Caltech Library

Research Data Management

Simple Ways to Make your Research Life Easier

Tom Morrell

BE/Bi 103

October 13, 2021

https://doi.org/10.5281/zenodo.5565508

Current Research Data Practices

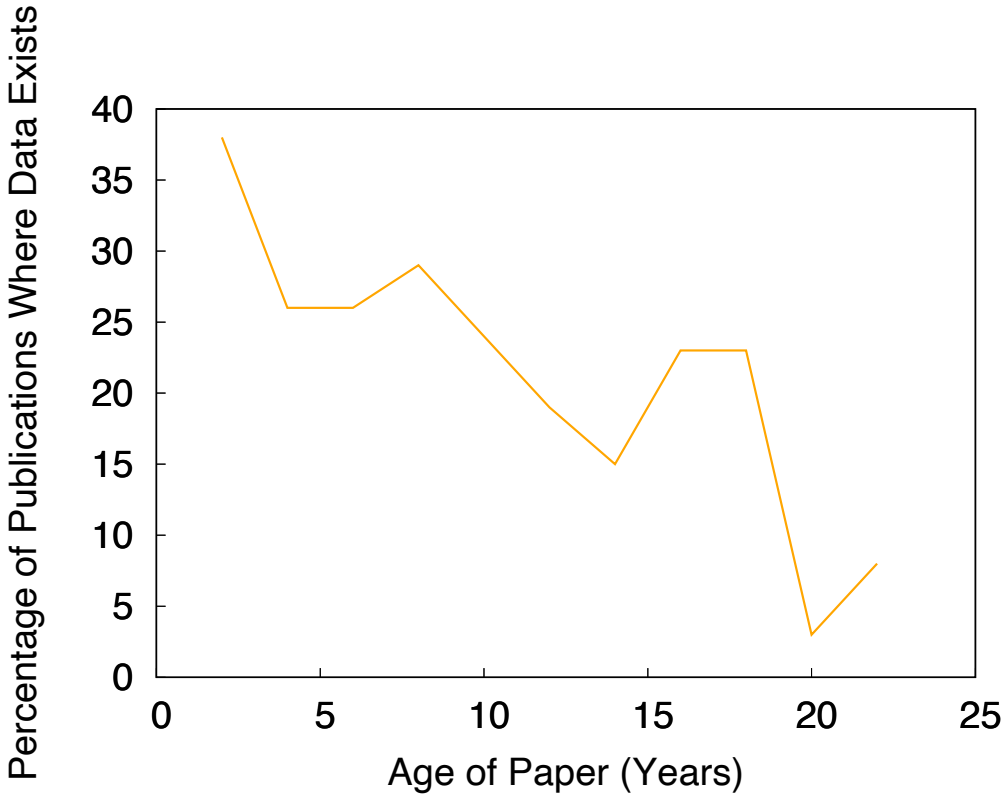
Most researchers store data on local computer hard drives

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). doi: 10.2218/ijdc.v8i2.263 (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). doi: 10.5860/crl.77.4.500

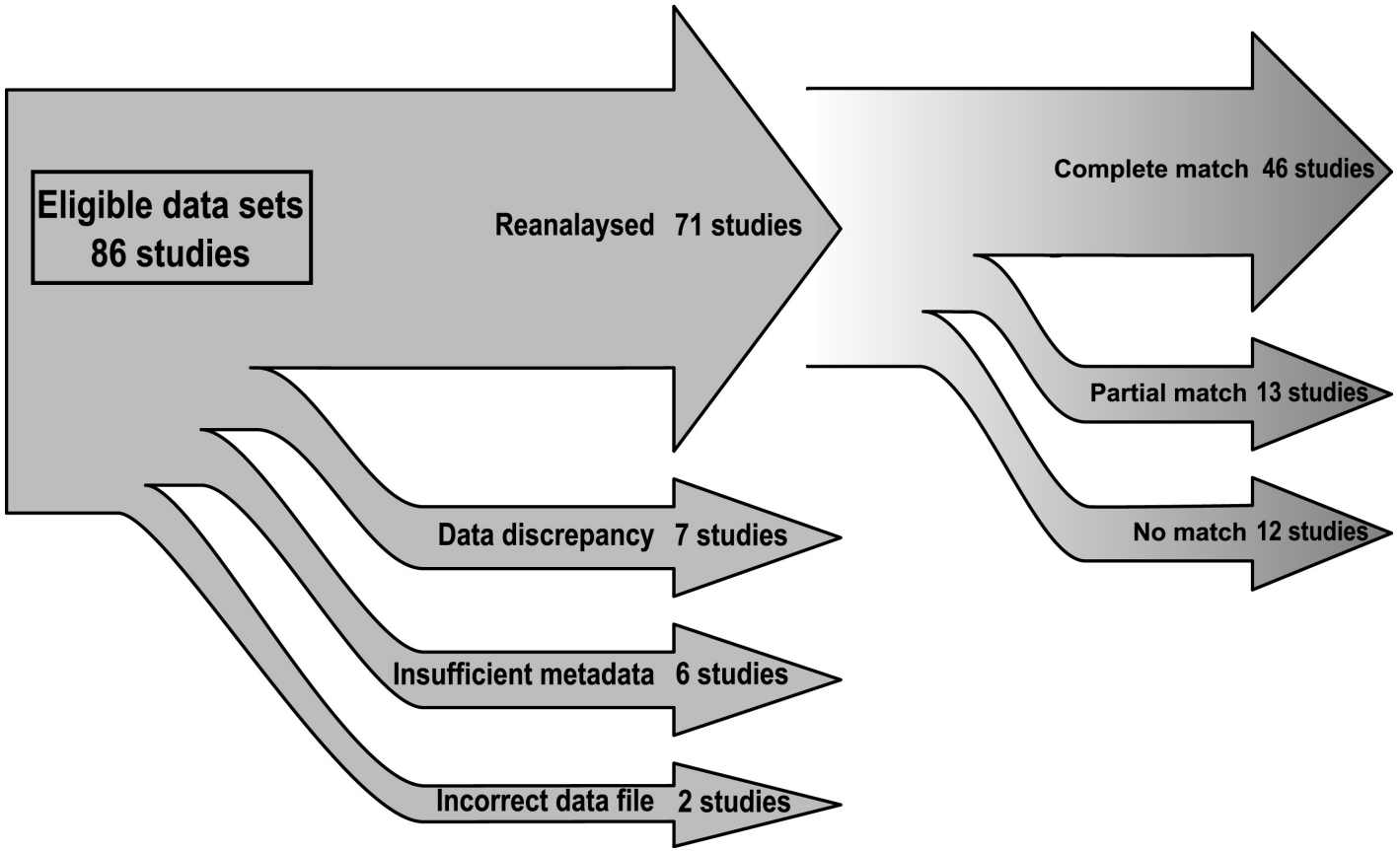
How Reusable is Research Data Today?

- 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



Vines, T. H. *et al.* The availability of research data declines rapidly with article age. *Curr. Biol.* **24,** 94–97 (2014). doi: 10.1016/j.cub.2013.11.014

Data Quality



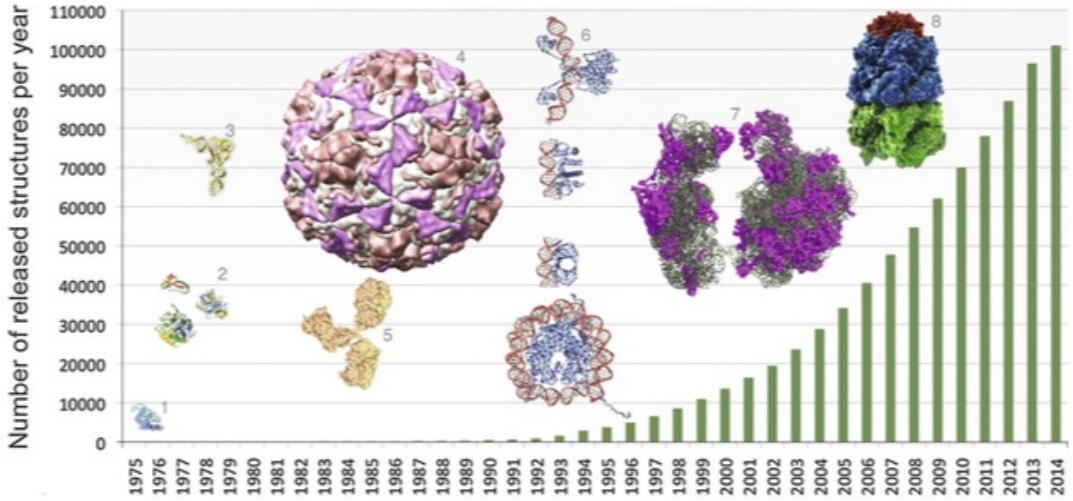
On Average, 13% of Papers Had Usable Data

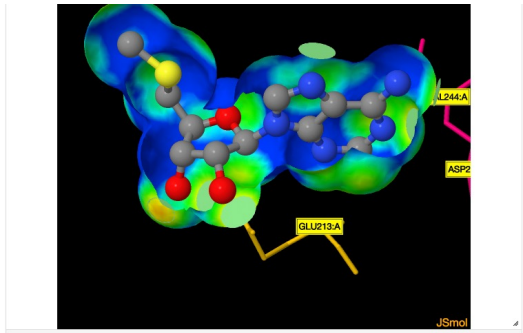
Andrew, R. L. *et al.* Assessing the reproducibility of discriminant function analyses. *PeerJ* (2015). doi:10.7717/peerj.1137

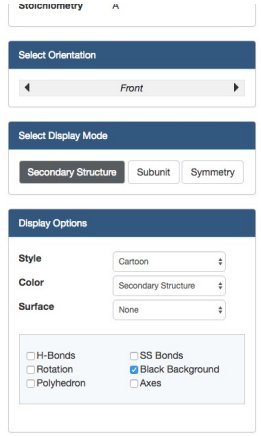
Why is it better to have data available?

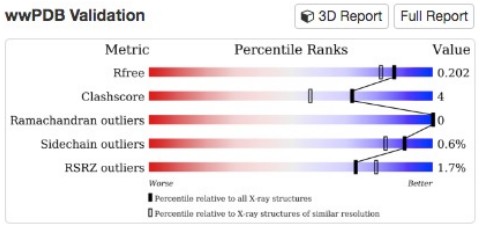
RCSB PDB
PROTEIN DATA BANK

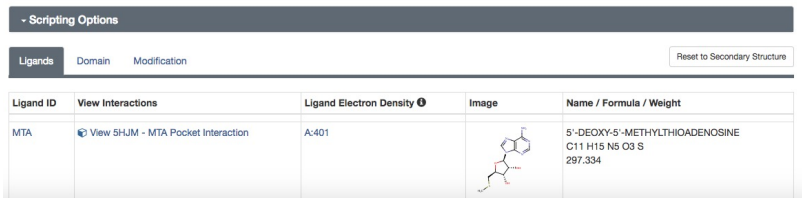
www.rcsb.org











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.science.gov/docs/ostp_public_access_memo_2013.pdf

Why is it better to have data available?

Journals require data availability:

Commitment Statement in the Earth, Space, and Environmental Sciences for depositing and sharing data

Caltech Library

https://copdess.org/enabling-fair-data-project/commitment-statement-in-the-earth-space-and-environmental-sciences/

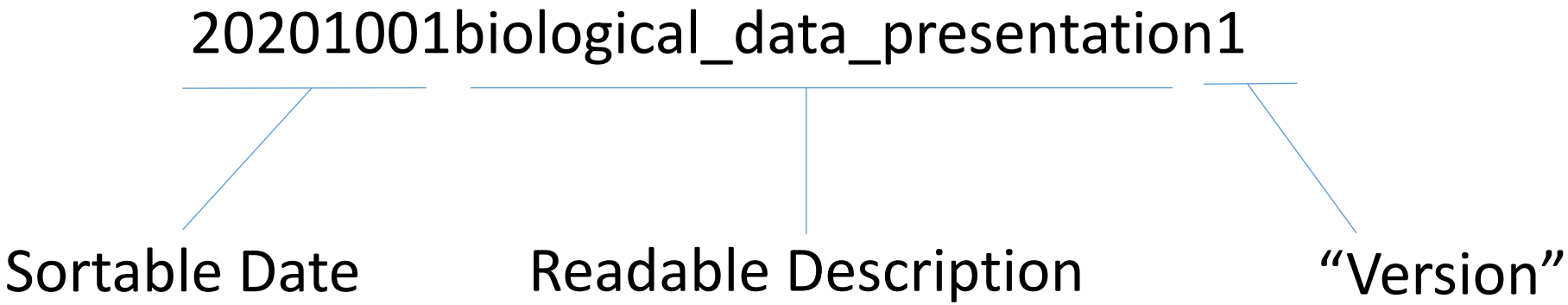
Stall, S. *et. al.* Make scientific data FAIR *Nature.* 570, 27-29 (2019). doi: 10.1038/d41586-019-01720-7

Simple Solutions

- Choose a file naming/organization scheme
- Save reasonable files
- Use reliable storage
- Plan for sharing

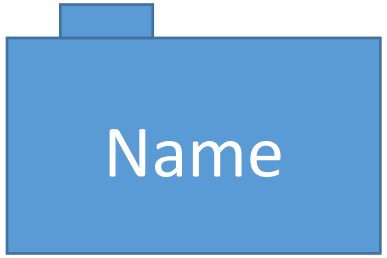
Naming

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



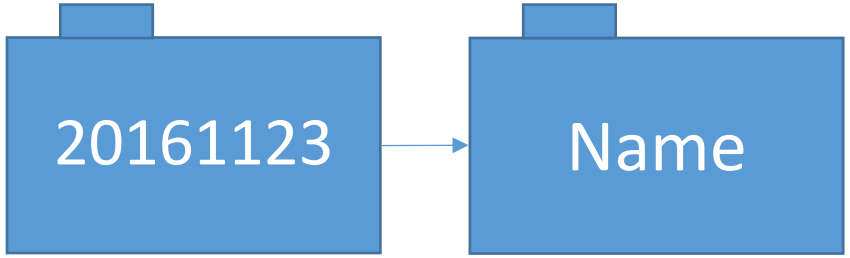
Data Architectures

Simple

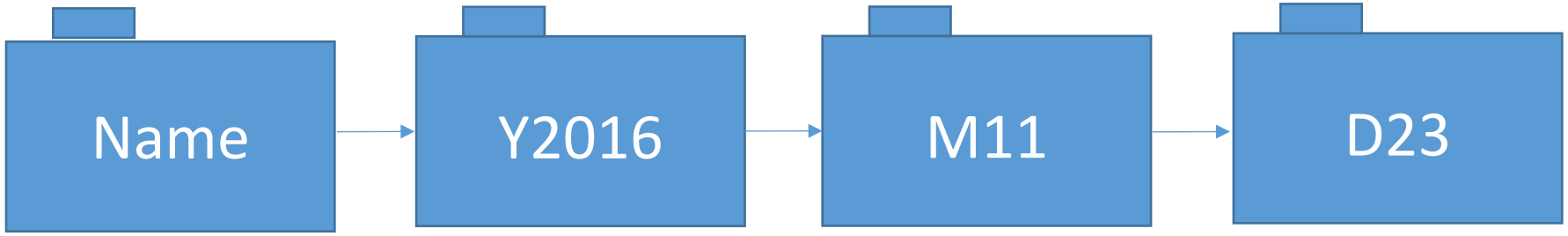


Full worksheet at:
https://resolver.caltech.edu/CaltechAUTHORS:20200601-161923247

Date Based



Complex



Caltech Library

Metadata

- What other information will be useful for reanalysis?
- Use standard terms (https://fairsharing.org/)
- Store with data
  - README template
  - JSON or XML document
  - Dataset – tool for managing collections of metadata and documents https://github.com/caltechlibrary/dataset

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

Document Software Dependencies

- Changing software dependencies can impact your analysis
- Document what version of software you're using
- Test with a reproducibility service like binder

requirements.txt

```
1  requests>=2.21
2  ames>=0.4.0
3  progressbar2
4  plotnine
5  pandas
6  numpy
7  scipy==1.2.2
```

mybinder.org

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

Local vs Cloud Storage

Local Storage

- RAID array can protect against data loss
- Reasonably 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 (AWS Glacier Deep Archive)

Earthquake damage a Cal State Northridge, 1994 by Stickpen, CC0 License

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 (Nonprofit-$120 per submission + Space)
- Figshare (20GB Max)
- Mendeley Data (Elsevier-Free)
- Dataverse (Harvard-Free)

CaltechDATA

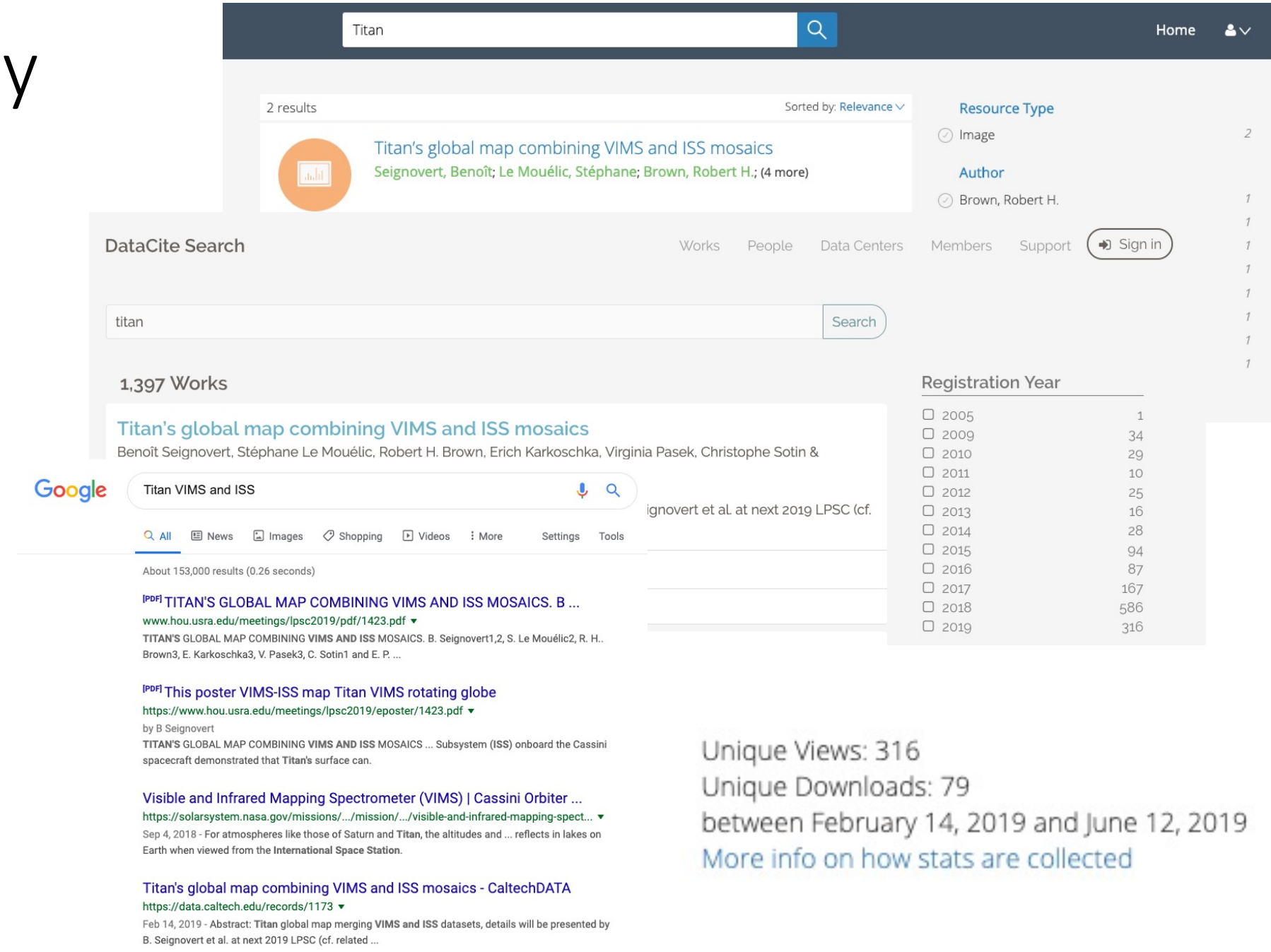
California Institute of Technology
Research Data Repository

- 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

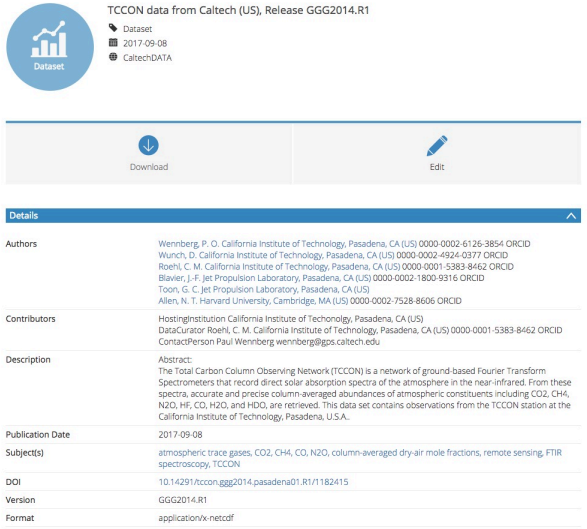
Discoverability

- CaltechDATA site search
- DOIs appear in DataCite search
- Broad discoverability
- We track standardized views and downloads

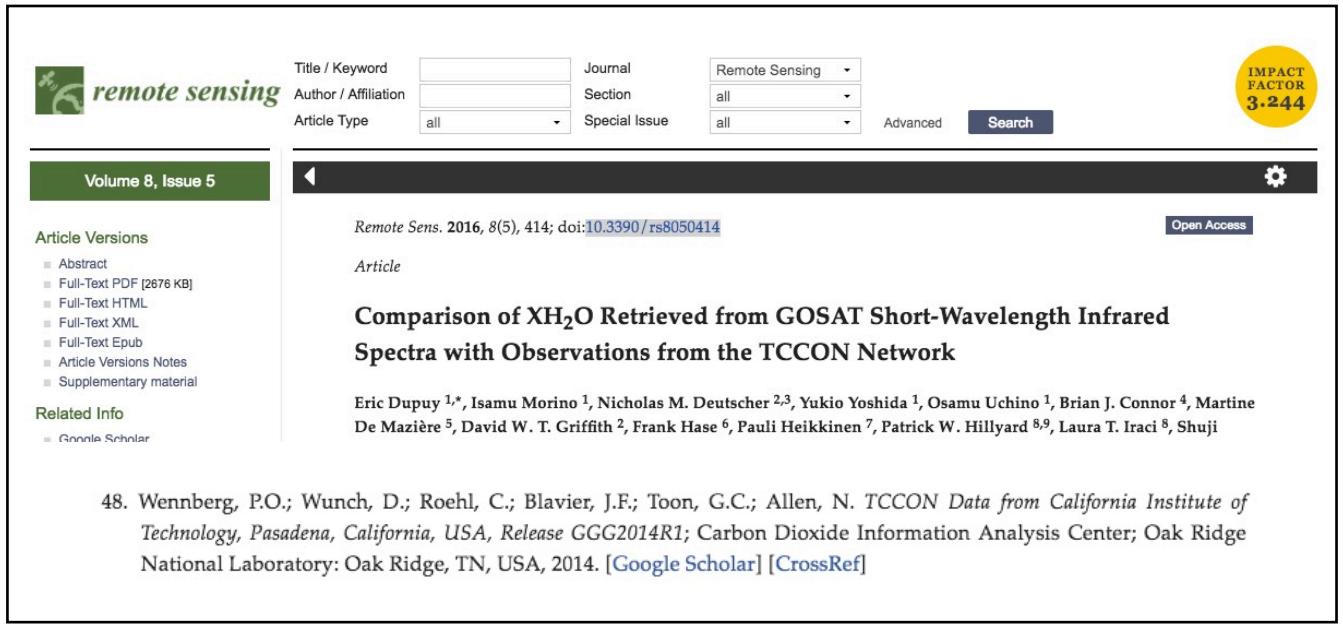


Caltech Library

Citations



California Institute of Technology
Research Data Repository



https://doi.org/10.3390/rs8050414

Citation

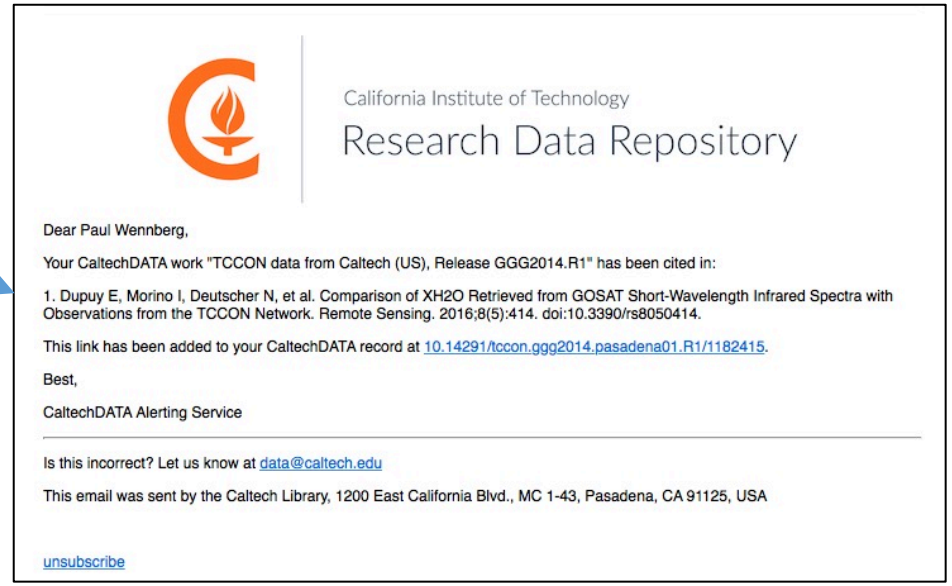
https://doi.org/10.14291/tccon.ggg2014.pasadena01.R1/1182415

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 - Theses

Upload files while writing

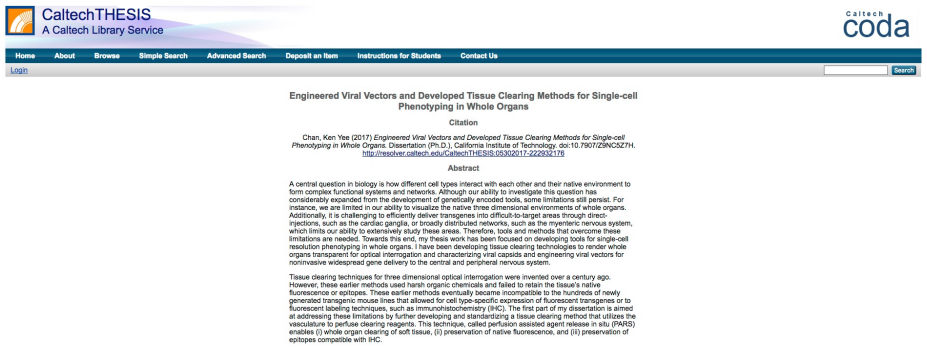
California Institute of Technology
Research Data Repository

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

Link in thesis

Link automatically added to CaltechDATA

https://doi.org/10.7907/Z9NC5Z7H


CaltechTHESIS
A Caltech Library Service

Caltech coda

Home About Browse Simple Search Advanced Search Deposit an Item Instructions for Students Contact Us

Login Search

Engineered Viral Vectors and Developed Tissue Clearing Methods for Single-cell Phenotyping in Whole Organs

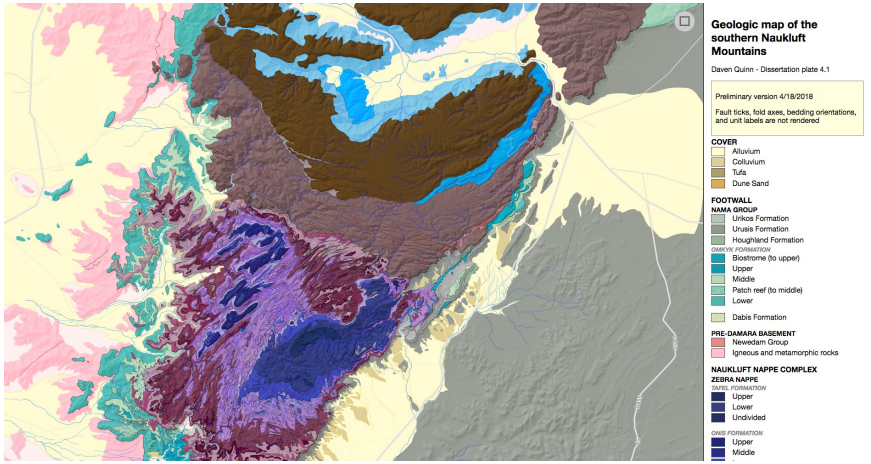
Citation

Chan, Ken Yee (2017) *Engineered Viral Vectors and Developed Tissue Clearing Methods for Single-cell Phenotyping in Whole Organs*. Dissertation (Ph.D.), California Institute of Technology. doi:10.7907/Z9NC5Z7H. http://resolver.caltech.edu/CaltechTHESIS:05302017-222932176

Abstract

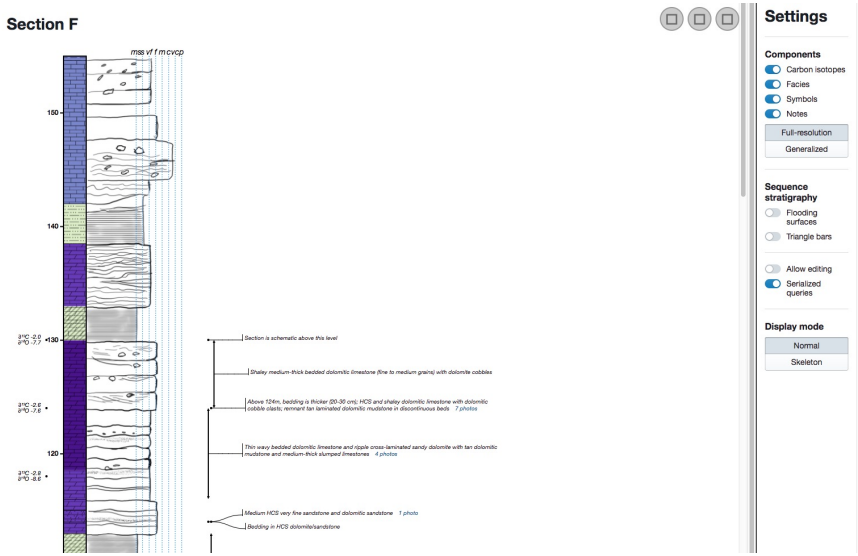
A central question in biology is how different cell types interact with each other and their native environment to form complex functional systems and networks. Although our ability to investigate this question has considerably expanded from the development of genetically encoded tools, some limitations still persist. For instance, we are limited in our ability to visualize the native three dimensional environments of whole organs. Additionally, it is challenging to efficiently deliver transgenes into difficult-to-target areas through direct-injections, such as the cardiac ganglia, or broadly distributed networks, such as the myenteric nervous system, which limits our ability to extensively study these areas. Therefore, tools and methods that overcome these limitations are needed. Towards this end, my thesis work has been focused on developing tools for single-cell resolution phenotyping in whole organs. I have been developing tissue clearing technologies to render whole organs transparent for optical interrogation and characterizing viral capsids and engineering viral vectors for noninvasive widespread gene delivery to the central and peripheral nervous system.

Tissue clearing techniques for three dimensional optical interrogation were invented over a century ago. However, these earlier methods used harsh organic chemicals and failed to retain the tissue's native fluorescence or epitopes. These earlier methods eventually became incompatible to the hundreds of newly generated transgenic mouse lines that allowed for cell type-specific expression of fluorescent transgenes or to fluorescent labeling techniques, such as immunohistochemistry (IHC). The first part of my dissertation is aimed at addressing these limitations by further developing and standardizing a tissue clearing method that utilizes the vasculature to perfuse clearing reagents. This technique, called perfusion assisted agent release in situ (PARS) enables (i) whole organ clearing of soft tissue, (ii) preservation of native fluorescence, and (iii) preservation of epitopes compatible with IHC.



https://doi.org/10.7907/5exk-mr58

https://doi.org/10.22002/D1.946



https://doi.org/10.7907/9kva-eq78

https://doi.org/10.22002/D1.947

Caltech Library

Use Cases

https://doi.org/10.1101/111013

Paper Website on GitHub

Data Files

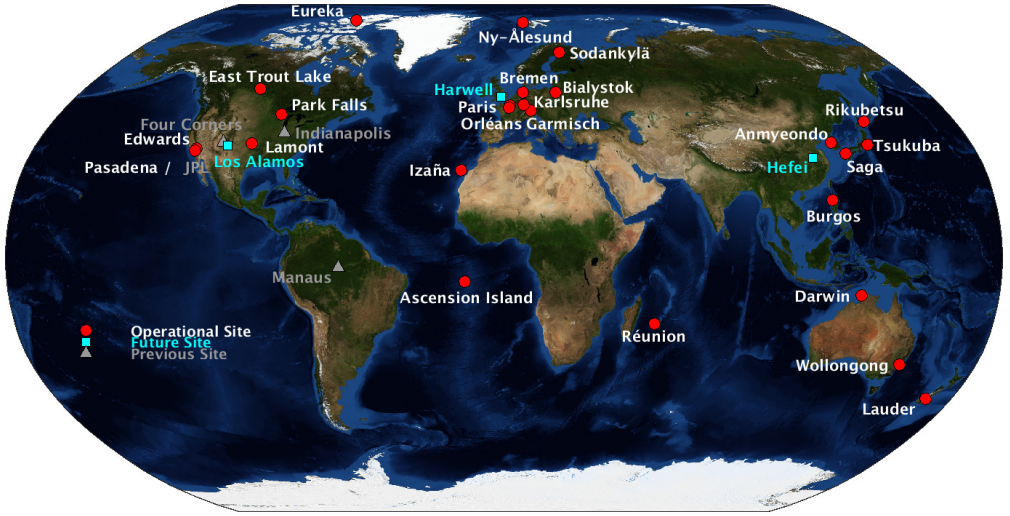
https://rpgroup-pboc.github.io/mwc_induction
https://doi.org/10.22002/D1.299

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

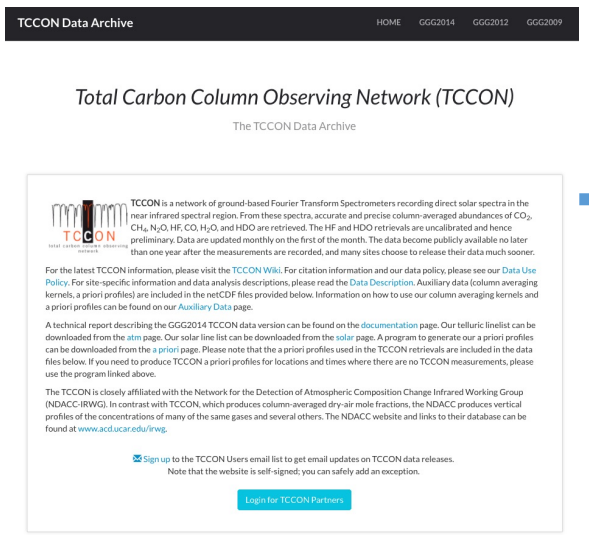
Caltech Library

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

Use Case - TCCON

OAK RIDGE National Laboratory

tccon.ornl.gov



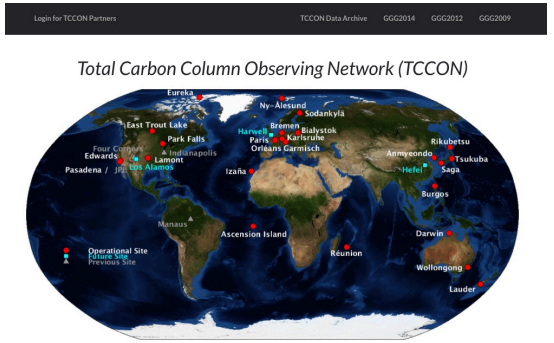
Private data files

Public data files

Automatically released 1x/month

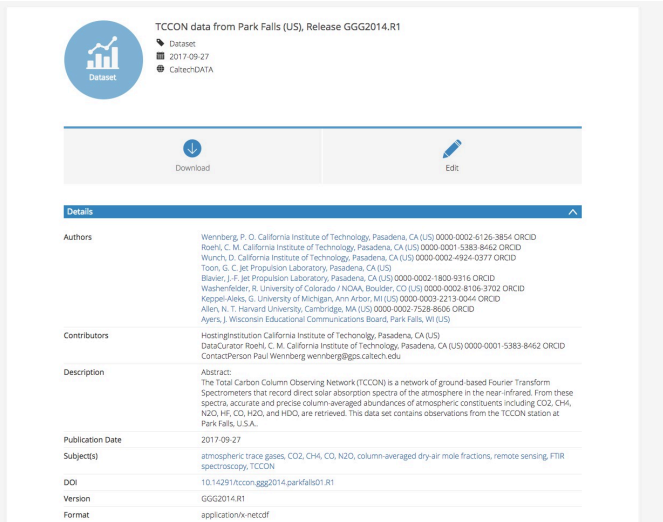
Migration

Departmental Server at Caltech



tccondata.org

CaltechDATA



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

Caltech Library

California Institute of Technology
Research Data Repository

Caltech Division of Geological and Planetary Sciences Theses

This map shows the coordinates of content in CaltechDATA associated with theses from the Geological and Planetary Science Division at Caltech. Data included from historic theses are supplemental pocket contents such as maps and drawings.

Scrolling inside the map will zoom and dragging will move the map. Click on any point or bounding box to see the original item in CaltechDATA.

Want your thesis to show up on the map?

Upload files associated with your thesis to CaltechDATA and include a geolocation point or area. You'll also have to include the keywords 'gps' and 'thesis' in the record. If you run into any problems just send us an email.

Did you complete your thesis in the Caltech GPS Division?

We haven't been able to assign locations for every thesis. Send us an email and we can get your thesis on the map.

Want to improve this map?

The code to generate the map is available on GitHub and we accept pull requests for improvements.

http://maps.library.caltech.edu/

https://doi.org/10.22002/D1.856

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

Caltech Library