Caltech Library

Research Data Management

Simple Ways to Make your Research Life Easier

Tom Morrell

BE/Bi 103

November 22, 2017

Current Research Data Practices

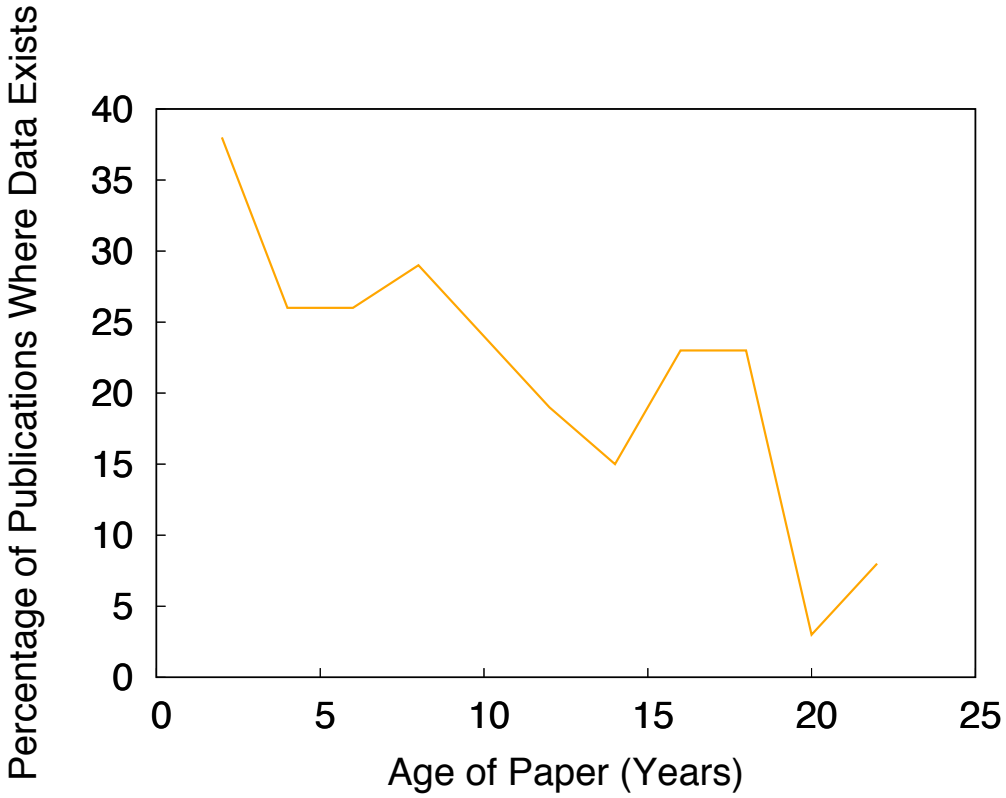
Most researchers store data on local computer hard drives

Researchers report that finding data is their biggest challenge

Akers, K. G. & Doty, J. Disciplinary differences in faculty research data management practices and perspectives. *Int. J. Digit. Curation* **8,** 5–26 (2013). (Emory)
Shen, Y. Strategic Planning for a Data-Driven, Shared-Access Research Enterprise: Virginia Tech Research Data Assessment and Landscape Study. *Coll. Res. Libr.* **77,** 500–519 (2016).

How Reusable is Research Data Today?

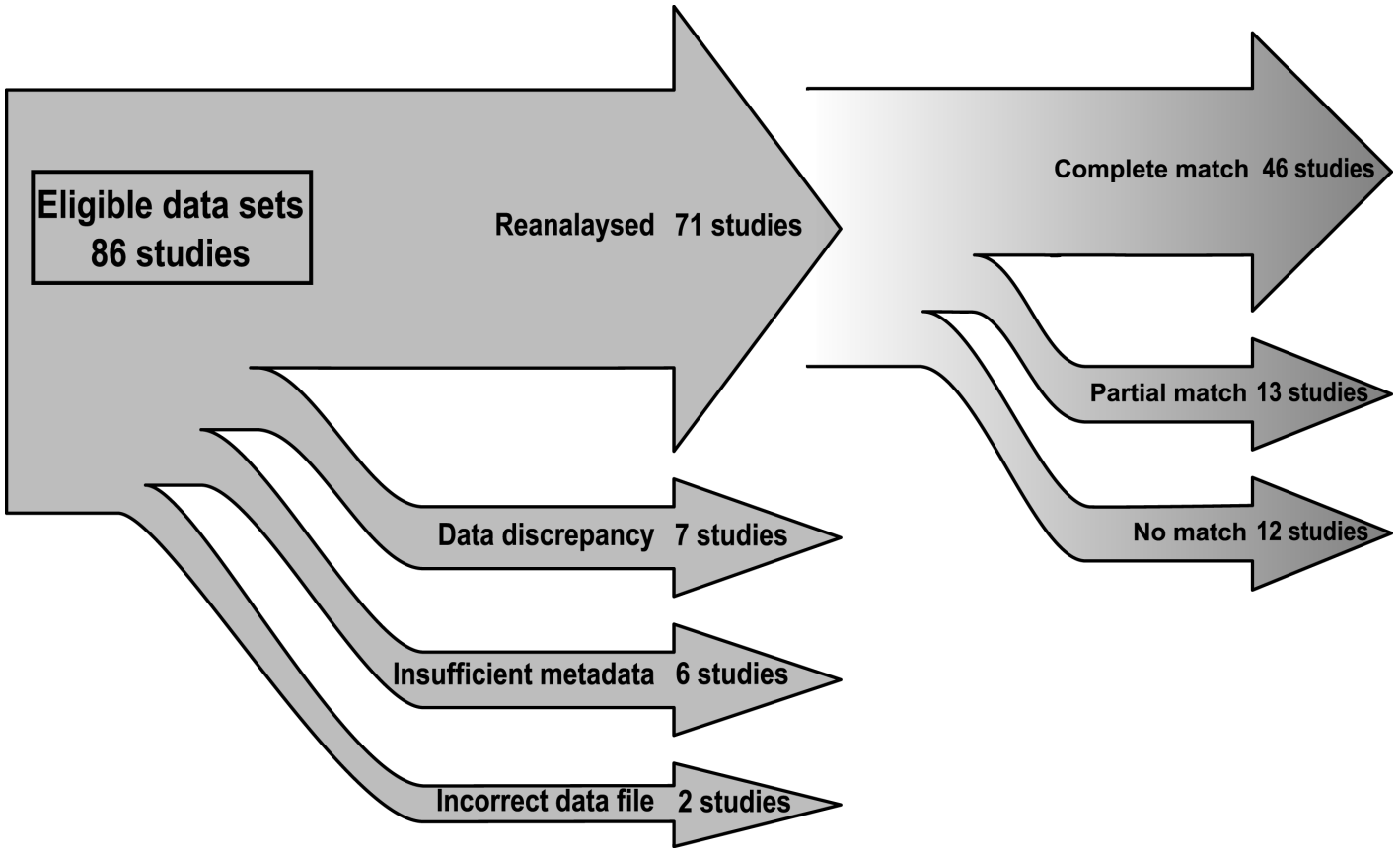
- Morphological characteristics of plants and animals
  - 516 publications using a specific analysis technique between 1991 and 2011
    - 25% of emails didn't work
    - 38% didn't respond to email
    - 13% didn't have data
    - 4% didn't want to share
    - Received 19% of data
    - Availability decreased with time



Caltech Library

Vines, T. H. *et al.* The availability of research data declines rapidly with article age. *Curr. Biol.* **24,** 94–97 (2014).

Data Quality



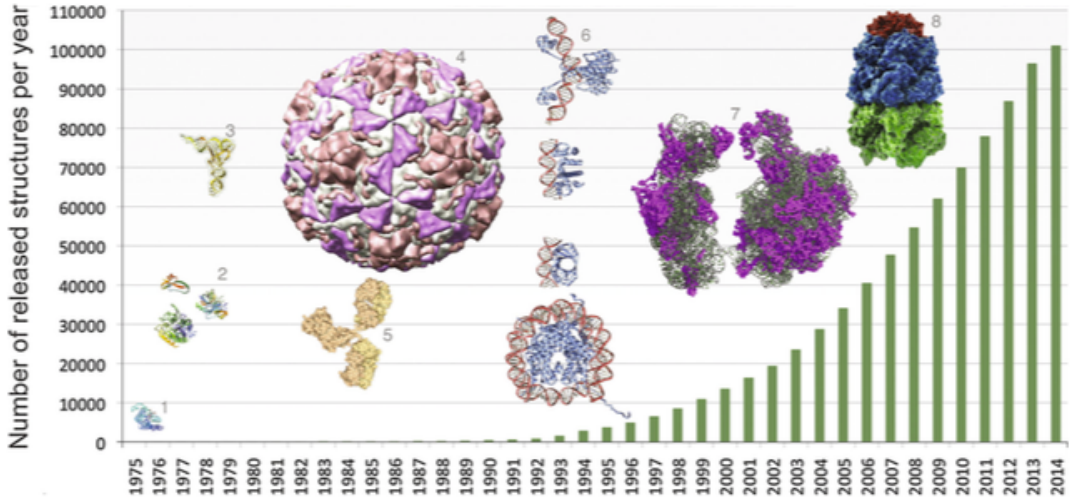
On Average, 13% of Papers Have Usable Data

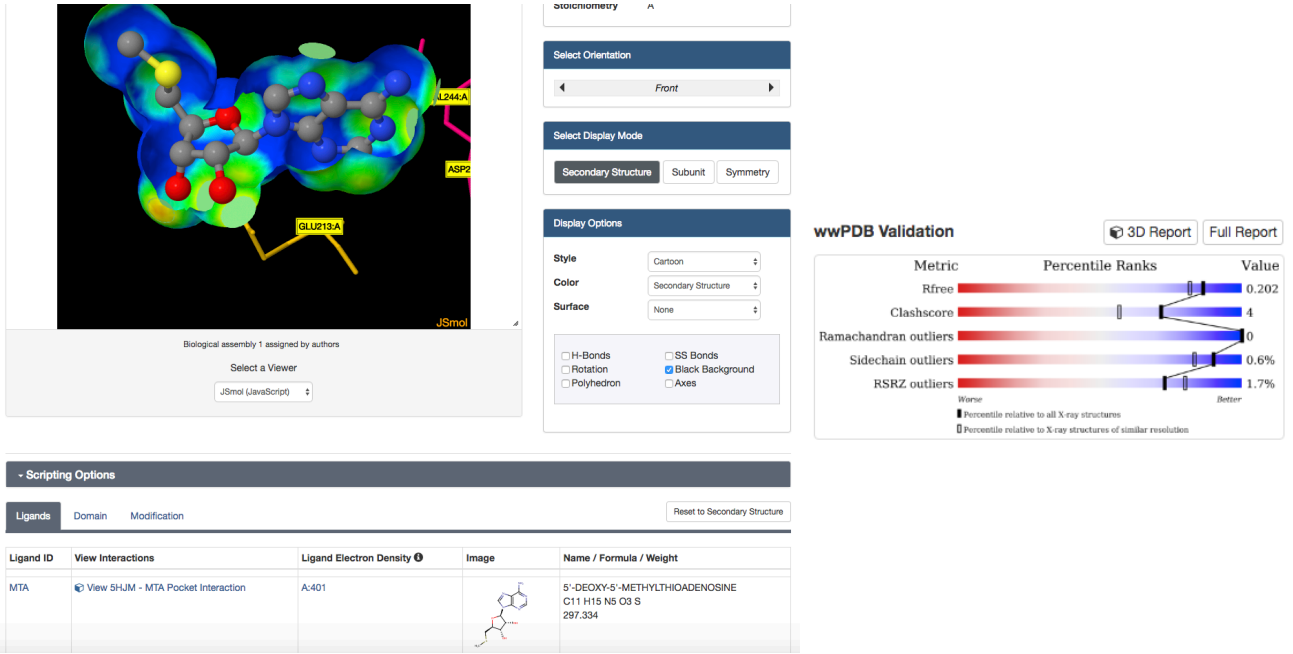
Andrew, R. L. *et al.* Assessing the reproducibility of discriminant function analyses. *PeerJ* (2015). doi:10.7717/peerj.1137

Why is it better to have data available?

RCSB PDB
PROTEIN DATA BANK

www.rcsb.org





Caltech Library

Berman, H. M., Kleywegt, G. J., Nakamura, H. & Markley, J. L. The Protein Data Bank archive as an open data resource. *J. Comput. Aided. Mol. Des.* 1009–1014 (2014). doi:10.1007/s10822-014-9770-y

Why is it better to have data available?

"Digitally formatted scientific data resulting from unclassified research supported wholly or in part by Federal funding should be stored and publicly accessible to search, retrieve, and analyze."

2013 OSTP Memo

Data Management Plans

- Expected Data
- Data Formats and Metadata
- Access to Data
- Data Archiving

https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

Why is it better to have data available?

Journals requiring complete data availability:

PLOS

The Review of Economics and Statistics

Journals requiring some data availability:

nature

Science AAAS

Cell PRESS

PNAS

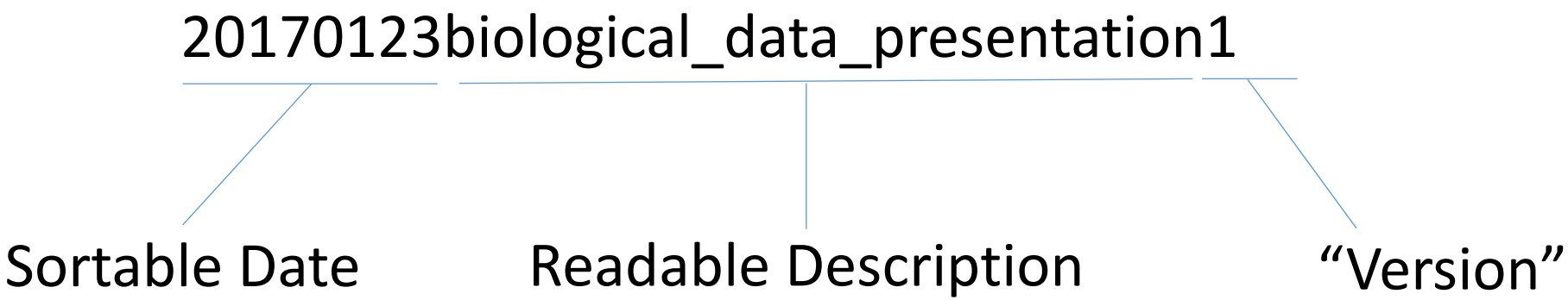
Caltech Library

Simple Solutions

- Choose a file naming/organization scheme
- Save reasonable files
- Use reliable storage
- Think about sharing

Naming

- Trying to recreate your work months/years later is hard
- Choosing a consistent naming system makes things easier



Data Architectures

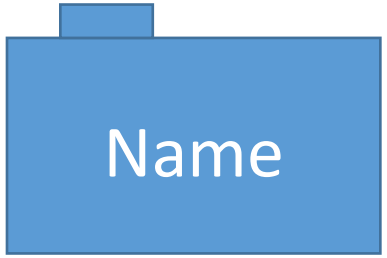
Dataset

Automatically Manage Metadata/Documents

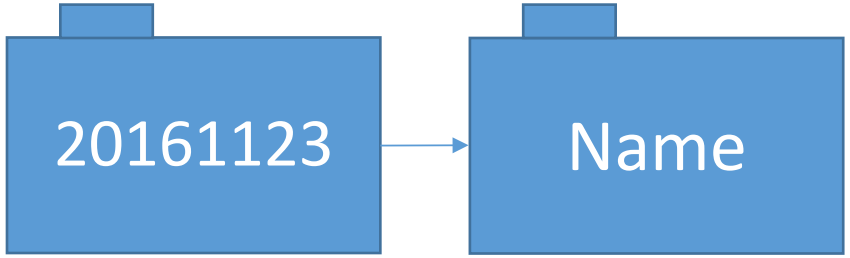
https://github.com/caltechlibrary/dataset

https://doi.org/10.22002/D1.297

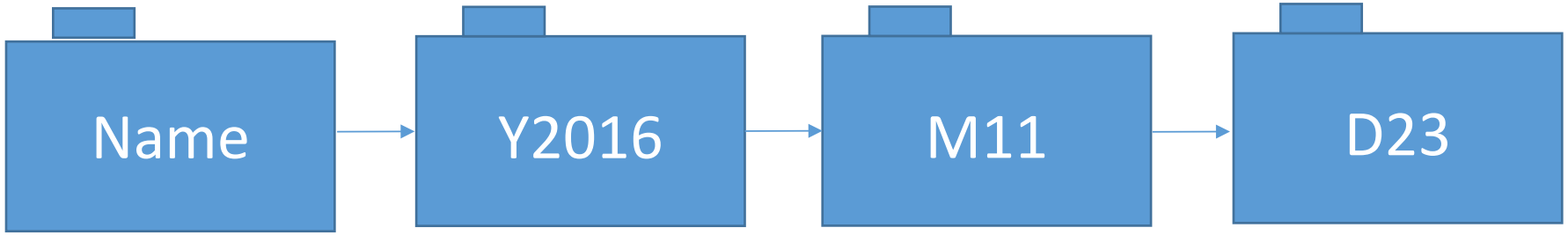
Simple



Date Based



Complex



Caltech Library

Save Reasonable Files

- Human-readable text files are best (.txt, .csv)
- Non-proprietary files are better than proprietary
- Do analysis with scripts if possible
- Save both input and output files as space allows

Active Data Storage

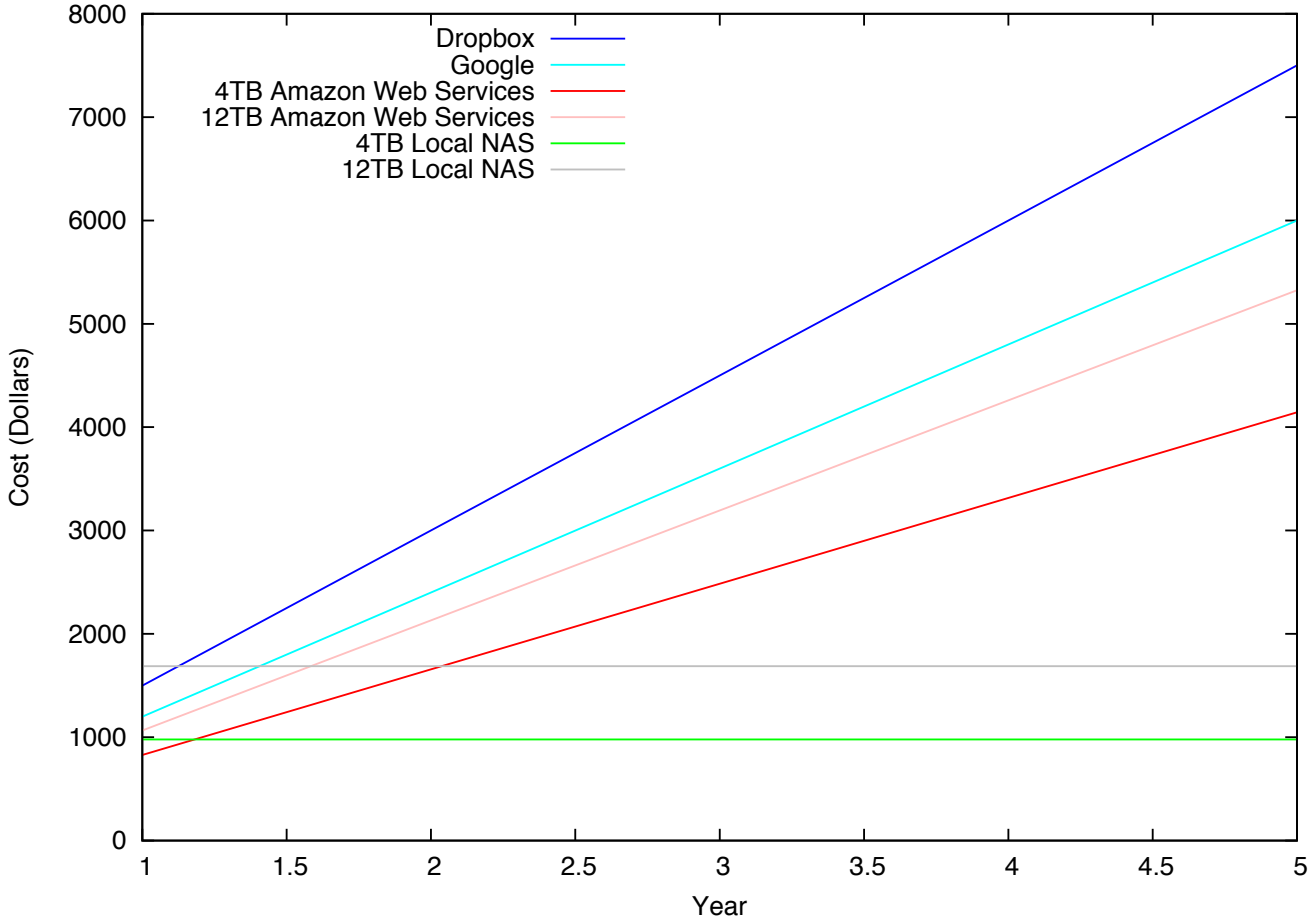
- Small amounts of data (GB) are easy
- TB-scale data require planning
  - Need a system that will be reliable
    - Network-Attached Storage (Local RAID array)
    - Cloud Storage

Network-Attached Storage

- Small computer with array of hard disks
- Consumer/Prosumer devices
- Low Cost (4 TB-$425; 42 TB-$3000)
- Need to plan space requirements
- Need to manage

Cloud Storage

- Defined or flexible storage
- Vendor Managed
- Continuous cost
- Limited by bandwidth
- Dependent on vendor



Disaster Recovery

- What Happens in a Disaster?
- Use 2 mirrored NAS units in 2 locations
- Mirror NAS to cloud storage (Box.com - imss.caltech.edu/box)

Data Sharing

- FAIR (Findability, Accessibility, Interoperability, Reusability)
  - Subject Repositories
  - General Repositories
  - Institutional Repositories

Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3:160018 doi: 10.1038/sdata.2016.18 (2016)

Subject Repositories

- Protein Data Bank
- GenBank
- Wormbase
- Pangaea
- Long Term Ecological Research Data Portal
- Good listing: journals.plos.org/plosone/s/data-availability
- Thousands more: www.re3data.org

General Repositories

- Zenodo (CERN-Free)
- Dryad (NCState-$120 per submission + Space)
- Figshare (20GB Max)
- Mendeley Data (Elsevier-Free)
- Dataverse (Harvard-Free)

Caltech Library

CaltechDATA

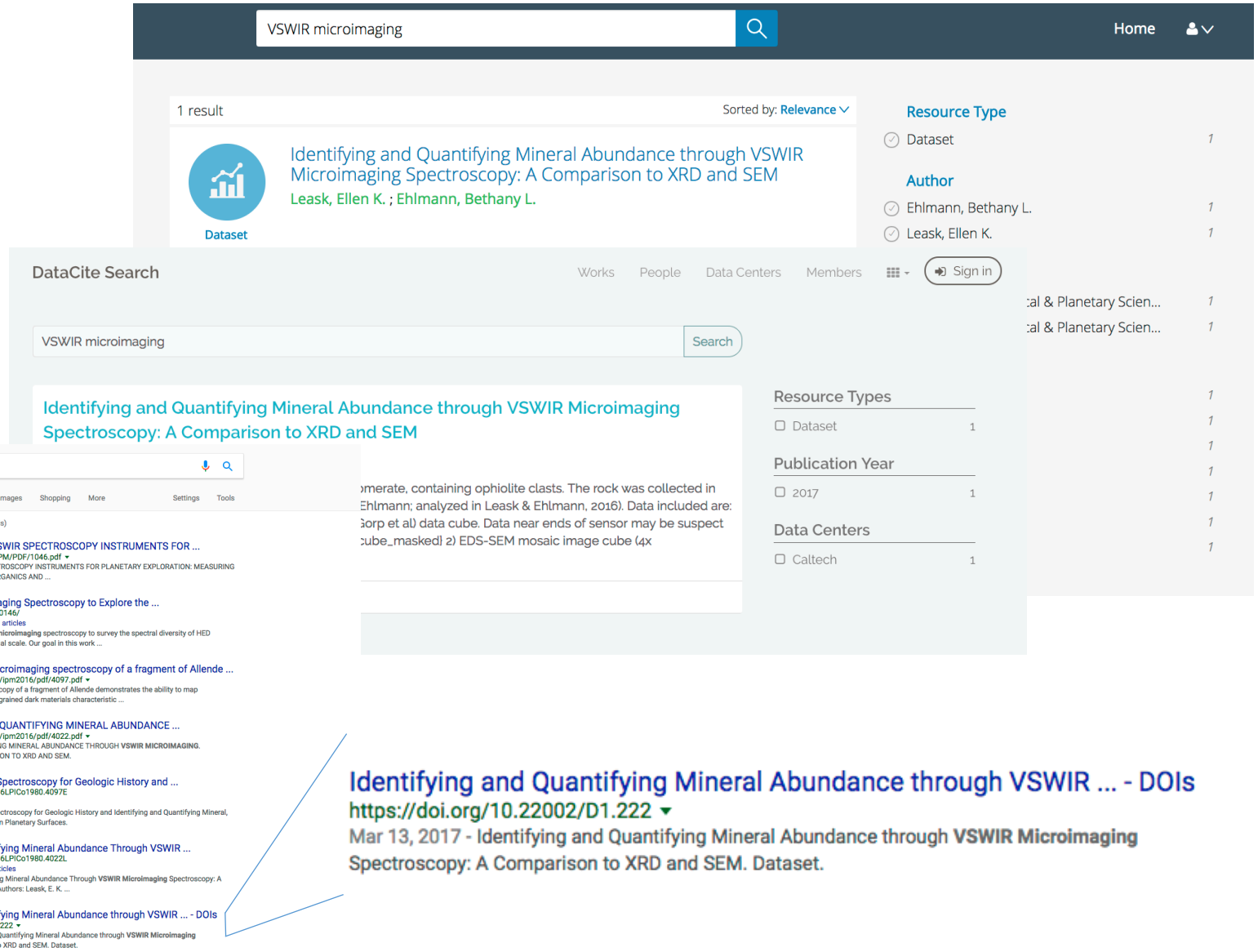
California Institute of Technology
Research Data Repository

- Available at data.caltech.edu
- Easy to describe and upload files
- All records get a DOI (permanent, registered link)
- Integration with Github
- API for accessing data
- Library takes care of preserving and maintaining access to files

Caltech Library

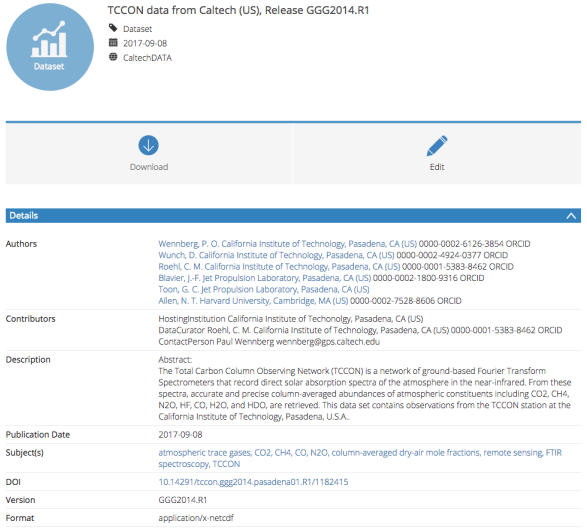
Discoverability

- CaltechDATA site search
- DOIs appear in DataCite search
- and Search Engines

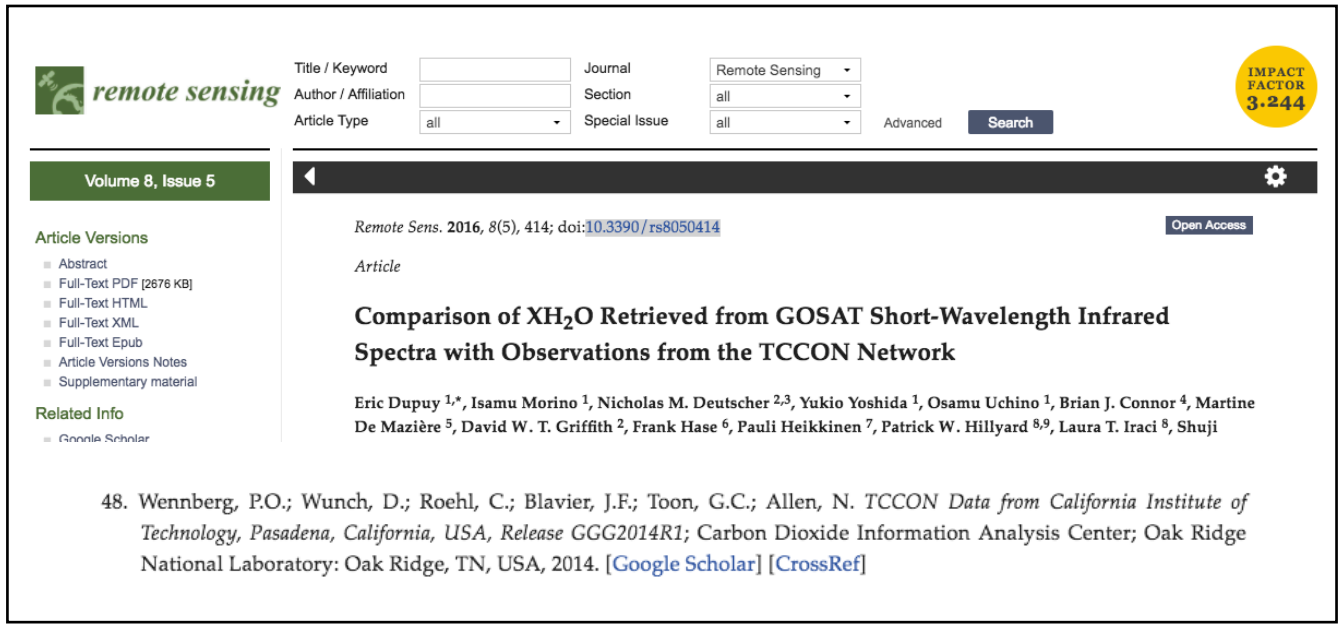


Caltech Library

Citations



California Institute of Technology
Research Data Repository



https://doi.org/10.3390/rs8050414

Citation

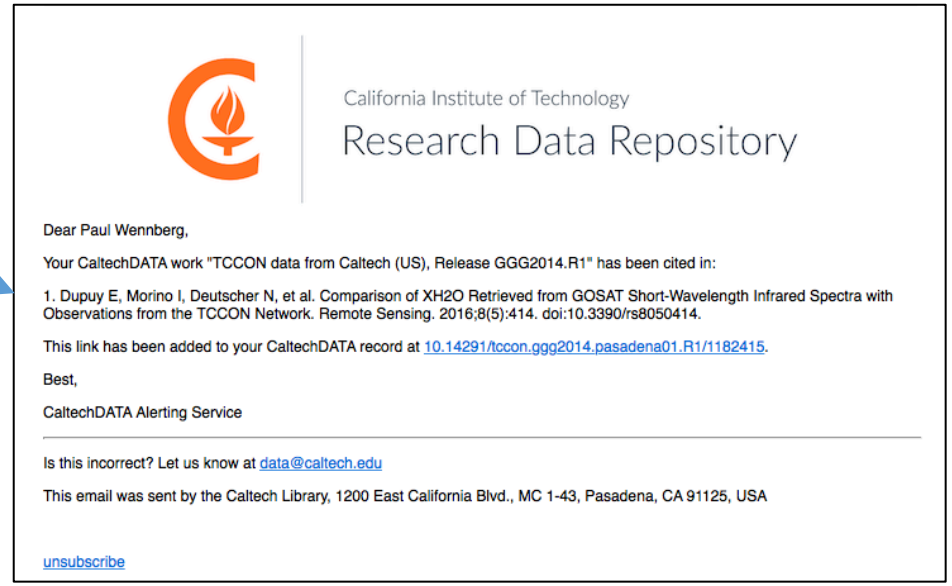
https://doi.org/10.14291/tccon.ggg2014.pasadena01.R1/1182415

Update Record

Email Alert

Related Identifier(s)

IsDocumentedBy (URL): https://tccon-wiki.caltech.edu/Network_Policy/Data_Use_Policy/Data_Description
IsDocumentedBy (URL): https://tccon-wiki.caltech.edu/Sites
IsPartOf (URL): http://tccondata.org
IsDocumentedBy (DOI): 10.14291/tccon.ggg2014.documentation.R0/1221662
IsCitedBy (DOI): 10.5194/amt-9-683-2016
IsCitedBy (DOI): 10.5194/amt-9-227-2016
IsCitedBy (DOI): 10.5194/amt-9-3491-2016
IsCitedBy (DOI): 10.5194/amt-9-3527-2016
IsNewVersionOf (DOI): 10.14291/tccon.ggg2014.pasadena01.R0/1149162
IsPartOf (DOI): 10.14291/TCCON.GGG2014
IsCitedBy (DOI): 10.3390/rs8050414



Caltech Library

California Institute of Technology

Research Data Repository

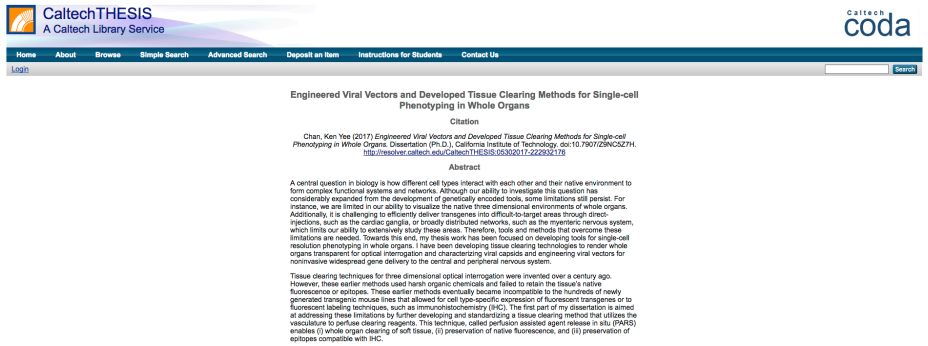
Demo

Caltech Library

Use Cases

Thesis Preparation

https://doi.org/10.7907/Z9NC5Z7H



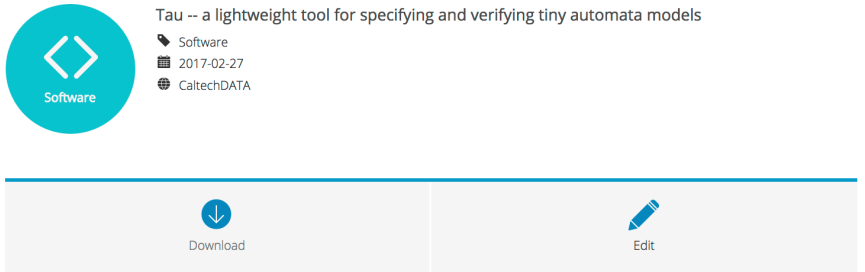
Data files uploaded during writing process

California Institute of Technology
Research Data Repository

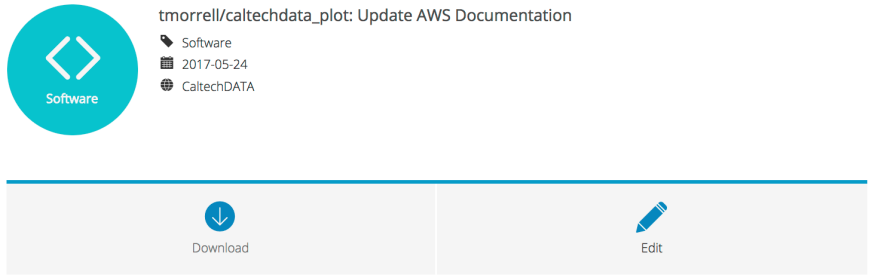
https://doi.org/10.22002/D1.234
https://doi.org/10.22002/D1.235
https://doi.org/10.22002/D1.236
https://doi.org/10.22002/D1.237

Software in CaltechDATA

https://doi.org/10.22002/D1.218



https://doi.org/10.22002/D1.240

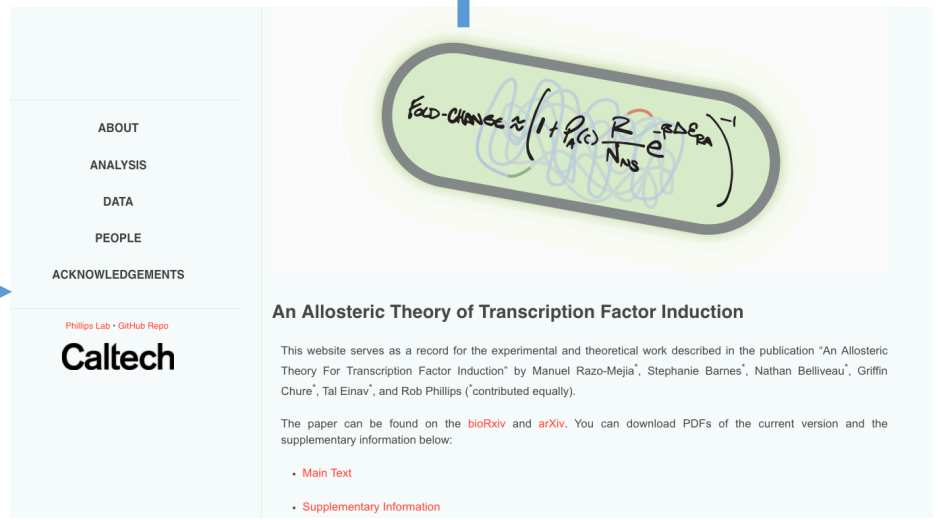


Caltech Library

Use Cases

https://doi.org/10.1101/111013

Paper Website on GitHub



Data Files

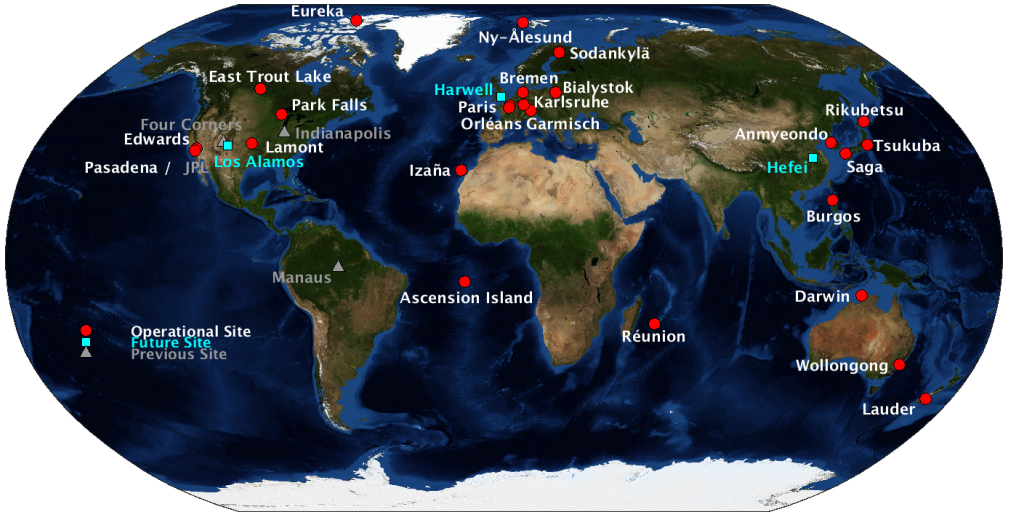
https://rpgroup-pboc.github.io/mwc_induction
https://doi.org/10.22002/D1.299

California Institute of Technology
Research Data Repository

https://doi.org/10.22002/D1.224
https://doi.org/10.22002/D1.227
https://doi.org/10.22002/D1.228
https://doi.org/10.22002/D1.229

Caltech Library

Use Case - TCCON



Total Carbon Column Observing Network (TCCON)
29 Data Collection Sites Around the World

Data files

~~CDIAC Carbon Dioxide Information Analysis Center~~

~~OAK RIDGE National Laboratory~~

Processed Data

Data Curation and Processing

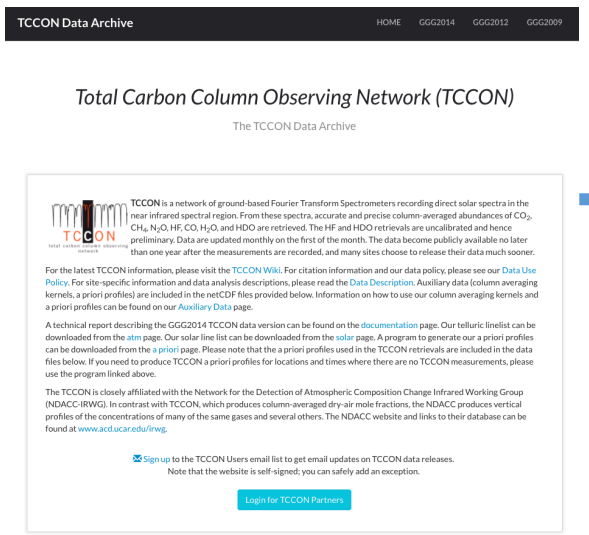
Caltech Library

https://tccon-wiki.caltech.edu/Sites/Park_Falls
http://tccondata.org/

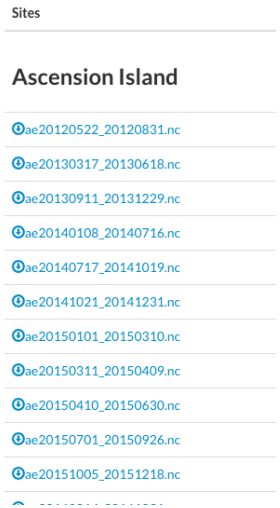
Use Case - TCCON

OAK RIDGE National Laboratory

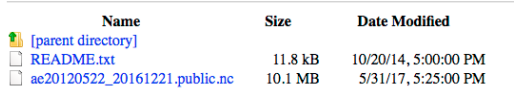
tccon.ornl.gov



Private data files



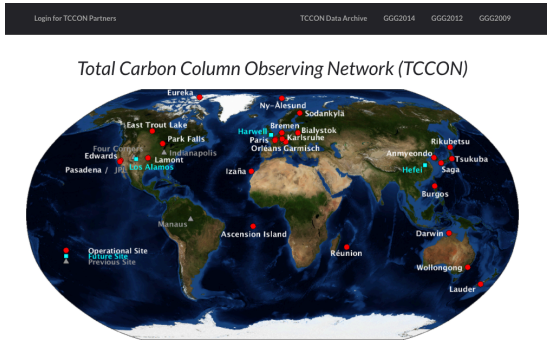
Public data files



Automatically released 1x/month

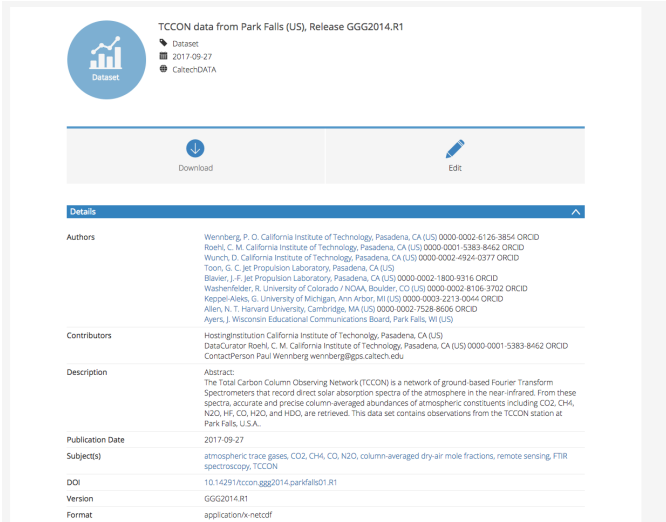
Migration

Departmental Server at Caltech



tccondata.org

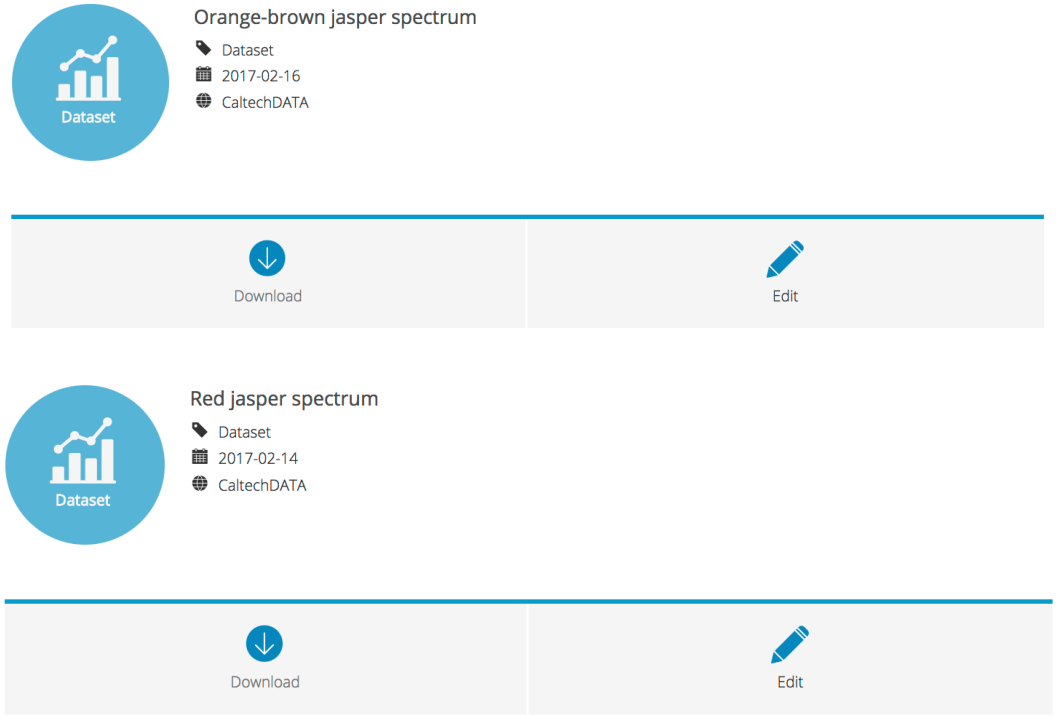
CaltechDATA

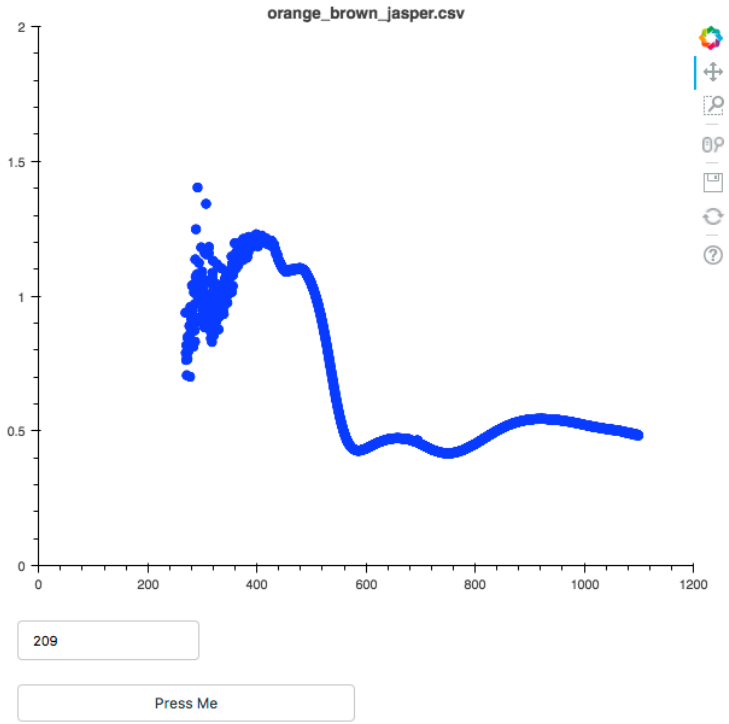


Caltech Library

https://doi.org/10.14291/tccon.ggg2014.parkfalls01.R1

Demo – CaltechDATA + Interactive Plotting





plots.caltechlibrary.org

doi.org/10.22002/D1.240

Caltech Library

Caltech Library Data Management Services

- Want to chat about data issues?
- Data management plan development
- Consultations on storage technologies or file organization

data@caltech.edu

tmorrell@caltech.edu

626-395-3827

Research Software Workshop

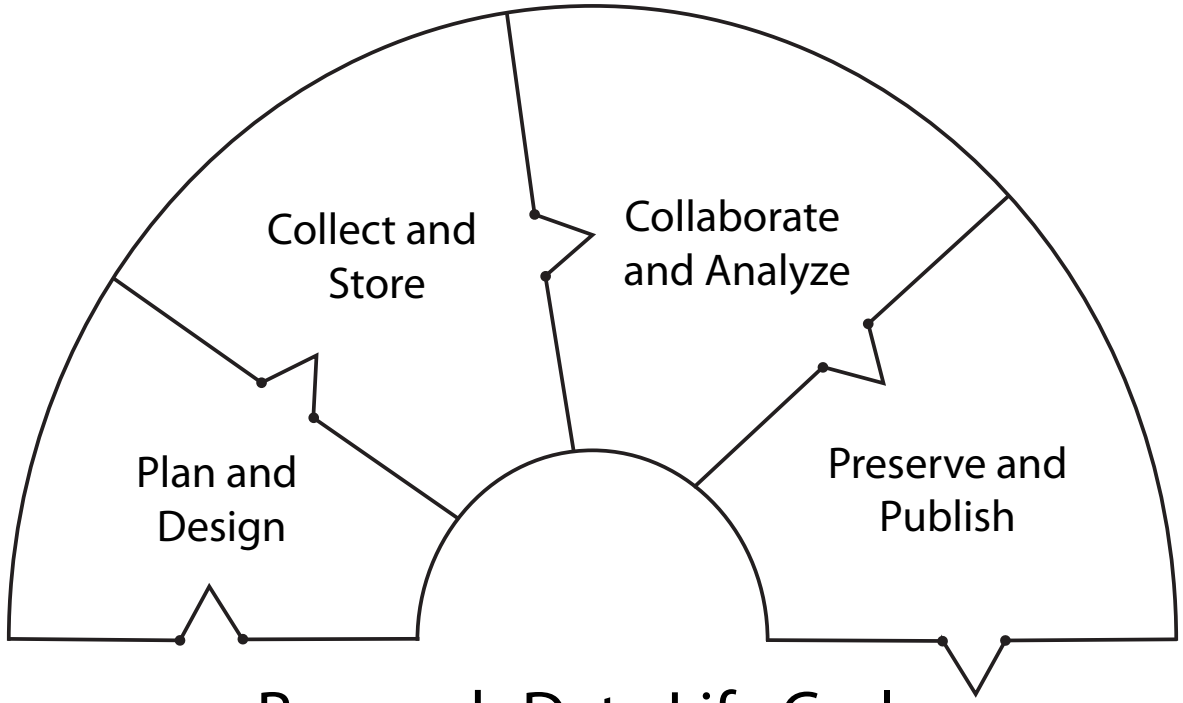
Dan Katz (NCSA)

January 31st 12-2

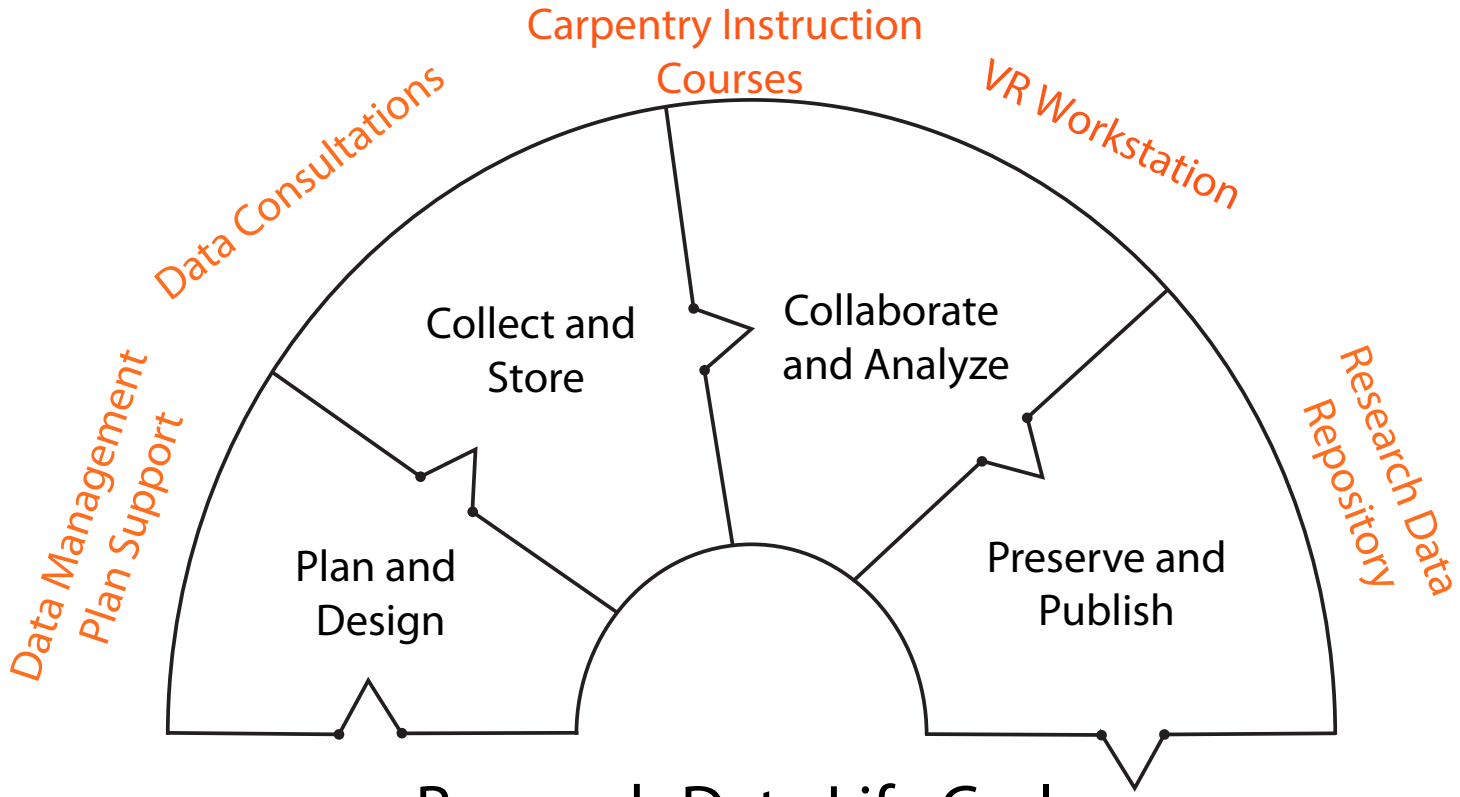
Lunch Provided

Registration Required

Caltech Library



Research Data Life Cycle



Research Data Life Cycle

and Caltech Library Services

Caltech Library

Things to Think About

- Choose a file naming/organization scheme
- Save reasonable files
- Use reliable storage
- Think about sharing