Presenter: Bob Cook

- Biogeochemist
- Chief Scientist, NASA's ORNL Distributed Active Archive Center for Biogeochemical Dynamics
- Associate Editor, *Biogeochemistry* (1995 – 2015)
- Editor, *Scientific Data*, a *Nature* Journal
- Member, PLOS Data Guidelines Board
- Oak Ridge National Laboratory, Oak Ridge, TN
- cookrb@ornl.gov

ORNL DAAC

ORNL, Oak Ridge, TN

Best Practices for Preparing Data to Share and Preserve

Bob Cook

Environmental Sciences Division

Oak Ridge National Laboratory

2015 Data Management Webinar Series

USGS and PNAMP

November 18, 2015

Agenda: Best Practices for Preserving Data

- Introduction
- Management for data sharing during a project
- Long-term data archiving

Preserving Data

USGS provides "crucial impartial data" to federal agencies and the public.

Sen. Joe Manchin (D-W.Va.) during
Suzette Kimball's confirmation hearing.
Eos October 29, 2015

Showstack, R. (2015), USGS acting director receives support at Senate hearing, Eos, 96, doi: 10.1029/2015EO038379

Benefits of Good Data Management Practices

Short-term

- Spend less time doing data management and more time doing research
- Easier to prepare and use data for yourself
- Collaborators can readily understand and use data files

Long-term

- Scientists outside your project can find, understand, and use your data to address broad questions
- You get credit for archived data products and their use in other papers
- Sponsors protect their investment

NAH, I'M NOT WORRIED ABOUT CLOUD SECURITY. MY STORED DATA IS SO DISORGANIZED THEY'D NEVER BE ABLE TO FIND ANYTHING!
© D.Fletcher for CloudTweaks.com

Metadata

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

– *descriptors*

–*documentation*

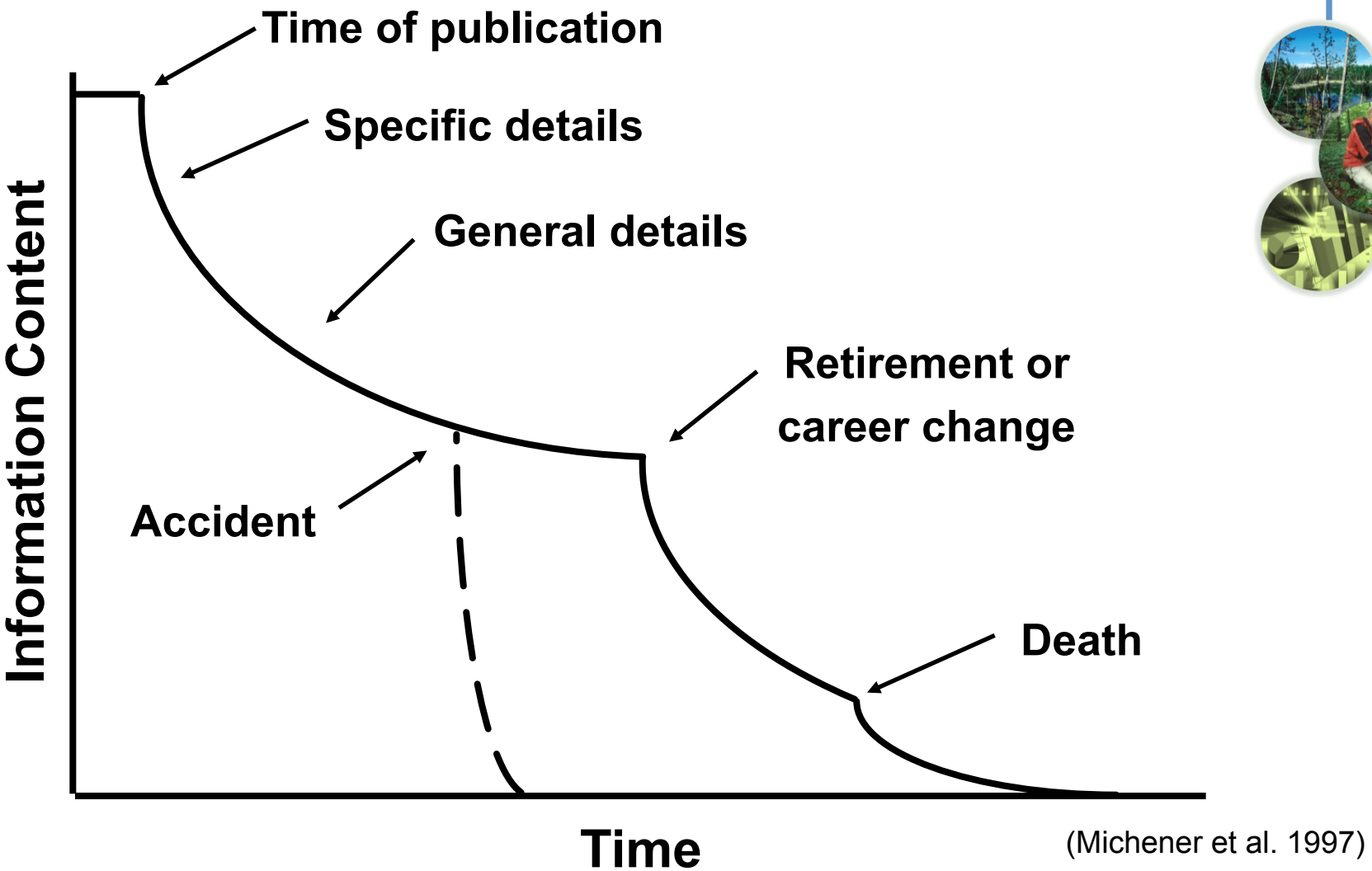
USGS / PNAMP Webinar: Metadata, Lisa Zolly

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

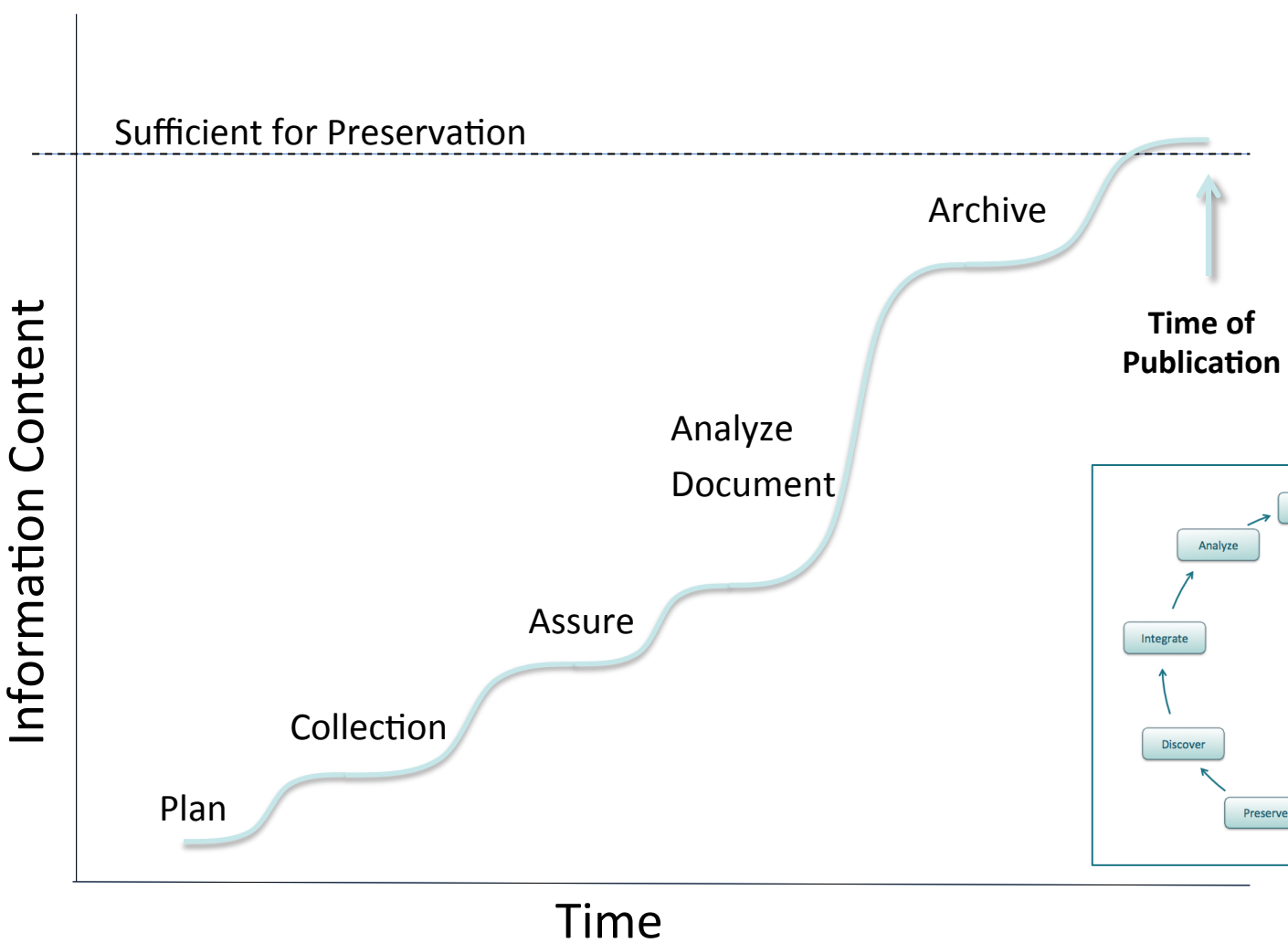
NRC (1991)

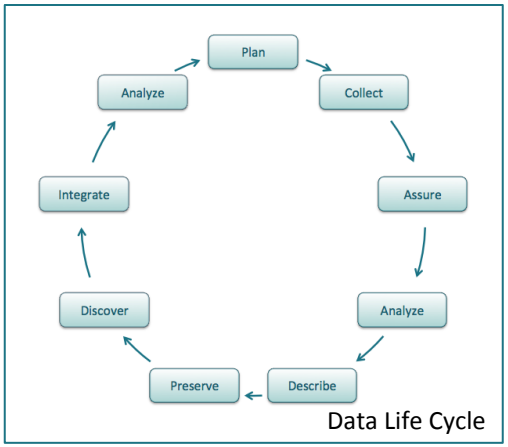
Data Entropy



(Michener et al. 1997)

Proper Curation Enables Data Reuse





Fundamental Data Practices

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

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
  - same investigator,
  - site,
  - methods,
  - time basis, and
  - instrument
  - No hard and fast rules about contents of each file.

2. Define the variables

1. Choose the units and format for each variable,
2. Explain the format in the metadata, and
3. Use that format consistently throughout the file

Date / Time Example

- e.g., use yyyymmdd; January 2, 1999 is 19990102
- Report in both local time and Coordinated Universal Time (UTC) and 24-hour notation (13:30 hrs instead of 1:30 p.m.)
- Use a code (e.g., -9999) for missing values

ISO 8601

Representation of dates and times

2. Define the variables (cont)

- Use commonly accepted variable names and units

International System of Units

UDUNITS

Unit database and conversion between units

CF Standard Name

Climate Forecast (CF) standards promote sharing

- ORNL DAAC Best Practices (Hook et al., 2010)
  - Additional examples of variable names, units, and their formats
- Next Generation Ecosystem Experiment – Arctic
  - Guidance for variable names and units
- FLUXNET
  - Guidance for flux tower variable names and units

2. Define the variables (cont)

Variable Table

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)

- Be consistent
- Explicitly state units
- Use ISO formats

Site Table

Site Name	Site Code	Latitude (deg)	Longitude (deg)	Elevation (m)	Date
Kataba (Mongu)	k	-15.43892	23.25298	1195	2000.02.21
Pandamatenga	p	-18.65651	25.49955	1138	2000.03.07
Skukuza Flux Tower	skukuza	-31.49688	25.01973	365	2000.06.15

……

Scholes, R. J. 2005. SAFARI 2000 Woody Vegetation Characteristics of Kalahari and Skukuza Sites. ORNL DAAC. doi:10.3334/ORNLDAAC/777

3. 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	20121001	12	0
HOGI	20121002	14	3
HOGI	20121003	19	-9999

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

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

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

This approach is used in relational databases.

Station	Date	Parameter	Value	Unit
HOGI	20121001	Temp	12	C
HOGI	20121002	Temp	14	C
HOGI	20121001	Precip	0	mm
HOGI	20121002	Precip	3	mm

3. 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

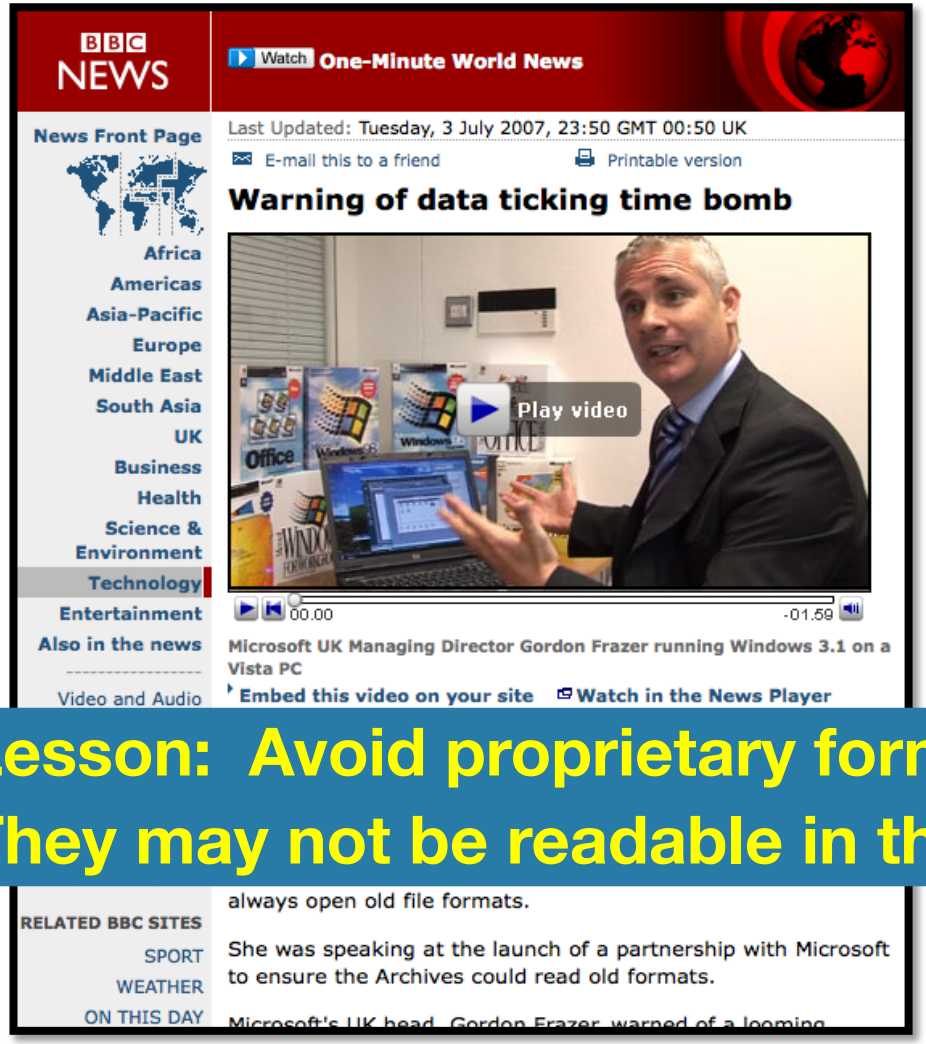
–USGS / PNAMP Webinar: Spreadsheet Mess

Stable Isotope Data at ORNL: tabular csv format

Aranabar and Macko. 2005. doi:10.3334/ORNLDAAC/783

SITE	COUNTRY	LAT	LONG	DATE	DISTURBANC	TAXONOMY	PLANT PART	NOTES	C	N	d13C	d15N
units	none	decimal degr	decimal degr	year-month	none	none	none	none	%	%	per mil	per mil
Mongu	Zambia	-15.44	23.52	2000-02	C0	Baphia mass	L	none	51.6	3	-27	1.4
Mongu	Zambia	-15.44	23.52	2000-02	C0	Bauhinia pet	L	none	47.6	2.31	-27	4.7
Mongu	Zambia	-15.44	23.52	2000-02	C0	Rubiaceae	L	none	51.8	1.9	-29	2.6
Mongu	Zambia	-15.44	23.52	2000-02	C0	Brachystegia	L	none	53.2	3.21	-25.7	4
Mongu	Zambia	-15.44	23.52	2000-02	C0	Brachystegia	L	none	53.2	3.08	-24.5	4.6
Mongu	Zambia	-15.44	23.52	2000-02	C0	Burkea africa	L	none	49.5	1.84	-27	-1.6
Mongu	Zambia	-15.44	23.52	2000-02	C0	Fabaceae	L	shrub	46.6	2.69	-28.1	3
Mongu	Zambia	-15.44	23.52	2000-02	C0	Combretum	L	none	48.7	-9999	-28	-9999
Mongu	Zambia	-15.44	23.52	2000-02	C0	Copaifera ba	L	none	57.5	1.63	-28	3
Mongu	Zambia	-15.44	23.52	2000-02	C0	Diospyrus ba	L	none	53.1	1.47	-27	4.4
Mongu	Zambia	-15.44	23.52	2000-02	C0	Hannoa chlo	L	none	49.2	1.92	-27.8	3.8
Mongu	Zambia	-15.44	23.52	2000-02	C0	Guibourtia c	L	none	53.3	2.73	-27.4	1.6
Mongu	Zambia	-15.44	23.52	2000-02	C0	Hannoa chlo	L	none	48.1	1.41	-27.4	3.8
Mongu	Zambia	-15.44	23.52	2000-02	C0	Indigofera sp	L	none	49.4	3.21	-28.6	0.7
Mongu	Zambia	-15.44	23.52	2000-02	C0	Indigofera sp	L	none	49.8	3.26	-27.5	0.4
Mongu	Zambia	-15.44	23.52	2000-02	C0	Ochna pulch	L	mature leaf	51.4	1.68	-26.4	3.1
Mongu	Zambia	-15.44	23.52	2000-02	C0	Parinari cura	L	none	51.3	1.55	-30.5	2.5
Mongu	Zambia	-15.44	23.52	2000-02	C0	Paropsia bra	L	none	52.8	2.55	-28.6	5.8
Mongu	Zambia	-15.44	23.52	2000-02	C0	Pseudolachn	L	none	47.2	1.74	-25.9	2.4
Mongu	Zambia	-15.44	23.52	2000-02	C0	Tephrosia sp	L	none	46.9	3.37	-28.7	0.5
Mongu	Zambia	-15.44	23.52	2000-02	C0	unidentified	Ll	litter	47.1	2.13	-25.8	4.4
Mongu	Zambia	-15.44	23.52	2000-02	C0	Basidiomyco	S	saprophytic f	41.2	2.39	-20.1	3.5
Mongu	Zambia	-15.44	23.52	2000-02	C0	Basidiomyco	S	ectomycorrh	24.7	2.15	-20.8	3.8
Mongu	Zambia	-15.44	23.52	2000-02	C0	Lichen	W	none	46.3	2.27	-23.8	-0.7
Mongu	Zambia	-15.44	23.52	2000-02	C0	Lichen	W	none	41.9	1.59	-19.6	-2.6
Mongu	Zambia	-15.44	23.52	2000-02	C0	Lichen	W	none	41	1.38	-20.5	-2.5
Mongu	Zambia	-15.44	23.52	2000-02	C0	Basidiomyco	S	ectomycorrh	39.9	3.77	-21.3	5
Mongu	Zambia	-15.44	23.52	2000-02	C0	Basidiomyco	S	ectomycorrh	28.7	4.04	-21.4	5.9

4. Use stable file formats



Los[e] years of critical knowledge because modern PCs could not always open old file formats.

Lesson: Avoid proprietary formats
They may not be readable in the future

http://news.bbc.co.uk/2/hi/6265976.stm

4. Use stable file formats (cont)

- Use text (ASCII) file formats for tabular data
  - (e.g., .txt or .csv (comma-separated values)

```
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
```

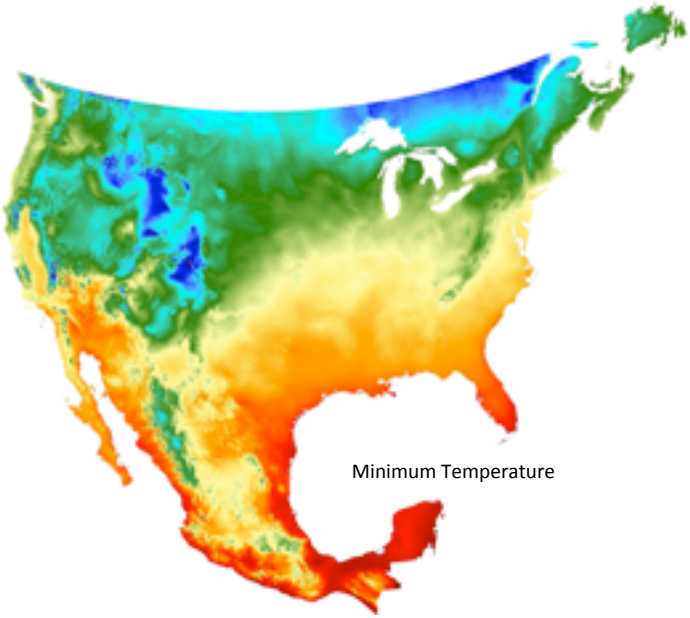
Aranibar, J. N. and S. A. Macko. 2005. SAFARI 2000 Plant and Soil C and N Isotopes, Southern Africa, 1995-2000. Data set. Available on-line [http://daac.ornl.gov/] from Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, U.S.A. doi:10.3334/ORNLDAAC/783

4. Use stable file formats (cont)

Suggested Geospatial File Formats

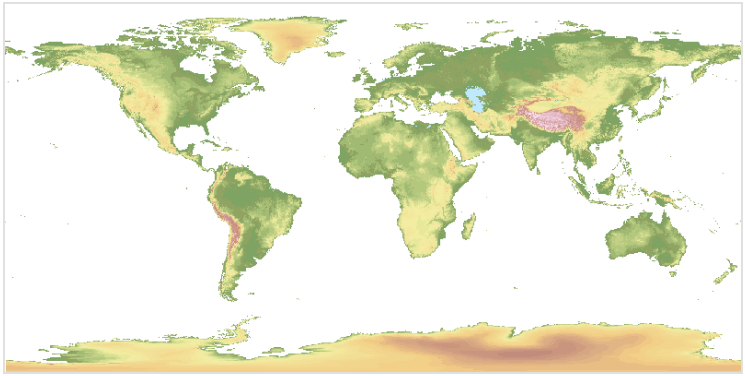
Raster formats

- Geotiff
- netCDF
  - with CF convention preferred
- HDF
- ASCII
  - plain text file gridded format with external projection information

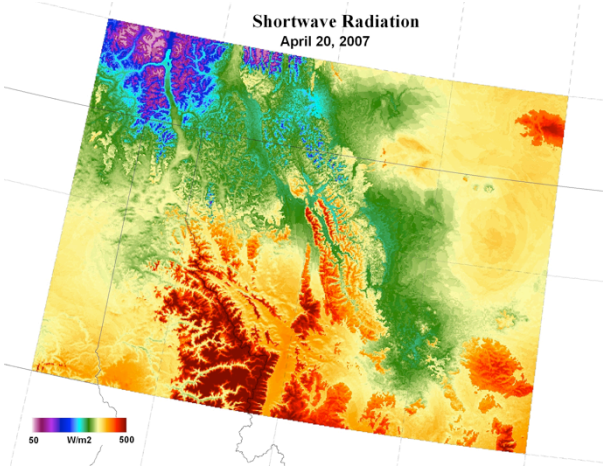


Vector

- Shapefile
- KML/GML



GTOPO30 Elevation



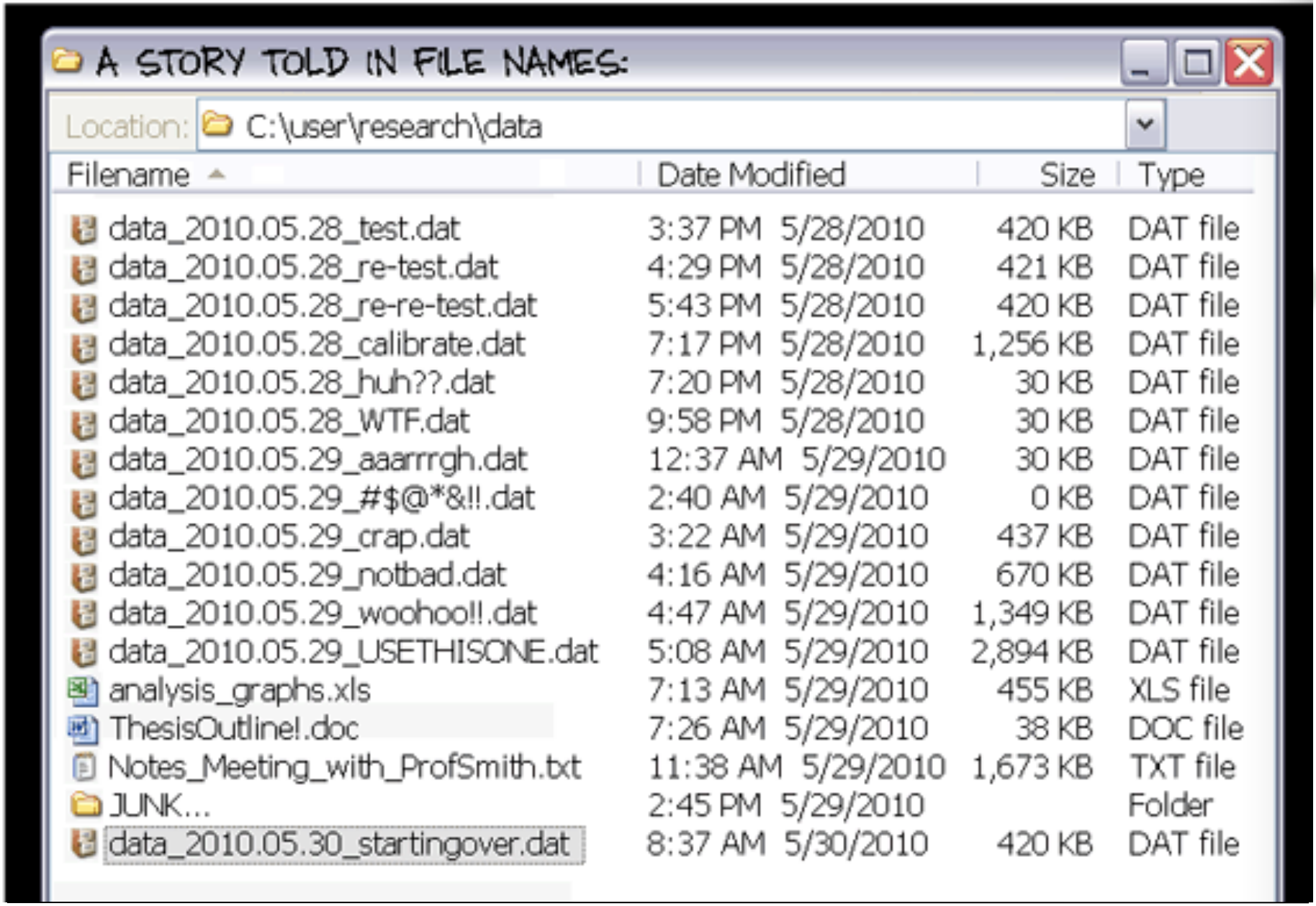
5. Assign descriptive file names

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

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

Better: bigfoot_agro_2000_gpp.tiff

- Project Name
- Site name
- Year
- What was measured
- File Format



Courtesy of PhD Comics

6. Preserve processing information

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
…	…	

```
### 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)
```

Keep raw data raw:

- Do not Include transformations, interpolations, etc in raw file
- Make your raw data "read only" to ensure no changes

When processing data:

- Use a programming language (e.g., R, SAS, MATLAB)
  - Code is a record of the processing done
  - Codes can be revised, rerun

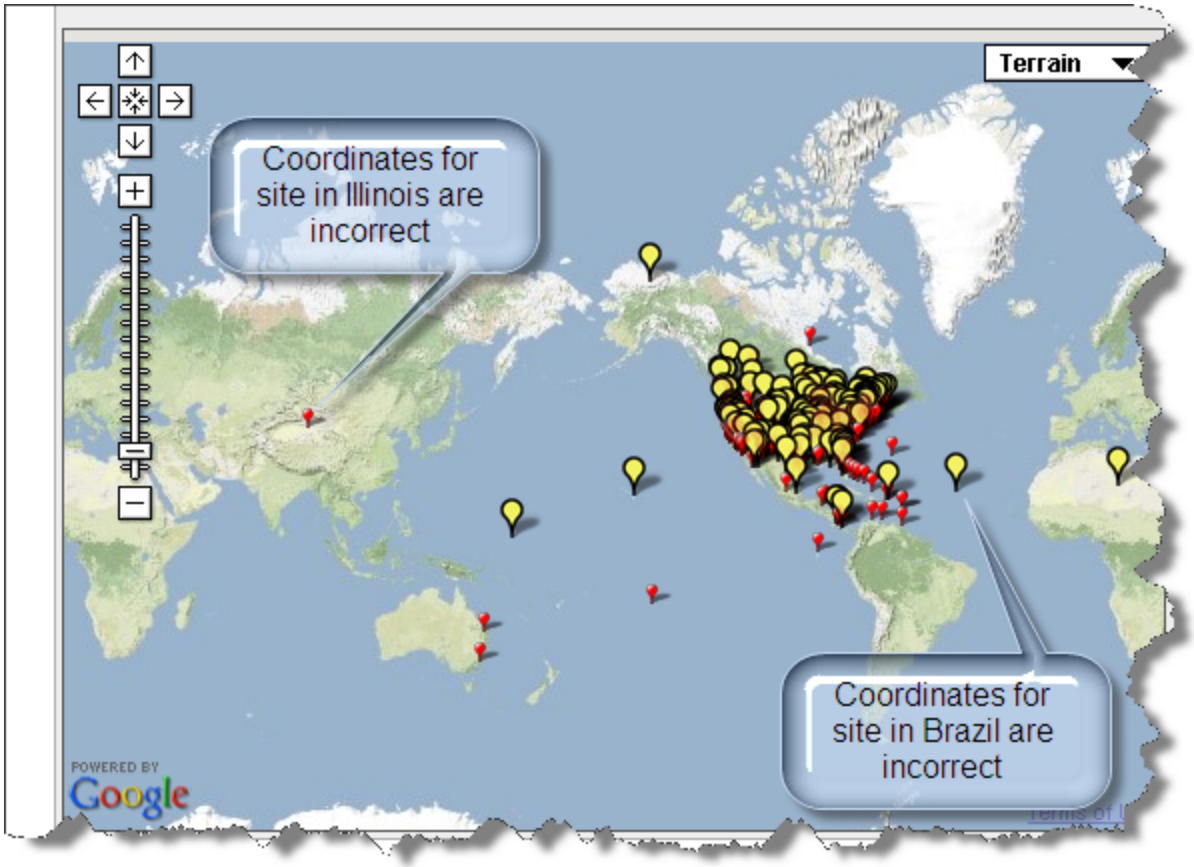
7. Perform basic quality assurance

- 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

7. Perform basic quality assurance (con't)

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

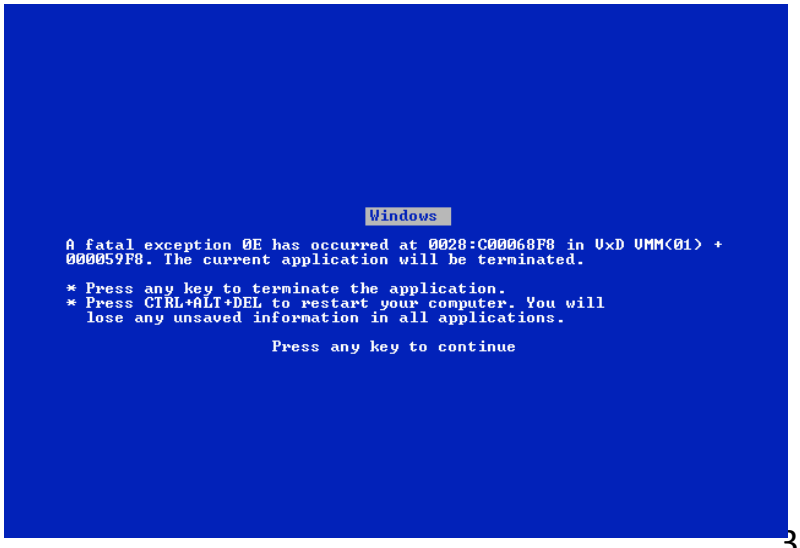


8. Provide Documentation / Metadata

- **What** does the data set describe?
- **Why** was the data set created?
- **Who** produced the data set and **Who** prepared the metadata?
- **When** and how frequently were the data collected?
- **Where** were the data collected and with what spatial resolution? (include coordinate reference system)
- **How** was each parameter measured?
- **How** reliable are the data?; what is the uncertainty, measurement accuracy?; what problems remain in the data set?
- **What** assumptions were used to create 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 three copies
    - original, one on-site (external), 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

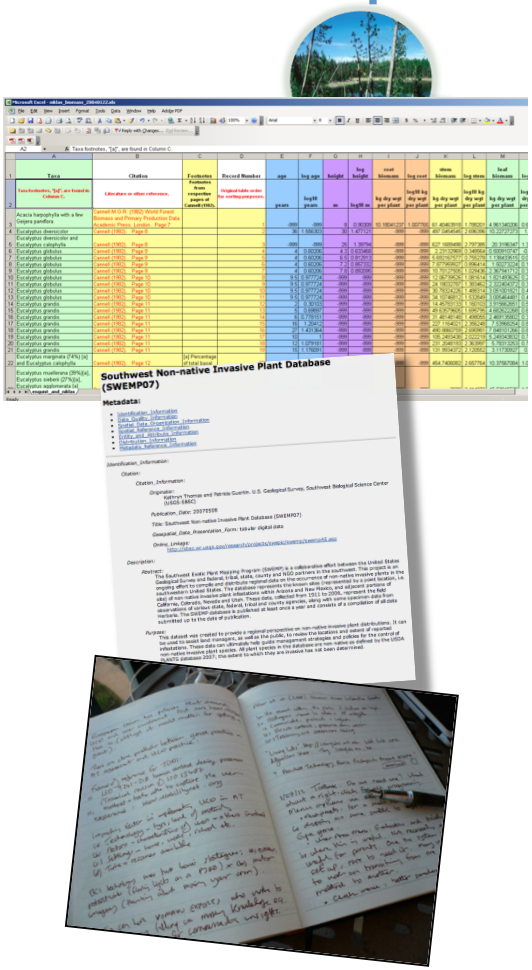
    http://www.pc-tools.net/win32/md5sums/

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

10. Preserve Your Data

What to preserve from the research project?

- Well-structured data files, with variables, units, and values defined
- Documentation and metadata record describing the data
- Additional information (provides context)
  - Materials from project wiki/websites
  - Files describing the project, protocols, or field sites (including photos)
  - Publication(s)



Fundamental Data Practices

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

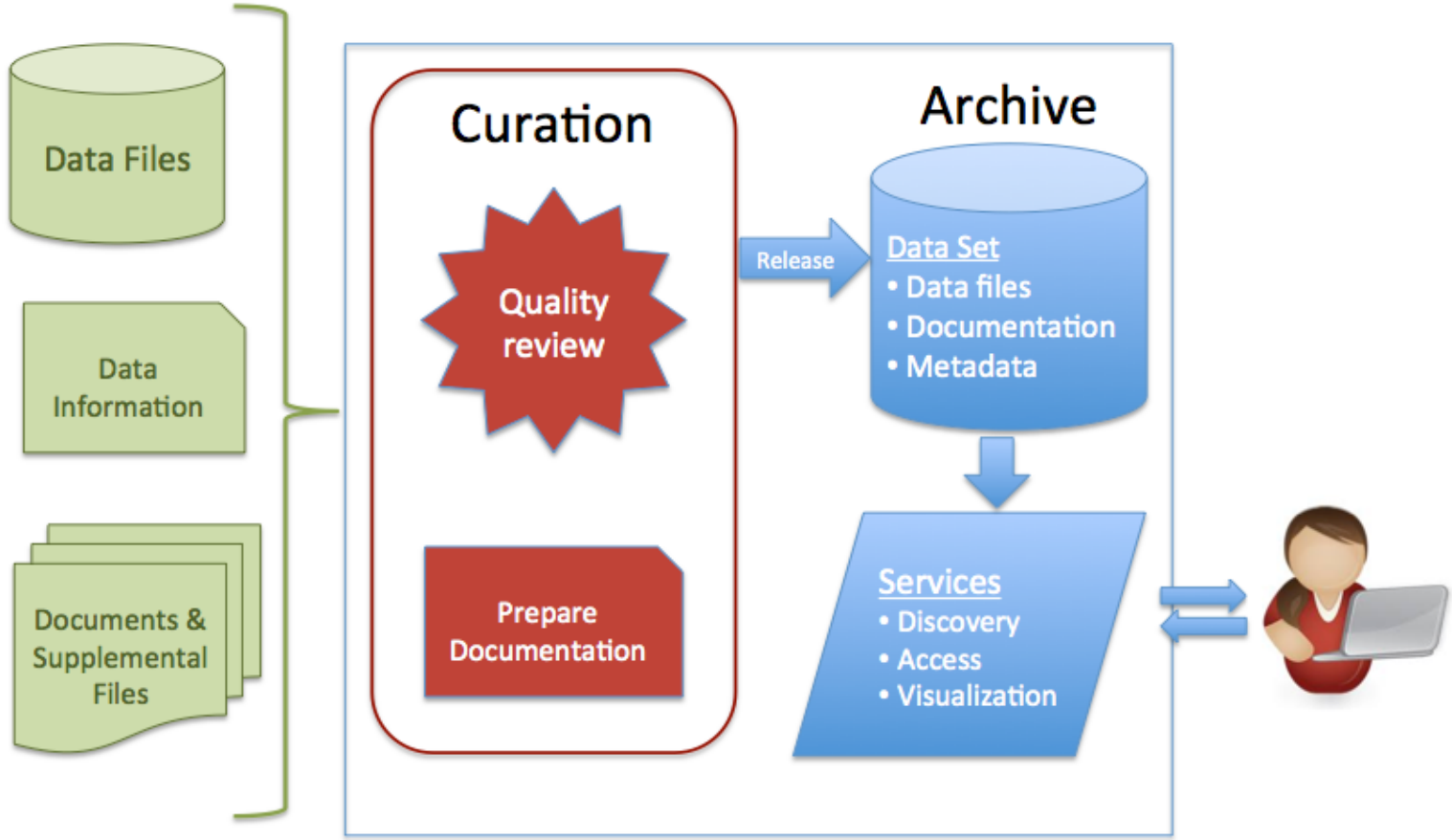
When your project is over, where should the data be archived?

- Part of project planning
  - "Begin with the end in mind"
  - Identified the Data Center
- Collaborated with data center during project
  - Volume and number of files
  - Special needs
  - Delivery dates
  - What additional data management steps would the data center like you to do?

USGS / PNAMP Webinar: Data Management Plan, Emily Fort

Curation and Archive

–Photos courtesy of S. Wullschleger, M. Mack, G. Shaver, and E. Kasischke



Data Providers

ORNL DAAC

for a user 20 years from now

Data Users

Data Quality Review Checklist

http://daac.ornl.gov/PI/qa_checklist.html

Data Files

1. Check for integrity of files (Checksum, file size, number of files)
2. Filenames are descriptive and consistent
   Action: Rename data files, if needed
3. Check if file format is appropriate and can be opened
   Action: Modify to archive format (non-proprietary) if needed
4. File organization is consistent and appropriate
5. Table header information complete and consistent with documentation
6. Properly versioned, if needed

Documentation files

1. Documentation matches files received.
2. Data set and its contents are clearly described
3. Geospatial and temporal information are complete and described
4. Variables and units follow standards or are well defined
5. Publication or manuscript describing the data is provided
6. Methodology, calibrations, and algorithms provided
7. Known issues/limitations clearly described
8. Statements are properly referenced

Data Quality Review Checklist (cont)

http://daac.ornl.gov/PI/qa_checklist.html

Parameter Values

1. Check to ensure valid range
2. Visualize (plot, map, or both)
3. Code(s) for missing values defined and used
4. Values for coded fields defined
5. Are accuracy and precision stated and reasonable?

Geospatial Information

View a webinar on geospatial information.

1. Spatial Reference System is well-defined for mapped data products
    a. Projection, datum, resolution, etc.
2. Spatial coordinates are well-defined, following standards
    a. lat / lon match description (geopolitical location, land vs water, correct hemisphere)

Temporal Information

1. Date and Time (calendar, time units and temporal extent, resolution, and boundary) are defined according to standards

Data Set Landing Page

doi: http://dx.doi.org/10.3334/ORNLDAAC/1290

NACP NAFD Project: Forest Disturbance History from Landsat, 1986-2010

Download Data

Data Set Overview

Data set	NACP NAFD Project: Forest Disturbance History from Landsat, 1986-2010
DOI	10.3334/ORNLDAAC/1290
Release date	2015-10-28
Project	North American Carbon Program (NACP)

Data set bounding box.
Lat: 51.68N to 22.69N, Long: 128.03W to 65.20W

Description

The North American Forest Dynamics (NAFD) products provided in this data set consist of 25 annual and two time-integrated forest disturbance maps for the conterminous United States (CONUS) derived from Landsat images for the period 1986-2010. Each annual map has classified pixels showing water, no forest cover, forest cover, no data available (data gaps) in present year, and forest disturbances that occurred in that year. The time-integrated maps are similarly classified, but over the entire 1986-2010 period with the first and last forest disturbance years identified and provided as separate maps.

Data set documentation

Data set reference document

Citation

Goward, S.N., C. Huang, F. Zhao, K. Schleeweis, K. Rishmawi, M. Lindsey, J.L. Dungan, and A. Michaelis. 2015. NACP NAFD Project: Forest Disturbance History from Landsat, 1986-2010. ORNL DAAC, Oak Ridge, Tennessee, USA. http://dx.doi.org/10.3334/ORNLDAAC/1290

Download citation from Datacite
RIS BibTex Other
Crosscite Citation Formatter

See our Data Citations and Acknowledgements policy for more information.

Data Set Landing Page (cont)

doi: http://dx.doi.org/10.3334/ORNLDAAC/1290

Citation

Goward, S.N., C.Q. Huang, J.G. Masek, W.B. Cohen, G.G. Moisen and K. Schleeweis. 2012. NACP North American Forest Dynamics Project: Forest Disturbance and Regrowth Data. Available on-line [http://daac.ornl.gov] from ORNL DAAC, Oak Ridge, Tennessee, U.S.A doi: http://dx.doi.org/10.3334/ORNLDAAC/1077

Download citation from Datacite
RIS BibTex Other
Crosscite Citation Formatter

See our Data Citations and Acknowledgements policy for more information.

Project

North American Carbon Program (NACP): The North American Carbon Program (NACP) is a multidisciplinary research program designed to obtain scientific understanding of North America's carbon sources and sinks and of changes in carbon stocks needed to meet societal concerns and to provide tools for decision makers.

Project data set list for North American Carbon Program (NACP).

Companion Files

- NAFD_Disturbance_guide.pdf
- NAFD_Disturbance_Image_Info.csv

Data Files

Check the 'Add to Cart' box and then click the 'Add Checked Items' button to order files. Click on the file link to see the file or save it. Display and save any companion files listed above. Click the 'Add Data Set' button to order the complete data set.

5524.2 MB in 112 Files

Show 50 entries Search:

Add to Cart	Data File (Granule)	File Size	Dates
☐	phase2_w2p026r037_disturbYear_flt_v1_last.tif	48.2 MB	1984-06-04 - 2009-09-16
☐	phase1_w2p025r029_disturbYear_flt_v1_last.tif	49.9 MB	1984-06-04 - 2009-09-16
☐	phase2_w2p021r030_disturbYear_flt_v1_last.tif	45.1 MB	1984-06-04 - 2009-09-16

Data Set Landing Page (cont)

doi: http://dx.doi.org/10.3334/ORNLDAAC/1290

☐	phase2_w2p027r027_disturbYear_flt_v1_first.tif	48.4 MB	1984-06-04 - 2009-09-16
☐	phase1_w2p021r037_disturbYear_flt_v1_first.tif	50.4 MB	1984-06-04 - 2009-09-16
☐	phase2_w2p042r029_disturbYear_flt_v1_first.tif	47.5 MB	1984-06-04 - 2009-09-16
☐	phase2_w2p042r035_disturbYear_flt_v1_first.tif	48.5 MB	1984-06-04 - 2009-09-16
☐	proto_w2p037r035_disturbYear_flt_v1_last.tif	51.8 MB	1984-06-04 - 2009-09-16
☐	phase1_w2p018r035_disturbYear_flt_v1_last.tif	51.0 MB	1984-06-04 - 2009-09-16
☐	phase1_w2p016r037_disturbYear_flt_v1_last.tif	48.6 MB	1984-06-04 - 2009-09-16
☐	phase2_w2p024r037_disturbYear_flt_v1_last.tif	48.1 MB	1984-06-04 - 2009-09-16
☐	phase1_w2p043r033_disturbYear_flt_v1_first.tif	48.5 MB	1984-06-04 - 2009-09-16
☐	phase2_w2p023r035_disturbYear_flt_v1_first.tif	45.7 MB	1984-06-04 - 2009-09-16
☐	phase2_w2p044r026_disturbYear_flt_v1_last.tif	50.3 MB	1984-06-04 - 2009-09-16
☐	phase2_w2p046r030_disturbYear_flt_v1_first.tif	49.1 MB	1984-06-04 - 2009-09-16
☐	proto_w2p014r032_disturbYear_flt_v1_last.tif	50.0 MB	1984-06-04 - 2009-09-16
☐	phase2_w2p023r028_disturbYear_flt_v1_first.tif	48.8 MB	1984-06-04 - 2009-09-16
☐	proto_w2p037r035_disturbYear_flt_v1_first.tif	51.8 MB	1984-06-04 - 2009-09-16

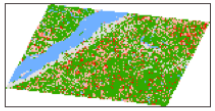
Showing 1 to 50 of 112 entries

Previous 1 2 3 Next

Add Checked Items to cart Add Data Set to cart Show Cart

Visualize and Subset Data

Download customized subsets in user-selected projection and format using the Spatial Data Access Tool.



Publications

Publications using this data set

Best Practices: Final Thoughts

- 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
- Include data management in your research workflow.
- Data Management should be a habit

Resources

Web Site

- Data Management for Data Providers

Workshops

- Workshops on Data Management

References

- 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, 82 (2): 138-141 http://www.jstor.org/stable/20168543
- Hook, L. A., T. W. Beaty, S. Santhana-Vannan, L. Baskaran, and R. B. Cook. 2010. Best Practices for Preparing Environmental Data Sets to Share and Archive. ORNL DAAC. doi: http://dx.doi.org/10.3334/ORNLDAAC/BestPractices-2010