USGS
science for a changing world

Data Management Training Modules:

Best Practices for Preparing Science Data to Share

U.S. Department of the Interior
U.S. Geological Survey

Version 1, October 2013

Narration: Introduction Slide 1

Welcome to the USGS Data Management Training Modules, a three part training series that will guide you on understanding and practicing good data management. Today we will discuss Best Practices for Preparing Science Data to Share.

Module objectives

- Describe the importance of maintaining well-managed science data.
- Outline 9 fundamental data management habits for preparing data to share.
- For each data management habit, list associated best practices.

USGS

Additional Resources

Glossary

Narration: Slide 2

Objectives

In this module, you'll learn:

- The importance of maintaining well-managed science data.
- Nine fundamental practices scientists should implement when preparing data to share.
- Associated best practices for each data management habit.

Problem Statement

- It is important to understand that good data management is crucial to achieving better and more streamlined data integration. There tends to be an underlying assumption that a majority of science data is available and poised for integration and re-use. Unfortunately, this is not the reality for most data. One problem scientists encounter when they discover data to integrate with other data, is the incompatibility of the data. Scientists can spend a lot of time trying to transform data to fit the needs of their project.

Narration: Slide 3

Problem Statement

It is important to understand that good data management is crucial to achieving better and more streamlined data integration. There tends to be an underlying assumption that a majority of science data is available and poised for integration and re-use. Unfortunately, this is not the reality for most data. One problem scientists encounter when they discover data to integrate with other data, is the incompatibility of the data. Scientists can spend a lot of time trying to transform data to fit the needs of their project. Some have estimated that researchers can spend up to 80% of their time finding, accessing, understanding, and preparing data and only 20% of their time actually analyzing the data. The habits described in this module will help scientists spend more time doing research and less time doing data management.

Importance of Well-Managed Data

- **Why manage our data?**
  - **Aids in the reproducibility of science.**
  - **Creates efficiencies in how science is done.**
  - **Makes sharing across groups more efficient.**
  - **Creates improved provenance in the science iteration process.**
  - **Aids responses to legal requirements such as Freedom of Information Act (FOIA) and Information Quality Act (IQA).**
  - **Supports scientific review and integrity.**

Narration: Slide 4 -- Contains video of Jeff Morisette talking

Importance of Well-Managed Data

I'm Jeff Morisette. I'm a Scientist with the USGS in Fort Collins, Colorado. Working as the director of the North Central Climate Science Center. There as climate is an important issue, that has a lot of stakeholders, we learned that it is very important to document our data as well as the analysis routines that stand behind our results. But today I wanted to provide two examples from the invasive species literature and some of the sciences that has been done here at the USGS in Fort Collins. These examples, I think, highlight the importance of well-managed data.

Some of the issues include reproducibility of science that you can go back when questioned or when updating your results, and reproduce the algorithms.

There's also efficiencies in how the science is done. If you have to spend a lot of time figuring out what was done last time you are losing some efficiencies in reproducing those results or updating analysis.

Along the line of those efficiencies is sharing across groups. Much of the work we do nowadays is collaborative, involves more than one agency or university or partner and if you can document the data and the analysis it helps to share the information and have everyone in that collaborative team understand what's being done.

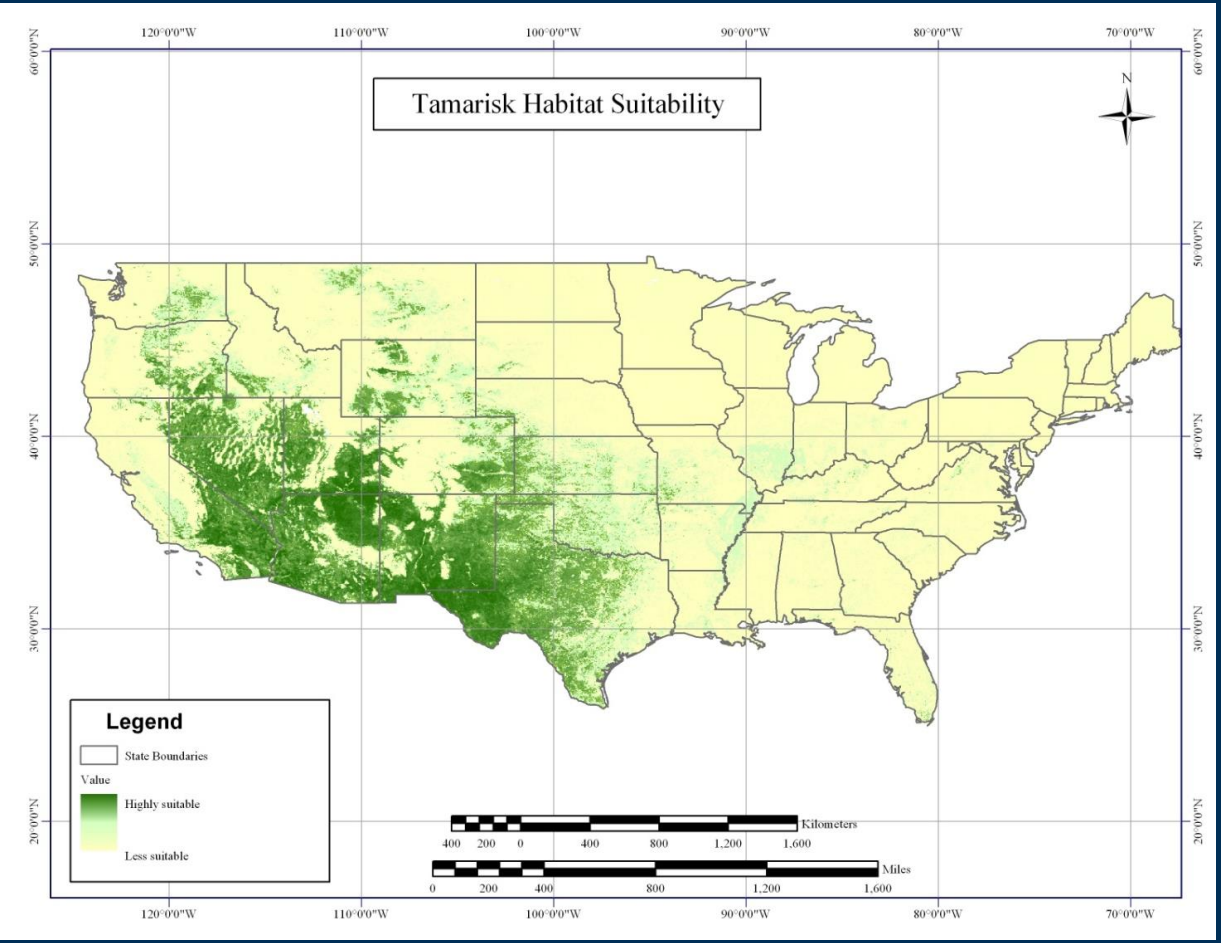
We also, like documenting the data and the analysis create a provenance that gives a full history of when the project was started how the analysis was done and how the final results were completed.

Provenance is important as you get into the scientific rigor of what you are doing as well as some of the legal implications of Freedom of Information request or Information Quality Act, the FOIA or IQA where we have the legal responsibility to describe what was done and to document thoroughly the results that were presented to the stakeholders. And so the provenance and well document data analysis is certainly helpful at meeting those requests.

Finally and I guess more in general is scientific integrity and just having results stand up to peer review and moving the community forward with results, dictates that you have it well documented and that those results can be reproducible.

Example: National Tamarisk Habitat Map: 2006

- This national map of habitat suitability for tamarisk was developed to aid land managers in early detection of this invasive plant species.
- The map produced by combining MODIS satellite products (vegetation indices and land cover type) with field observations to predict suitable habitat for tamarisk.



Morisette, J.T., C. S. Jernevich, A. Ullah, W. Cai, J.A. Pedelty, J. Gentle, T.J. Stohlgren, J.L. Schnase, A tamarisk habitat suitability map for the continental US., *Frontiers in Ecology and the Environment*, Volume 4, Issue 1 (February 2006) pp. 11–17 .

USGS

Additional Resources

Glossary

Narration: Slide 5 -- Contains video of Jeff Morisette talking

Example: National Tamarisk Habitat Map: 2006

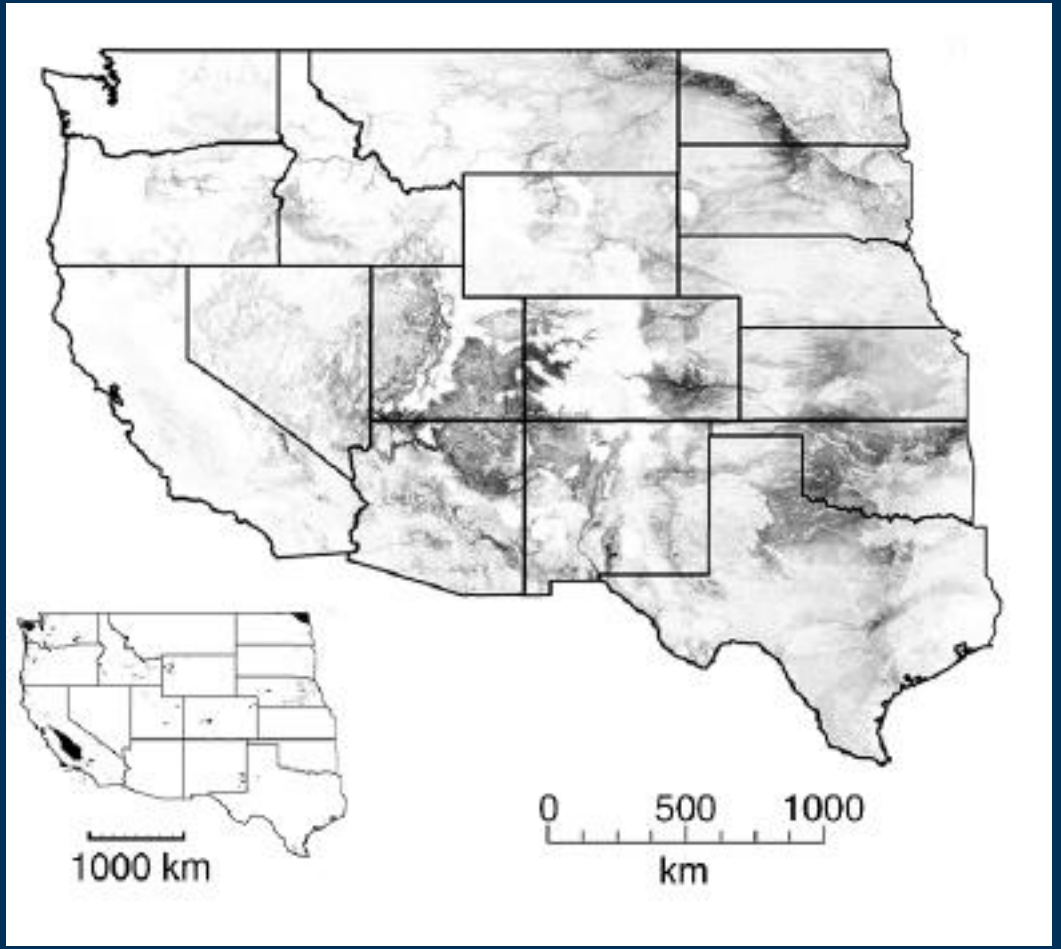
So two of the examples that we could walk through coming from the invasive species work at the USGS Fort Collins Science Center involve first a map of tamarisk where we had created a national map of tamarisk habitat suitability in 2006 and with that publication we were showing this invasive shrub that grows along riparian areas trying to update and prioritize which states had the most suitable habitat help managers get at primary areas of concern to either search for and try to remove tamarisks or watch for it establishing itself.

The map was helpful in that regard but almost as soon as it was published and actually in the publication we highlighted that it ought to be a living map. Invasives come in and spread rapidly so the species might find new places to go we might want to change our habitat suitability map of that species. So almost as soon as that map was produced various land managers and agencies started providing us with additional data and as those data came in … (Continued on next slide).

Example: National Tamarisk Habitat Map: 2013

“Perhaps one of the most important implications of this work is first proof that invasive species habitat suitability mapping should be seen as an ongoing and iterative process…”

“Careful and continued iteration between the model development and model-use communities can help ensure the most prudent use of modeling tools."



Crall et al., 2013. Using habitat suitability models to target invasive plant species surveys. Ecological Applications, 23(1), 2013, pp. 60–72.

USGS

Additional Resources

Glossary

Narration: Slide 6 -- Contains video of Jeff Morisette talking

Example: National Tamarisk Habitat Map: 2013

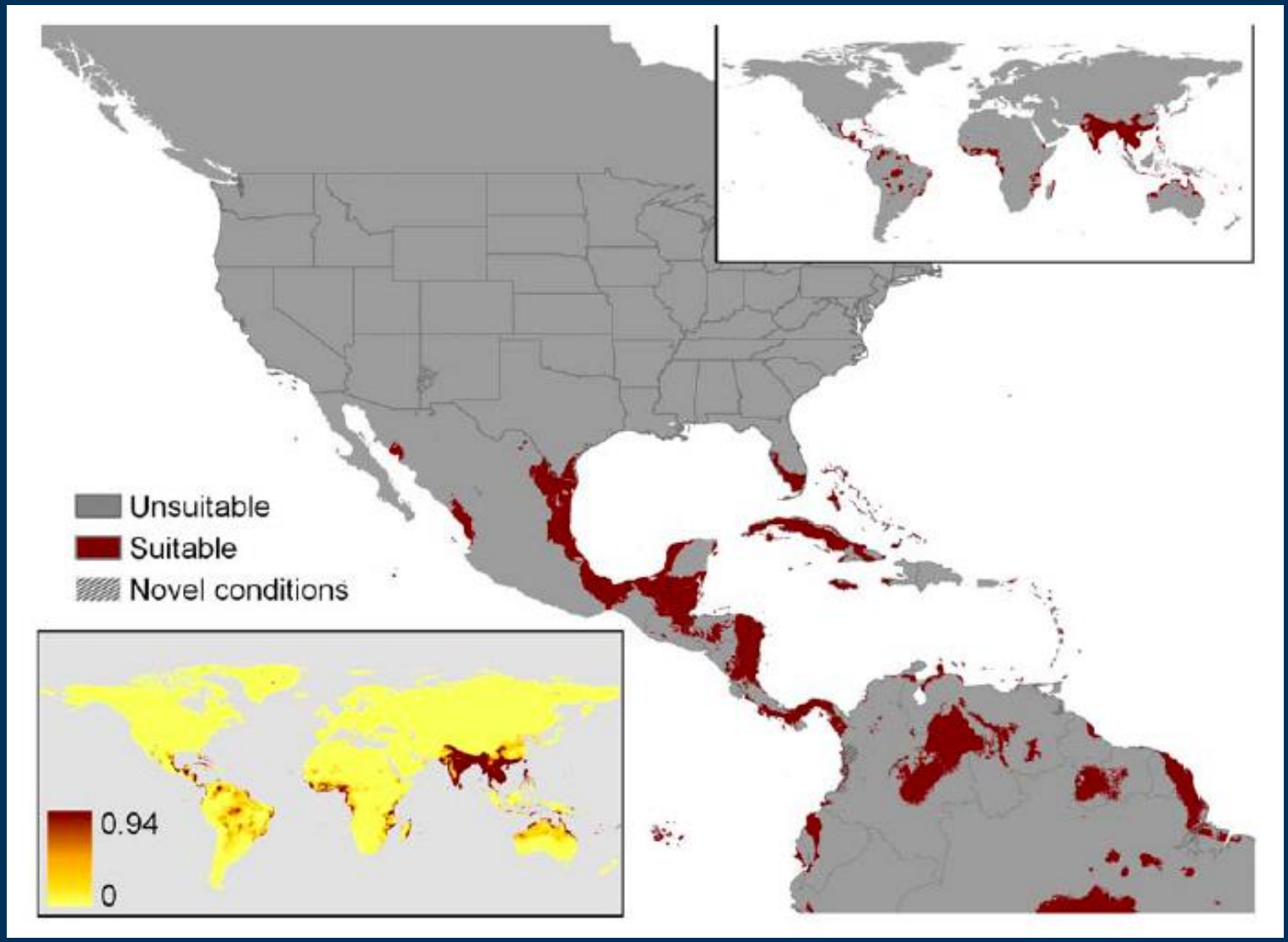
(continued from Previous slide)

… in 2010 we accumulated those new data as well as other data that followed to create an update that showed as those data come in this is what we can do and produce a modified updated version of that tamarisk habitat suitability map.

And in going back to analysis that was done nearly five years previously it became aware that we need a better way to document those routines that went into the original map. There may be different people working on a project, there may be just a time lapse, different computer systems come and go, so we realized with this update of the national tamarisk map that we need to do a better job and work towards documenting not just the data but the analysis routine that went into generating the results. We, at that time then, started to focus on scientific workflow.

The particular software we use is VisTrails as a provenance in scientific workflow management system and there's others out there but I think that the bottom line is that there are tools available to help document data and routines and the algorithms. And that it makes sense to use those for in this case, updating results as new data came in.

Example: Python Maps



Rodda GH, Jarnevich CS, Reed RN (2011) Challenges in Identifying Sites Climatically Matched to the Native Ranges of Animal Invaders. PLoS ONE 6(2):e14670. doi:10.1371/journal.pone.0014670.
Jarnevich, C.S., G.H. Rodda, and R.N. Reed. 2011. Data for giant constrictors—Biological management profiles and an establishment risk assessment for nine large species of pythons, anacondas, and the boa constrictor: U.S. Geological Survey Data Series 579.

USGS

Additional Resources

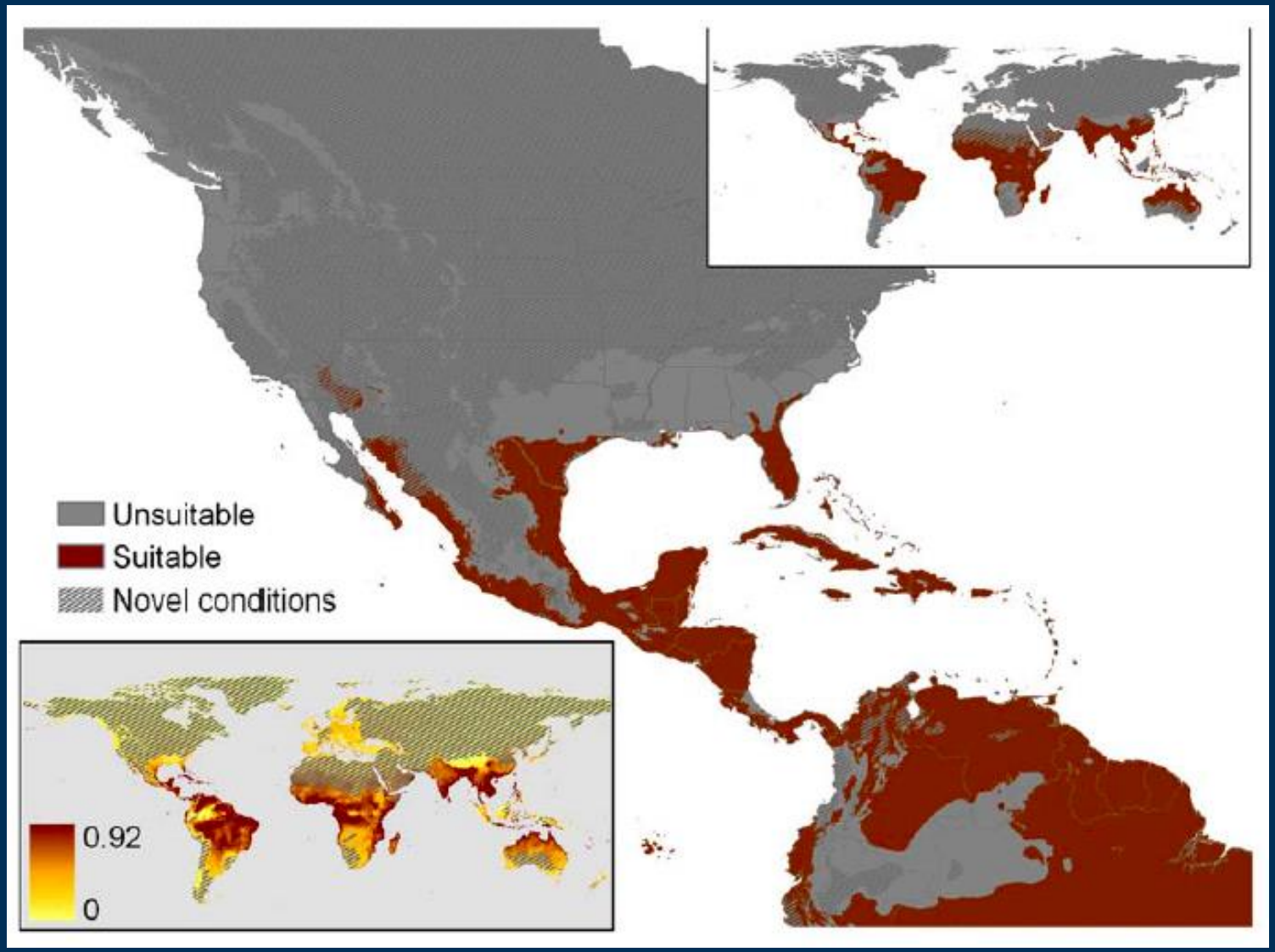
Glossary

Narration: Slide 7 -- Contains video of Jeff Morisette talking

Example: Python Maps

Another example was with invasive giant pythons in Florida. And in this paper the authors demonstrated that even with the same dataset there can be a significant difference in model results. And therefore the results of which areas would be suitable habitat, where those differences are a function of the settings used in the model. So in one particular setting, type of setting, the configuration for the model was limited to the southern and southwest part of Florida as suitable habitat.

Example: Python Maps



Rodda GH, Jarnevich CS, Reed RN (2011) Challenges in Identifying Sites Climatically Matched to the Native Ranges of Animal Invaders. PLoS ONE 6(2):e14670. doi:10.1371/journal.pone.0014670.
Jarnevich, C.S., G.H. Rodda, and R.N. Reed. 2011. Data for giant constrictors—Biological management profiles and an establishment risk assessment for nine large species of pythons, anacondas, and the boa constrictor: U.S. Geological Survey Data Series 579.

USGS

Additional Resources

Glossary

Narration: Slide 8 -- Contains video of Jeff Morisette talking

Example: Python Maps

There is another setting or different settings for the model that showed nearly all of Florida being suitable habitat and as there are stakeholders on both sides of this issue, it was important to document how one could come at the analysis with the same data but yet have different results and to document those decisions so the stakeholders are left with the information they need to understand those discrepancies between the models and then to use the map that makes the most scientific sense and makes the most sense in the context of what they are trying to accomplish.

Importance of Well-Managed Data

- Down the road, you may need the documentation for your data to be readily available.
- It's not just well defined data, but also the analysis that needs to be documented.
- You may have to justify the conclusions and defend the results.

Narration: Slide 9 -- Contains video of Jeff Morisette talking

Importance of Well-Managed Data

So these two examples show reproducible and updating science and documenting decisions in the analysis I think some good examples of where data and documenting the analysis were useful. I think that anyone doing science or working in the ecological community may down the road be asked to document the data and the results and to justify the conclusion they came to and I think that if they have well defined data, we defined well documented analysis that it would be helpful to justify the conclusions they came to and defend the results as they were produced.

Benefits of Good Data Management Practices

- **Short-term benefits:**
  - Spend less time doing data management and more time doing research.
  - Easier to prepare and use data.
  - Collaborators can readily understand and use data files.

- **Long-term benefits:**
  - Scientists outside your project can find, understand, and use your data to address broad questions.
  - You get credit for preserving data products and for their use in other papers.
  - Sponsors protect their investment.

USGS

Additional Resources

Glossary

Benefits of Good Data Management Practices are numerous.

What are the short-term benefits for implementing basic data management principles?

- First, scientists spend less time doing data management and more time doing research.
- Second, it is easier to prepare and use data.
- Third, collaborators can readily understand and use data files. For example, a researcher can turn her data over to a colleague and not have to spend time explaining the data (such as format, units, etc.)

What are some long-term benefits for following best practices for data management?

- First, scientists outside your project can find, understand, and use your data to address broad questions.
- Second, you get credit for preserving data products and for their use in other papers.
- Third, your sponsors protect their investment.

Fundamental 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 processing information.
6. Perform basic quality assurance.
7. Provide documentation.
8. Protect your data.
9. Preserve your data.

Narration: Slide 11

Fundamental Practices

The following are nine basic data habits that will help improve the information content of your data and make it easier to share data with others:

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 processing information
6. Perform basic quality assurance
7. Provide documentation
8. Protect your data
9. Preserve your data

The rest of this module will examine these principles more in depth.

Fundamental Practice #1

Define the contents of your data files:

Example of a Parameter 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)

- Use commonly accepted parameter names, descriptions, and units.
- Be consistent.
- Explicitly state units.
- Choose a format for each parameter, explain the format in the metadata, and use that format throughout the file.
- Use ISO formats.

Examples:

Use yyyymmdd; January 2, 1999 is 19990102

Use 24-hour notation (13:30 hrs instead of 1:30 p.m.)

USGS

Additional Resources

Glossary

Fundamental Practice #1: Define the contents of your data files

To implement basic data management practices, scientists should begin by using commonly accepted parameter names, descriptions, and units for the dataset. Some key aspects to keep in mind include being consistent in naming conventions and how the data is organized. It is critical to explicitly state the units reflected in the data.

Choose a format for each parameter, explain the format in the metadata, and use that format throughout the file.

It is important to select and use standardized formats for the dataset. An example of an ISO standard format can be found in the use of dates and times.

- Use the format “yyyymmdd”, where January 2, 1999 is represented as 19990102
- Use 24-hour notation, in which 1:30 p.m. is expressed as “13:30” hours.

Fundamental Practice #2

Use consistent data organization:

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

- Don't change or re-arrange columns
- Include header rows (first row contains file name, data set title, author, date, and companion file names).
- Column headings should describe content of each column.
  - Include one row for parameter names and one for parameter units.

Example 1: Each row in a file represents a complete record, and the columns represent all the parameters that make up the record.

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

Example 2: Parameter name, value, and units are placed in individual rows. This approach is used in relational databases.

USGS

Additional Resources

Glossary

Fundamental Practice #2: Use consistent data organization

When you are initially organizing your dataset, a best practice is to choose one way to organize your data and remain consistent with that method throughout the file.

For tabular data, each separate line or row should represent an observation, one complete record. The columns represent all the parameters that make up the record. As shown in our first example, for records that do not have measurements for most parameters, each parameter can be a column header, which is defined and the values recorded in two columns. Here, columns are defined as “station”, “date”, “temp”, and “precip”. Additional information (soil moisture, humidity, etc.) will be added as additional columns.

In our second example, there are more parameters to define in the data collection, thus the parameter is the column header and requires the actual parameter name be identified in the data column. Values and units are each placed in individual rows in this example, representing a relational database model approach. When additional variables are added they are added as rows.

It is a good idea to keep a set of similar measurements together (same investigator, methods, time basis, and instruments) in one data file. Many small files are more difficult to process than one larger file. There are exceptions, however. For example, observations using different types of measurements might be placed into separate data files, and data collected on different time scales or temporal resolution might be handled more efficiently in separate files. Use similar data organization, parameter formats, and common site names across the data set.

Include data set organization and provide definitions for all coded values or abbreviations, including spatial coordinates, in the documentation.

Some basic principles to keep in mind when designing your tabular data: Don’t change or re-arrange the columns once you begin. Include header rows (a first row contains the file name, data set title, author, date, and companion file names). Column headings should describe content of each column, and include one row for parameter names and one for parameter units.

Example of Poor Data Practice for Collaboration and Sharing

C:\Documents and Settings\hampton\My Documents\NCEAS Distributed Graduate Seminars\[Wash Cres Lake Dec 15 Dont_Use.xls]Sheet1

Stable Isotope Data Sheet

Sampling Site / Identifier: Wash Cresc Lake
Sample Type: Algal
Date: Dec. 16
Tray ID and Sequence: Tray 004

Reference statistics: SD for delta ^{13}C = 0.07 SD for delta ^{15}N = 0.15

Peter's lab Don't use - old data
Washed Rocks

Position	SampleID		Weight (mg)	%C	delta 13C	delta 13C_ca	%N	delta 15N	delta 15N_ca	Spec. No.	
A1		ref	0.98	38.27	-25.05	-24.59	1.96	4.12	3.47	25354	
A2		ref	0.98	39.78	-25.00	-24.54	2.03	4.01	3.36	25356	
A3		ref	0.98	40.37	-24.99	-24.53	2.04	4.09	3.44	25358	
A4		ref	1.01	42.23	-25.06	-24.60	2.17	4.20	3.55	25360	
A5	ALG01		3.05	1.88	-24.34	-23.88	0.17	-1.65	-2.30	25362	c
A6	Lk Outlet Alg		3.06	31.55	-30.17	-29.71	0.92	0.87	0.22	25364	
A7	ALG03		2.91	6.85	-21.11	-20.65	0.48	-0.97	-1.62	25366	c
A8	ALG05		2.91	35.56	-28.05	-27.59	2.30	0.59	-0.06	25368	
A9	ALG07		3.04	33.49	-29.56	-29.10	1.68	0.79	0.14	25370	
A10	ALG06		2.95	41.17	-27.32	-26.86	1.97	2.71	2.06	25372	
B1	ALG04		3.01	43.74	-27.50	-27.04	1.36	0.99	0.34	25374	c
B2	ALG02		3	4.51	-22.68	-22.22	0.34	4.31	3.66	25376	
B3	ALG01		2.99	1.59	-24.58	-24.12	0.15	-1.69	-2.34	25378	c
B4	ALG03		2.92	4.37	-21.06	-20.60	0.34	-1.52	-2.17	25380	c
B5	ALG07		2.9	33.58	-29.44	-28.98	1.74	0.62	-0.03	25382	
B6		ref	1.01	44.94	-25.00	-24.54	2.59	3.96	3.31	25384	
B7		ref	0.99	42.28	-24.87	-24.41	2.37	4.33	3.68	25386	
B8	Lk Outlet Alg		3.04	31.43	-29.69	-29.23	1.07	0.95	0.30	25388	
B9	ALG06		3.09	35.57	-27.26	-26.80	1.96	2.79	2.14	25390	
B10	ALG02		3.05	5.52	-22.31	-21.85	0.45	4.72	4.07	25392	
C1	ALG04		2.98	37.90	-27.42	-26.96	1.36	1.21	0.56	25394	c
C2	ALG05		3.04	31.74	-27.93	-27.47	2.40	0.73	0.08	25396	
C3		ref	0.99	38.46	-25.09	-24.63	2.40	4.37	3.72	25398	
				23.78			1.17				

Shore	Avg Con
-1.26	-27.22
1.26	0.32

Courtesy of S. Hampton, National Center for Ecological Analysis and Synthesis)

USGS

Additional Resources

Glossary

Example of Poor Data Practice for Collaboration and Sharing

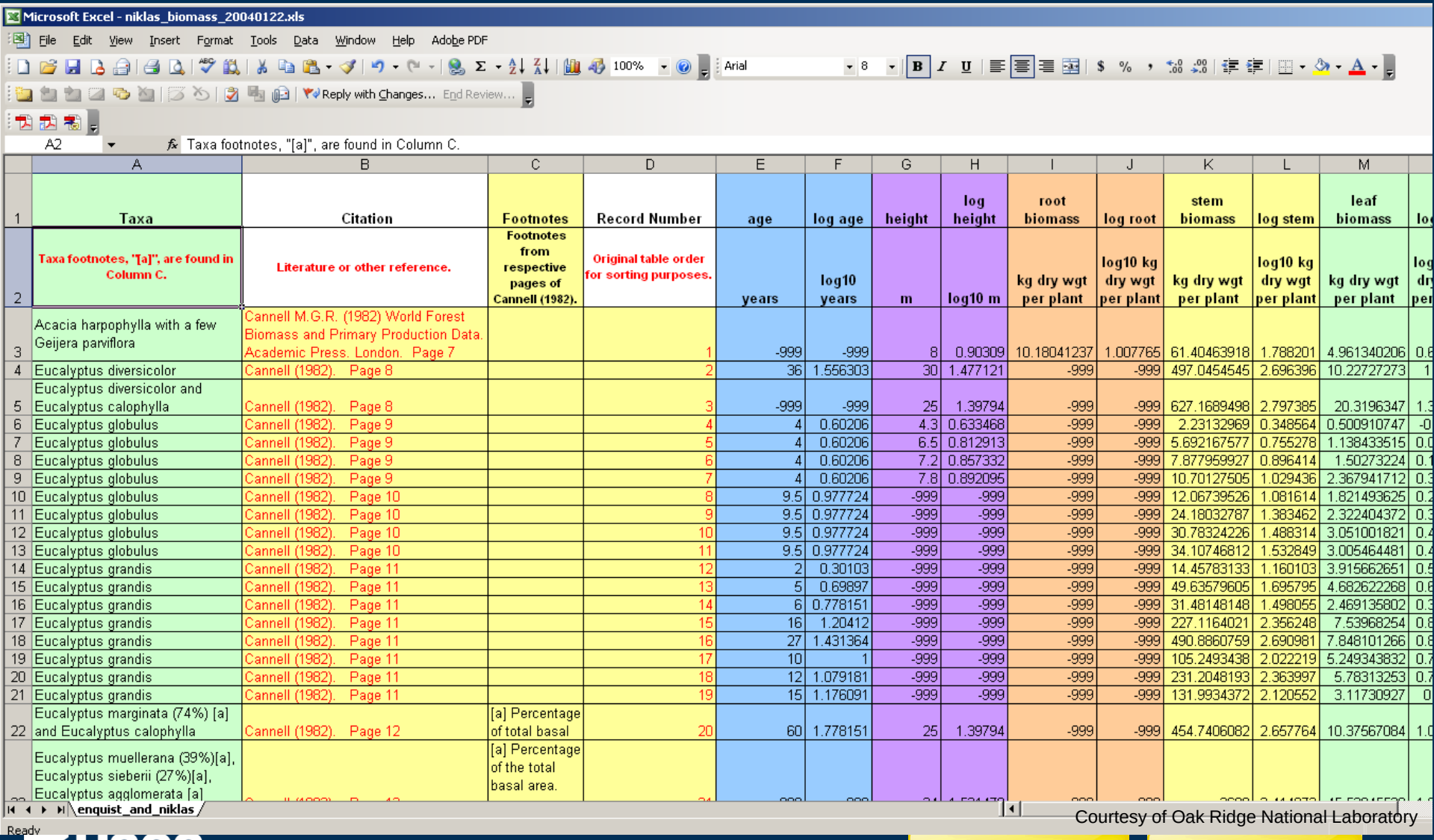
This illustration shows an example of poor practice for working in spreadsheets for data collection. At first glance, it may appear this data is well formulated, but a closer look reveals a number of practices that will make it difficult to re-use in its present state. For example, there are calculations in the far right columns that appear to have been made during a data analysis phase but that do not represent valid data entries.

Notice in the upper right corner a comment stating “Don’t use – old data”, and “Peter’s lab”. These remarks leave the viewer wondering about who Peter is and which lab he was located in, as well as why this data may not be the most accurate data spreadsheet. One also may wonder what the “c” located in the far right column represents, and what the numbers at the bottom of the spreadsheet represent, since they are unaffiliated with a particular row of data in the spreadsheet.

Notice there are numbers added in inconsistent places (two numbers at the bottom of the chart) and the letter “C” appears in an unlabeled column.

An improved approach to data is illustrated on the next slide.

Example of Good Data Practice for Collaboration and Sharing



Microsoft Excel - niklas_biomass_20040122.xls

File Edit View Insert Format Tools Data Window Help Adobe PDF

Reply with Changes... End Review...

A2 Taxa footnotes, "[a]", are found in Column C.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Taxa	Citation	Footnotes	Record Number	age	log age	height	log height	root biomass	log root	stem biomass	log stem	leaf biomass
2	Taxa footnotes, "[a]", are found in Column C.	Literature or other reference.	Footnotes from respective pages of Cannell (1982).	Original table order for sorting purposes.	years	log10 years	m	log10 m	kg dry wgt per plant	log10 kg dry wgt per plant	kg dry wgt per plant	log10 kg dry wgt per plant	kg dry wgt per plant
3	Acacia harpophylla with a few Geijera parviflora	Cannell M.G.R. (1982) World Forest Biomass and Primary Production Data. Academic Press. London. Page 7		1	-999	-999	8	0.90309	10.18041237	1.007765	61.40463918	1.788201	4.961340206
4	Eucalyptus diversicolor	Cannell (1982). Page 8		2	36	1.556303	30	1.477121	-999	-999	497.0454545	2.696396	10.22727273
5	Eucalyptus diversicolor and Eucalyptus calophylla	Cannell (1982). Page 8		3	-999	-999	25	1.39794	-999	-999	627.1689498	2.797385	20.3196347
6	Eucalyptus globulus	Cannell (1982). Page 9		4	4	0.60206	4.3	0.633468	-999	-999	2.23132969	0.348564	0.500910747
7	Eucalyptus globulus	Cannell (1982). Page 9		5	4	0.60206	6.5	0.812913	-999	-999	5.692167577	0.755278	1.138433515
8	Eucalyptus globulus	Cannell (1982). Page 9		6	4	0.60206	7.2	0.857332	-999	-999	7.877959927	0.896414	1.50273224
9	Eucalyptus globulus	Cannell (1982). Page 9		7	4	0.60206	7.8	0.892095	-999	-999	10.70127505	1.029436	2.367941712
10	Eucalyptus globulus	Cannell (1982). Page 10		8	9.5	0.977724	-999	-999	-999	-999	12.06739526	1.081614	1.821493625
11	Eucalyptus globulus	Cannell (1982). Page 10		9	9.5	0.977724	-999	-999	-999	-999	24.18032787	1.383462	2.322404372
12	Eucalyptus globulus	Cannell (1982). Page 10		10	9.5	0.977724	-999	-999	-999	-999	30.78324226	1.488314	3.051001821
13	Eucalyptus globulus	Cannell (1982). Page 10		11	9.5	0.977724	-999	-999	-999	-999	34.10746812	1.532849	3.005464481
14	Eucalyptus grandis	Cannell (1982). Page 11		12	2	0.30103	-999	-999	-999	-999	14.45783133	1.160103	3.915662651
15	Eucalyptus grandis	Cannell (1982). Page 11		13	5	0.69897	-999	-999	-999	-999	49.63579605	1.695795	4.682622268
16	Eucalyptus grandis	Cannell (1982). Page 11		14	6	0.778151	-999	-999	-999	-999	31.48148148	1.498055	2.469135802
17	Eucalyptus grandis	Cannell (1982). Page 11		15	16	1.20412	-999	-999	-999	-999	227.1164021	2.356248	7.53968254
18	Eucalyptus grandis	Cannell (1982). Page 11		16	27	1.431364	-999	-999	-999	-999	490.8860759	2.690981	7.848101266
19	Eucalyptus grandis	Cannell (1982). Page 11		17	10	1	-999	-999	-999	-999	105.2493438	2.022219	5.249343832
20	Eucalyptus grandis	Cannell (1982). Page 11		18	12	1.079181	-999	-999	-999	-999	231.2048193	2.363997	5.78313253
21	Eucalyptus grandis	Cannell (1982). Page 11		19	15	1.176091	-999	-999	-999	-999	131.9934372	2.120552	3.11730927
22	Eucalyptus marginata (74%) [a] and Eucalyptus calophylla	Cannell (1982). Page 12	[a] Percentage of total basal	20	60	1.778151	25	1.39794	-999	-999	454.7406082	2.657764	10.37567084
	Eucalyptus muellerana (39%)[a], Eucalyptus sieberii (27%)[a], Eucalyptus agglomerata [a]		[a] Percentage of the total basal area.										

enquist_and_niklas

Ready

Courtesy of Oak Ridge National Laboratory

USGS

Additional Resources

Glossary

Example of Good Data Practice for Collaboration and Sharing

This spreadsheet is much better organized and thus entirely more useful. Each parameter is separated into its own column, with definitions provided and units defined for each. Color coding helps with ease of reading the data. -999 is used as a standard placeholder representing unknown values.

We suggest saving this file as a csv (comma separated value) so that people can read this well into the future.

Fundamental Practice #3

Use stable file formats:

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 et al. (2005)

(Chase et al., 2016)

Use text (ASCII) file formats for tabular data.

- **Examples: .txt or .csv (comma-separated values)**

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

USGS

Additional Resources

Glossary

Fundamental Practice #3: Use stable file formats

Data re-use depends on the ability to return to a dataset, perhaps long after the proprietary software you used to develop it, is available. Remember floppy disks? It is difficult to find a computer that will read a floppy disk today. We must think of digital data in a similar way. Select a consistent format that can be read well into the future and is independent of changes in applications. If your data collection process used proprietary file formats, converting those files into a stable, well-documented, and non-proprietary format to maximize others' abilities to use and build upon your data is a best practice.

When possible, convert your tabular dataset into ASCII text format. (For example, ".txt" or **c**omma-**s**eparated **v**alue(s), ".csv".

Within the ASCII file, delimit fields using commas, pipes (|), tabs, or semicolons (in order of preference).

Some suggested stable geospatial file formats include the following:

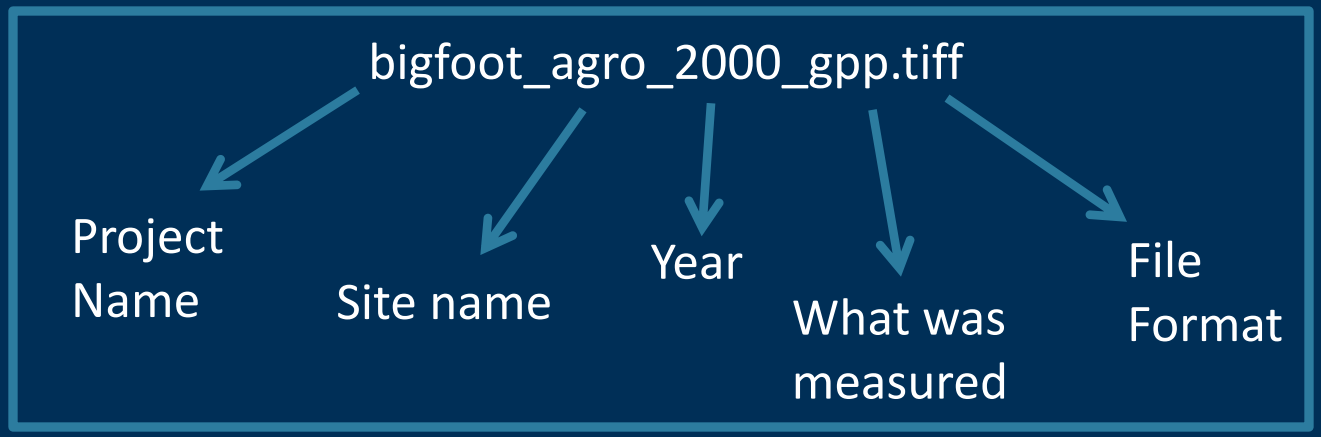
For raster formats, geotiffs, netCDF preferably with the CF convention, HDF. and ASCII. (The ASCII is a plain text file gridded format with external projection information).

For vector formats, preferred file formats include: shapefiles and KML/GML.

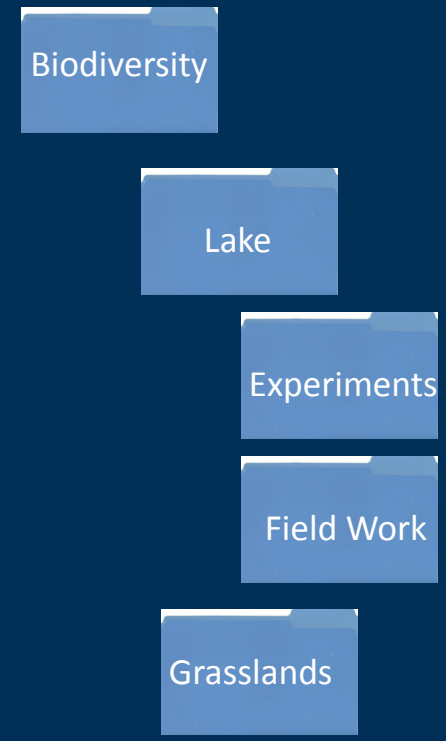
Storing data in recommended formats with detailed documentation will allow your data to be easily read many years into the future. Easy access means improved usability of your data and more researchers using and citing your data. Users can spend more time analyzing the data and spend less time in preparing the data.

Fundamental Practice #4

Assign descriptive file names



Aranibar and Macko (2005)



File names should:

- **Be unique.**
- **Reflect contents.**
- **Use ASCII characters only.**
- **Use lower case letters, numbers, dashes, and underscores.**
- **Avoid spaces and special characters.**

Make sure your file system is logical and efficient.

USGS

Additional Resources

Glossary

Narration: Slide 17

Fundamental Practice #4: Assign descriptive file names

File names should reflect the contents of the file and uniquely identify the data file. File names may contain information such as project acronym, study title, location, investigator, year(s) of study, data type, version number, and file type. Avoid using file names such as “mydata.dat” or “1998.dat”. A well-constructed file name will include information such as who, what, date, location. For example, “bigfoot_agro_2000_gpp.tiff” reveals the project name called bigfoot, the site name called agro, the year in 2000, what was measured, gpp, and the file format, a tiff file.

File names should be constructed to contain only lower-case letters, numbers, dashes, and underscores, with no spaces or special characters. This allows easy management by various data systems and to decrease software and platform dependency. File names should not be more than 64 characters in length and, if well-constructed, could be considerably less. Similar logic is useful when designing file directory structures and names, which you should ensure is logical and efficient in design.

Fundamental Practice #5

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.**
- **Consider making your raw data "read only" to ensure no changes.**

When processing data:

- **Use a scripted language (e.g., R, SAS, MATLAB).**
  - **Processing scripts are records of the processing done.**
  - **Scripts can be revised and rerun.**

USGS

Additional Resources

Glossary

Fundamental Practice #5: Preserve processing information

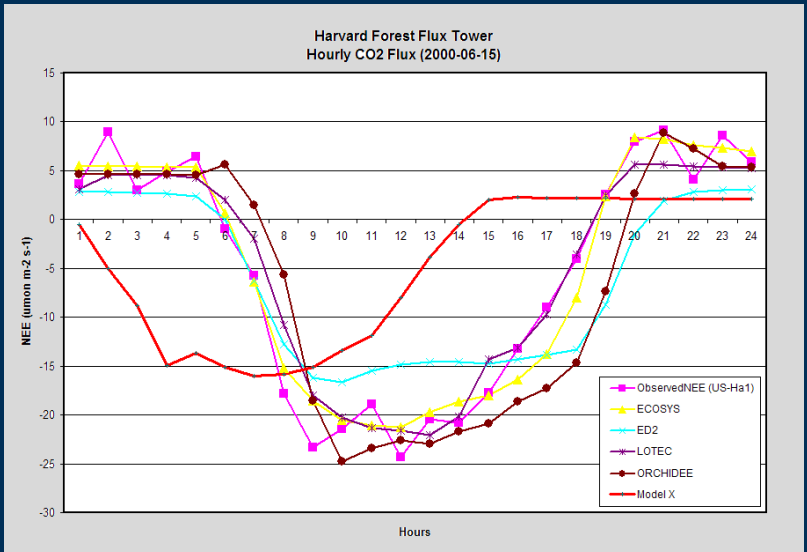
To preserve your data and its integrity, save a "read-only" copy of your raw data files with no transformations, interpolation, or analyses. Use a scripted language such as “R”, “SAS” or “MATLAB” to process data in a separate file, located in a separate directory. The scripts you have written are an excellent record of data processing, they can also easily and quickly be revised and rerun in the event of data loss or requests for edits, and have the added benefit of allowing a future worker to follow-up or reproduce your processing. Keep in mind that while GUI-based programs are easy on the front end, they do not keep a record of changes to your data and make reproducing results difficult.

In this example, an “R” to call is made on the data set to plot the data and perform a log transform – this way, changes are not retained in the original, raw data file.

Fundamental Practice #6

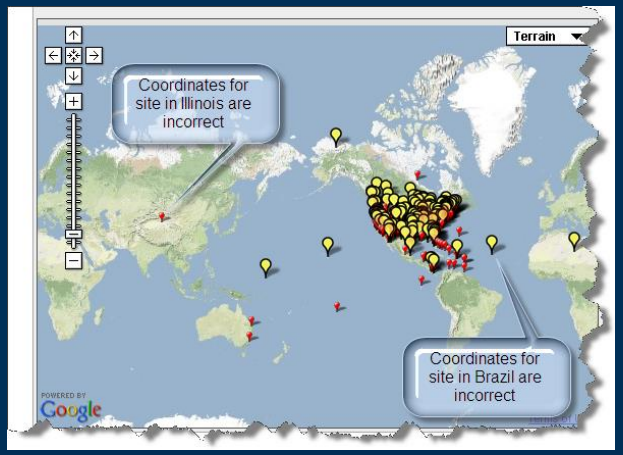
Perform Basic Quality Assurance

Photo Courtesy of Dan Ricciuto & Yaxing Wei, ORNL



Example: Model X in red uses UTC time, all others use Eastern Time.

- Assure data are delimited and line up in proper columns.
- Check for missing values in key parameters.
- Scan for impossible and anomalous values.
- Perform and review statistical summaries.
- Map location data and assess any errors.



USGS

Additional Resources

Glossary

Fundamental Practice #6: Perform Basic Quality Assurance

Quality Assurance is an essential and critical activity in ensuring integrity of your dataset. Perform frequent checks on your data to assess any errors.

Some key checkpoints include assuring data are delimited and line up in proper columns, checking for missing values in key parameters, scanning for impossible and anomalous values, performing and reviewing statistical summaries, and mapping location data to assess any errors.

Our example shows Model X using UTC time in one plot illustration, where all other plots use Eastern Time. The second example shows a map plot of all latitude and longitudes, such that errors can be identified more easily – some observations are shown some distance from where they should be appearing on a map.

Fundamental Practice #7

Provide Documentation / Metadata that follows standards

Who
- Who collected the data?
- Who processed the data?
- Who wrote the metadata?
- Who to contact for questions?
- Who to contact to order?
- Who owns the data?

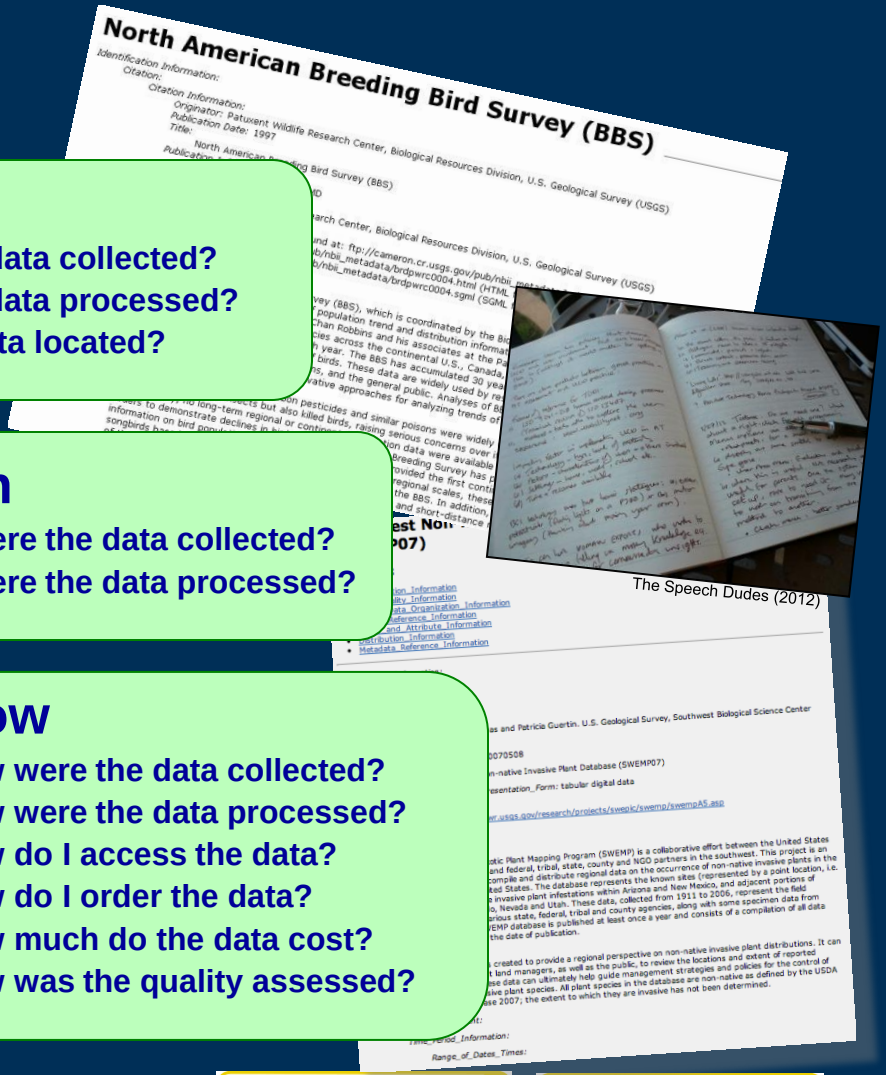
Where
- Where were the data collected?
- Where were the data processed?
- Where are the data located?

What
- What are the data about?
- What project were they collected under?
- What are the constraints on their use?
- What is the quality?
- What are appropriate uses?
- What parameters were measured?
- What format are the data in?

When
- When were the data collected?
- When were the data processed?

How
- How were the data collected?
- How were the data processed?
- How do I access the data?
- How do I order the data?
- How much do the data cost?
- How was the quality assessed?

Why
- Why were the data collected?



Additional Resources

Glossary

Narration: Slide 20

Fundamental Practice #7: Provide Documentation and formal Metadata that follows standards

Metadata and associated documentation is absolutely crucial for any potential use or reuse of data; no one can responsibly re-use or interpret data without accompanying compliant and standardized metadata. Metadata can be used for analysis of data, maintaining the longevity of a dataset, and tracking the progress of a research project.

Metadata describe your data so that others can understand what your data set represents; they are thought of as "data about the data" or the "who, what, where, when, and why" of the data. Metadata documentation can be in the form of a document or a formatted list of descriptors that include keywords, spatial and temporal extent, investigators, and other information about the data set. Metadata should be written from the standpoint of someone reading it who is unfamiliar with your project, methods, or observations. What does a user, 20 years into the future, need to know to use your data properly?

As with data, associated documentation should be saved using stable, non-proprietary formats. Images, figures, and pictures should be individual GIF or JPEG files. Documents should be in separate PDF or PS files identified in the data file. Names of documentation files should be similar to the name of the data set and the data file(s). The documentation is most useful when structured as a user's guide for the data product. Documentation can never be too complete. Users who are not familiar with your data will need more detailed documentation to understand your data set. Long-term experimental activities require more documentation because personnel change over time.

Fundamental Practice #8

Protect Your Data

Create back-up copies often:

- Ideally three copies: original, one on-site (external), and one off-site.
- Frequency based on need / risk.

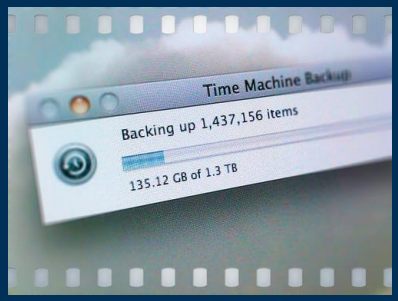
From Flickr The Commons

Ensure that you can recover from a data loss:

- Periodically test your ability to restore information.

Ensure file transfers are done without error:

- Compare checksums before and after transfers.



From Flickr by Brian J Matis

USGS

Additional Resources

Glossary

Fundamental Practice #8: Protect your data

Create and test back-up copies often to prevent the disaster of lost data. Maintain at least three copies of your data: the original, an on-site but external backup, and an off-site backup in case of a disaster. The advent of cloud storage allows for remote file storage that can be accessed from virtually anywhere. To ensure you can recover from data loss, periodically test your ability to recover your data.

It is also important to ensure that file transfers are done without error by using checksums. Checksums are numerical values calculated from the number of bytes of data. If the current value matches a previous checksum, your data has likely not changed.

Fundamental Practice #9

Preserve Your Data

What to preserve from the research project?

- **Well-structured data files, with variables, units, and values well-defined.**
- **Metadata record describing the data structured using Federal standards.**
- **Additional information (provides context):**
  - **Materials from project wiki/websites.**
  - **Files describing the project, protocols, or field sites (including photos).**
  - **Publication(s).**

Courtesy of ORNL

The Speech Dudes (2012)

USGS

Additional Resources

Glossary

Narration: Slide 22

Fundamental Practice #9 : Preserve your data

You may wonder what should you be preserving from your research project. Generally you should preserve your well-structured data files, with variables, units, and values well-defined. Additionally, preserve your standardized metadata record that describes the data, and, finally, any additional information that provides context such as materials from project wikis or websites, files describing the project, protocols, or field sites, including photos, and any publications using the data.

Key Points

- Data Management is important and critical in today's science.
- Well-organized and documented 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. Make it a habit to manage your data well.

Narration: Slide 23

Key Points

Data Management is important and critical in today's science.

Well-organized and documented data enables researchers to work more efficiently, can be shared easily by collaborators, and can potentially be re-used in ways not imagined when originally collected.

Include data management in your research workflow. Make it a habit to manage your data well.