

# *Services for Big Data... and the Long Tail*



From Flickr by  
rahen z

Patricia Cruse  
Stephen Abrams  
Carly Strasser  
Perry Willett

*University of California Curation Center  
California Digital Library  
<http://www.cdlib.org/uc3>*

# Big Data



OSTP March 2012  
Big Data Effort Launched

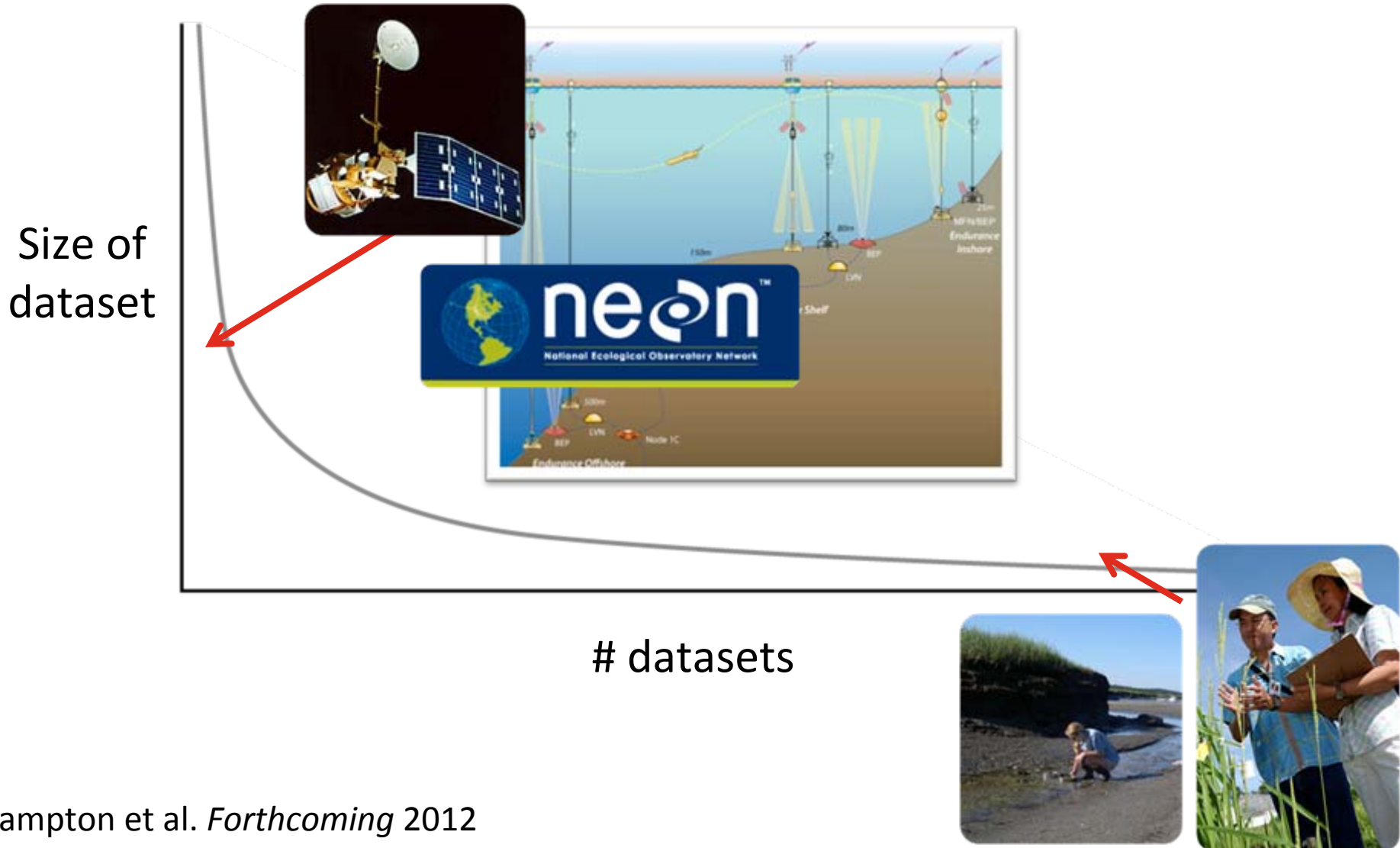


# *The Little Guys?*



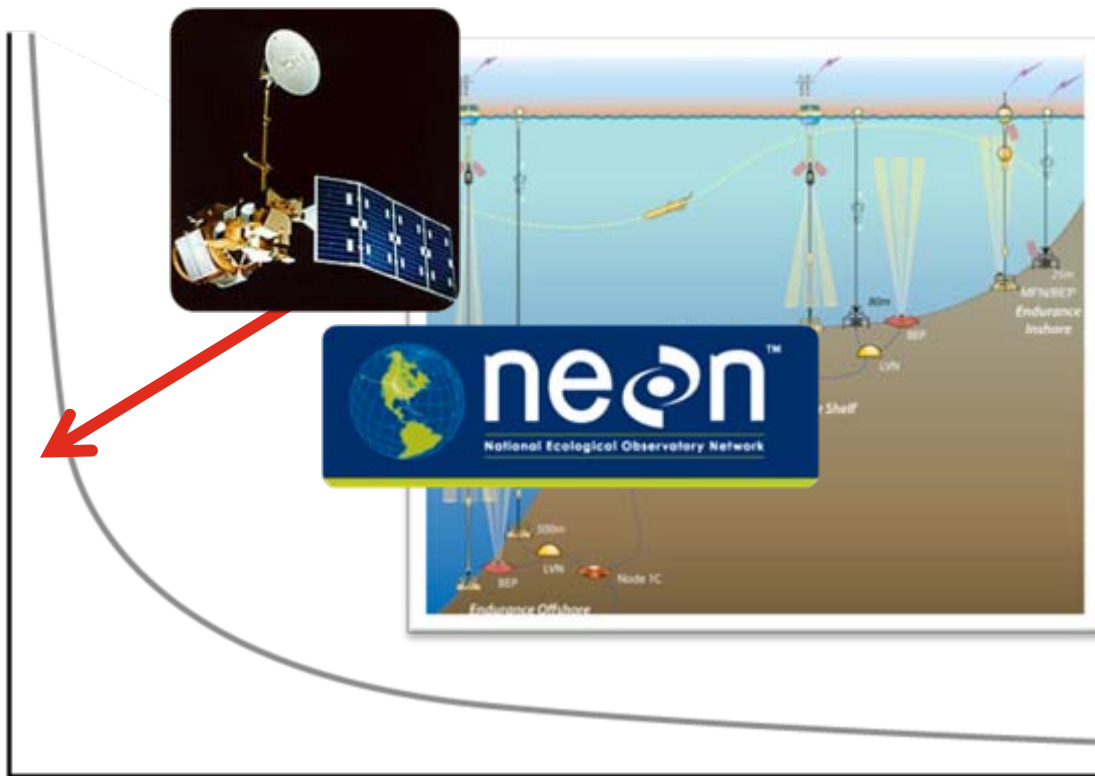
*From Flickr by jason tinder*

# The Long Tail



# The Long Tail

Size of dataset



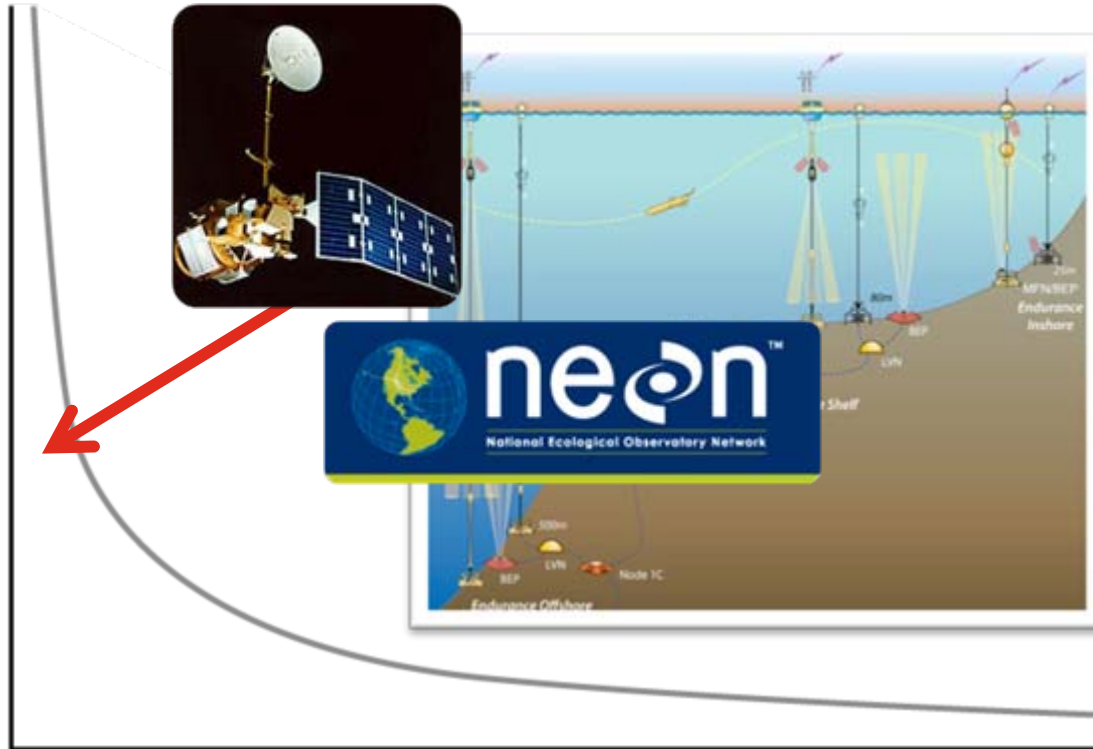
# datasets

# researchers



# The Long Tail

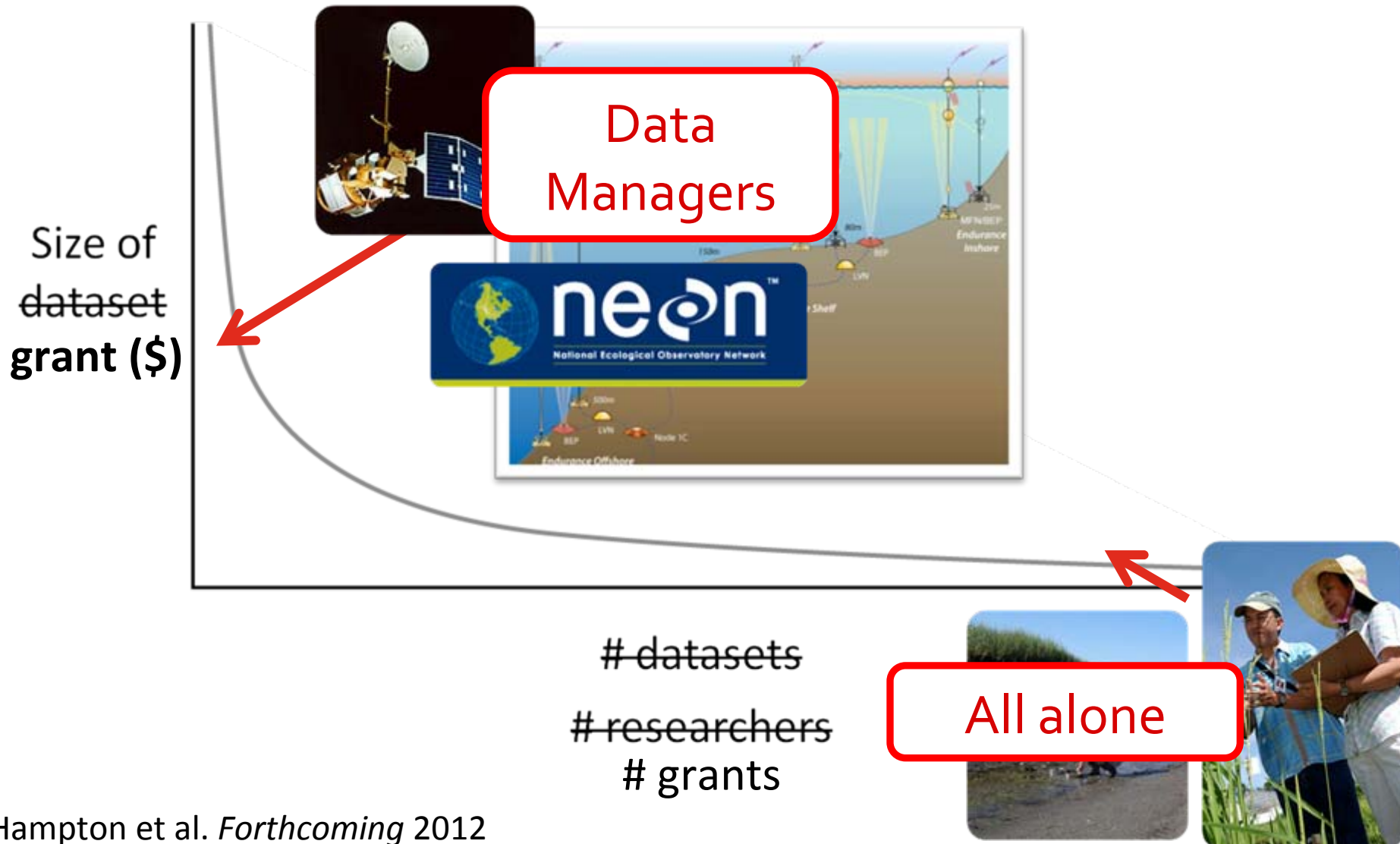
Size of dataset



# datasets  
# researchers  
# grants



# The Long Tail





# UGLY TRUTH

Many (most?)  
researchers...

are not taught data management

don't know what metadata are

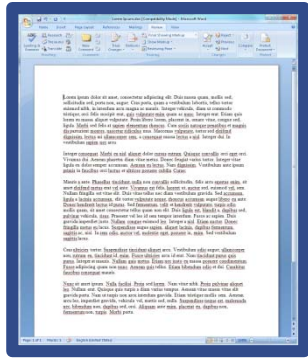
can't name data centers or repositories

don't share data publicly or store it in an archive

aren't convinced they should share data



# The Fallout: Where data end up



From Flickr by diyibrarian



blog.order2disorder.com



From Flickr by cseussms

**Data**  
**Metadata**



From Flickr by cseussms

# Barriers

- Cost: time, personnel, software & hardware
- Confusion about standards, how to start
- Disparate datasets
- Lack of training
- Fear of lost rights/benefits
- No incentives



From Flickr by Christina Ann VanMeter



From Flickr by bthomso

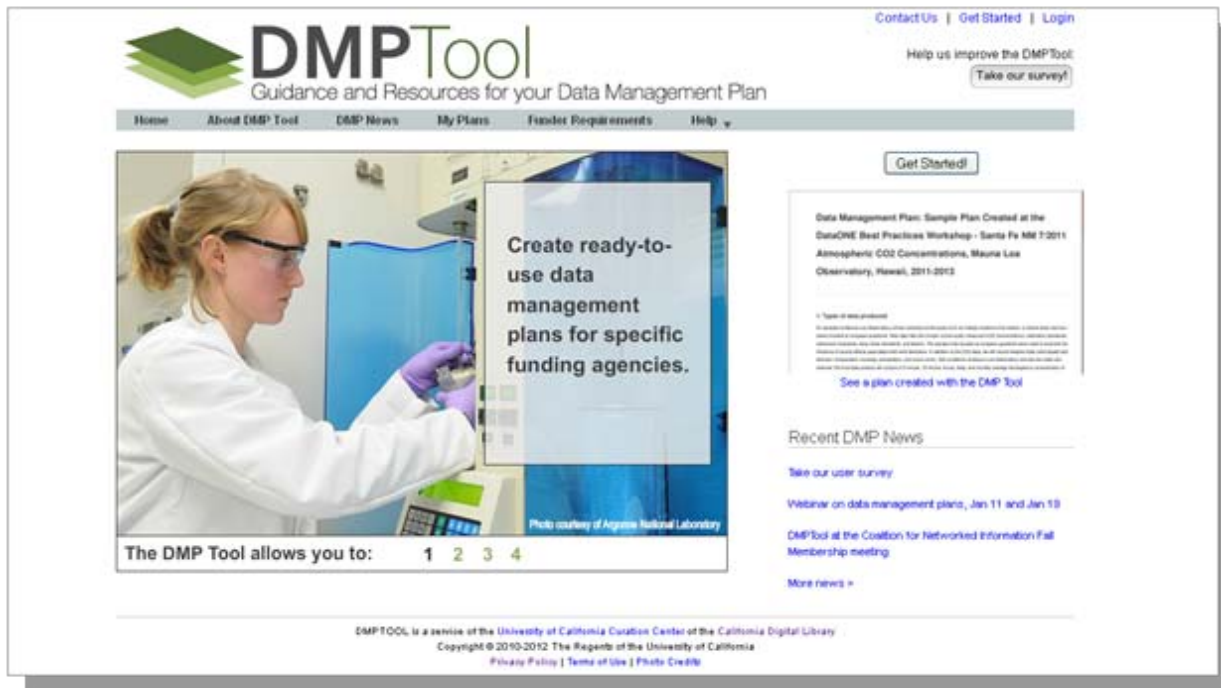
## *Two Services to Mitigate Long Tail Challenges*

1.  **DMPTool**  
Build your Data Management Plan

2.  DataUp

# *Data Management Planning – DMPTool*

## *Meeting funding agencies data management plan requirements*



The screenshot shows the DMPTool website homepage. At the top left is the logo for DMPTool, which consists of three green squares of varying sizes to the left of the text "DMPTool". Below the logo is the tagline "Guidance and Resources for your Data Management Plan". To the right of the logo are links for "Contact Us", "Get Started", and "Login". Below these links is a survey prompt: "Help us improve the DMPTool: [Take our survey!]" and a "Get Started!" button. A navigation menu includes "Home", "About DMPTool", "DMP News", "My Plans", "Funder Requirements", and "Help". The main content area features a large image of a scientist in a lab coat and safety glasses working with a piece of equipment. Overlaid on this image is a text box that reads "Create ready-to-use data management plans for specific funding agencies." Below the image is a section titled "The DMP Tool allows you to:" followed by a numbered list (1, 2, 3, 4). To the right of the image is a "Recent DMP News" section with links to "Take our user survey", "Webinar on data management plans, Jan 11 and Jan 13", and "DMPTool at the Coalition for Networked Information Fall Membership meeting". At the bottom of the page, there is a footer with copyright information: "DMPTOOL is a service of the University of California Curation Center of the California Digital Library. Copyright © 2010-2012 The Regents of the University of California. Privacy Policy | Terms of Use | Photo Credits".

- Seven institutions
- Helps researchers meet funding agency requirements
- Guide researchers through the process of creating a data management plan
- Connects researchers to local resources
- Open to everyone

<http://dmptool.org>

# Customize resources for your institution



You are logged in as Trisha Cruse

- Home
- About DMP Tool
- DMP News
- My Plans
- Funder Requirements
- Help

## NSF-BIO: Biological Sciences: 2. Data Storage and Preservation

What physical and/or cyber resources and facilities (including third party resources) will be used to store and preserve the data?

### Progress

Click on a section below to edit it at any time.

✔ = complete

#### Plan description

- ✔ 1. Products of Research
- ✔ 2. **Data Storage and Preservation**
- ✔ 3. Data Formats and Metadata

**Suggested answer text** box size: [small](#) | [medium](#) | [full](#)  
(copy and paste as needed)

All public data will be deposited in the Merritt Repository Service from the University of California Curation Center (UC3) that has capabilities to manage, archive and share digital content. Merritt allows access to the public via persistent URLs, provides tools for long term

**Help** box size: [small](#) | [medium](#) | [full](#)

Use this section to describe your long-term strategy for storing, archiving and preserving your data. Describe how and where you will make these data and metadata available to the community. Consider these questions:

- What is the long-term strategy for maintaining, curating and archiving the data?

Rich text editor toolbar: **B** *I* U abc  $x_2$   $x^2$  [font color] [font size] [font style] [bulleted list] [numbered list] [link] [unlink] [undo] [redo]

### Resources

- University of California, Office of the President**
  - [UC3: Data management plans](#)
  - [UC3: Merritt Repository Service](#)
  - [UC3: Security, storage and backups](#)

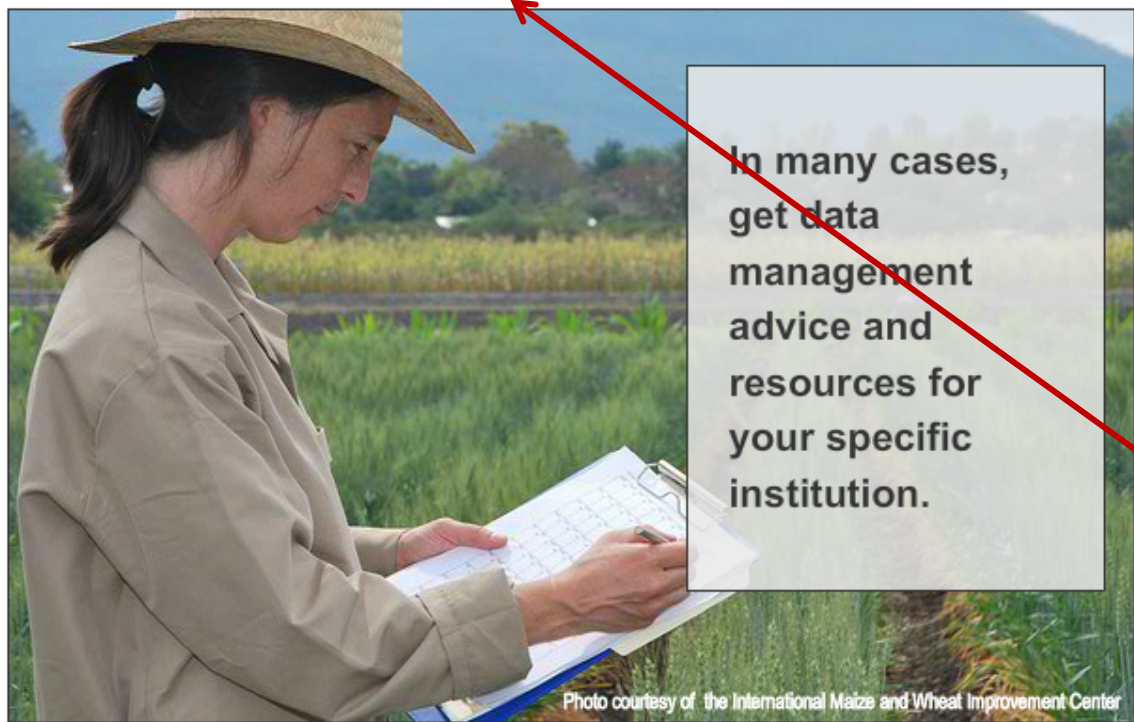
### General

- [NSF Data Sharing Policy](#)
- [NSF Data Management Plan Requirements](#)
- [NSF BIO Directorate Data Management Plan Guidelines](#)

# DMPTool

Guidance and Resources for your Data Management Plan

- Home
- About DMP Tool
- DMP News**
- My Plans
- Funder Requirements
- Help



**In many cases, get data management advice and resources for your specific institution.**

Photo courtesy of the International Maize and Wheat Improvement Center

**The DMP Tool allows you to:**    1   2   3   4

[Get Started!](#)

**Data Management Plan  
Atmospheric CO2 Concentrations,  
Mauna Loa Observatory, 2011-2013**

1. Types of data produced

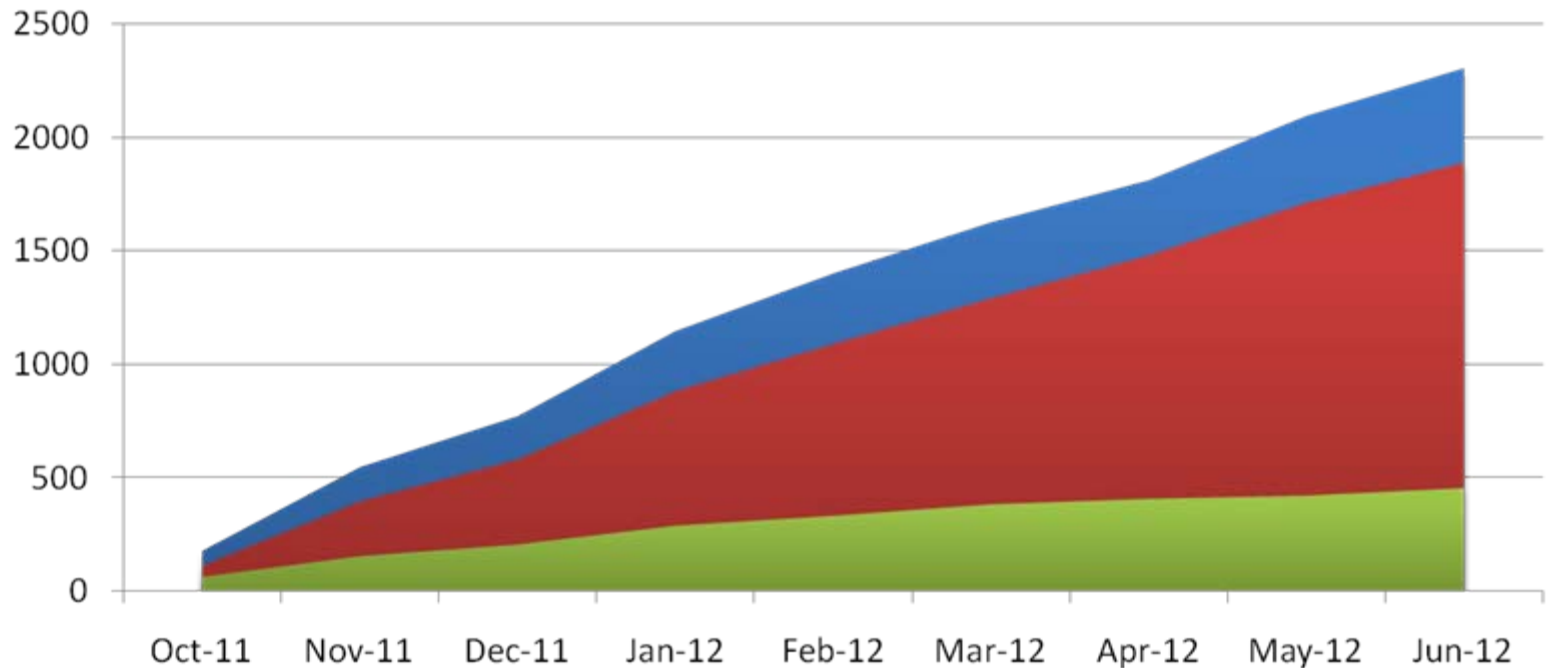
An samples at Mauna Loa Observatory will be collected continuously from air intakes located at five towers – a central tower and four towers located at compass quadrants. Raw data files will contain continuously measured CO2 concentrations, calibration standards, reference standards, daily check standards, and blanks. The sample lines located at compass quadrants were used to examine the influence of source effects associated with wind directions (SW). In addition to the CO2 data, we will record weather data (wind speed and direction, temperature, humidity, precipitation, and cloud cover). Site conditions at Mauna Loa Observatory will also be noted and related.

[See a plan created with the DMP Tool](#)

## Recent DMP News

- [DMPTool workshop at the DLF Fall Forum](#)
- [DMPTool demo: Wed Oct 19](#)
- [Importance of Data Management Education](#)
- [More news >](#)

# *Growth rate since launch in Oct 2011*



	Oct-11	Nov-11	Dec-11	Jan-12	Feb-12	Mar-12	Apr-12	May-12	Jun-12
Unique Users	180	548	772	1146	1399	1624	1809	2092	2302
Plans	114	398	582	882	1090	1291	1479	1711	1887
Institutions	65	155	206	288	331	381	405	419	452

## *New additions*

- Launched an open organization to build a community around the data management issues!
- On the horizon
  - ▶ Adding new functionality
  - ▶ Tighter community integration and educational components
  - ▶ Integration with other systems and initiatives
  - ▶ Routing of DMPs to support personnel
  - ▶ Add analytics



# DataUp

Open Source  
Tool

Add-in & Web  
Application

Facilitate sharing,  
preserving, using,  
citing, re-using  
tabular data



# DataUp

## Features

Best practices check

Generate metadata

Generate citation

Post data to repository

# Best Practices Check

The screenshot shows a Microsoft Excel window titled "copes\_log4.xlsx". The ribbon includes "Home", "Insert", "Page Layout", "Formulas", "Data", "Review", "View", and "DataUp". The "DataUp" tab is active, showing buttons for "Post to Repository", "Check", "Data Description", "Column Description", and "Generate".

The spreadsheet data is as follows:

	A	L	M	N	O	P	Q
1							
2		S8T15_R3		S12T15_R1		S12T15	
3							
12	23-Oct		300		380		200+
13	25-Oct		450		900		200+
14	27-Oct		500		810	765	450+
15	29-Oct		400		1450	1100	350
16	31-Oct		350	500	200	1150	550
17	2-Nov	1000	200	100	800	840	850
18	4-Nov	2425	250	1600	550	2080	250
19	6-Nov	1300	165	800	760	1640	300
20	8-Nov	1450	300	1600	600	1440	180
21	10-Nov	1050	250	2120	500	2200	160
22	12-Nov	1320	300	1400	950	1500	50
23	14-Nov	1240	300	680	400	1640	50
24	16-Nov	720	630	840	450	760	180
25	18-Nov	1000	420	880	700	600	80
26	20-Nov	0	400	440	200	480	<50
27	22-Nov	90	350	510	250	180	200
28	24-Nov	30	350	480	100	60	200
29	26-Nov	90	180	390	100	240	<100
30	28-Nov	60	<100	330	<100	0	<100
31	30-Nov	30	<50	120	<50	30	<50
32	2-Dec	0	<50	90	<25	0	<25
33	4-Dec	0	<25	0	<10	0	0
34	6-Dec		0	0	0	0	
35	8-Dec						
36	10-Dec						
37	12-Dec						
38	14-Dec						
39	16-Dec						

The Error Details pane on the right shows a "Remove Selected" button and a legend for "Fixable Errors" (green dot) and "Non fixable Errors" (red dot). Below the legend, there are several error categories with checkboxes and "Remove" and "Advice" buttons:

- Oct\_7to15 (6)
- Oct17to26(4)
- Oct27toNov16(6)
- scope\_data\_No...(6)
- Notes (6)
- red\_days
- Embedded comments
- Embedded charts, tables, pictures
- Color coded text or cell shading
- Non-contiguous data
- Blank Cells
- Special Characters

# Generate Metadata

	A	B	C	D	E	F	G	H	I	J	K	L
1	<b>Name</b>	<b>Value</b>										
2	Creator:First name	Carly										
3	Creator:Last name	Strasser										
4	Creator:Address											
5	Creator:City											
6	Creator: State/province											
7	Creator:Postal code											
8	Creator:Country											
9	Creator:Email	cah.1208@										
10	Creator:Phone											
11	Creator:Organization											
12	Title of dataset	Title here										
13	Today's date											
14	Abstract	abstract										
15	Repository name and contact information											
16	url for data											
17	Data Contact Person: First name											
18	Data Contact Person:Last name											
19	Data Contact Person:Address											
20	Data Contact Person:City											
21	Data Contact Person:State/province											
22	Data Contact Person:Postal code											
23	Data Contact Person:Country											
24	Data Contact Person: Phone											
25	Data Contact Person:Email											
26	Data Contact Person:Organization											
27	keyword(s)	keyword										
28	Keyword thesaurus used											
29	Geographic coverage:Description											
30	Geographic coverage:West bounding coordinate											
31	Geographic coverage:East bounding coordinate											
32	Geographic coverage:North bounding coordinate											
33	Geographic coverage:South bounding coordinate											
34	Temporal coverage:Description											
35	Temporal coverage:Beginning date											
36	Temporal coverage:Ending date											
37	Project title											
38	Project description											
39	Funding											
40	Intellectual rights											
41	Data table name											
42	Data table description											
43	Identifier											
44	Citation											
45												
46												
47												

**CREATE METADATA**

It generates and applies metadata based on a pre-defined schema. Some of the metadata values are automatically populated by the system – the remaining must be supplied by you. Some of the metadata fields are required, designated by red asterisk, and some are optional. Please make sure you provide at least all the required metadata values.

**Data descriptions** | Column descriptions

Creator:First name \*

Creator:Last name \*

Creator:Address

Creator:City

Creator: State/province

Creator:Postal code

Creator:Country

Creator:Email \*

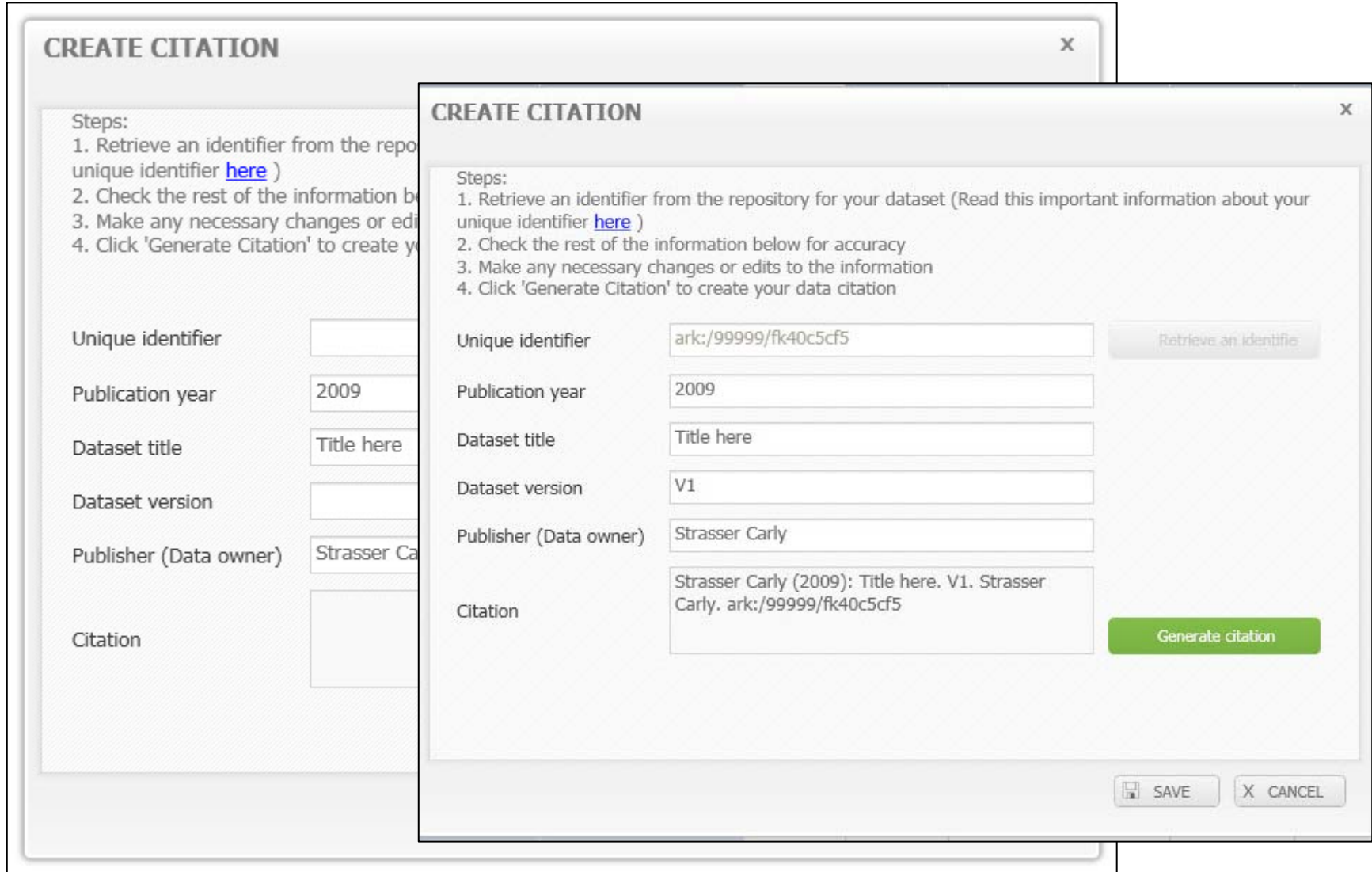
Creator:Phone

Creator:Organization

Title of dataset \*

scope\_data\_Nov17toDec8 | Notes | red\_days | pink\_days | scope\_data\_Dec9to... | phytoplankton&counting | **Metadata**

# Create a Data Citation



The image shows a 'CREATE CITATION' dialog box with a light gray background and a white content area. The dialog has a title bar with 'CREATE CITATION' and a close button (X). The content area is divided into two main sections: instructions and input fields. The instructions section, titled 'Steps:', lists four steps: 1. Retrieve an identifier from the repository for your dataset (Read this important information about your unique identifier [here](#)), 2. Check the rest of the information below for accuracy, 3. Make any necessary changes or edits to the information, and 4. Click 'Generate Citation' to create your data citation. The input fields section contains: 'Unique identifier' (text input with 'ark:/99999/fk40c5cf5'), 'Publication year' (text input with '2009'), 'Dataset title' (text input with 'Title here'), 'Dataset version' (text input with 'V1'), and 'Publisher (Data owner)' (text input with 'Strasser Carly'). A 'Retrieve an identifier' button is located to the right of the unique identifier field. Below these fields is a 'Citation' text area containing the generated citation: 'Strasser Carly (2009): Title here. V1. Strasser Carly. ark:/99999/fk40c5cf5'. A green 'Generate citation' button is located to the right of the citation text area. At the bottom right of the dialog are 'SAVE' and 'CANCEL' buttons.

**CREATE CITATION** [X]

Steps:

1. Retrieve an identifier from the repository for your dataset (Read this important information about your unique identifier [here](#) )
2. Check the rest of the information below for accuracy
3. Make any necessary changes or edits to the information
4. Click 'Generate Citation' to create your data citation

Unique identifier:

Publication year:

Dataset title:

Dataset version:

Publisher (Data owner):

Citation:

[SAVE] [CANCEL]

# *Upload to a Repository*

**FILE POST** x  
Post as XLSX

✓ issues ... descriptions ... ✓ citation ... post

You are now ready to post your curated document to the repository of your choice. Select the repository from the list, and provide your credentials for the repository.

Repository Name\*

Repository Type

I accept [User Agreement](#)

**Launching in  
September 2012**

← BACK    X CANCEL    ↑ POST

# *Meeting the Long Tail Challenge*

- ✓ Engage with the research community early in the lifecycle?
- ✓ Deploy simple and flexible infrastructures and services that can be used in diverse ways
- ✓ Start small: provide simple solutions that build up to solve more complex problems
- ✓ Collaborate now more than ever!

# *UC3's Service Landscape*

UC Curation Center

<http://www.cdlib.org/uc3>  
[uc3@ucop.edu](mailto:uc3@ucop.edu)

*Stephen Abrams*

*Patricia Cruse     Joan Starr*

*Scott Fisher     Tracy Seneca*

*Erik Hetzner     Carly Strasser*

*Greg Janée     Marisa Strong*

*John Kunze     Adrian Turner*

*David Loy     Perry Willett*

*Mark Reyes*

