MBLWHOI LIBRARY

MBLWHOI LIBRARY
Marine Biological Laboratory
Woods Hole Oceanographic Institution

DATA MANAGEMENT:

CREATING, MAINTAINING, AND PROTECTING QUALITY RESEARCH DATA

September 24th, 2014

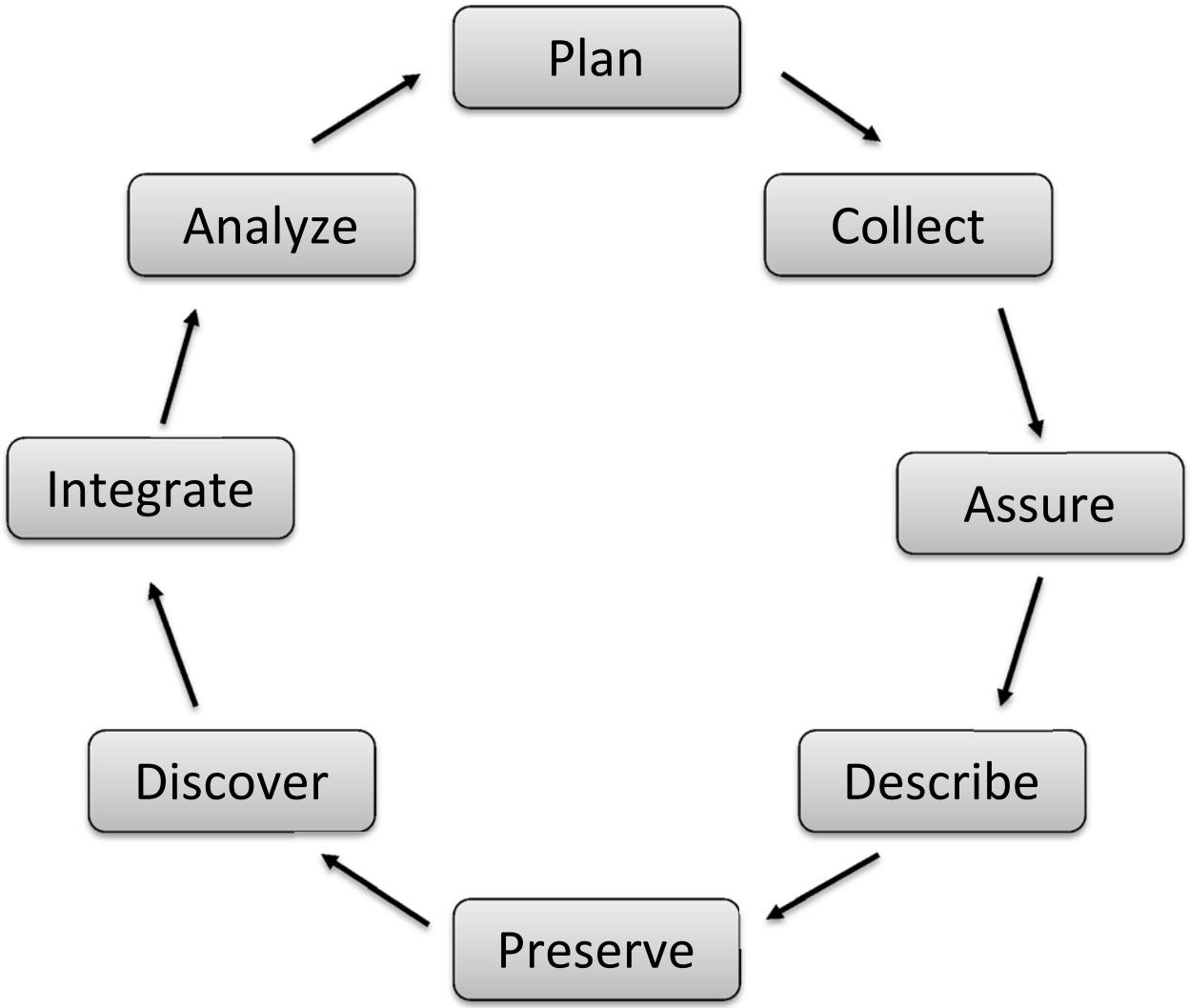
Audrey Mickle
Data/Systems Librarian
MBLWHOI Library

Research Data Management (RDM)

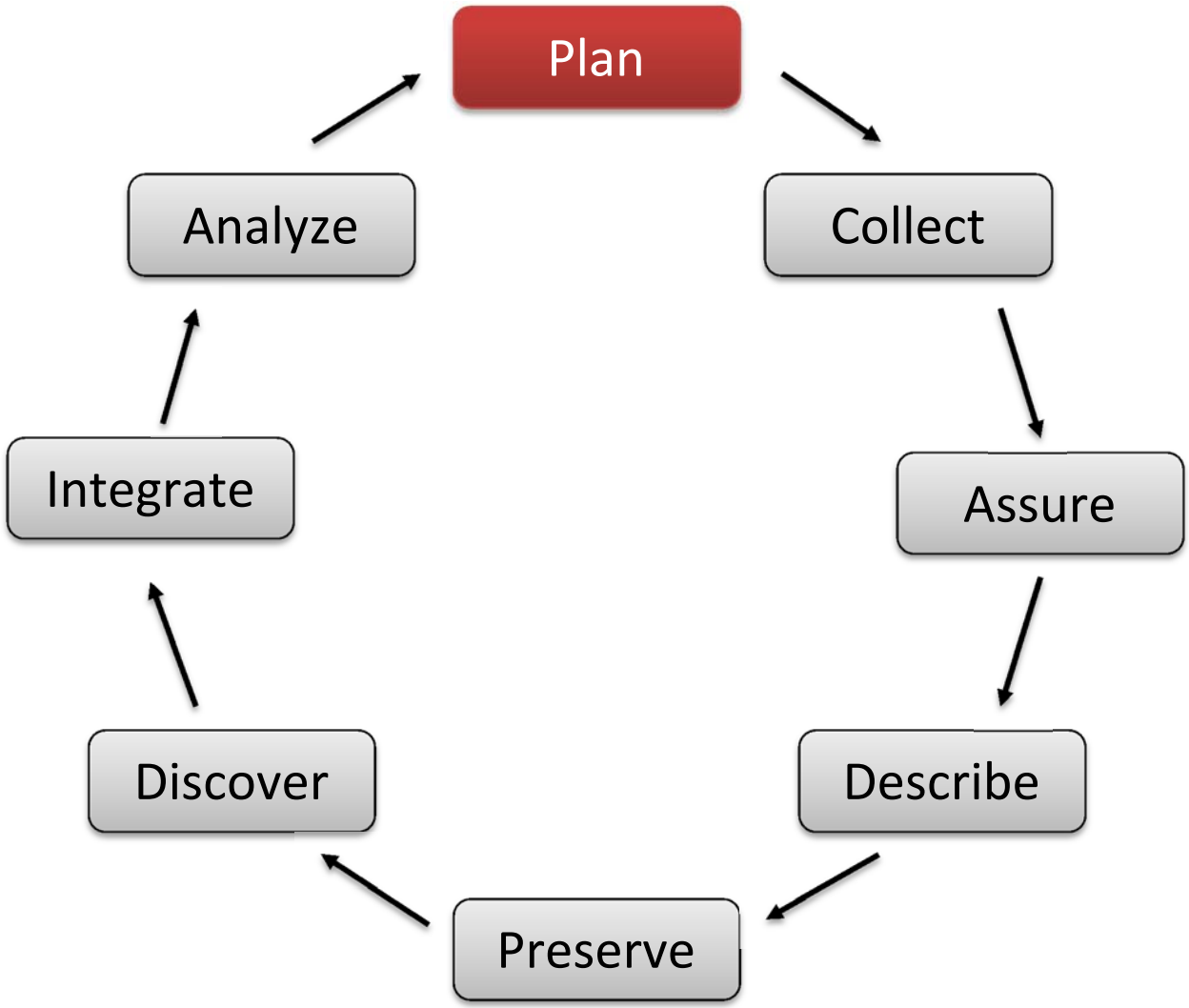
"Research data management concerns the organisation of data, from its entry to the research cycle through to the dissemination and archiving of valuable results. It aims to ensure reliable verification of results, and permits new and innovative research built on existing information."

Whyte, A., Tedds, J. (2011). 'Making the Case for Research Data Management'. DCC Briefing Papers. Edinburgh: Digital Curation Centre.

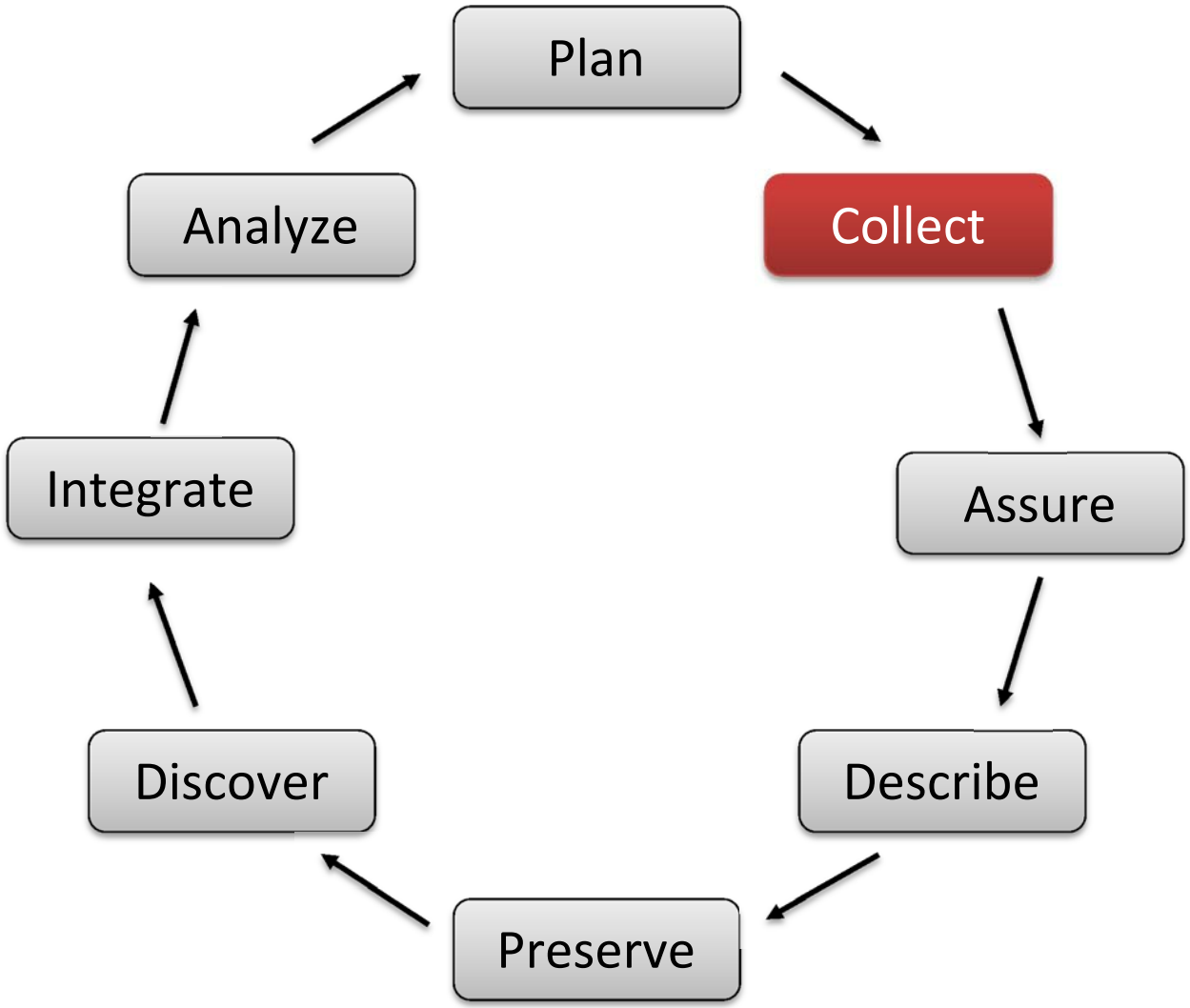
Research Cycle



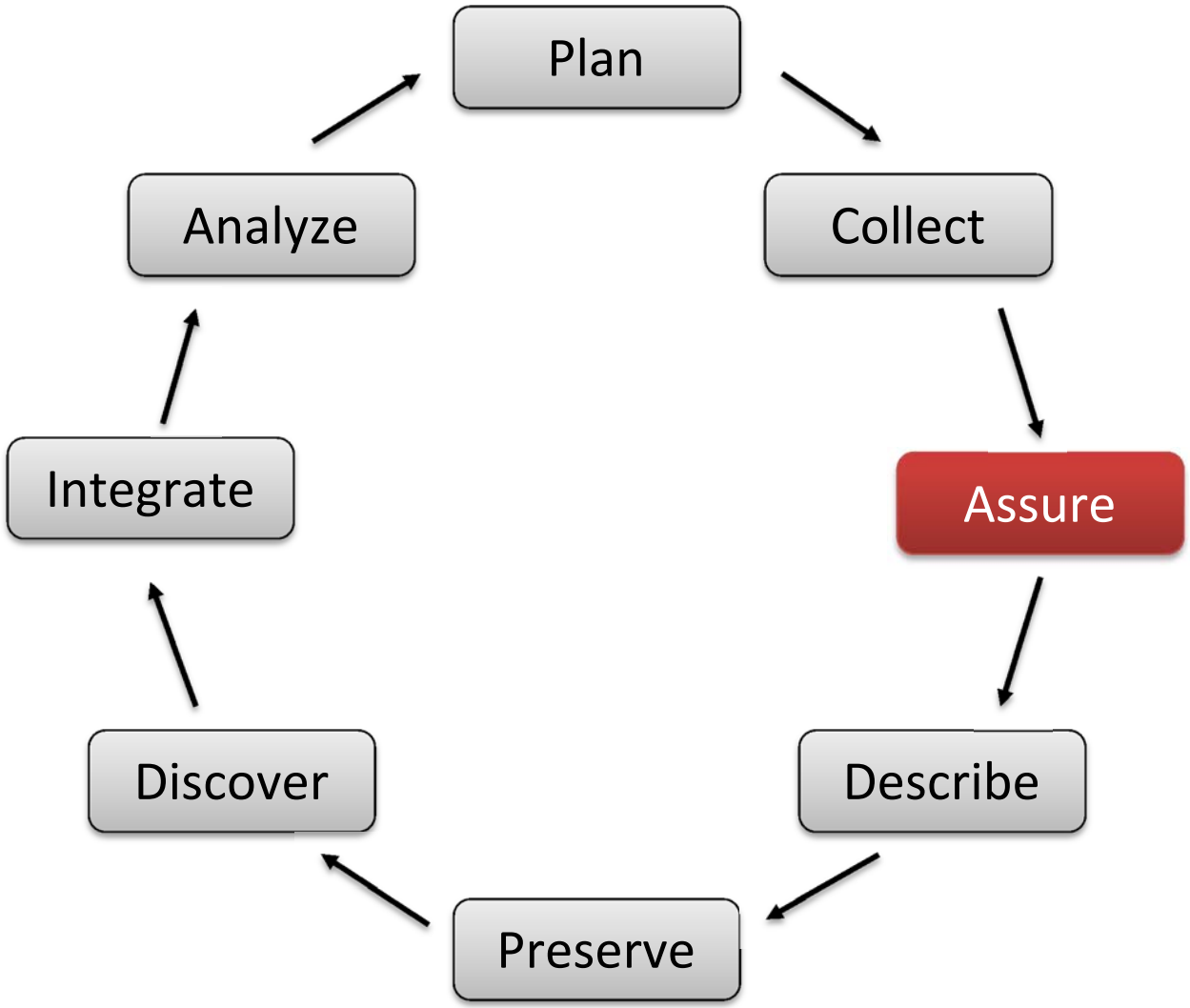
Research Cycle



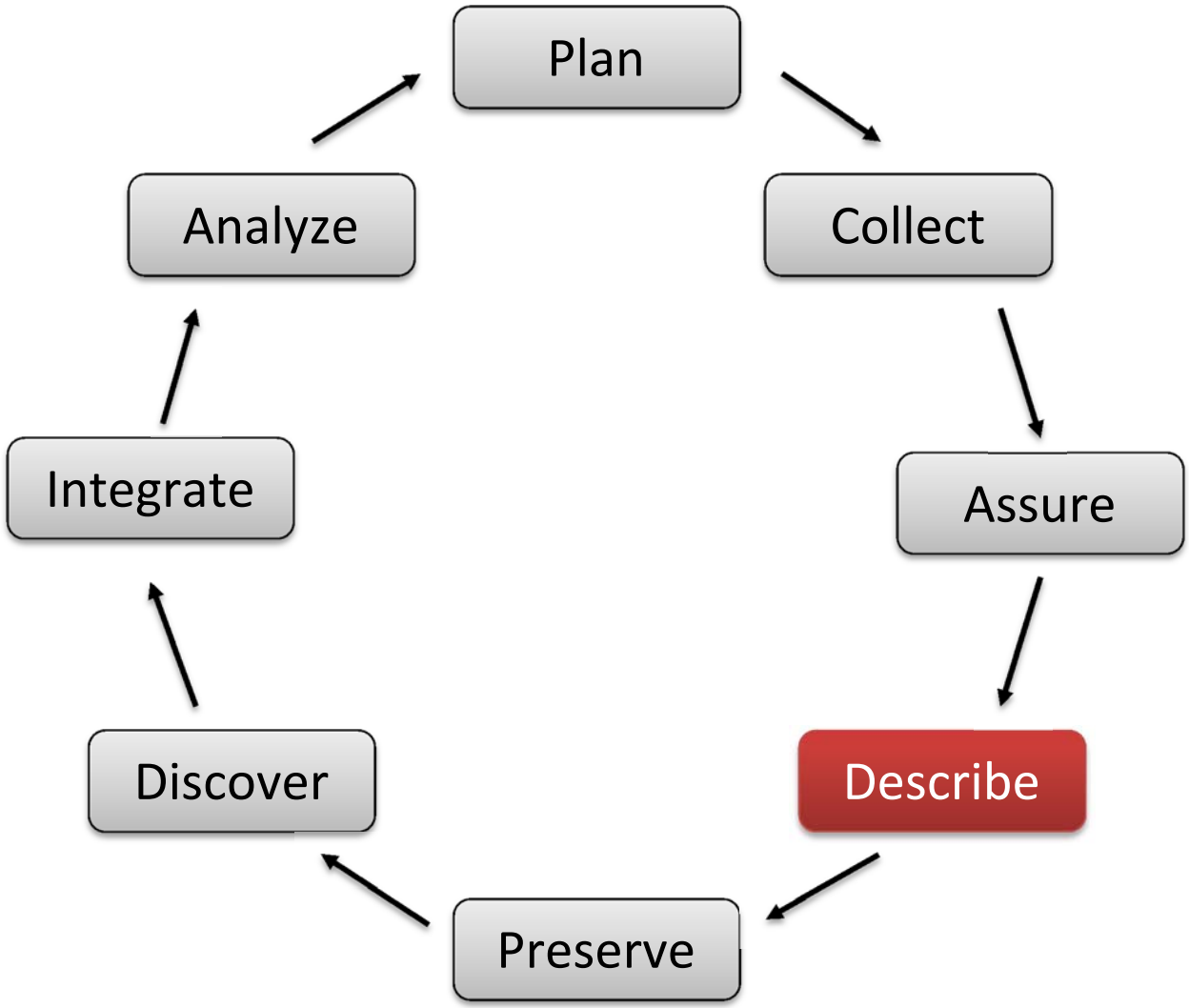
Research Cycle



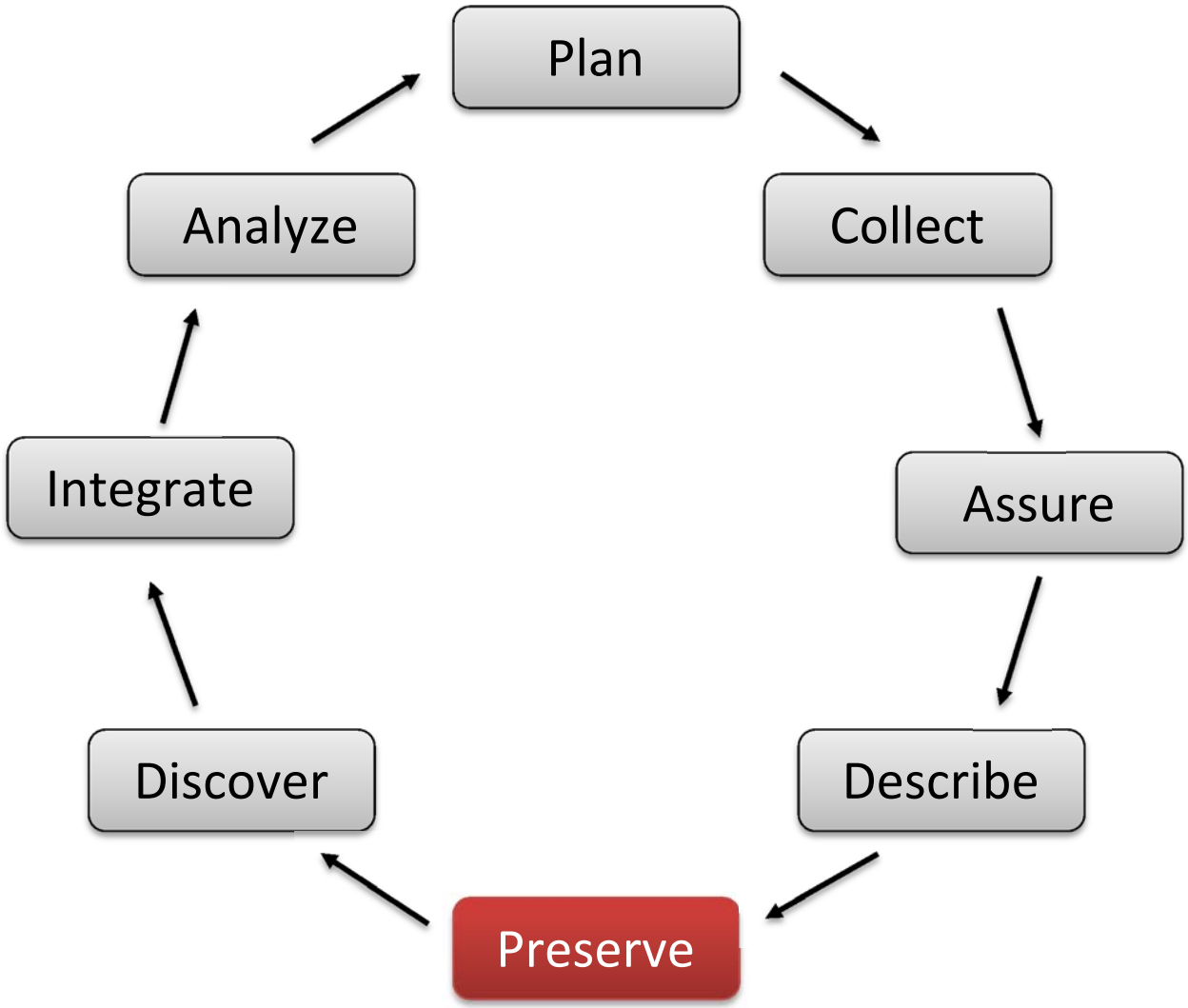
Research Cycle



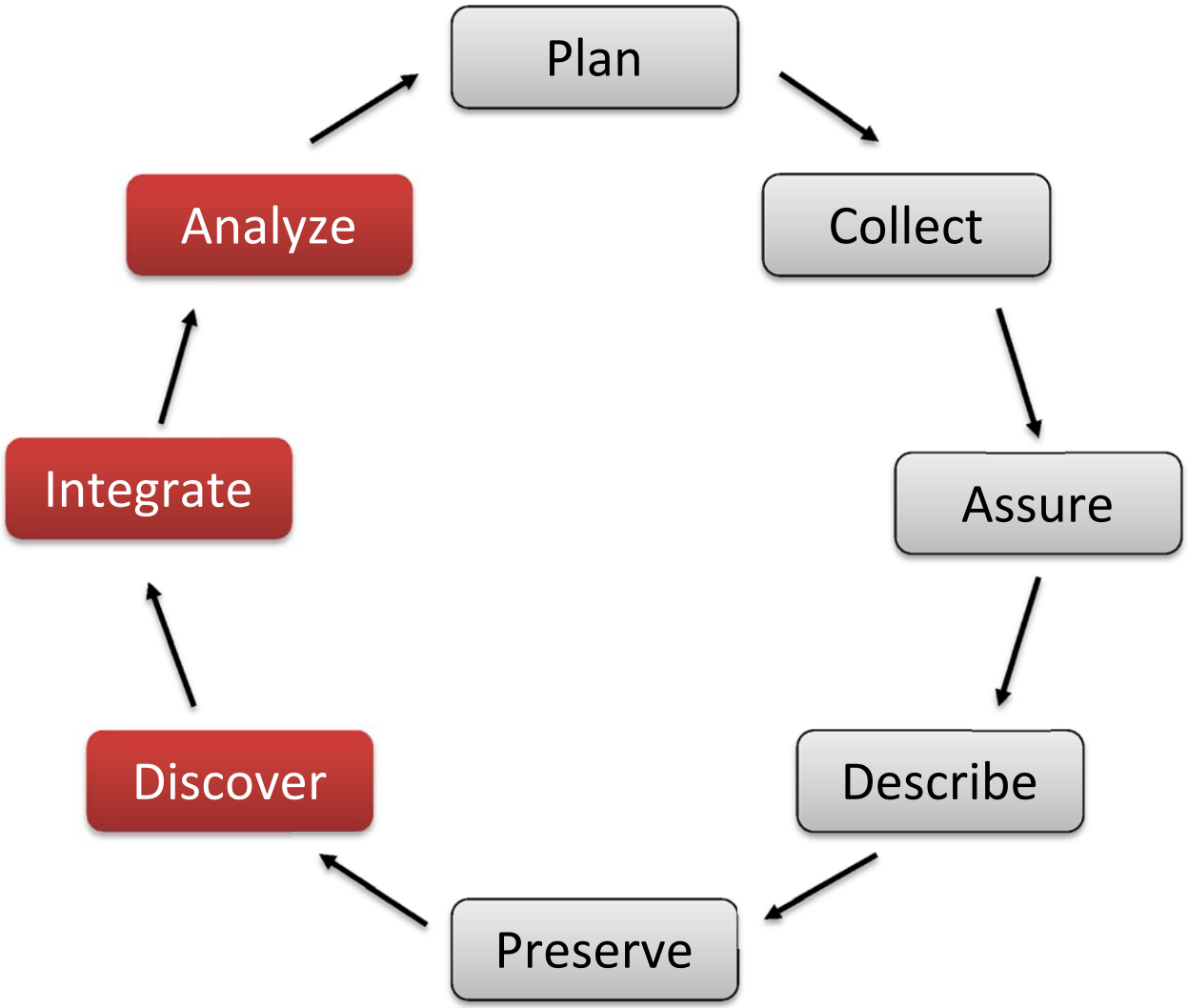
Research Cycle



Research Cycle

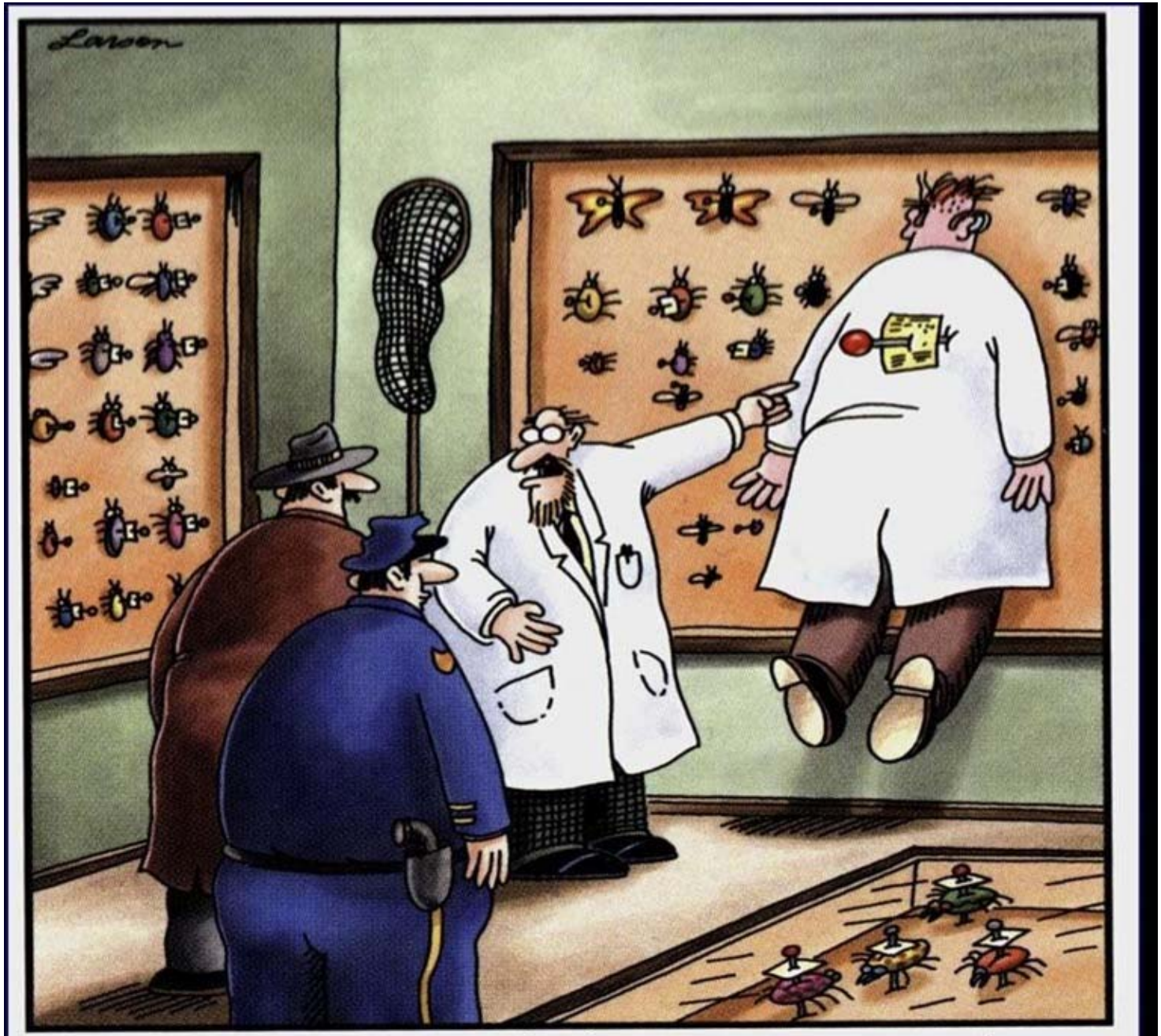


Research Cycle



What is Metadata

- Collection Metadata
- Processing Metadata
- Publication Metadata



"Professor LaVonne had many enemies in the entomological world, detective, but if you examine that data label, you'll find exactly when and where he was—shall we say—'collected.'"

What is Metadata

- Collection Metadata
  - Date, method, and location of data collection
  - Type and unit of measurement being taken
  - Flags, notes, and errors in data collection
- Processing Metadata
- Publication Metadata

What is Metadata

- Collection Metadata
- Processing Metadata
  - Key to coding system
  - Author, creation date, revision dates of code
  - How has data been processed, using what methods
- Publication Metadata

What is Metadata

- Collection Metadata
- Processing Metadata
- Publication Metadata
  - Copyright information for data or paper
  - DOI of Paper
  - DOI of Data

Why RDM?

Why **Good** RDM?

- Science is *fast*
- Science is *collaborative*
- Science is *valuable*

RDM Responsibilities

"Roles and responsibilities should be clearly defined, rather than assumed; this is especially important for collaborative projects that involve many researchers, institutions, and/or groups."

– Data Observation Network for Earth (DataONE)

Hard Decisions about Storage

- Local, networked, or cloud storage
  - Edit locally, share locally - Local files or Shared Drive
  - Edit online - Google Docs
  - Edit locally, share online - Amazon, Dropbox
- Back it up!
  - Local Backup
  - Offsite Backup
  - Daily, Weekly, Monthly Backup

RDM Makes Good Science

Communication is *hard:* Create a data dictionary, keep all reference docs in one place, inform all participants if something has changed

Data_40HZ/Waveform/Characteristics

Label	Datatype (Dimensions)	long_name (standard_name)	units	description	source	coordinates
i_waveformType	INTEGER (UNLIMITED)	i_waveformType (NOT_SET)	NOT_SET	Indicates number of valid samples in waveform; 0 = missing;1 = Long waveform (544 samples); 2 =Short waveform (200 samples) flag values: 0, 1, 2 flag_meanings: missing long short	Rel 33 GLAS Binary Data	DS_UTCTime_40
i_LastThrXingT	INTEGER (UNLIMITED)	Last Threshold Crossing Location for Selected Filter (NOT_SET)	ns	Address, in digitizer counts, of the detected last (i.e. last in time) threshold crossing (as measured from the start of Acquisition Memory, i.e. Start of digitization). Also called the trailing edge. Set to 0 if threshold crossing was NOT detected. From APID12/13, Offset 84.	Rel 33 GLAS Binary Data	DS_UTCTime_40
i_NextThrXing	INTEGER	Next to Last Threshold Crossing Location for Selected Filter	ns	Address (in dititizer counts) of the detected next to last threshold crossing (as measured from the start of Acquisition Memory, i.e. Start of digitization. Also called the leading edge. Set to 0 if a threshold crossing was NOT detected. From APID12/13 offset 88.	Rel 33 GLAS Binary Data	DS_UTCTime_40

FieldFlowers2001 ← Was this the latest draft? I had two…

low	1	7	3	177	3 C	1	45.7
low	1 14c		4	196	3 D	5	62.5
low	1 1d		3	175	3 D	2	51.8
high	1 209a		1	201	2 C	4	30.9
high	1 209d		4	165	1 B	5	40.3
low	1 36b		3	182	2 A	3	45.2
low	1 36d		3	138	1 D	3	47.5
low	1 47d		2	154	1 C	1	84
high	1 50a		2	190	3 A	1	74.5
high	1 50c		4	192	3 A	5	41.3
control	1 A16.1		3	206	2 C	6	49
high	2 A196b		2	204	1 D	5	43.8
high	2 A196c		2	169	2 D	4	34.9
high	2 A244		1	173	1 B	2	32.2
control	1 A245		2	152	3 C	6	25.4
control	1 A254.2		4	187	3 A	3	43.4
control	2 A266		3	198	2 A	1	50.3
control	2 A286		2	202	3 B	5	45.3
control	2 A292		1	188	3 C	4	44
control	2 A31		3	158	2 B	1	50.1
low	2 A313		4	199	1 C	6	52.8
low	2 A327b		2	137	3 D		
low	2 A342.2		1	189	1 D		
high	2 A343		2	184	3 C		
control	2 A364		1	186	2 B		
control	1 B154.2		2	181	2 A		
low	2 B189.1a		2	143	1 D		
low	2 B189.1d		2	141	2 A		

No headers!

I really should remember this…

Date2	Num_p	Num_p2	TB	SF_2	LF_2	PI_2
23-Jun	8.48	6.8	11.63	9.1	11.35	14.61
23-Jun	8.41	6.84	9.71	7.14	9.21	14.84
23-Jun	7.1	7.36	9.99	6.57	9.3	11.12
22-Jun	8.44	7.14	9.89	9.08	11.58	14.86
25-Jun	10.65	6.76	10.67	9.09	11.5	14.45
22-Jun	10.34	9.17	11.49	9.09	10.93	17.99
25-Jun	9.57	7.05	10.99	6.79	9.21	16.09
25-Jun	8.49	7.78	10.08	7.46	9.51	13.06
23-Jun	9.1	6.36	9.75	7.51	9.66	10.87

RDM Makes Good Science

Communication is *hard:* Create a data dictionary, keep all reference docs in one place, inform all participants if something has changed

Standardize: Set standard values and units, make sure this is noted in the documentation

The Joys of Collaboration

Birth Group	First Born	Mom Id	Sex	Entry Date
KK	U	?	M	12-Dec-01
KK	U	OL	M	15-Jan-12
null	U	null	F	15-May-12
KK	N	FLO	M	17-May-12
KK	U	null	M	24-May-12
Unknown	No	OL	Female	7-Jun-12
KK	U	null	M	13-Jun-12
NULL	Unknown	NULL	Male	15-Jun-12
			M	15-Jun-12
KK	U	null	M	15-Jun-12

The Joys of Collaboration

Birth Group	First Born	Mom Id	Sex	Entry Date
KK	U	?	M	12-Dec-01
KK	U	OL	M	15-Jan-12
null	U	null	F	15-May-12
KK	N	FLO	M	17-May-12
KK	U	null	M	24-May-12
Unknown	No	OL	Female	7-Jun-12
KK	U	null	M	13-Jun-12
NULL	Unknown	NULL	Male	15-Jun-12
			M	15-Jun-12
KK	U	null	M	15-Jun-12

These researchers need help with empty fields

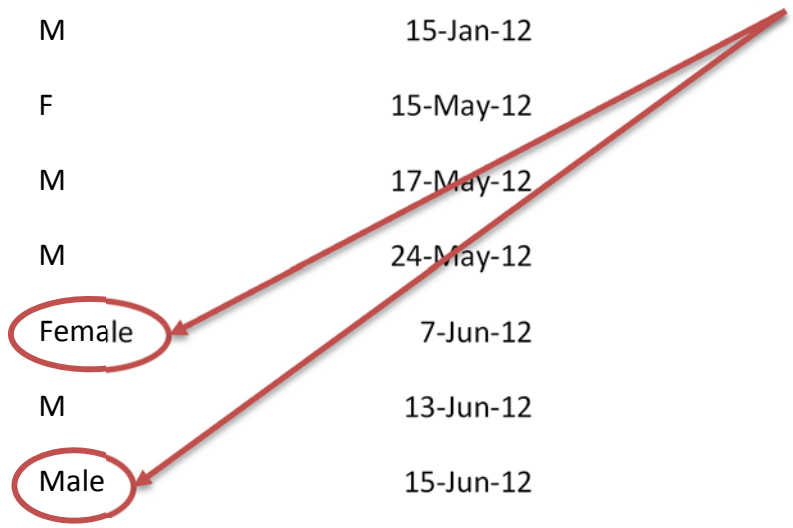
The Joys of Collaboration

Birth Group	First Born	Mom Id	Sex	Entry Date
KK	U	?	M	12-Dec-01
KK	U	OL	M	15-Jan-12
null	U	null	F	15-May-12
KK	N	FLO	M	17-May-12
KK	U	null	M	24-May-12
Unknown	No	OL	Female	7-Jun-12
KK	U	null	M	13-Jun-12
NULL	Unknown	NULL	Male	15-Jun-12
			M	15-Jun-12
KK	U	null	M	15-Jun-12

This researcher has formatted the date incorrectly, Jan 12th?

The Joys of Collaboration

Birth Group	First Born	Mom Id	Sex	Entry Date
KK	U	?	M	12-Dec-01
KK	U	OL	M	15-Jan-12
null	U	null	F	15-May-12
KK	N	FLO	M	17-May-12
KK	U	null	M	24-May-12
Unknown	No	OL	Female	7-Jun-12
KK	U	null	M	13-Jun-12
NULL	Unknown	NULL	Male	15-Jun-12
			M	15-Jun-12
KK	U	null	M	15-Jun-12



Without guidance researchers will make individual decisions

RDM Makes Good Science

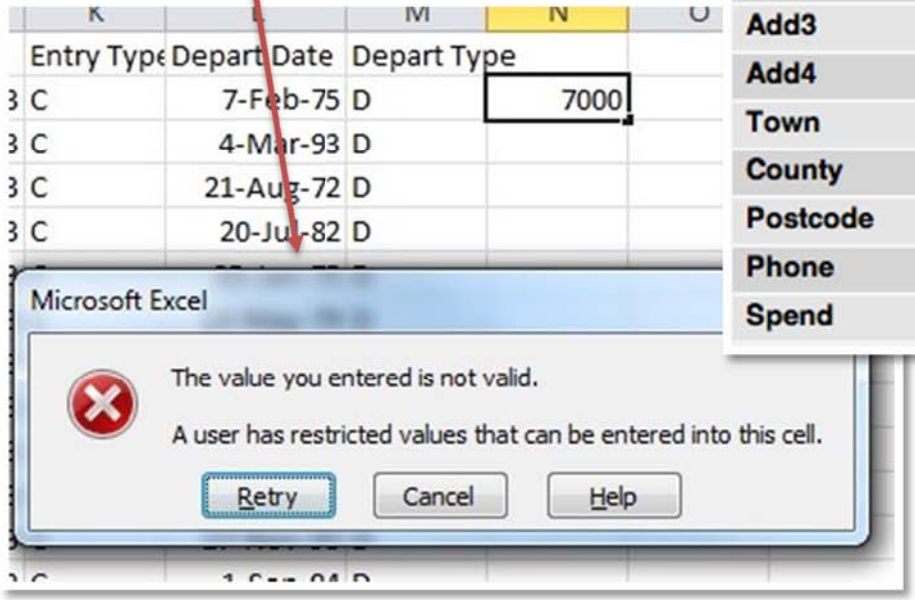
Communication is *hard:* Create a data dictionary, keep all reference docs in one place, inform all participants if something has changed

Standardize: Set standard values and units, make sure this is noted in the documentation

Enforce: Limit allowed values in your spreadsheet or database, require proper date format or correct spelling

Values from list in Excel

Range restriction in Excel



Field	Type	Collation	Attributes	Null	Default
id	int(11)			No	*None*
mailinglistID	int(11)			No	*None*
Prefix	varchar(100)	latin1_swedish_ci		Yes	*NULL*
Forename	varchar(50)	latin1_swedish_ci		Yes	*NULL*
Surname	varchar(50)	latin1_swedish_ci		Yes	*NULL*
Email	varchar(50)	latin1_swedish_ci		Yes	*NULL*
Company_Name	varchar(50)	latin1_swedish_ci		Yes	*NULL*
Jobrole	varchar(100)	latin1_swedish_ci		Yes	*NULL*
Add1	varchar(200)	latin1_swedish_ci		Yes	*NULL*
Add2	varchar(200)	latin1_swedish_ci		Yes	*NULL*
Add3	varchar(200)	latin1_swedish_ci		Yes	*NULL*
Add4	varchar(200)	latin1_swedish_ci		Yes	*NULL*
Town	varchar(200)	latin1_swedish_ci		Yes	*NULL*
County	varchar(200)	latin1_swedish_ci		Yes	*NULL*
Postcode	varchar(200)	latin1_swedish_ci		Yes	*NULL*
Phone	varchar(18)	latin1_swedish_ci		Yes	*NULL*
Spend	int(11)			No	0

Database table attributes

Data Collection: Best Practices

Do math later: Reduce the data to separate parts

Don’t crowd the data: Make sure everything has its own field

Plan for problems: Make sure there is a free-text field, or document

Account for problems: Make sure your researchers know what to do

File Names: Best Practices

University of Washington: Data Management Guide

- Use names that are **brief but descriptive**
- Avoid spaces and special characters (like *, #, % etc.)
- Come up with a **naming convention** adhered to by everyone using the files
- Identify versions of files using **dates and version numbering** in file name
- Use **three letter file extensions** to ensure backwards compatibility (ex: .doc, .tif, .txt)
- Do **not** use letter case to identify different files (ex. datasetA.txt vs. dataseta.txt)

File Structure: Best Practices

University of Washington: Data Management Guide

- Folder structure for your files can assist in the unique identification of the files contained within them. Consider the structure of the folders containing your data files before you begin to collect your data. Ideas for how to organize your folders include:
  - Data type (text, images, models, etc.)
  - Time (year, month, session, etc.)
  - Subject characteristic (species, age grouping, etc.)
  - Research activity (interview, survey, experiment, etc.)

Examples: Best Practices

University of Washington: Data Management Guide

Consider these examples of file naming and folder structure:

File001.txt vs.
201206blood_ID0234.txt

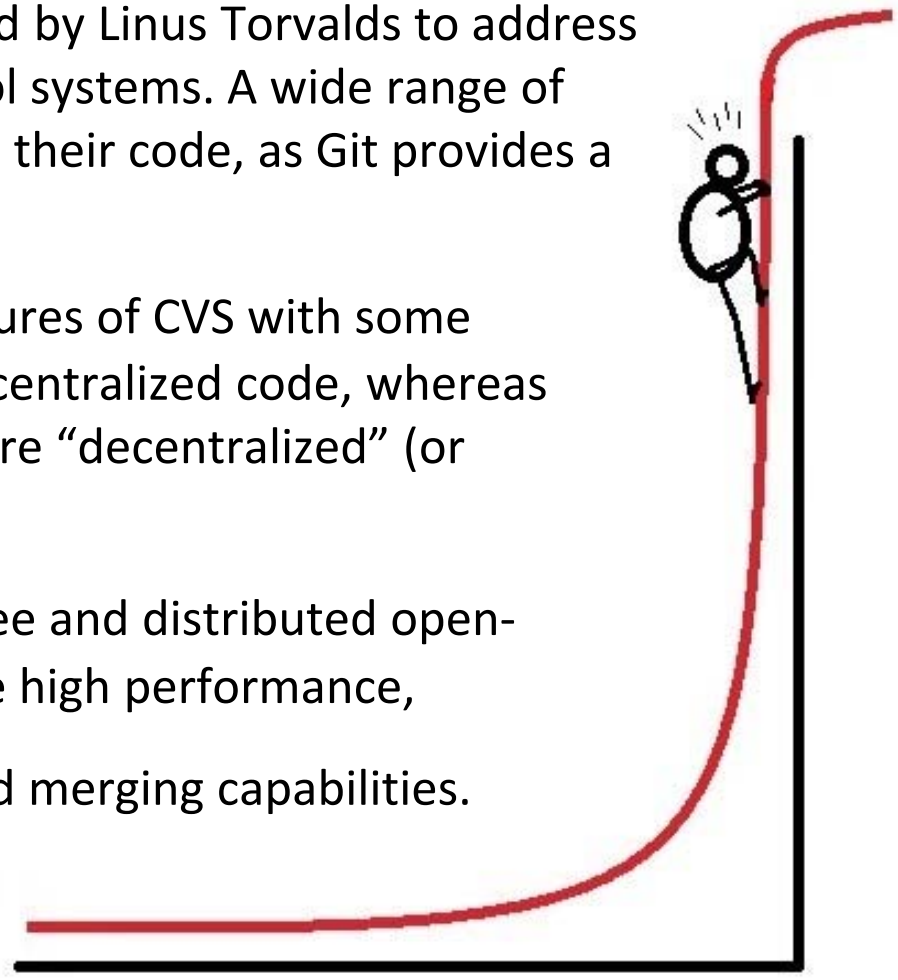
MyDocuments\Research\Sample12.jpg vs
C:\\NEHGrant01234\WWI\Images\London_001.jpg

Low Tech Revisions Options

- File naming system, text log
- Track changes: MS Products, Google Docs, OpenOffice
- Backups and system software

High Tech Revisions Options

- Git: A distributed version control system, is arguably the most popular version control system today. It was developed by Linus Torvalds to address the issue of speed with existing version control systems. A wide range of organizations worldwide prefer Git to manage their code, as Git provides a huge range of features.
- Subversion: Subversion offers the best features of CVS with some improvements. Subversion puts emphasis on centralized code, whereas other popular version control systems today are “decentralized” (or distributed).
- Mercurial: Mercurial, much like Git, is a free and distributed open-source system. Mercurials main objectives are high performance, scalability, along with advanced branching and merging capabilities.



Long Term RDM

Curating your data - Deciding what makes the cut

Repositories - Finding a forever home for your data

Encouraging Use - Ensuring your data lives a long and happy life

Resources

Strasser, Carly. (June 18, 2014). “The Research Data Life Cycle” [PowerPoint slides]. California Digital Library. Retrieved from http://www.slideshare.net/carlystrasser.

Strasser, Carly. (April 10, 2014). “Coping With Your Data: Tips and Tools” [PowerPoint slides]. California Digital Library. Retrieved from http://www.slideshare.net/carlystrasser.

“Data Management.” MIT Libraries. Retrieved from http://libraries.mit.edu/data-management/.

“Data Management.” NYU Health Sciences Library. Retrieved from http://hslguides.med.nyu.edu/data_management.

“Data Management Guide.” University of Washington Libraries. Retrieved from http://guides.lib.washington.edu/content.php?pid=259952&sid=2350038.

“GLAH01 Product Data Dictionary.” National Snow and Ice Data Center. Retrieved from http://nsidc.org/data/docs/daac/glas_altimetry/data-dictionary-glah01.html.

Resources

"Define roles and assign responsibilities for data management." DataOne. Retrieved from https://www.dataone.org/best-practices/define-roles-and-assign-responsibilities-data-management.

Michener, Bill. "Metadata." DataOne. Retrieved from https://www.dataone.org/sites/all/documents/ESA11_SS3_Metadata_WKM_Final.pdf.

Whyte, Angus & Tedds, Jonathan. (2011, September 1). "Making the Case for Research Data Management." The Digital Curation Centre. Retrieved from http://www.dcc.ac.uk/resources/briefing-papers/making-case-rdm.

"What is Research Data Management." The University of Leicester. Retrieved from http://www2.le.ac.uk/services/research-data/rdm/what-is-rdm.

Daityari, Shaumik. (2014, April 22). "Version Control Software in 2014: What are Your Options?" Retrieved from http://www.sitepoint.com/version-control-software-2014-what-options/.

"Versioning file system. " (n.d.). Retrieved September 23, 2014 from Wikipedia: http://en.wikipedia.org/wiki/Versioning_file_system.

"List of revision control software. " (n.d.). .). Retrieved September 23, 2014 from Wikipedia: http://en.wikipedia.org/wiki/List_of_revision_control_software.