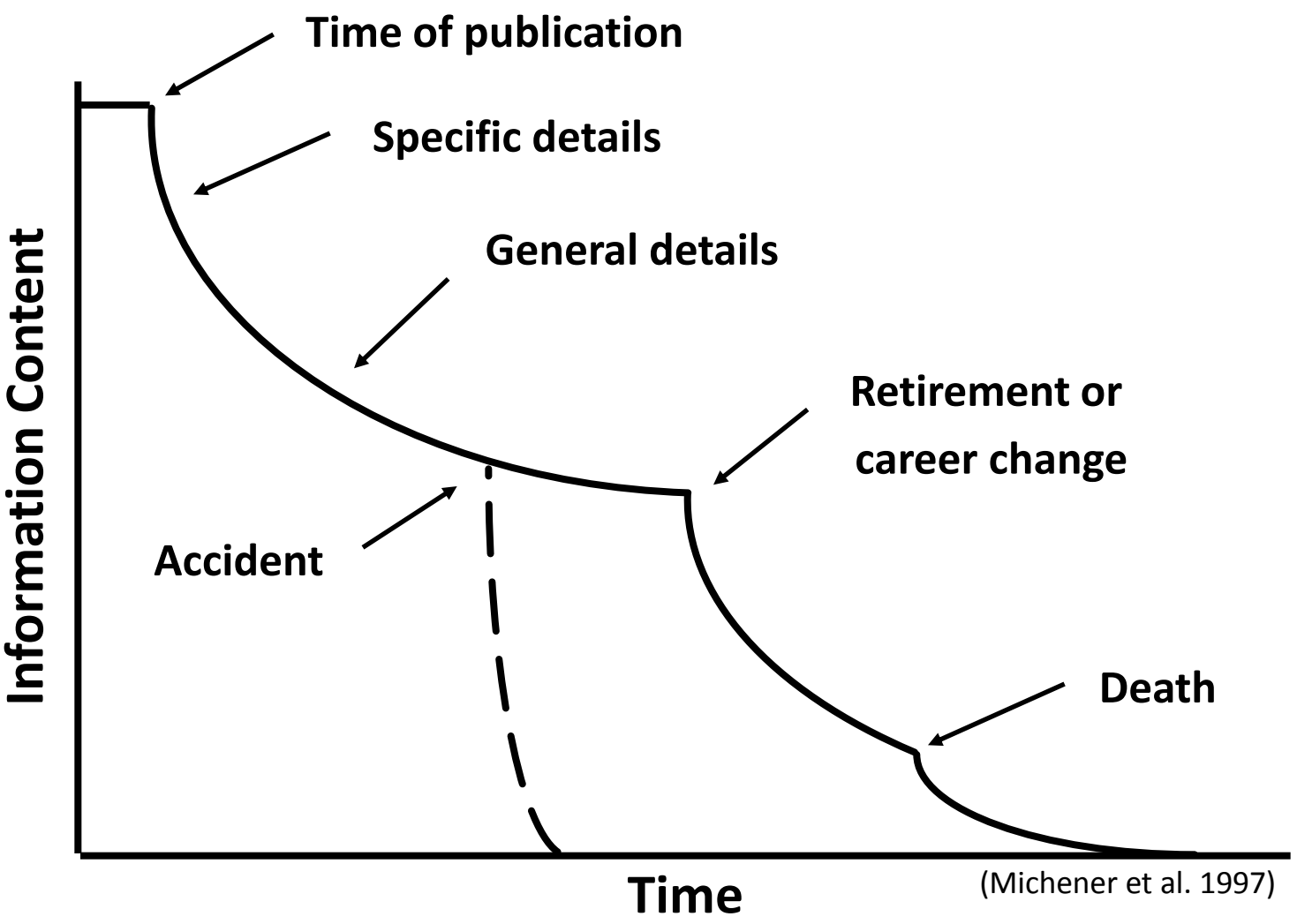
Metadata

Information to let you find, understand, and use the data

– *descriptors*

–*documentation*

Poor data practice results in loss of information (data entropy)



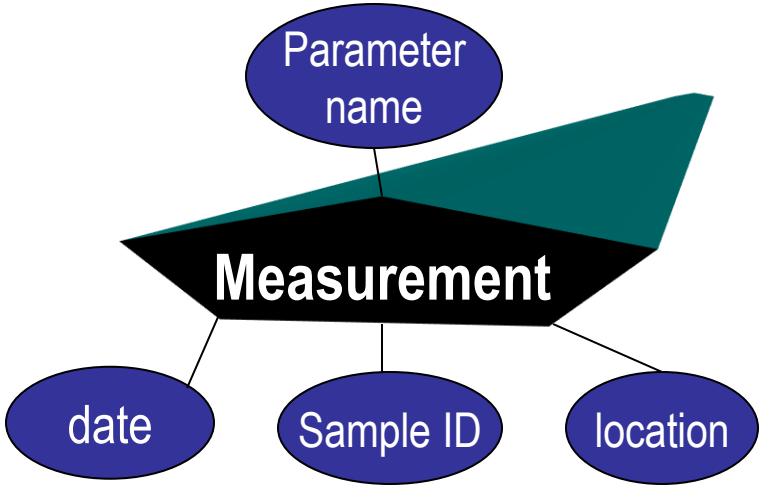
(Michener et al. 1997)

The 20–Year Rule

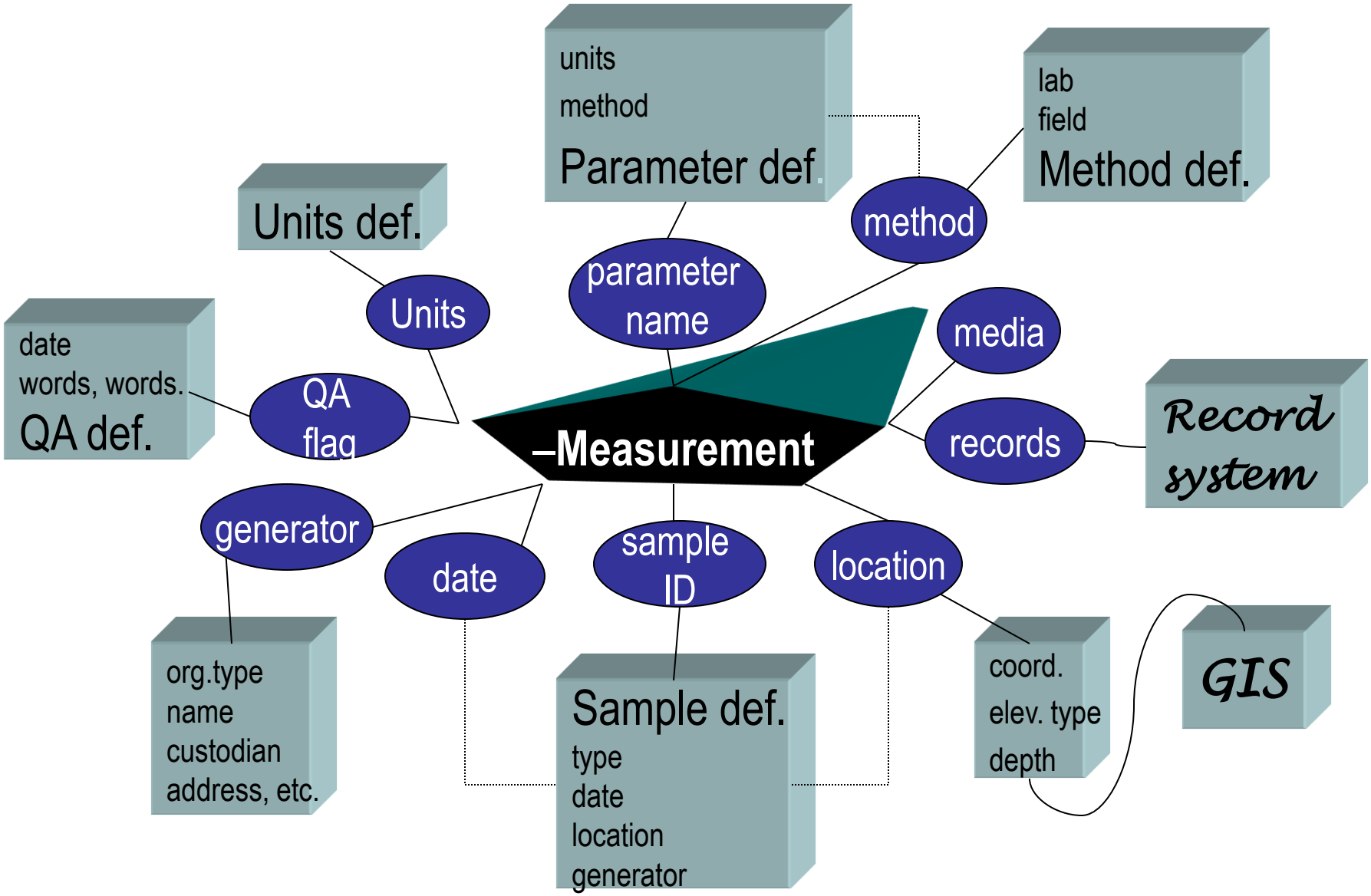
- The metadata accompanying a data set should be written for a user 20 years into the future--what does that investigator need to know to use the data?
- Prepare the data and documentation for a user who is unfamiliar with your project, methods, and observations

Metadata needed to Understand Data

–The details of the data ….



Metadata Needed to Understand Data



Fundamental Data Practices

1. Define the contents of your data files
2. Use consistent data organization
3. Use stable file formats
4. Assign descriptive file names
5. Preserve information
6. Perform basic quality assurance
7. Assign descriptive data set titles
8. Provide documentation
9. Protect your data
10. Acknowledge contributions

1. 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 (e.g., same investigator, methods, time basis, and instruments)
  - No hard and fast rules about contents of each files

Define the parameters

- Use commonly accepted parameter names that describe the contents (e.g., precip for precipitation)
- Use consistent capitalization (e.g., not temp, Temp, and TEMP in same file)
- Explicitly state units of reported parameters in the data file and the metadata
  - SI units are recommended

Define the parameters (cont)

- Choose a format for each parameter, explain the format in the metadata, and use that format throughout the file
  - e.g., use yyyymmdd; January 2, 1999 is 19990102
  - Use 24-hour notation (13:30 hrs instead of 1:30 p.m. and 04:30 instead of 4:30 a.m.)
  - Report in both local time and Coordinated Universal Time (UTC)
  - See Hook et al. (2007) for additional examples of parameter formats
    - http://daac.ornl.gov/PI/bestprac.html#prac3

1. Define the Contents of Your Data Files (cont)

Column	Description	Units/Format
SITE	k=Kataba forest, p=Pandamatenga, m=Near Maun, e=HOORC/MPG Maun tower, o=Okwa river crossing, t=Tshane, skukuza=Skukuza Flux Tower	text
SPECIES	Scientific name up to 25 characters	text
DATE	Date of measurement	yyyymmdd
BA	Woody plant basal area	m2/ha
SEBA	Standard error of BA	m2/ha
DENSITY	Woody plant density (number of trees per hectare)	number/ha
SEDEN	Standard error of DENSITY (n=42 for KT, n=49 for Skukuza)	number/ha
STEMS	Number of stems per hectare (/ha)	number/ha
HEIGHT	Basal area-weighted average height	m2/ha
WOOD	Aboveground woody plant wood dry biomass	kg/ha
LEAF	Aboveground woody plant leaf dry biomass	kg/ha
LAI	Leaf Area Index calculated by allometry	m2/m2

Scholes (2005)

Site Table

Site Name	Site Code	Latitude (deg)	Longitude (deg)	Elevation (m)	Date
Kataba (Mongu)	k	-15.43892	23.25298	1195	29-Feb-00
Pandamatenga	p	-18.65651	25.49955	1138	7-Mar-00
Skukuza Flux Tower	skukuza	-31.49688	25.01973	365	15-Jun-00

.....

Scholes, R. J. 2005. SAFARI 2000 Woody Vegetation Characteristics of Kalahari and Skukuza Sites. Data set. Available on-line [http://daac.ornl.gov/] from Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, U.S.A. doi:10.3334/ORNLDAAC/777

2. Use consistent data organization (one good approach)

Each row in a file represents a complete record, and the columns represent all the parameters that make up the record.

Station	Date	Temp	Precip
Units	YYYYMMDD	C	mm
HOGI	19961001	12	0
HOGI	19961002	14	3
HOGI	19961003	19	-9999

Note: -9999 is a missing value code for the data set

2. Use consistent data organization (a 2nd good approach)

Parameter name, value, and units are placed in individual columns.

This approach is used in relational databases.

Station	Date	Parameter	Value	Unit
HOGI	19961001	Temp	12	C
HOGI	19961002	Temp	14	C
HOGI	19961001	Precip	0	mm
HOGI	19961002	Precip	3	mm

2. Use consistent data organization (cont)

- Be consistent in file organization and formatting
  - 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

3. Use stable file formats

- Use text (ASCII) file formats for tabular data
  - (e.g., .txt or .csv (**c**omma-**s**eparated **v**alues)
  - within the ASCII file, delimit fields using commas, pipes (|), tabs, or semicolons (in order of preference)
- Use GeoTiffs / shapefiles for spatial data

- Avoid proprietary formats
  - They may not be readable in the future

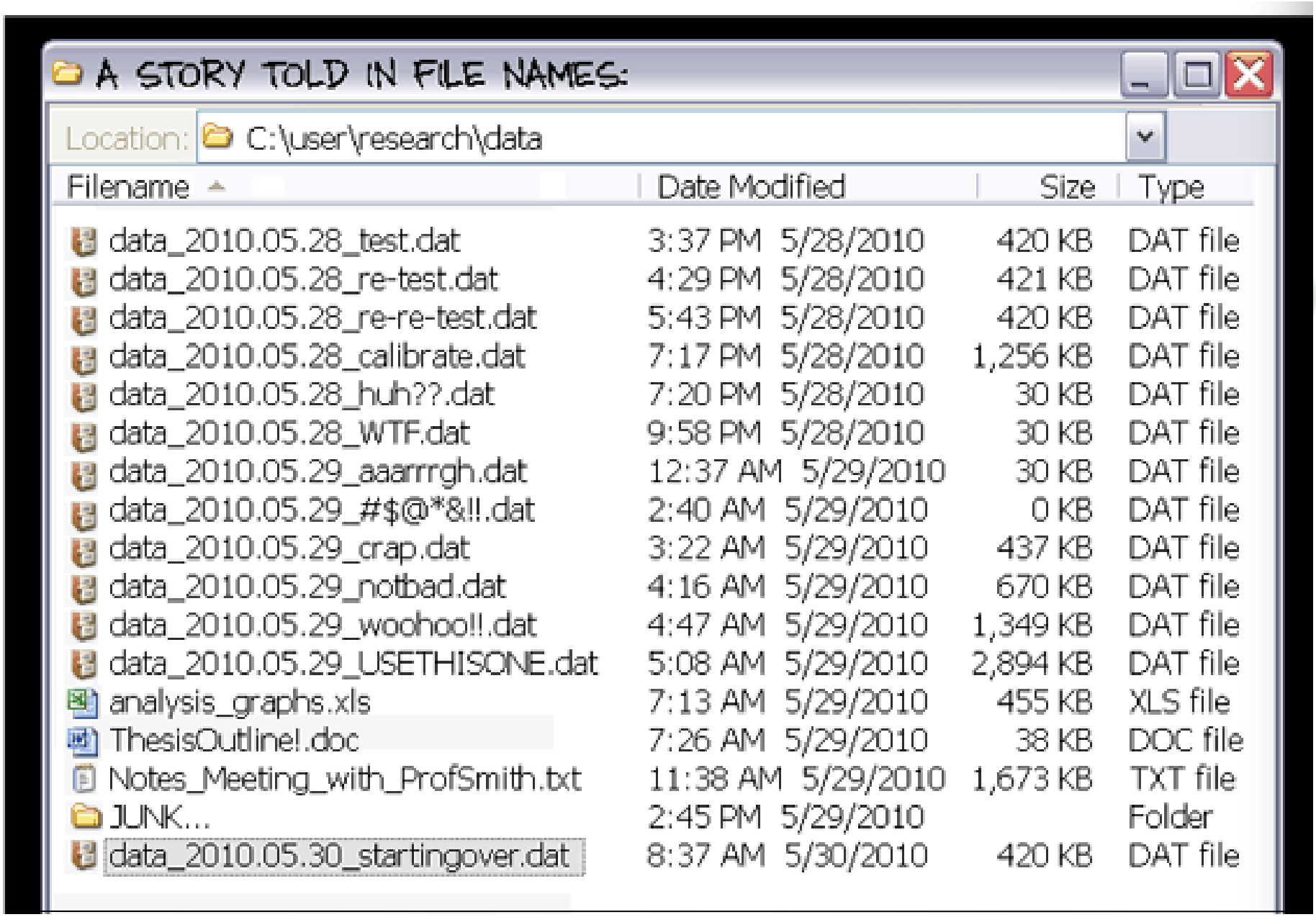
3. Use consistent and stable file formats (cont)

```
SAFARI 2000 Plant and Soil C and N Isotopes, Southern Africa, 1995-2000
SITE,COUNTRY,LAT,LONG,DATE,START_DEPTH,END_DEPTH,CHARACTERISTICS,C,N,d13C,d15N
units,none,decimal degrees,decimal
degrees,yyyy/mm/dd,cm,cm,none,percent,percent,per mil,per mil
USGS-1,Botswana,-21.62,27.37,1999/07/12,5,20,Hardveld,0.67,0.052,-17,8.9
USGS-2,Botswana,-21.07,27.42,1999/07/12,5,20,Hardveld,0.68,0.063,-18.3,8
USGS-3,Botswana,-20.72,26.83,1999/07/12,5,20,Hardveld,0.94,0.087,-17,6.8
USGS-4,Botswana,-20.52,26.41,1999/07/12,5,20,Hardveld,0.53,0.04,-19.9,5.5
USGS-5,Botswana,-20.55,26.15,1999/07/12,5,20,Lacustrine,2.11,0.162,-15.2,5.9
...
USGS-30,Botswana,-19.81,23.63,1999/07/18,5,20,Alluvium,0.67,0.063,-19.2,11.8
USGS-31,Botswana,-20.62,22.74,1999/07/18,5,20,Hardveld,0.23,0.014,-16.8,16.2
```

Aranibar, J. N. and S. A. Macko. 2005. SAFARI 2000 Plant and Soil C and N Isotopes, Southern Africa, 1995-2000. Data set. Available on-line [http://daac.ornl.gov/] from Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, U.S.A. doi:10.3334/ORNLDAAC/783

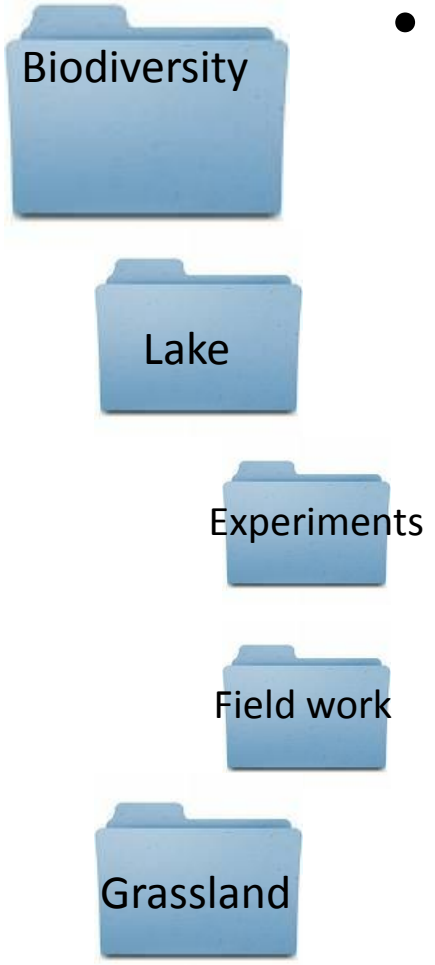
4. Assign descriptive file names

- File names should be unique and reflect the file contents
- Bad file names
  - Mydata
  - 2001_data
- A better file name
  - 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


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_aaarrrgh.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

4. Assign descriptive file names

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

…

From S. Hampton

5. Preserve information

- Keep your raw data raw
  - No transformations, interpolations, etc, in raw file

Raw Data File

Giles_zoopCount_Diel_2001_2003.csv

TAX	COUNT	TEMPC
C	3.97887358	12.3
F	0.97261354	12.7
M	0.53051648	12.1
F	0	11.9
C	10.8823893	12.8
F	43.5295571	13.1
M	21.7647785	14.2
N	61.6668725	12.9
–	…	

Processing Script (R)

```
–### Giles_zoop_temp_regress_4jun08.r
–### Load data
–Giles<-
read.csv("Giles_zoopCount_Diel_2001_2003.csv")
–### Look at the data
–Giles
–plot(COUNT~ TEMPC, data=Giles)
–### Log Transform the independent variable (x+1)
–Giles$Lcount<-log(Giles$COUNT+1)
–### Plot the log-transformed y against x
–plot(Lcount ~ TEMPC, data=Giles)
```

From S. Hampton

5. Preserve information (cont)

- Use a scripted language to process data
  - R Statistical package (free, powerful)
  - SAS
  - MATLAB
- Processing scripts are records of processing
  - Scripts can be revised, rerun
- Graphical User Interface-based analyses may seem easy, but don't leave a record

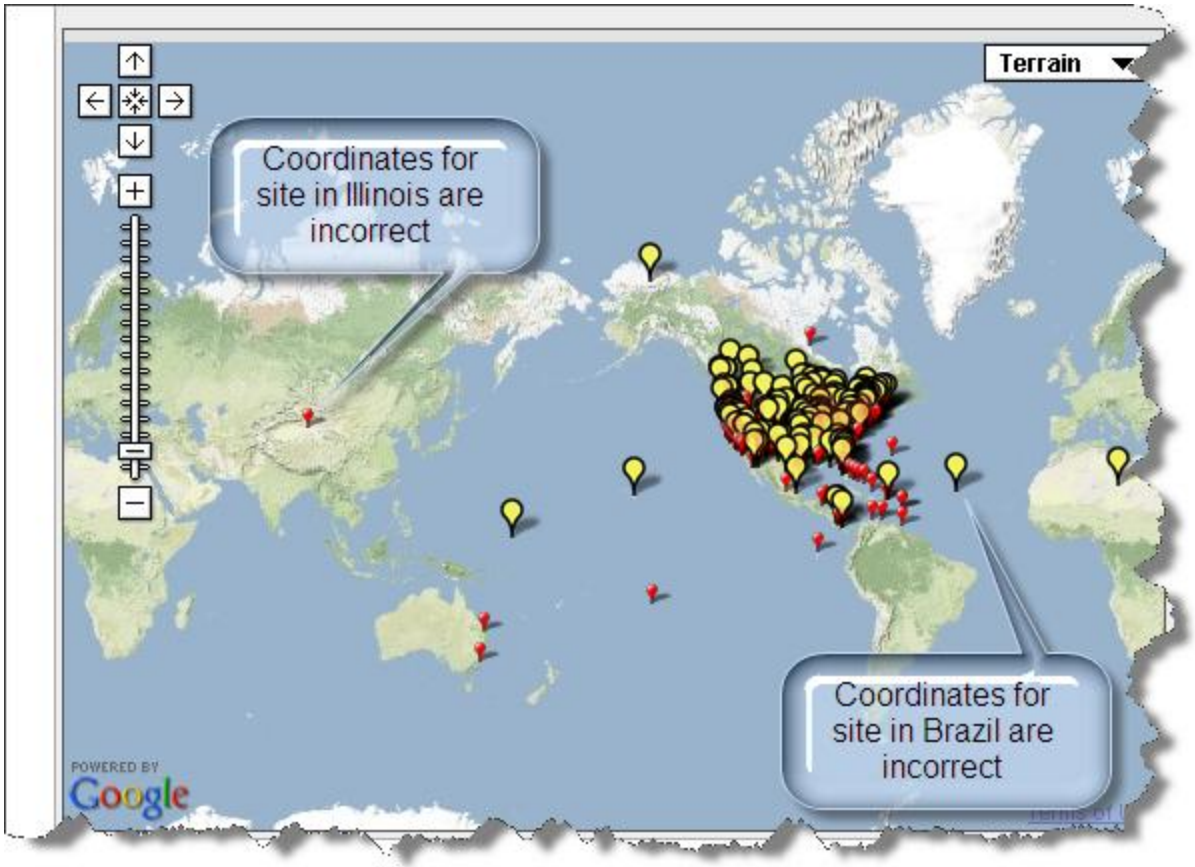
6. Perform basic quality assurance

- Assure that data are delimited and line up in proper columns
- Check that there no missing values (blank cells) for key parameters
- Scan for impossible and anomalous values
- Perform and review statistical summaries
- Map location data (lat/long) and assess errors

- *No better QA than to analyze data*

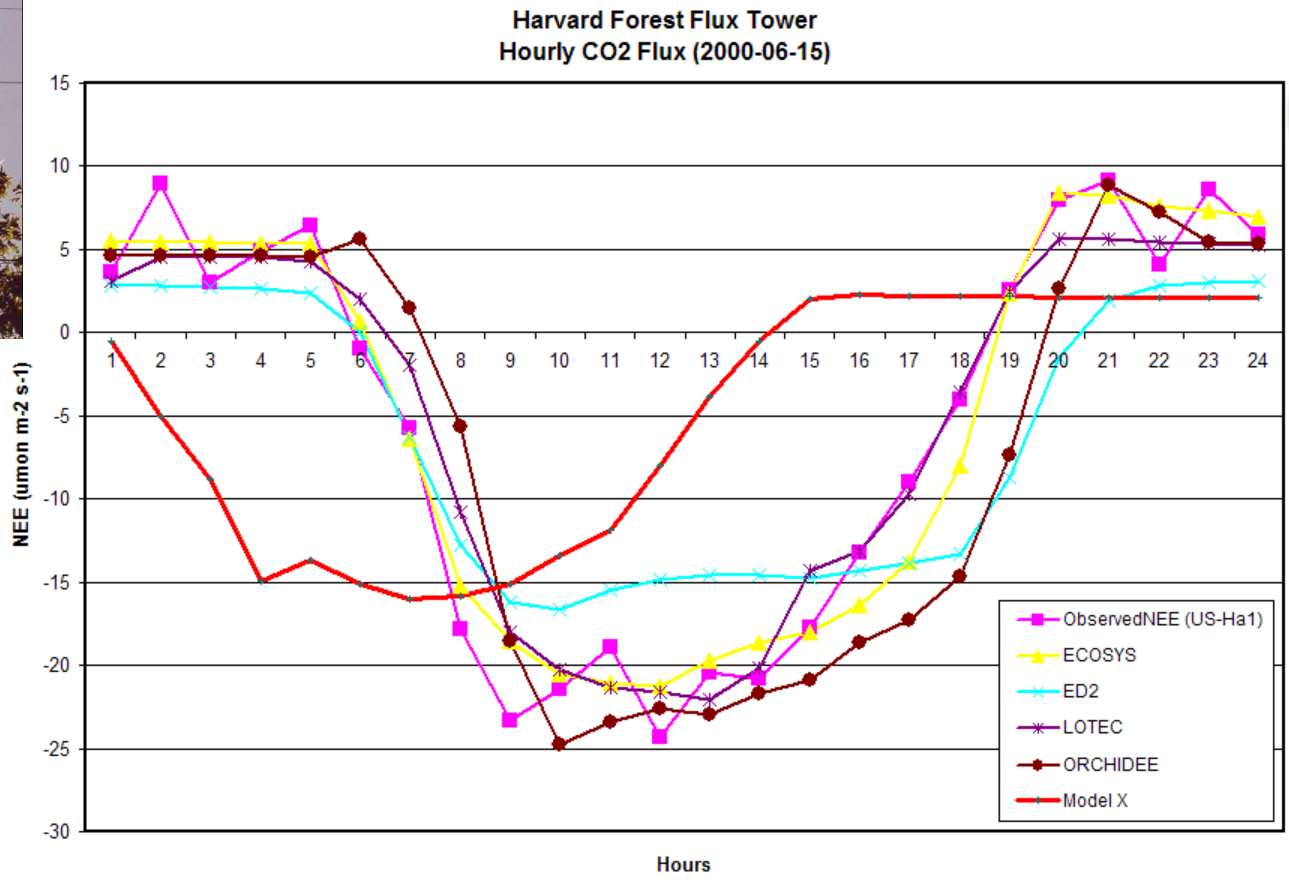
6. Perform basic quality assurance (con't)

Place geographic data on a map to ensure that geographic coordinates are correct.



Plot information to examine outliers

Model-Observation Intercomparison

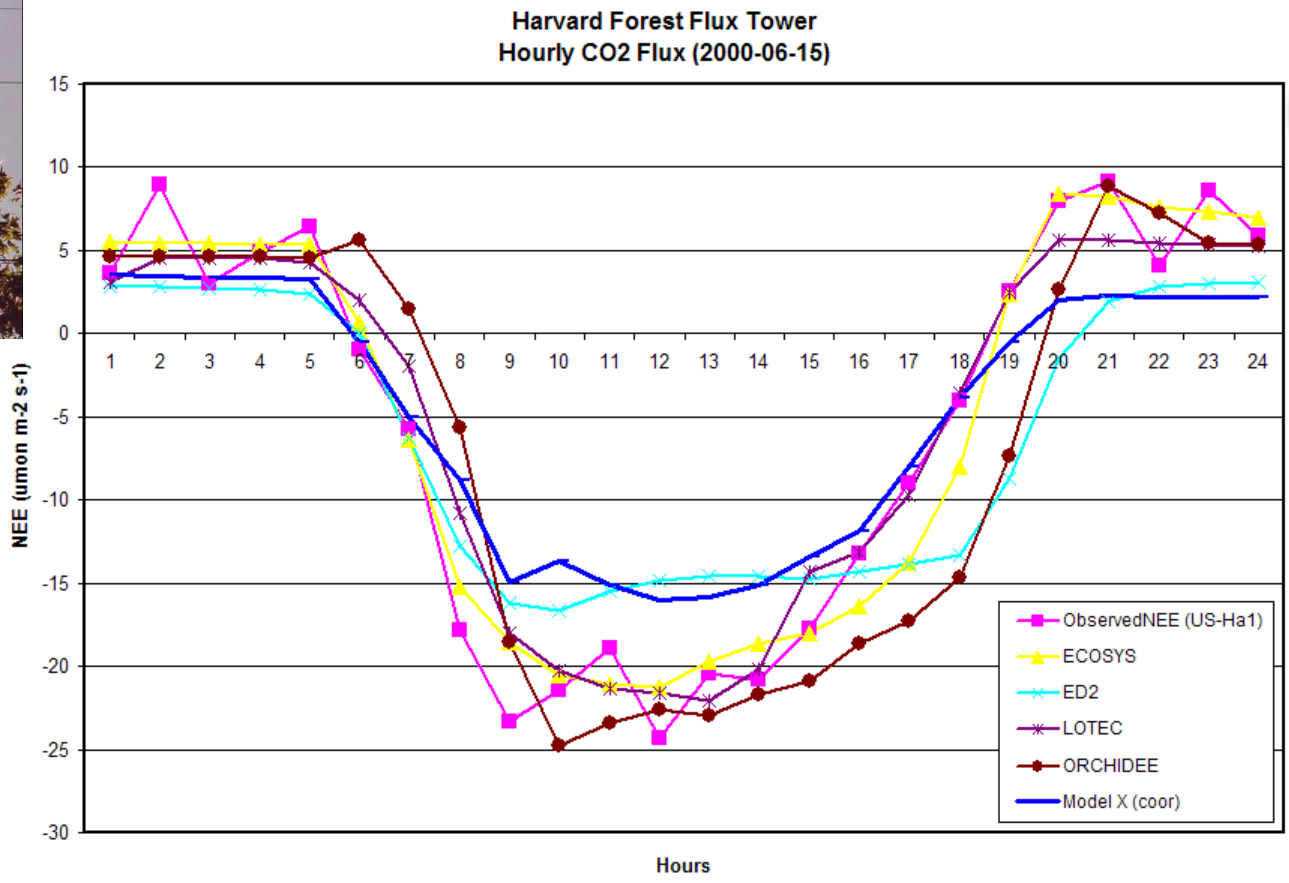


Model X uses UTC time, all others use Eastern Time

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

Plot information to examine outliers

Model-Observation Intercomparison



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

7. Assign descriptive data set titles

- Data set titles should ideally describe the type of data, time period, location, and instruments used (e.g., Landsat 7).
- Titles should be concise (< 85 characters)
- Data set title should be similar to names of data files
  - Good: SAFARI 2000 Upper Air Meteorological Profiles, Skukuza, Dry Seasons 1999-2000"
  - Bad: “Productivity Data”

8. Provide Documentation / Metadata

- **What** does the data set describe?
- **Why** was the data set created?
- **Who** produced the data set and **Who** prepared the metadata?
- **How** was each parameter measured?
- **What** assumptions were used to create the data set?
- **When** and how frequently were the data collected?
- **Where** were the data collected and with what spatial resolution? (include coordinate reference system)
- **How** reliable are the data; what is the uncertainty; what problems remain in the data set?
- **What** is the use and distribution policy of the data set? **How** can someone get a copy of the data set?
- **Provide** any references to use of data in publication(s)?

9. Protect data

- Create back-up copies often
  - Ideally two copies, one on-site and one off-site
  - Frequency based on need / risk
- Know that you can recover from a data loss
  - Periodically test your ability to restore information

9. Protect data (cont)

- Ensure that file transfers are done without error
  - Compare checksums before and after transfers
    - Example tools to generate checksums

http://www.pc-tools.net/win32/md5sums/

http://corz.org/windows/software/checksum/

10. Acknowledge contributions

- **Who contributed data to your data set?**
- **How should your data be acknowledged / used by others?**

All data have been collected and curated by Dr. C. Hippocrepis at All Rotifers University.

Data may be freely downloaded for non-commercial uses. Please contact Dr. C. Hippocrepis (chippocrepis@allrotifers.edu) to let us know if you use these data. We report uses of the public data to our funders, and it is extremely helpful to know if others have been able to use these data in teaching or research.

From S. Hampton

When your project is over, where should the data be archived?

- Part of project planning
- Contact archive / data center early to find out their requirements
  - What additional data management steps would they like you to do?
- Suggested data centers / archives:
  - Dryad
  - *Ecological Archives*
  - Knowledge Network for Biocomplexity (KNB)
  - ORNL DAAC
  - DataONE (Observation Network for Earth)
  - National Biological Information Infrastructure

Fundamental Data Practices

1. Define the contents of your data files
2. Use consistent data organization
3. Use stable file formats
4. Assign descriptive file names
5. Preserve information
6. Perform basic quality assurance
7. Assign descriptive data set titles
8. Provide documentation
9. Protect your data
10. Acknowledge contributions

Best Practices: Conclusions

- Data management is important in today's science
- Well organized data:
  - enables researchers to work more efficiently
  - can be shared easily by collaborators
  - can potentially be re-used in ways not imagined when originally collected

Bibliography

- Ball, C. A., G. Sherlock, and A. Brazma. 2004. Funding high-throughput data sharing. Nature Biotechnology 22:1179-1183. doi:10.1038/nbt0904-1179.
- Borer, ET., EW. Seabloom, M.B. Jones, and M. Schildhauer. 2009. Some Simple Guidelines for Effective Data Management ,Bulletin of the Ecological Society of America. 90(2): 205- 214.
- Christensen, S. W. and L. A. Hook. 2007. NARSTO Data and Metadata Reporting Guidance. Provides reference tables of chemical, physical, and metadata variable names for atmospheric measurements. Available on-line at: http://cdiac.ornl.gov/programs/NARSTO/
- Cook, Robert B, Richard J. Olson, Paul Kanciruk, and Leslie A. Hook. 2001. Best Practices for Preparing Ecological Data Sets to Share and Archive. Bulletin of the Ecological Society of America, Vol. 82, No. 2, April 2001.
- Hook, L. A., T. W. Beaty, S. Santhana-Vannan, L. Baskaran, and R. B. Cook. June 2007. Best Practices for Preparing Environmental Data Sets to Share and Archive. http://daac.ornl.gov/PI/pi_info.shtml
- Kanciruk, P., R.J. Olson, and R.A. McCord. 1986. Quality Control in Research Databases: The US Environmental Protection Agency National Surface Water Survey Experience. In: W.K. Michener (ed.). Research Data Management in the Ecological Sciences. The Belle W. Baruch Library in Marine Science, No. 16, 193-207.
- Michener, W K. 2006. Meta-information concepts for ecological data management. Ecological Informatics. 1:3-7.
- Michener, W.K. and J.W. Brunt (ed.). 2000. Ecological Data: Design, Management and Processing, Methods in Ecology, Blackwell Science. 180p.
- Michener, W. K., J. W. Brunt, J. Helly, T. B. Kirchner, and S. G. Stafford. 1997. Non-Geospatial Metadata for Ecology. Ecological Applications. 7:330-342.
- U.S. EPA. 2007. Environmental Protection Agency Substance Registry System (SRS). SRS provides information on substances and organisms and how they are represented in the EPA information systems. Available on-line at: http://www.epa.gov/srs/
- USGS. 2000. Metadata in plain language. Available on-line at: http://geology.usgs.gov/tools/metadata/tools/doc/ctc/