Data Management

Courtesy of Calisphere: http://content.cdlib.org/ark:/13030/kt22901871/

SUSAN BORDA
DIGITAL CURATION LIBRARIAN

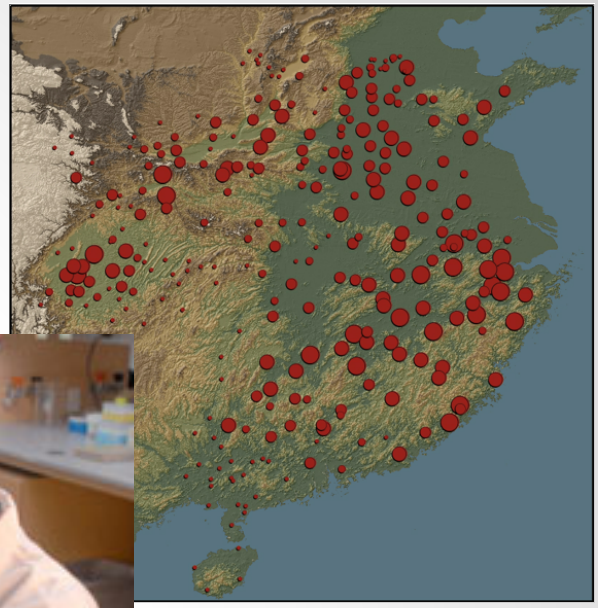
CC BY

Data

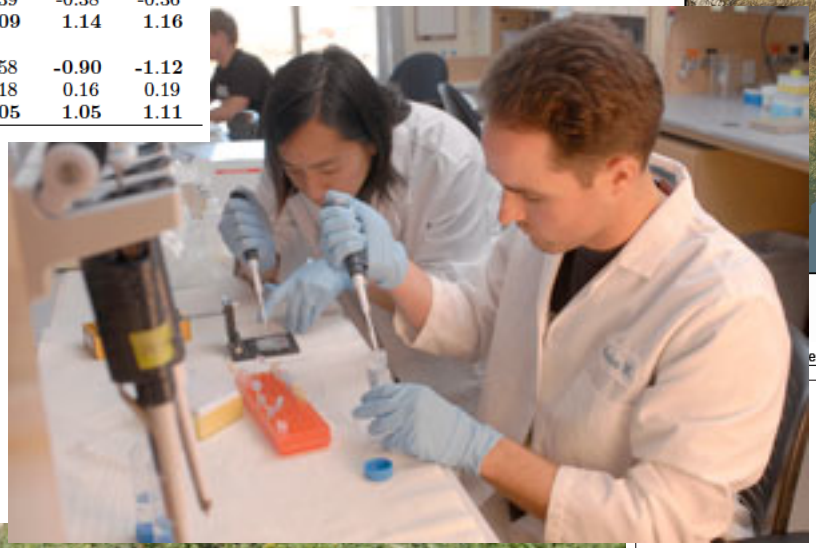
	33/33/33			70/20/10		
Parameter	10%	20%	50%	10%	20%	50%
C1 Proportion	(1.02)	(1.02)	(1.02)	(0.73)	(0.69)	(0.66)
C2 Proportion	(1.01)	(1.01)	(1.01)	(1.87)	(1.98)	(1.07)
C3 Proportion	(1.00)	(1.00)	(1.00)	(1.17)	(1.20)	(1.21)
Growth Parameters: C1						
Intercept	0.60	0.73	**0.90**	**-1.09**	-0.20	0.72
Slope	-0.27	-0.41	-0.70	**0.94**	0.41	-0.28
Quadratic	-0.17	-0.14	-0.17	-0.56	-0.44	-0.18
Growth Parameters: C2						
Intercept	-0.35	-0.46	**-0.82**	**-3.78**	**-4.13**	**-4.52**
Slope	0.31	0.44	0.74	**1.98**	**2.31**	**2.65**
Quadratic	-0.05	-0.10	-0.14	**1.03**	**0.98**	**0.81**
Growth Parameters: C3						
Intercept	-0.09	-0.13	-0.16	-0.69	**-0.82**	**-0.89**
Slope	0.06	0.09	0.06	-0.39	-0.38	-0.36
Quadratic	0.07	0.06	0.11	**1.09**	**1.14**	**1.16**
Variances: All Classes						
Intercept Variance	-0.72	-0.78	**-0.83**	-0.58	**-0.90**	**-1.12**
Slope Variance	0.04	0.09	0.04	0.18	0.16	0.19
Quadratic	0.64	0.64	0.63	**1.05**	**1.05**	**1.11**

http://faculty1.ucmerced.edu/sdepaoli/docs/Remaining%20Study%202%20tables%20for%20the%20web.pdf

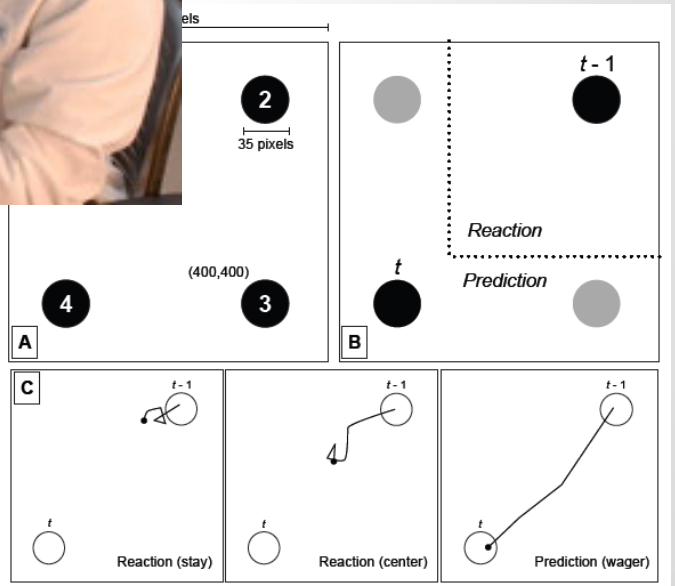
http://qsb.ucmerced.edu



http://songgis.ucmerced.edu/?page_id=4



https://eng.ucmerced.edu/people/rbales/mhrg



http://cognaction.org/rick/pdfs/papers/dale_duran_morehead_2012.pdf

WHY DATA MANAGEMENT?

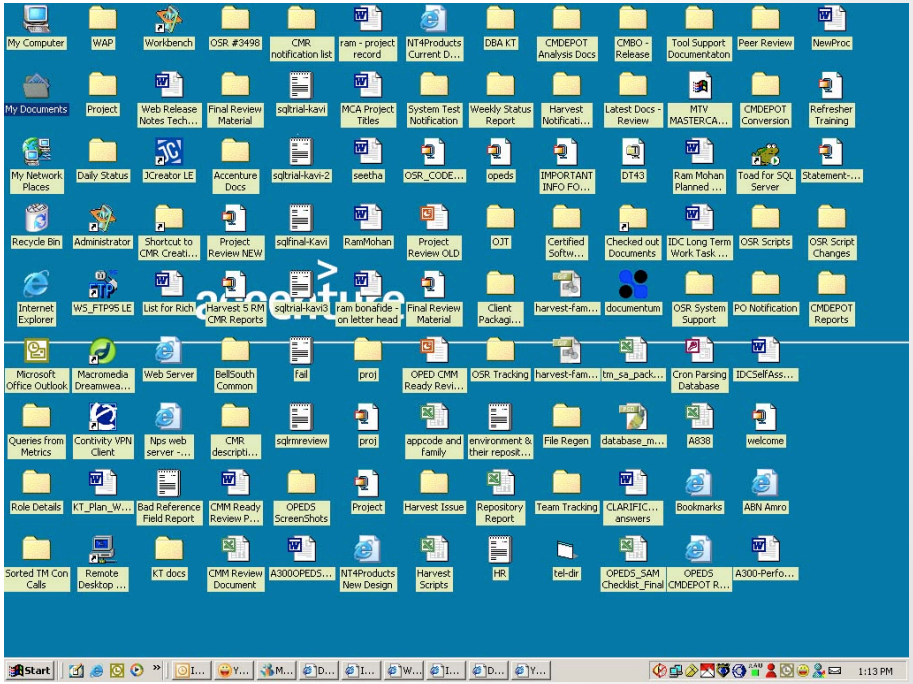
Funders often require it

Makes for easier transitions with lab staff turnover

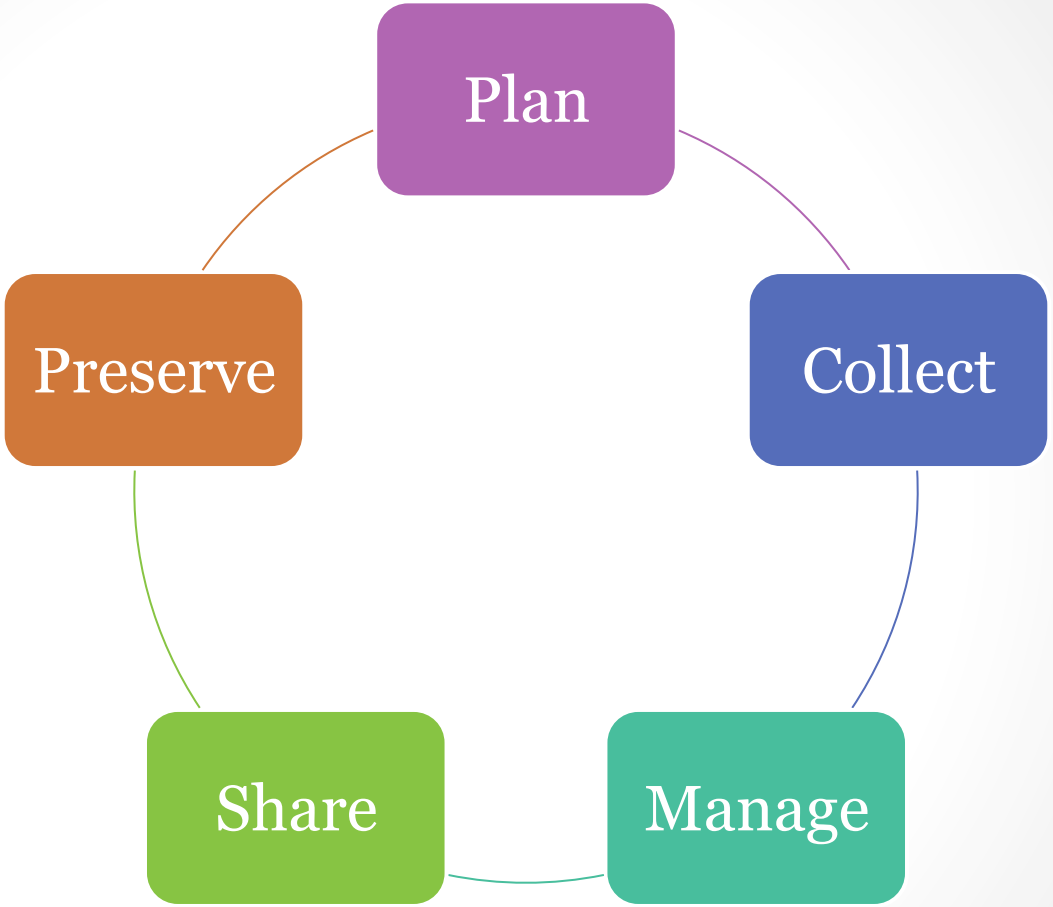
Helps better document research output

Ensures usefulness of data over its entire life cycle

Facilitates data sharing



THE RESEARCH DATA LIFE CYCLE



Plan

WHAT IS A DATA MANAGEMENT PLAN?

A document that describes what you will do with your data both

during your research

and after you complete your project

From Flickr by spanaut

Plan

THE DATA MANAGEMENT PLAN

NSF guidance: http://www.nsf.gov/bfa/dias/policy/dmpfaqs.jsp
NIH guidance: http://grants.nih.gov/grants/policy/data_sharing/data_sharing_faqs.htm

Plan

DMPS FOR FUNDERS

A short plan submitted alongside grant applications

An outline of

- what will be created/collected
- methods
- Standards
- Metadata
- sharing/access
- long-term storage

Includes how and why

NSF DMP REQUIREMENTS

From Grant Proposal Guidelines:

DMP supplement may include:

1. the **types of data**, samples, physical collections, software, curriculum materials, and other materials to be produced in the course of the project
2. the **standards to be used for data and metadata** format and content (where existing standards are absent or deemed inadequate, this should be documented along with any proposed solutions or remedies)
3. **policies for access and sharing** including provisions for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights or requirements
4. policies and **provisions for re-use**, re-distribution, and the production of derivatives
5. **plans for archiving** data, samples, and other research products, and for preservation of access to them

NATIONAL ENDOWMENT FOR THE Humanities

NEH DMP REQUIREMENTS

- **Should be short (no more than 2 pages)**
- **Will need to address 2 main topics:**
  - What data are generated by your research?
  - What is your plan for managing the data?
- **Specific components of the DMP:**
  - Expected data
  - Period of data retention
  - Data formats and dissemination
  - Data storage and preservation of access
- **Post-award monitoring is through:**
  - Interim reports, final performance reports, white papers

NIH DATA SHARING

- Grants of $500,000 or more in direct costs in any single year are expected to include a plan for data sharing or state why data sharing is not possible.

DOE DMP REQUIREMENT

- No more than two pages
- At a minimum, DMPs must describe:
  - how data sharing and preservation will enable validation of results
  - or how results could be validated if data are not shared or preserved.

Collect

BEST PRACTICES FOR DATA COLLECTION

Consider future preservation and context, not just what is needed right now

Use non-proprietary file formats

- Plain text/RTF vs. Microsoft .doc/.docx
- PDF/PNG vs. BMP/TIF
- CSV vs. Microsoft .xls/.xlsx

Collect good metadata

Storage and Backups

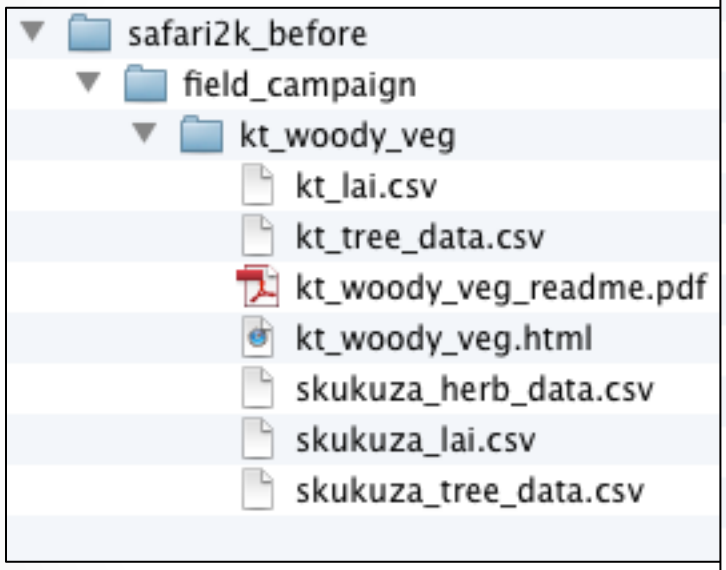
- SDSC Cloud - $390 TB/yr
- https://cloud.sdsc.edu/

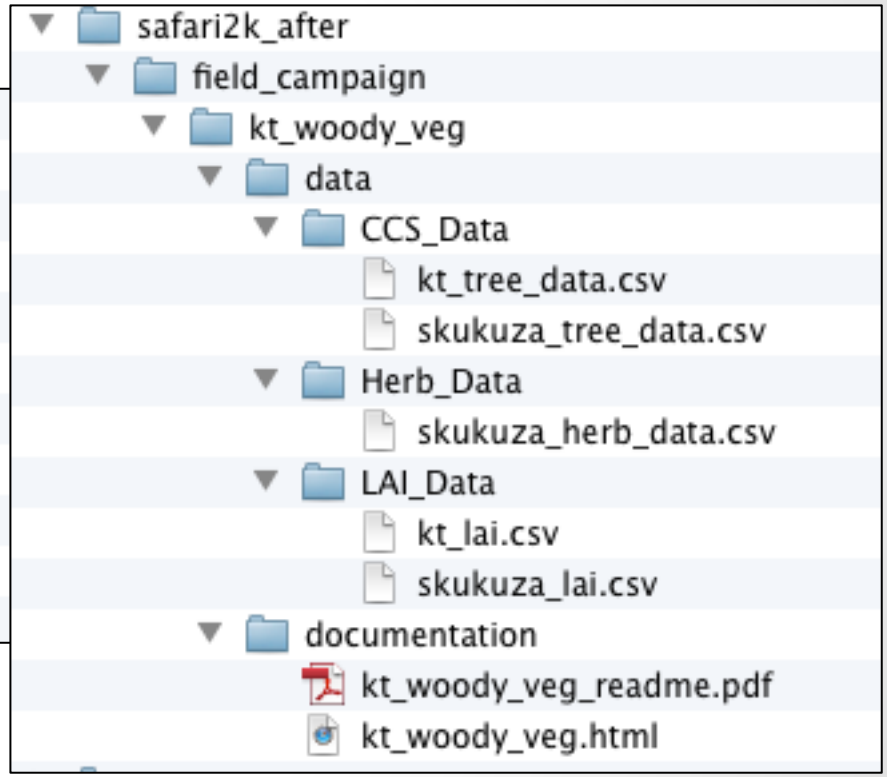
Additional info on open formats: http://www.openformats.org

Collect

BEST PRACTICES FOR DATA COLLECTION

File and Folder Organization





Collect

BEST PRACTICES FOR DATA COLLECTION (DON'T DO THIS!)

2 tables

Random notes

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
Washed Rocks

Don't use - old data

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		Shore	Avg Con
A5	ALG01		3.05	1.88	-24.34	-23.88	0.17	-1.65	-2.30	25362	c	-1.26	-27.22
A6	Lk Outlet Alg		3.06	31.55	-30.17	-29.71	0.92	0.87	0.22	25364		1.26	0.32
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						

Collect

DO THIS!

ITEM	Description	VALUE_UNIT
System	Costech Elemental Combustion System; using an open sample carousel and tin capsules	text
Date of Creation	7/26/12	mm/dd/yyyy
Anal.	analysis	NC elemental analysis by combustion
File	file name from ECS software	text
Signal	signal used by ECS	text
Sample	sample identification	internal lab sample ID, unique by date; ex. "a05-P304-A" a - archive, d - sediment, l - soil; watershed or profile location; A/B/C is the repetition ID
Sample Weight	Weight of sample in tin	mg
N RetenTime	Reten. Time	minutes
N Response	Measured area of response peak	number
N wt mg	Weight	mg
N wt pct	Weight	%
N peak	Peak	Type
Nitrogen	Element	Name
NtoC response ratio	Carbon Response Ratio	ratio
C RetenTime	Reten. Time	minutes
C Response	Measured area of response peak	number
C wt mg	Weight	mg
C wt pct	Weight	%
C peak	Peak	Type
Carbon	Element	Name
C Response Ratio	Carbon Response Ratio	ratio
Notes	Missed 1 sample in the data table but sample still ran - messed up the calculations for all later samples. Recalculated values using weights and responses in the standard calibration equations from that day. There was only enough sample for 1 run on E16-1,2,3 OP(w)53; lost it on attempt run on 20120807	

MetaData | 20120726 | 20120726INITIAL | ECS_20120726_calcurve | +

Data courtesy of Erin Stacy

Manage

DATA DOCUMENTATION

Document, Document, Document

- Data collection methods
- Context of data collection
- Variable names and descriptions
- Algorithms used to transform data from raw to analyzed

Many fields have existing metadata standards

- EML: Ecological Metadata Language
- Darwin Core: biology and biodiversity

Additional info: http://www.dcc.ac.uk/resources/briefing-papers/standards-watch-papers/what-are-metadata-standards

Share

BEST PRACTICES FOR DATA SHARING

Different communities of practice have different cultures of sharing

- Institutional/subject repositories
- Person-to-person sharing of data
- Publication in data journals

Data sharing is associated with increased citations[1]

1. Piwowar HA, Day RS, Fridsma DB (2007) Sharing Detailed Research Data Is Associated with Increased Citation Rate. PLoS ONE 2(3): e308. doi:10.1371/journal.pone.0000308

Share

UC TOOLS FOR SHARING

creates unique identifiers

http://www.cdlib.org/services/uc3/ezid/

cost-effective repository

for sharing and storage

//www.cdlib.org/services/uc3/merritt/

Preserve

PLANNING FOR LONG TERM PRESERVATION

Especially important for unique and non-replicable datasets

Use institutional and subject specialized repositories to preserve data for the long-term

LOCKSS: Lots of Copies Keep Stuff Safe

Talk to the library about your data

Institutional repositories at the UC:

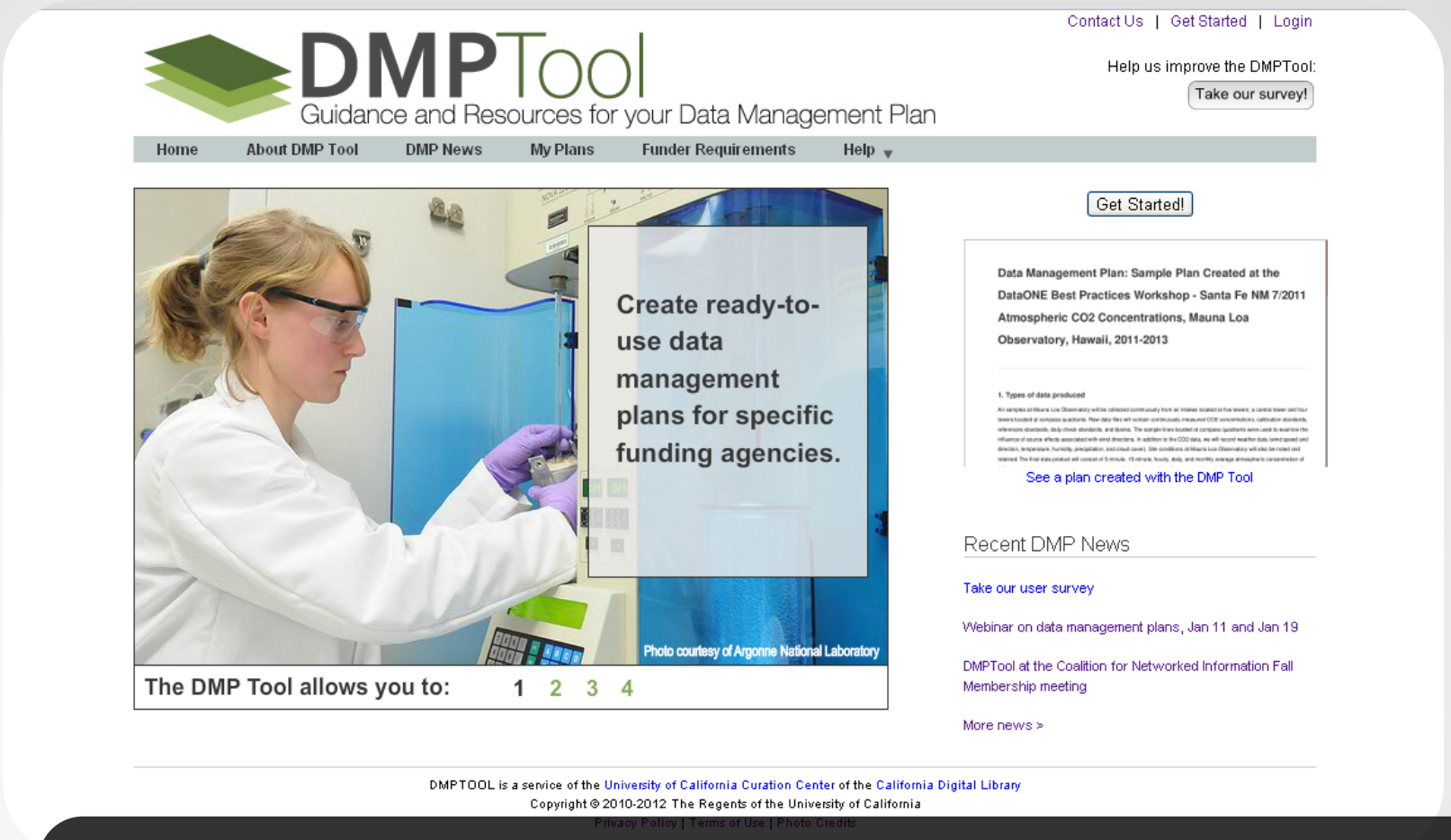
- Merritt https://merritt.cdlib.org/
- eScholarship http://www.escholarship.org/uc/ucla

Other repositories:

- On our Data Management LibGuide: http://libguides.ucmercedlibrary.info/content.php?pid=266626&sid=2200811

dmptool.org

- **Free**
- **Guides through creating a DMP**
- **Helps meet funder requirements**
  - Supplies questions
  - Includes explanation/context provided by the agency
  - Provides links to the agency website



Step-by-step wizard for generating DMP

Create | edit | re-use | share | save | generate

QUESTIONS?

Susan Borda: sborda@ucmerced.edu

http://libguides.ucmercedlibrary.info/data-management

dmptool.org

Special thanks to: Carly Strasser (CDL) and Lisa Federer (UCLA) for the use of their slides.