

# Data Curation

## The Current Landscape

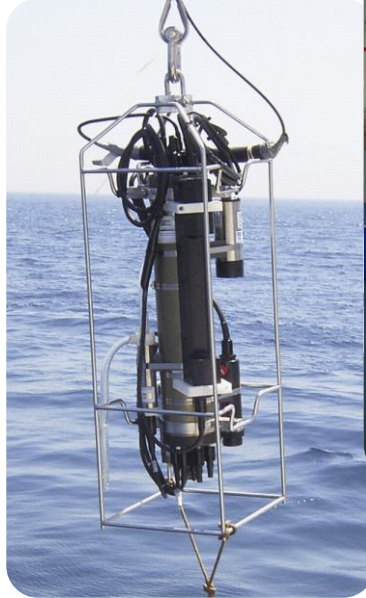
**Carly Strasser, PhD**

California Digital Library  
@carlystrasser

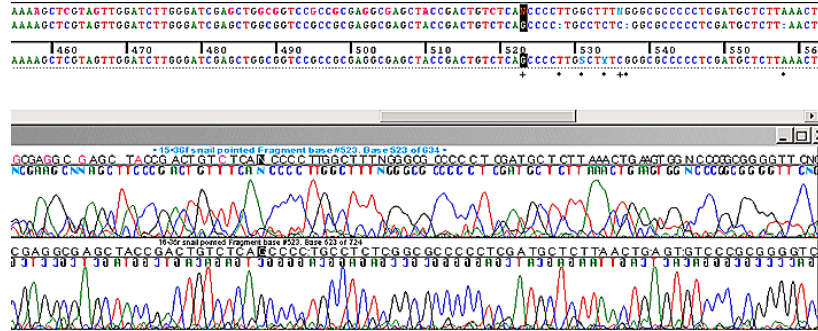
Microsoft eScience Workshop  
Oct 2012



# Digital data



www.woodrow.org

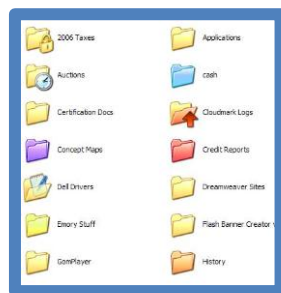
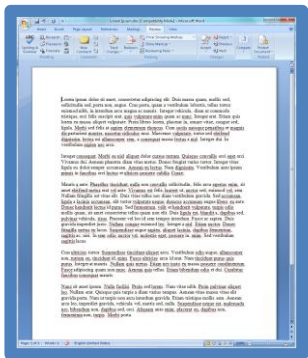
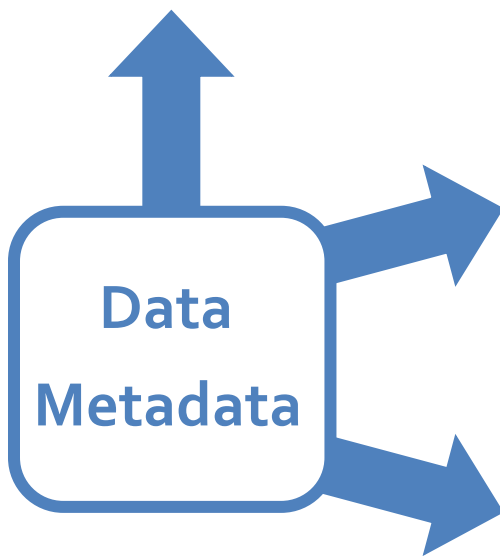


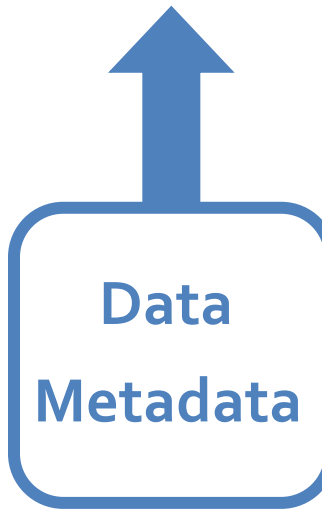
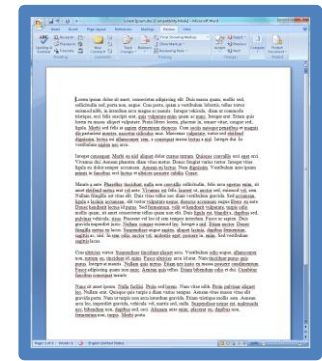
# Digital data



From Calisphere via San Jose Public Library

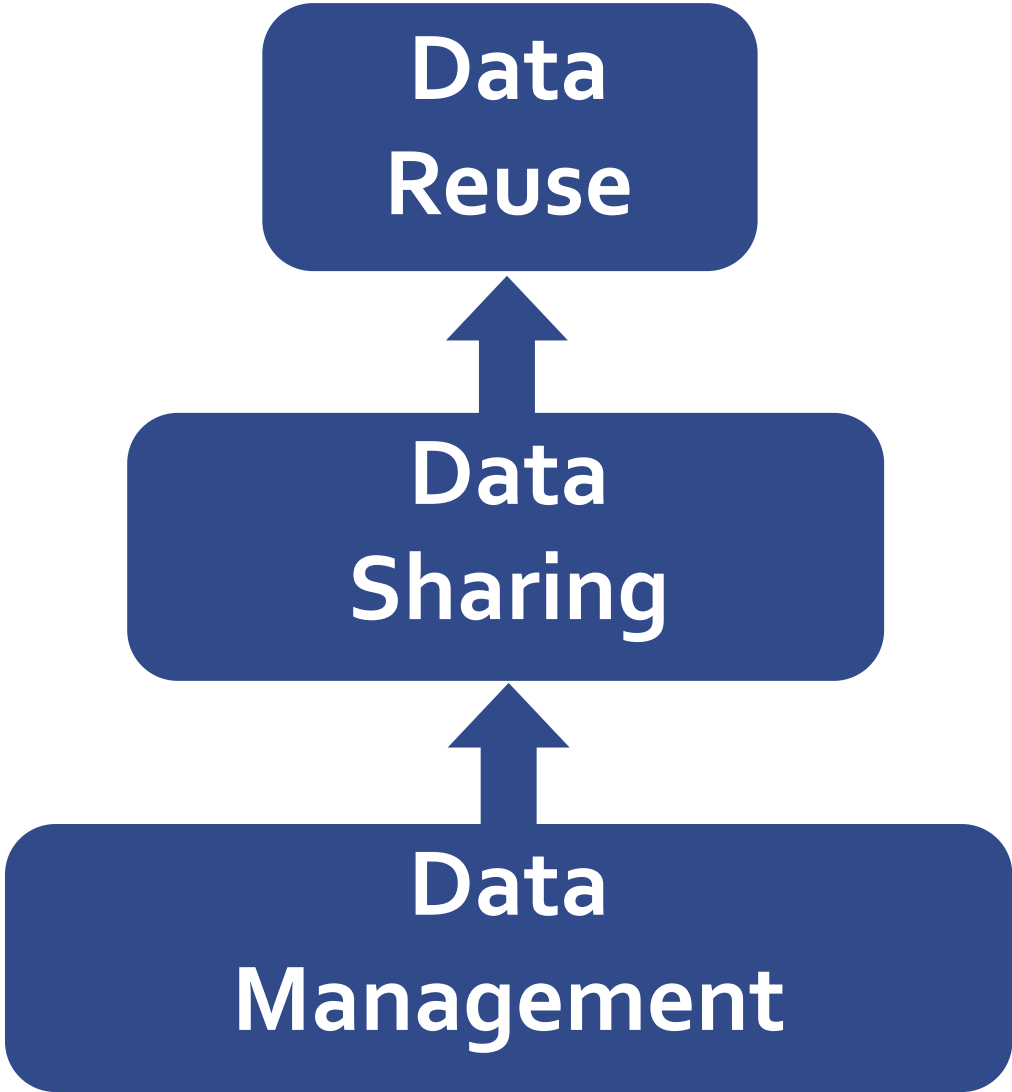






From Flickr by torkildr







# UGLY TRUTH

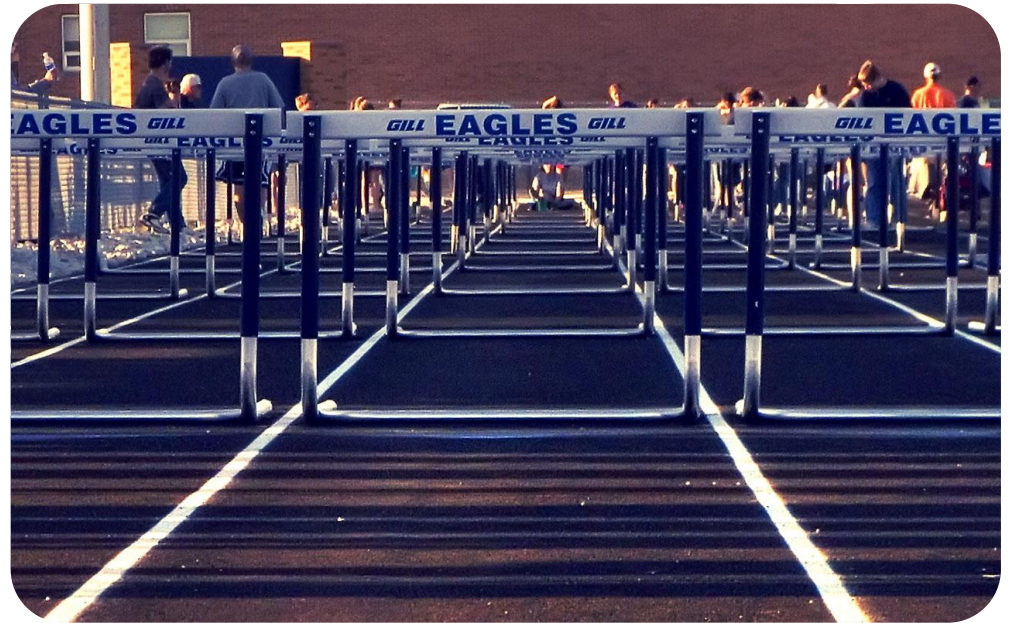
Many (most?)  
researchers...



are not taught  
are not effective  
don't share data  
aren't convinced they should  
don't know about tools that can help them

their data  
archive

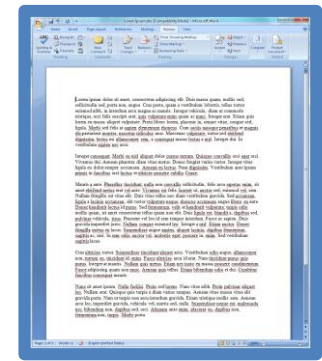
# Hurdles to Data Stewardship



From Flickr by iowa\_spirit\_walker

- Cost
- Confusion about standards
- Disparate datasets
- Lack of training
- Fear of lost rights or benefits
- No incentives





From Flickr by diylibrarian



Data  
Metadata



From Flickr by torkildr



# Who cares?



GORDON AND BETTY  
**MOORE**  
FOUNDATION

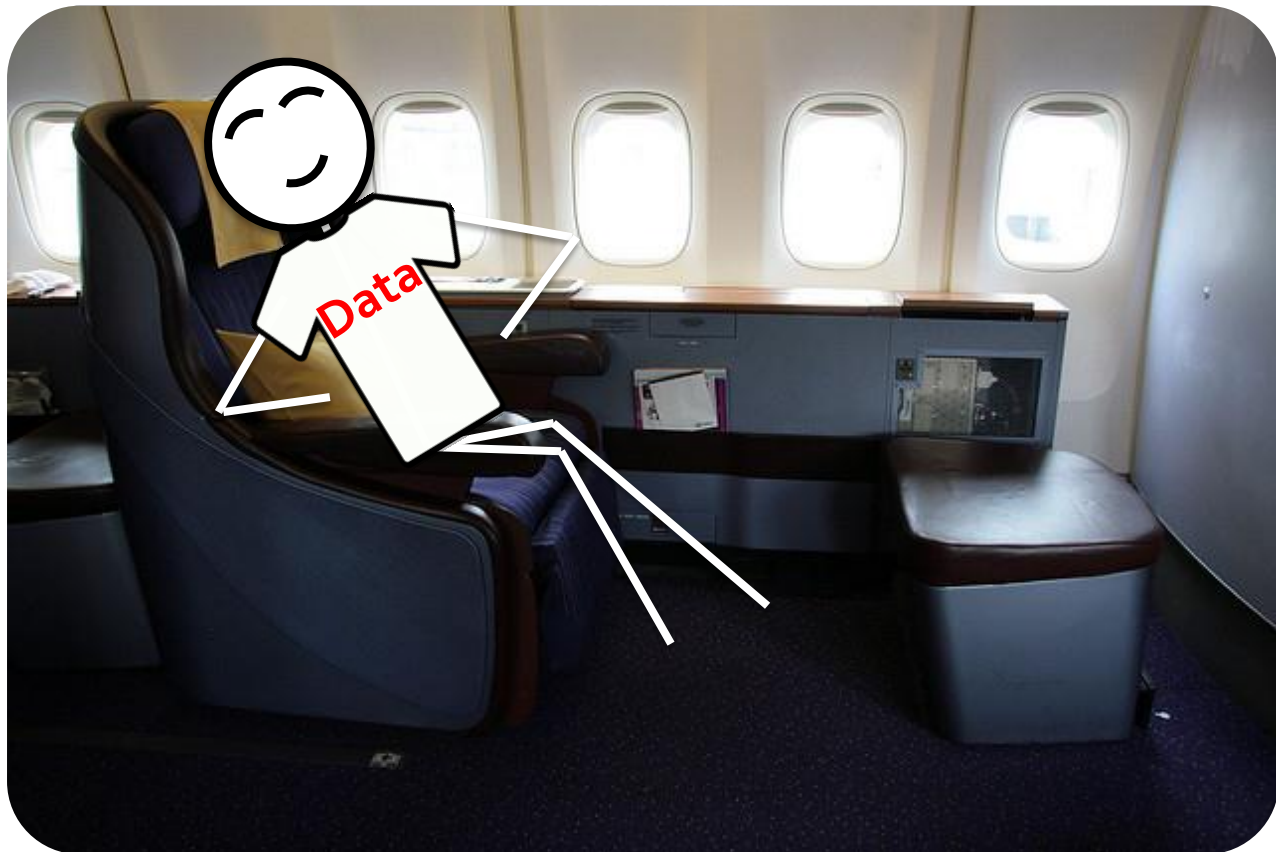


*From Flickr by Redden-McAllister*



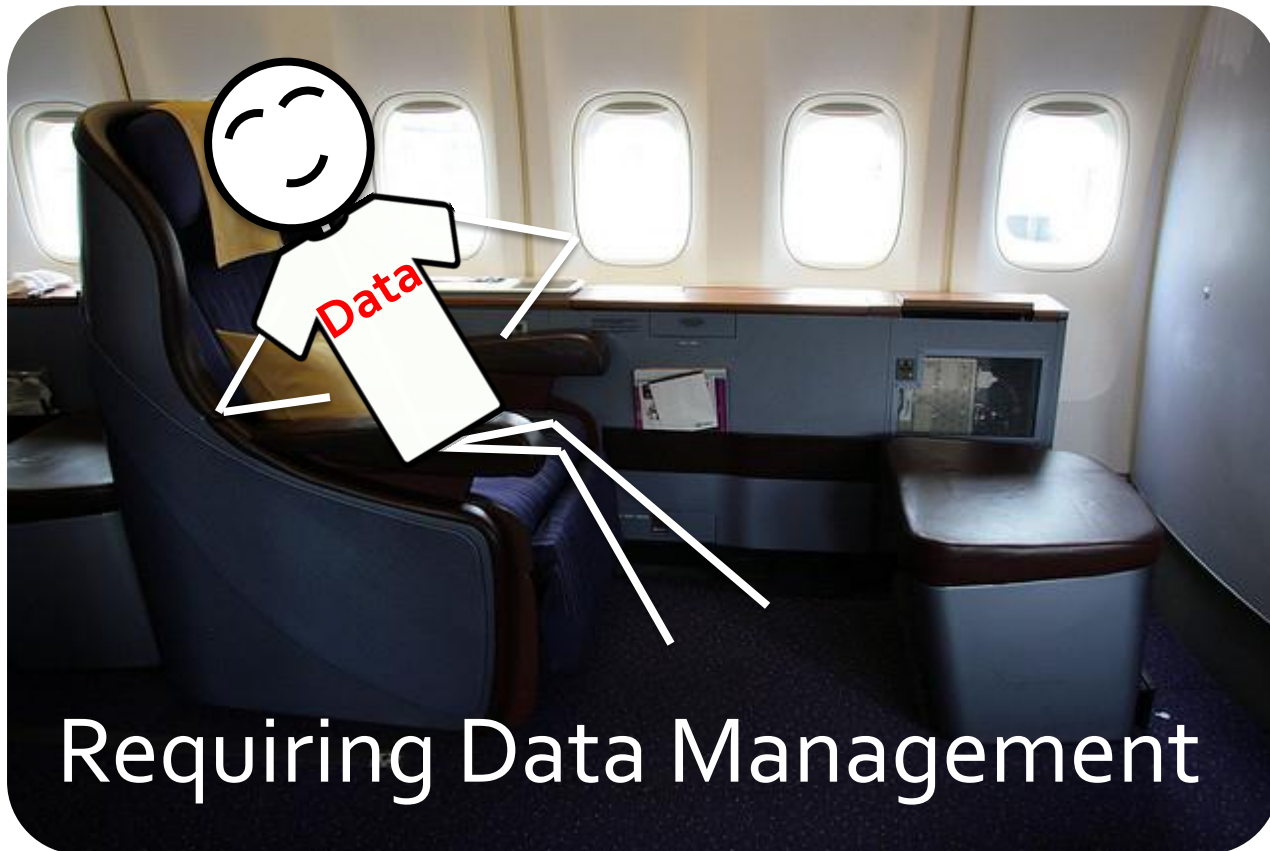
*www.rba.gov.au*

# Data are first class products of research



*From Flickr by Richard Moross*

# Data are first class products of research



Requiring Data Management

## CURRENT ISSUES

Volume 490 Number 7418



About the cover >

## American Naturalist [Publication Info](#)

[About Journal](#) | [News & Announcements](#) | [Feedback](#)

Coverage: 1867-2012 (Vols. 1-180)

Published by: [The University of Chicago Press](#)  
[Naturalists](#)

### Availability of data and materials

The policy outlined on this page applies to *Nature* journals (those with the word "*Nature*" in their title). NPG publishes many other journals, each of which has separate publication policies described on its website. A current list of these journals, with links to each journal's homepage [is available](#).

### Availability of data and materials

An inherent principle of publication is that others should be able to replicate and build upon the authors' published claims. Therefore, a condition of publication in a *Nature* journal is that **authors are required to make materials, data and associated protocols promptly available to readers without undue qualifications**. Any restrictions on the availability of materials or information must be disclosed to the editors at the time of submission. Any restrictions must **also** be disclosed in the submitted manuscript, including details of how readers can obtain materials and information. If materials are to be distributed by a for-profit company, this must be stated in the paper.

**Supporting data** must be made available to editors and peer-reviewers at the time of submission for the purposes of evaluating the manuscript. **Peer-reviewers** may be asked to

repository. See [Data Policy](#).

Authors are strongly encouraged to deposit their data in the

official Data Registry at [data.es](#)

ESA submission site itself will feature a

will serve as a mechanism for "data discovery," leading to communication (and possibly collaboration) between

researchers and to meta-analyses.

- Integrative and Comparative Biology (Society for Integrative and Comparative Biology)
- Journal of Fish and Wildlife Management
- Journal of Paleontology (The Paleontological Society)
- Oxford University Press
- Pensoft Publishers

## Supporting

The *American* phylogenetic tree data set, including TreeBASE, and provide a link. impediments to worked out. For

When you use your article as



## Grant Proposal Guide

### PAPP - Introduction

A. About the NSF

B. Foreword

C. Acronym List

D. Definitions

E. NSF Organizations

NSF 13-1 January 2013

### GPG Summary of Changes

**Significant Changes to Implement the Recommendations of the National Science Board's Report entitled, "[National Science Foundation's Merit Review Criteria: Review and Revisions](#)"**

**Chapter II, Introduction**, has been supplemented with information regarding the Foundation's core strategies from the NSF 2011-2016 Strategic Plan. Similar language

certifications were added to implement provisions included in the *Commerce, Justice, and Related Agencies Appropriations Act of 2012*.

**Chapter II.C.2.f(i)(c), Biographical Sketch(es)**, has been revised to rename the "Publications" section to "Products" and amend terminology and instructions accordingly. This change makes clear that products may include, but are not limited to, publications, data sets, software, patents, and copyrights.

**Chapter II.C.2.g(viii), Indirect Costs**, has been modified to clarify that, except as noted in GPG II.C.2.g(v) and II.D.9 or in an NSF program solicitation, the applicable indirect cost rate is established by the organization with the most recent certification

# Data are first class products of research



*From Flickr by Richard Moross*

# Data Publication



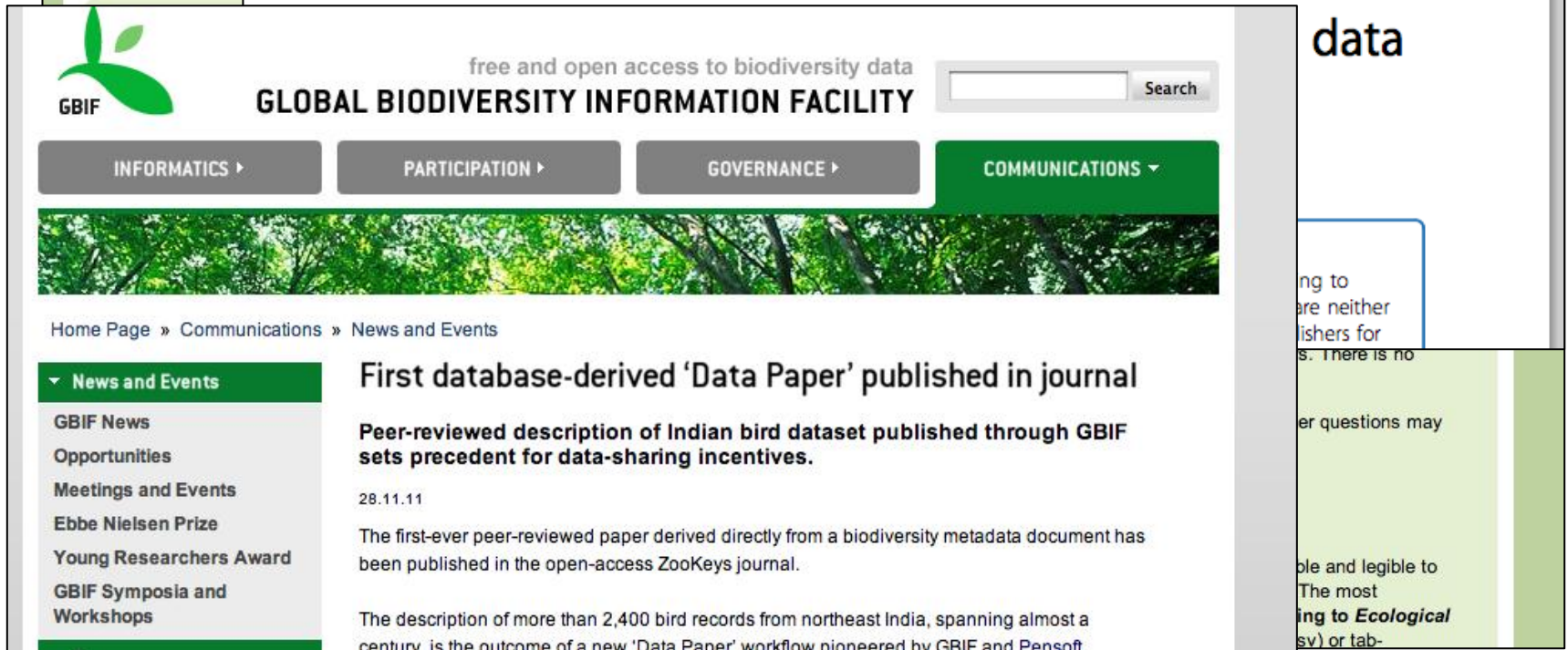
ECOLOGICAL SOCIETY OF AMERICA

Chavan and Penev *BMC Bioinformatics* 2011, **12**(Suppl 15):S2  
<http://www.biomedcentral.com/1471-2105/12/S15/S2>

BMC Bioinformatics

RESEARCH Open Access

esa pubs ho



GBIF free and open access to biodiversity data

## GLOBAL BIODIVERSITY INFORMATION FACILITY

INFORMATICS PARTICIPATION GOVERNANCE COMMUNICATIONS

Home Page » Communications » News and Events

### First database-derived 'Data Paper' published in journal

Peer-reviewed description of Indian bird dataset published through GBIF sets precedent for data-sharing incentives.

28.11.11

The first-ever peer-reviewed paper derived directly from a biodiversity metadata document has been published in the open-access ZooKeys journal.

The description of more than 2,400 bird records from northeast India, spanning almost a century, is the outcome of a new 'Data Paper' workflow pioneered by GBIF and Pensoft

data

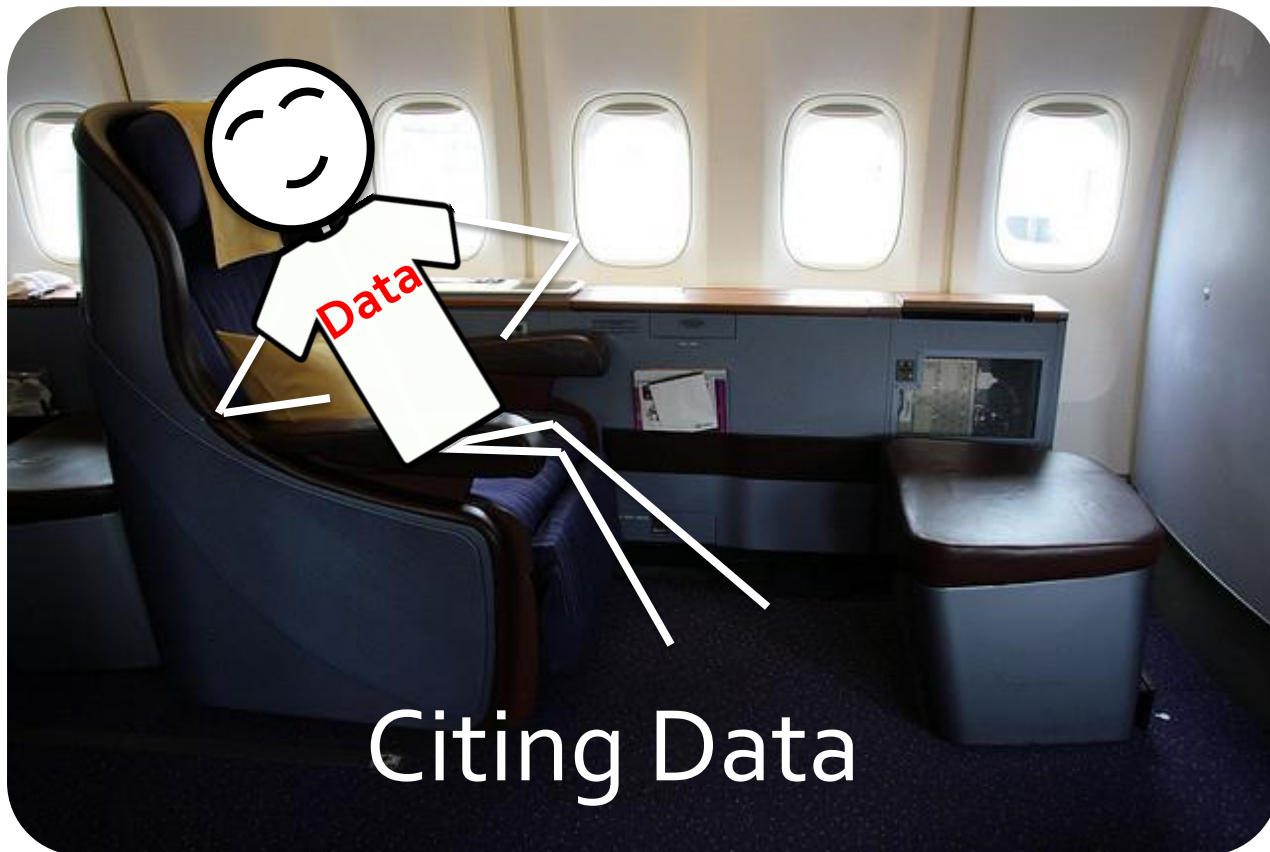
ing to  
are neither  
ishers for  
s. There is no

er questions may

ple and legible to  
The most  
ing to *Ecological*  
sv) or tab-



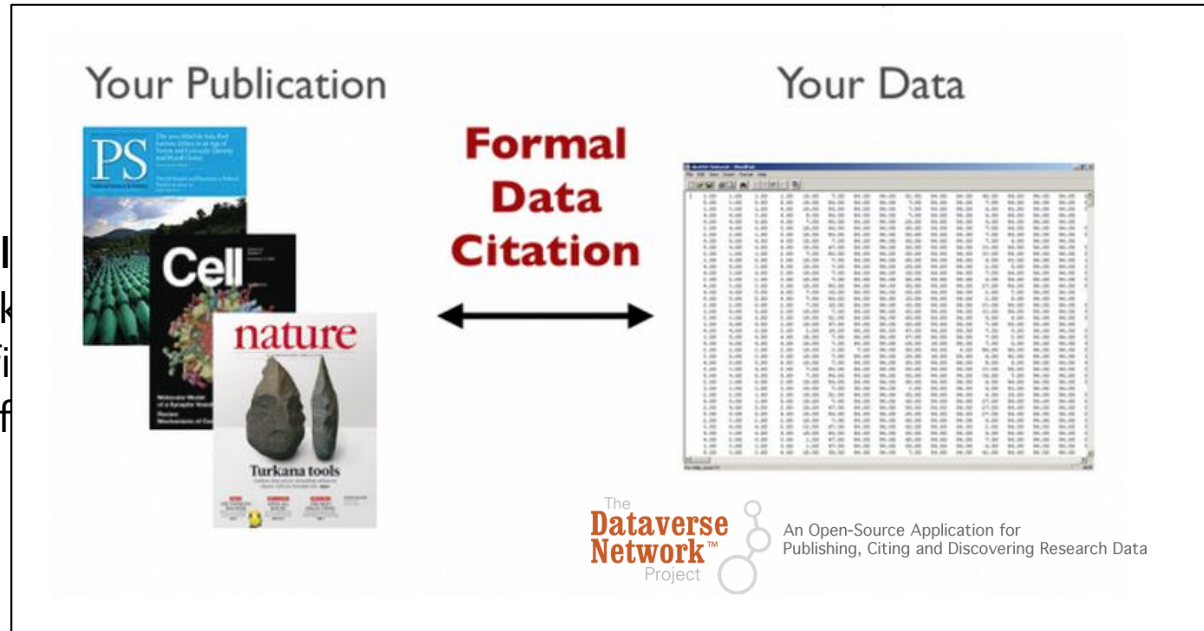
# Data are first class products of research



*From Flickr by Richard Moross*

# Data Citation

Example  
Sidlausk  
diversifi  
characif



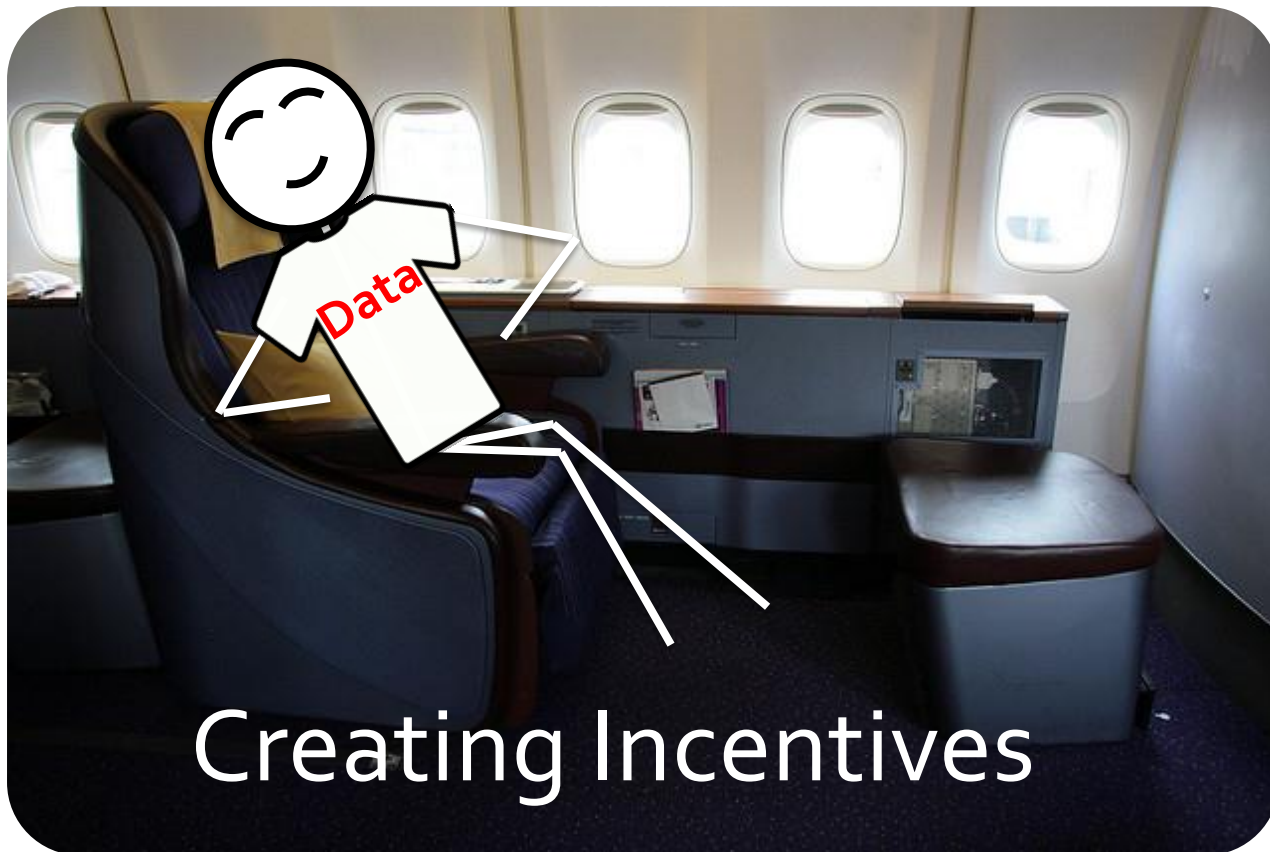
UC3EZID  
long-term identifiers made easy



DataCite

Helping you to find,  
access, and reuse data

# Data are first class products of research

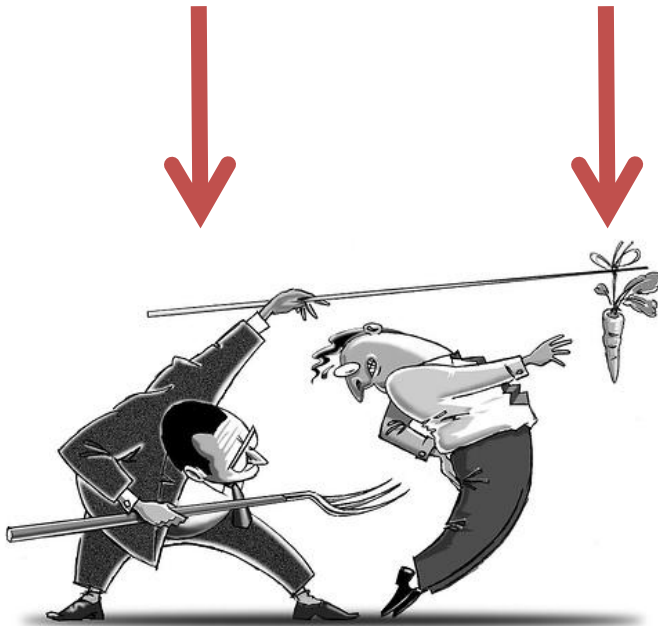


Creating Incentives

# Incentivizing

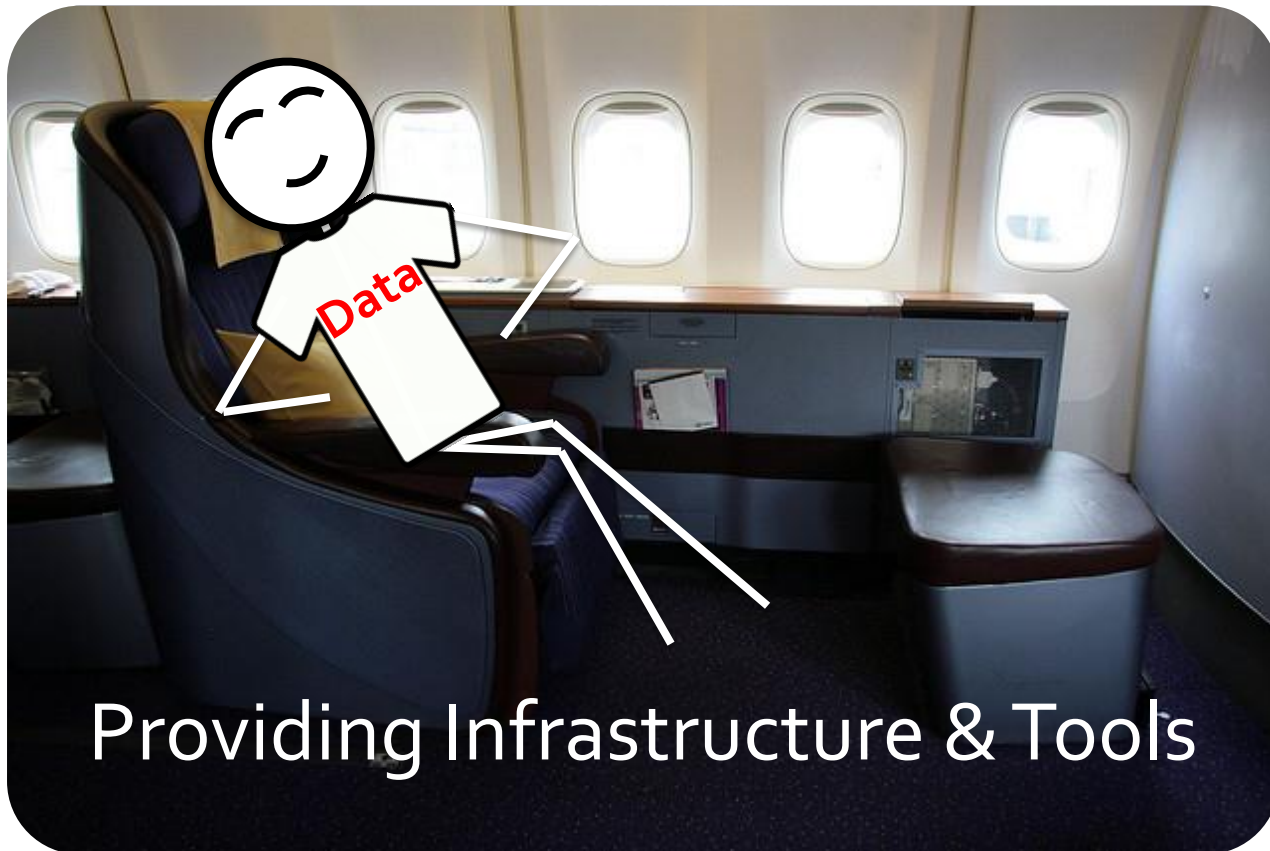


[www.rba.gov.au](http://www.rba.gov.au)



From Flickr by bthomso

# Data are first class products of research



Providing Infrastructure & Tools



NSF funded DataNet Project  
Office of Cyberinfrastructure

www.dataone.org

The screenshot shows the DataONE website homepage. At the top left is the DataONE logo. To its right is a search bar with 'ONEMercury' entered and a 'Go' button. Further right are social media icons for Facebook, Twitter, LinkedIn, and RSS. Below the search bar is a large banner for 'DataONE NEWS Volume 1 Issue 1'. The banner features a 'Message from the DataONE Project Director' with a photo of a man and a 'Click to Search' button. Below the banner is a 'Latest News' section with a dark background and white text. At the bottom, there are five columns of navigation links: About, Participate, Resources, Education, and Data.

**DataONE** Search  For   [f](#) [t](#) [in](#) [RSS](#)

**DataONE NEWS** Volume 1 Issue 1  
©2012 DataONE 1310 South St. University of New Mexico Albuquerque NM 87131

**Message from the DataONE Project Director**  
I am pleased to announce that on July 23 2012 DataONE was officially launched. As the earth and environmental sciences evolve to be more data-intensive discovering integrating and analyzing massive amounts of heterogeneous information becomes critical to enable researchers to address complex questions about our environment and our health risks? Are climate models sufficiently predictive? DataONE addresses this need by providing a single search interface that queries data repositories distributed globally. These data centers individually store and manage digital scientific data holdings and DataONE now enables scientists around the world to download the Fall edition of DataONE News research enabled by this widespread access to data will range from studies that illuminate fundamental environmental processes to take the grand challenges facing science and society. Data held by South Africa National Parks

**ONE Mercury**  
A DataONE Search Tool for Scientific Data  
Click to Search

**Safety Data Challenge**  
SAFETY FIRST  
data.gov/safety

**Latest News** California Digital Library and Partners Laund

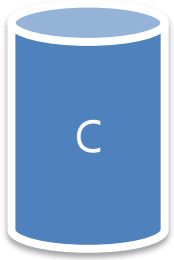
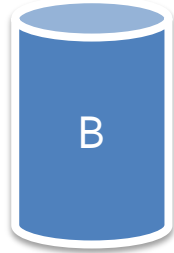
**About**  
What is DataONE?  
DataONE Organization  
Working Groups  
Partners  
Communication  
Videos  
Contact Us

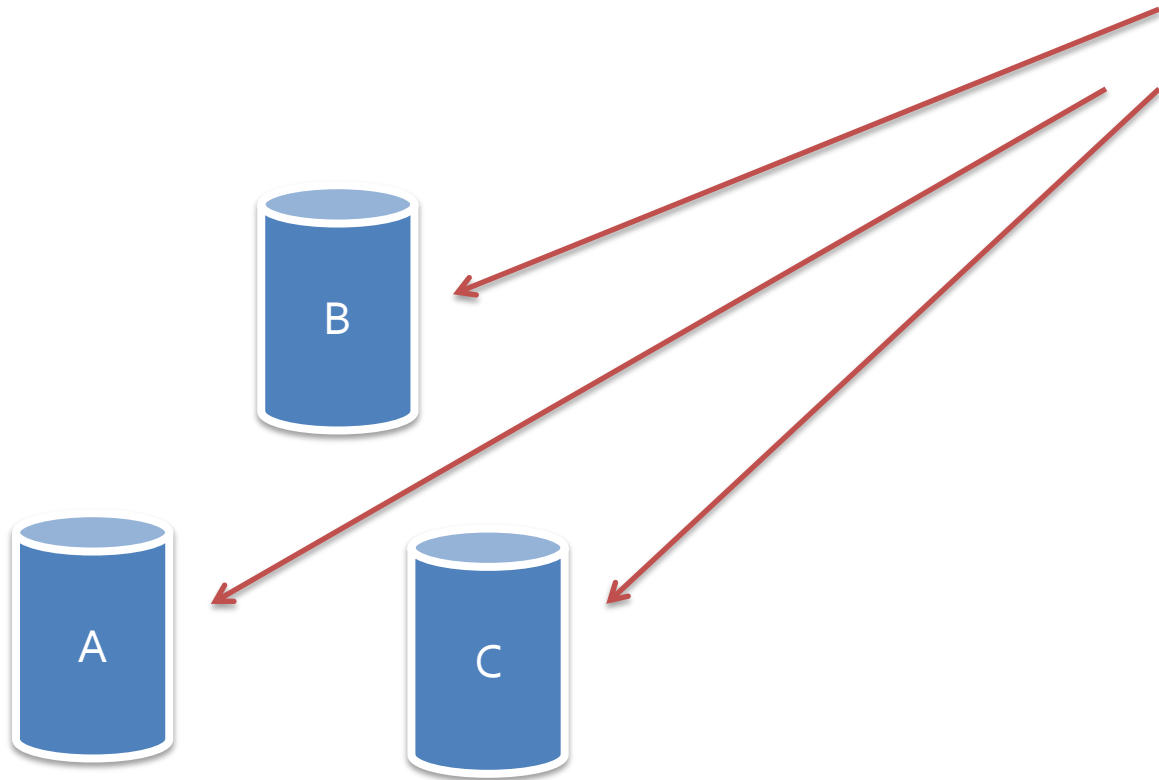
**Participate**  
DataONE Users Group  
Member Nodes  
Internships  
Developer Resources  
Open Positions  
Events Calendar

**Resources**  
Investigator Toolkit  
Data Management Planning  
Best Practices  
Software Tools Catalog  
Publications

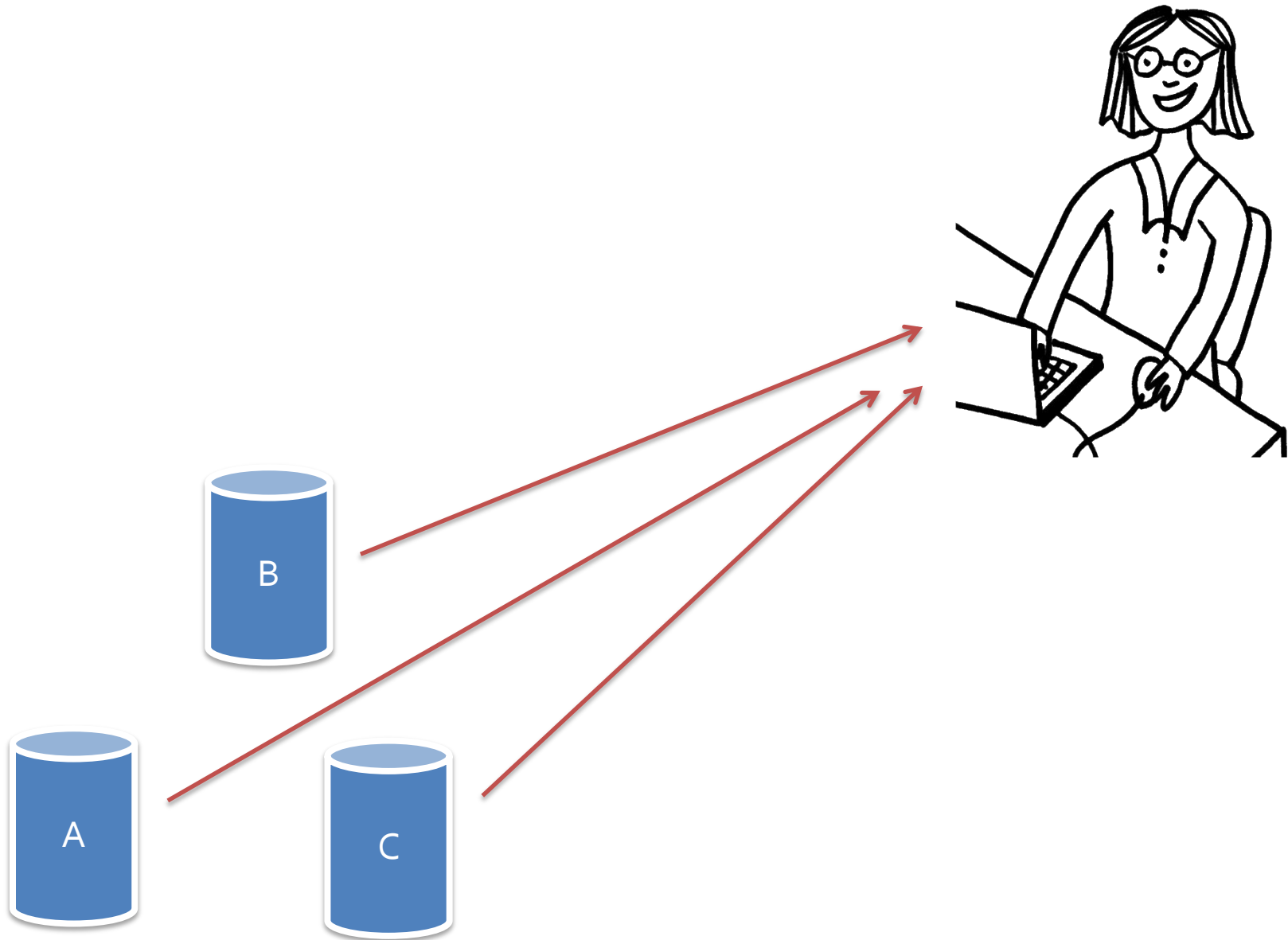
**Education**  
Training Activities  
Education Modules  
Graduate Courses

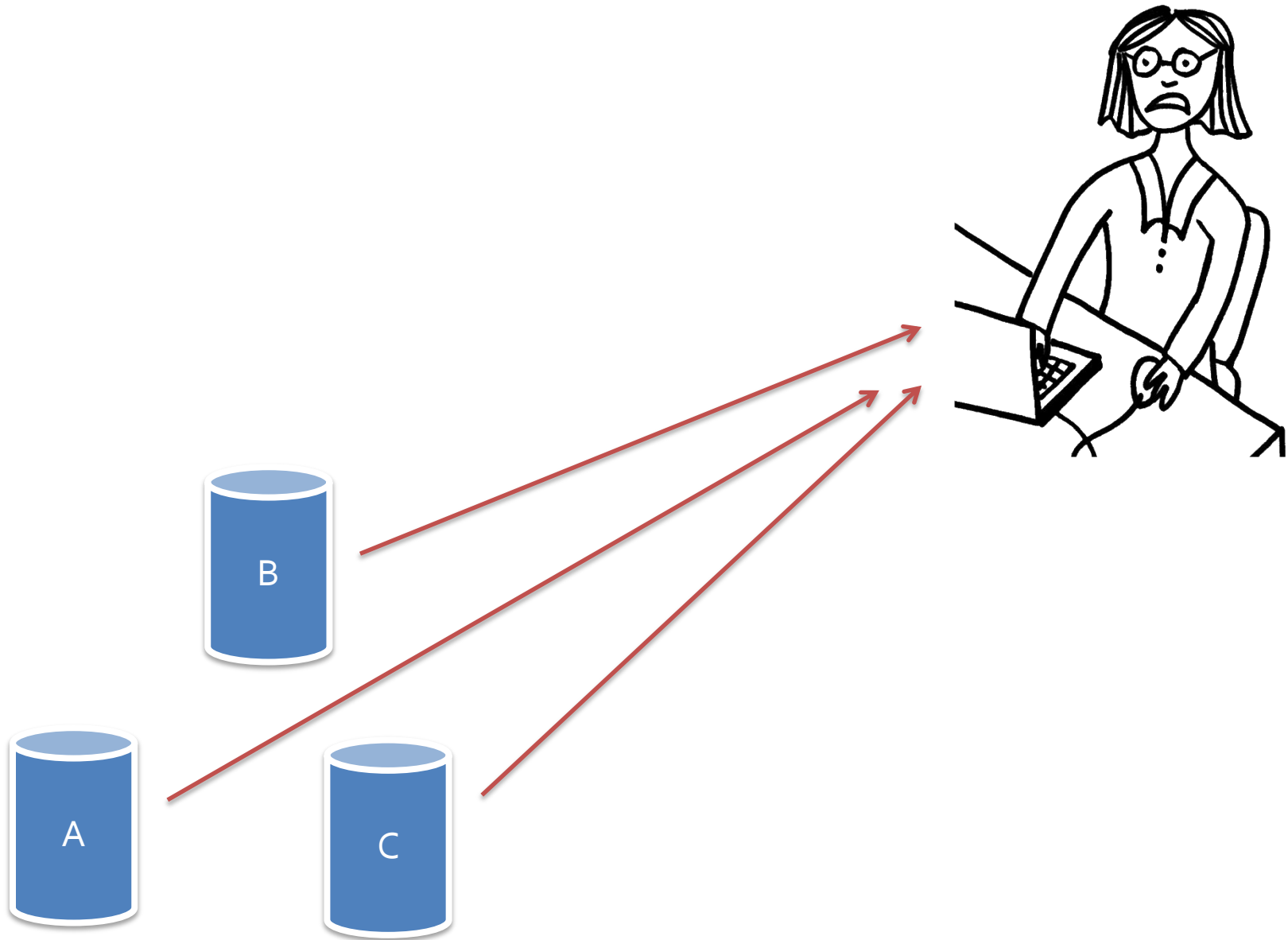
**Data**  
Find  
Contribute  
Cite  
Use  
Data Holdings  
Safety Data Challenge

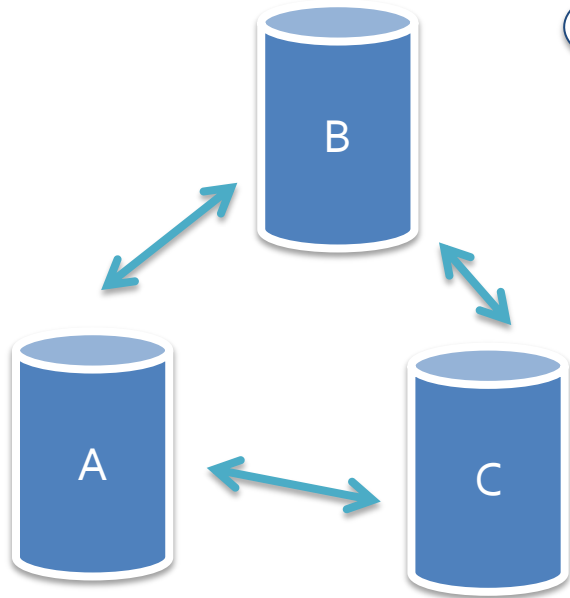


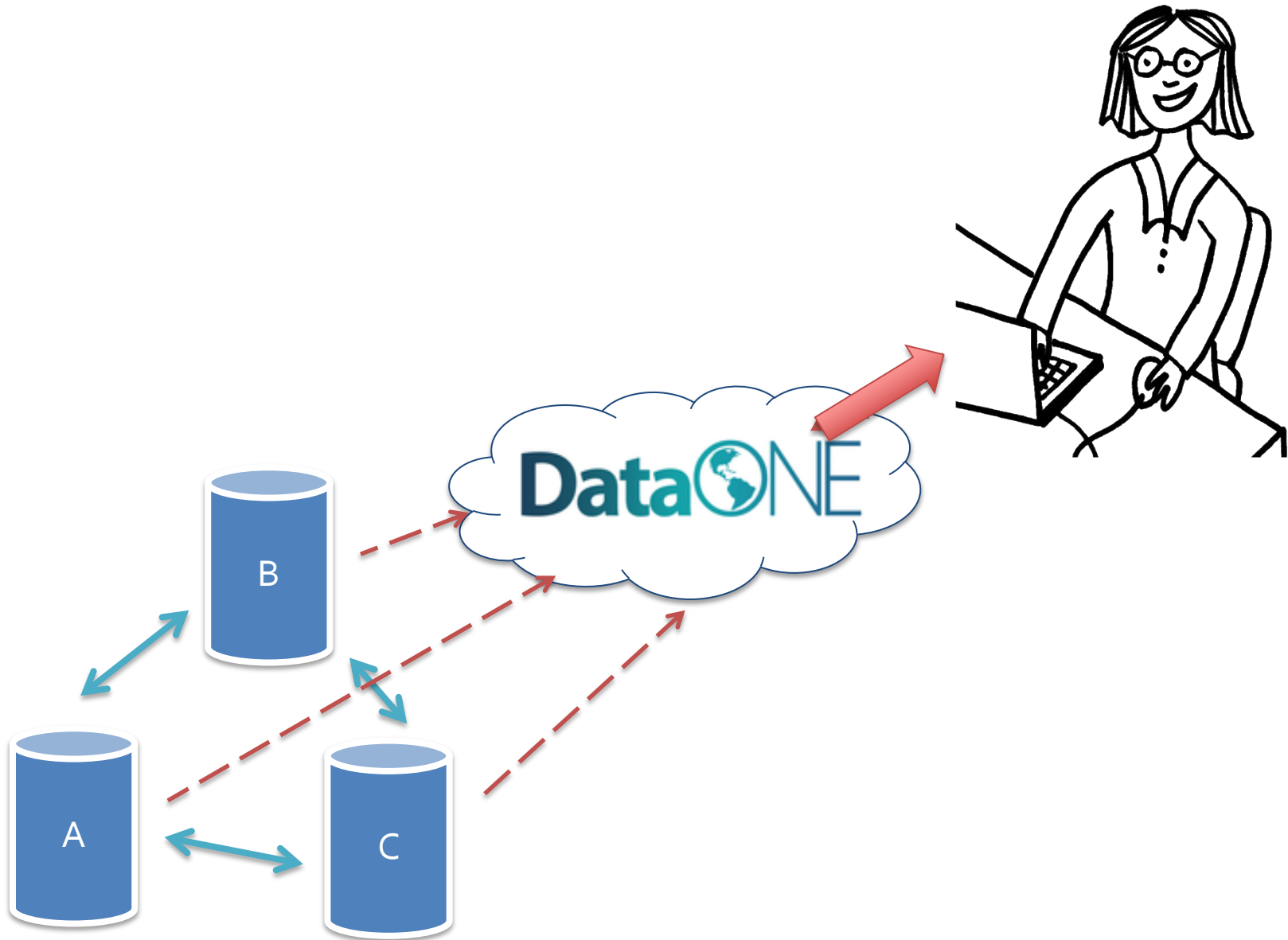








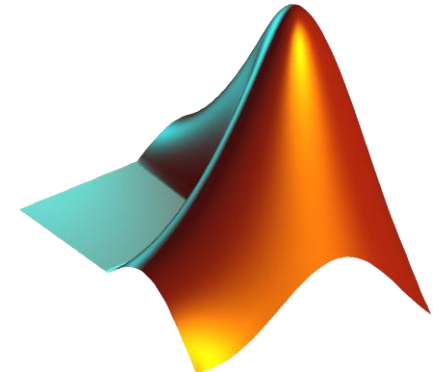




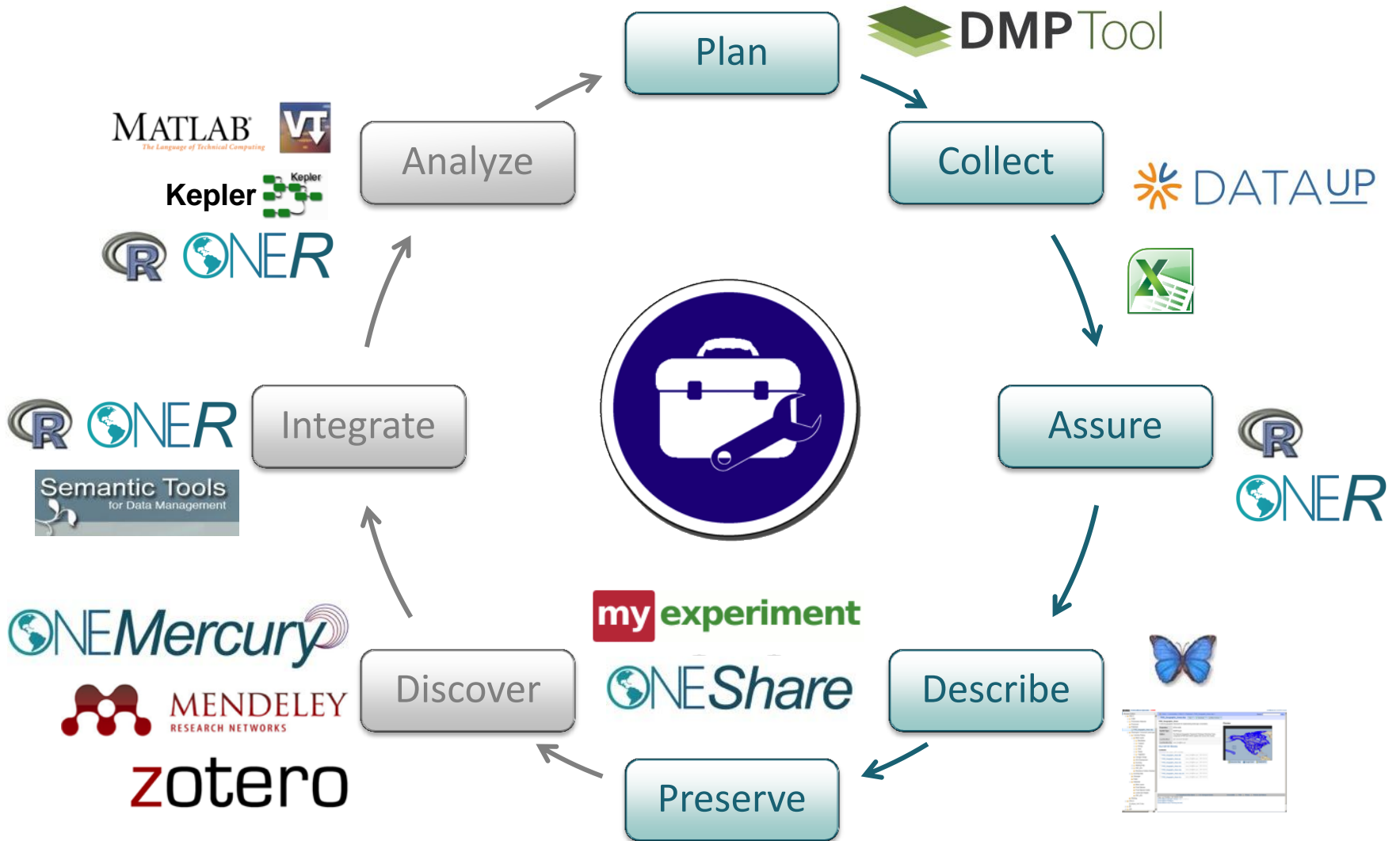
Intercept researchers  
where **they already**

**work**

**zotero**



# Investigator Toolkit Support

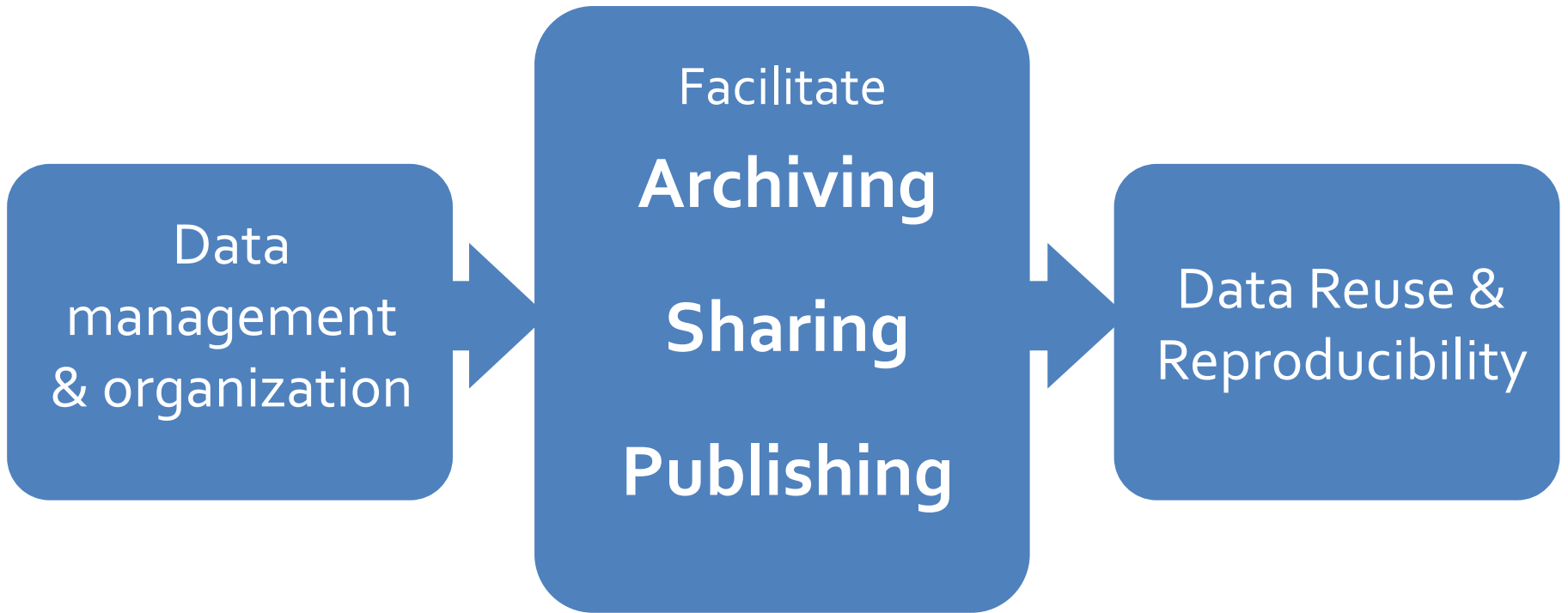




# DATAUP

Describe, Manage, & Share Your Data

A Service of the University of California Curation Center (UC3)



University of California

**CDL**

California Digital Library

Microsoft®

**Research**

GORDON AND BETTY

**MOORE**

FOUNDATION



# DATAUP

**Describe, Manage, & Share Your Data**

A Service of the University of California Curation Center (UC3)

Open Source  
Tool

Add-in & Web  
Application

Earth,  
environmental,  
ecological  
researchers



# Requirements



## Features

Best practices check

Generate metadata

Generate citation

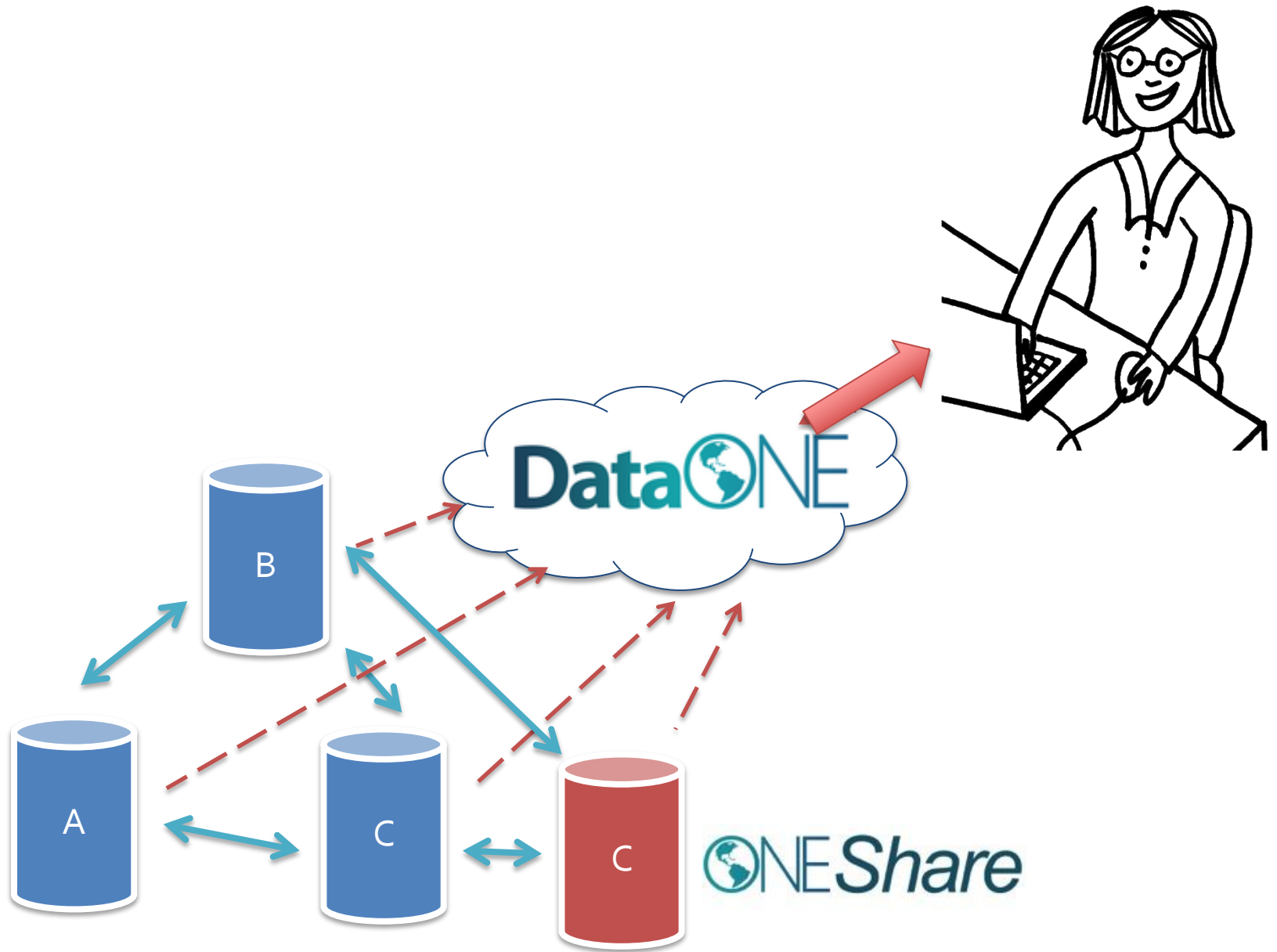
Post data to repository

 UC3 Merritt

Data  ONE

 ONE *Share*

Data Repository for  
**Anyone | Anywhere**



# dataup.cdlib.org



An open source tool helping researchers document, manage, and archive their tabular data, DataUp operates within the scientist's workflow and integrates with Microsoft® Excel.

Looking for the DataUp blog?

[» Go to DataPub](#)

DataUp Features

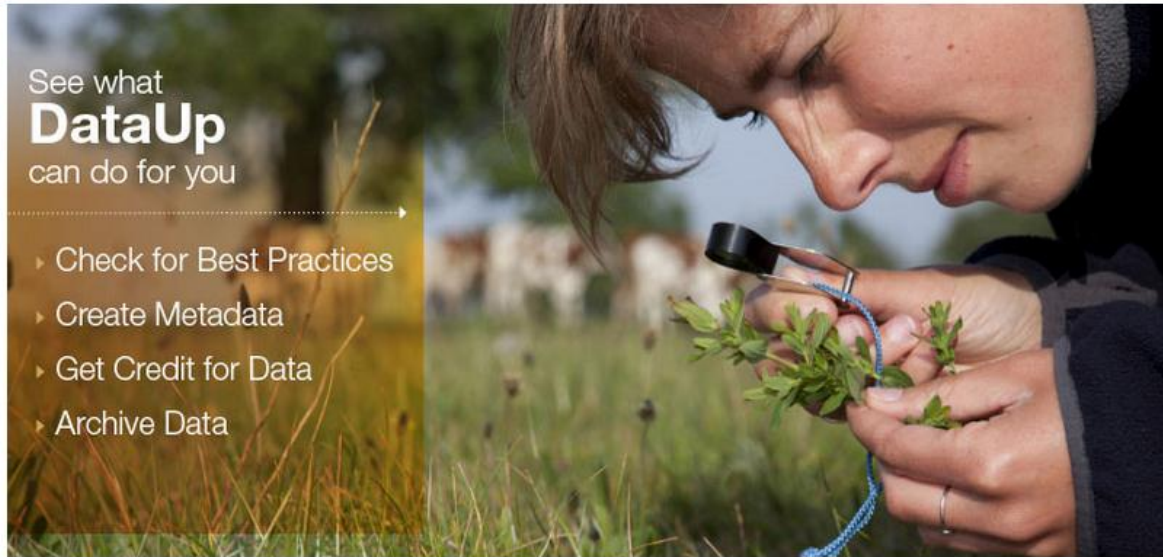
Start Using DataUp

Customize DataUp

Contact Us

See what  
**DataUp**  
can do for you

- ▶ Check for Best Practices
- ▶ Create Metadata
- ▶ Get Credit for Data
- ▶ Archive Data



#### About the Project:

Funders, partners, history and evolution of DataUp »

#### Resources:

CDL resources, data management help Presentations about DataUp »

#### News & Events:

Stay informed about DataUp, including upcoming presentations and demos »

#### Discover UC3:

Connect to the UC Curation Center's services for data »

#### Tweets

[Follow](#)

 **DataUp at CDL** 24m  
@DataUpCDL  
It's not too late to join the free #DataUp webinar fun! Today at 2pm PST. Pre-registration necessary:  
[cc.readytalk.com/cc/s/registrat](http://cc.readytalk.com/cc/s/registrat).

 **DataUp at CDL** 15h  
@DataUpCDL  
My interview with @OpenAtMicrosoft about #Opensource @DataUpCDL for #Excel (available today) [bit.ly/QL7eEJ](http://bit.ly/QL7eEJ)

Tweet to @DataUpCDL

Add your email to the DataUp listserv for updates

Email Address

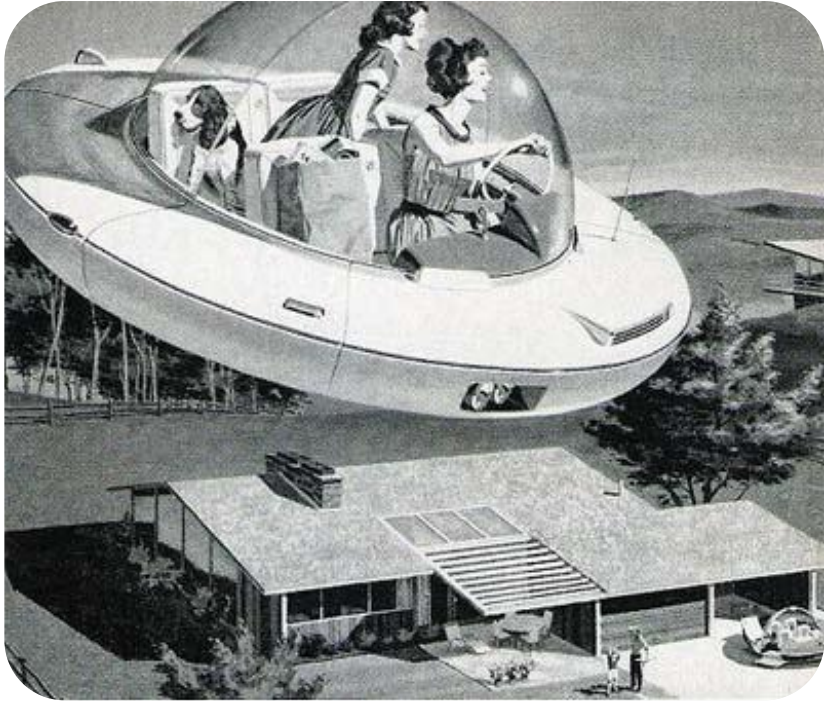
[» Submit](#)



Powered by The California Digital Library [Terms and Conditions](#) [Privacy Policy](#) [Accessibility Policy](#) [Contact Us](#) © 2012 The Regents of the University of California



From [animationresources.org](http://animationresources.org)



Build community

Add repositories


Add metadata  
schema

[dataup.cdlib.org](http://dataup.cdlib.org)

@DataUpCDL

[facebook.com/DataUpCDL](https://facebook.com/DataUpCDL)

# Data Management Plans: [dmp.cdlib.org](http://dmp.cdlib.org)




**DMPTool**  
Guidance and Resources for your Data Management Plan

[Contact Us](#) | [Sign Up](#) | [Login](#)

**NOTE: Pre-Production Release Version - Data may not be saved**

[Home](#) [About DMP Tool](#) [DMP News](#) [My Plans](#) [Funder Requirements](#) [Help](#)



**Create ready-to-use data management plans for specific funding agencies.**

Photo courtesy of Argonne National Laboratory

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

[Get Started!](#)

**Data Management Plan  
Atmospheric CO<sub>2</sub> Concentrations,  
Mauna Loa Observatory, 2011-2013**

1. Type of data produced

All 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 CO<sub>2</sub> 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 direction (SW). In addition to the CO<sub>2</sub> 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 retained.

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

**Recent DMP News**

- [UC Funding at Risk without Good DMPs](#)
- [Video demo now available](#)
- [Test Drive the DMP Tool at the ESA](#)
- [More news >](#)

---

DMPTOOL is a service of the University of California Curation Center of the California Digital Library  
Copyright © 2010-2011 The Regents of the University of California  
[Privacy Policy](#) | [Terms of Use](#) | [Photo Credits](#)



University of California

**CDL**

California Digital Library

carlystrasser.net

@carlystrasser

carlystrasser@gmail.com

Jeff Dozier

*Researcher*



UNIVERSITY OF CALIFORNIA  
SANTA BARBARA

Bill Michener

*Project Lead*



Chris Mentzel

*Funder*



Dave Vieglais

*Cyberinfrastructure dude*



Stephanie Wright

*Librarian*

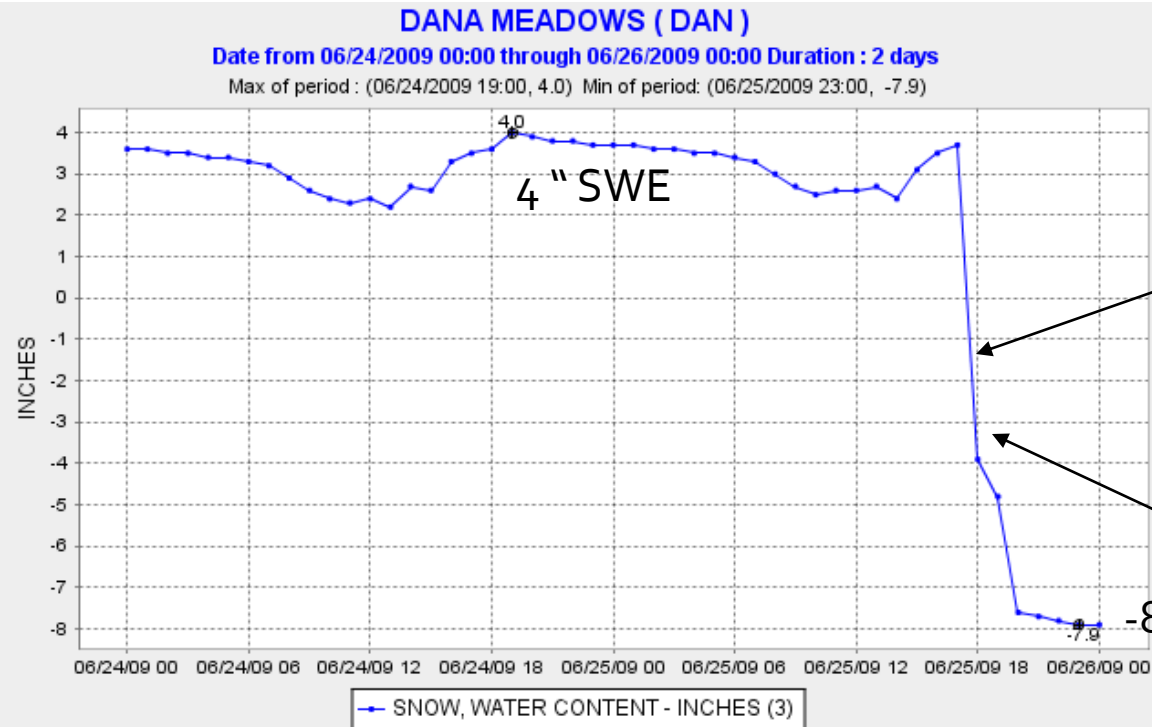




# Data quality problem:

## How to identify questionable measurements

Dana Meadows snow pillow, evening of June 25 2009



12" SWE decline after removal

(J. Lundquist)