

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

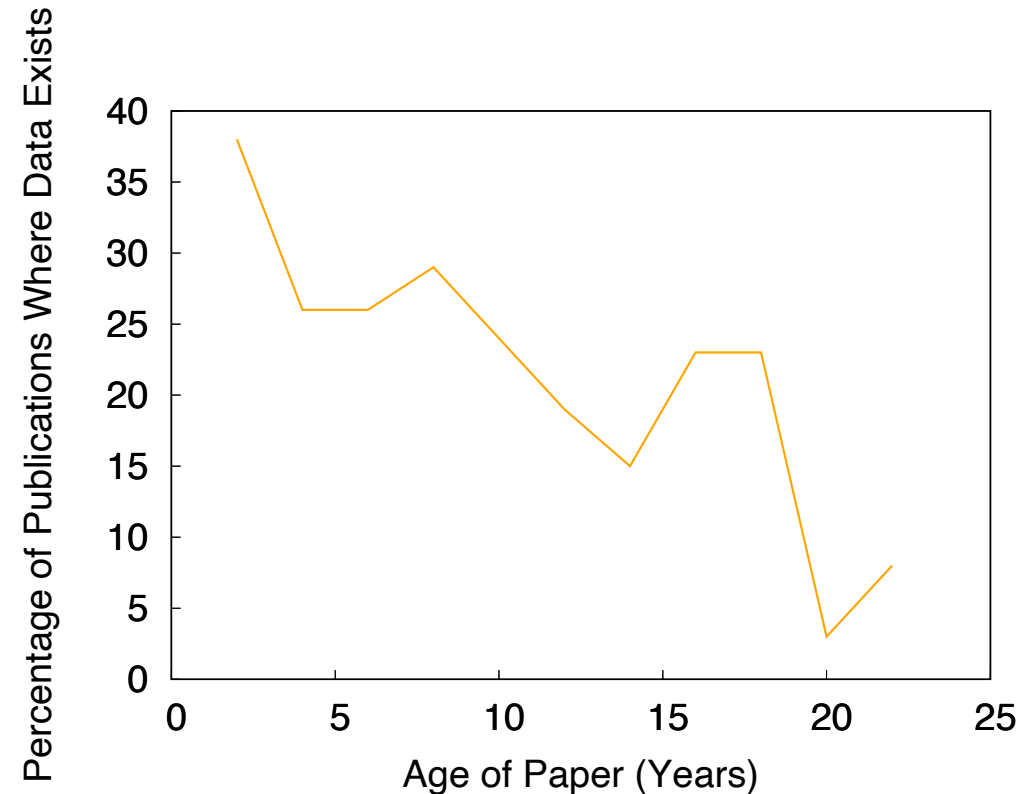
Researchers report that finding data is their biggest challenge

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](https://doi.org/10.2218/ijdc.v8i2.263) (Emory)

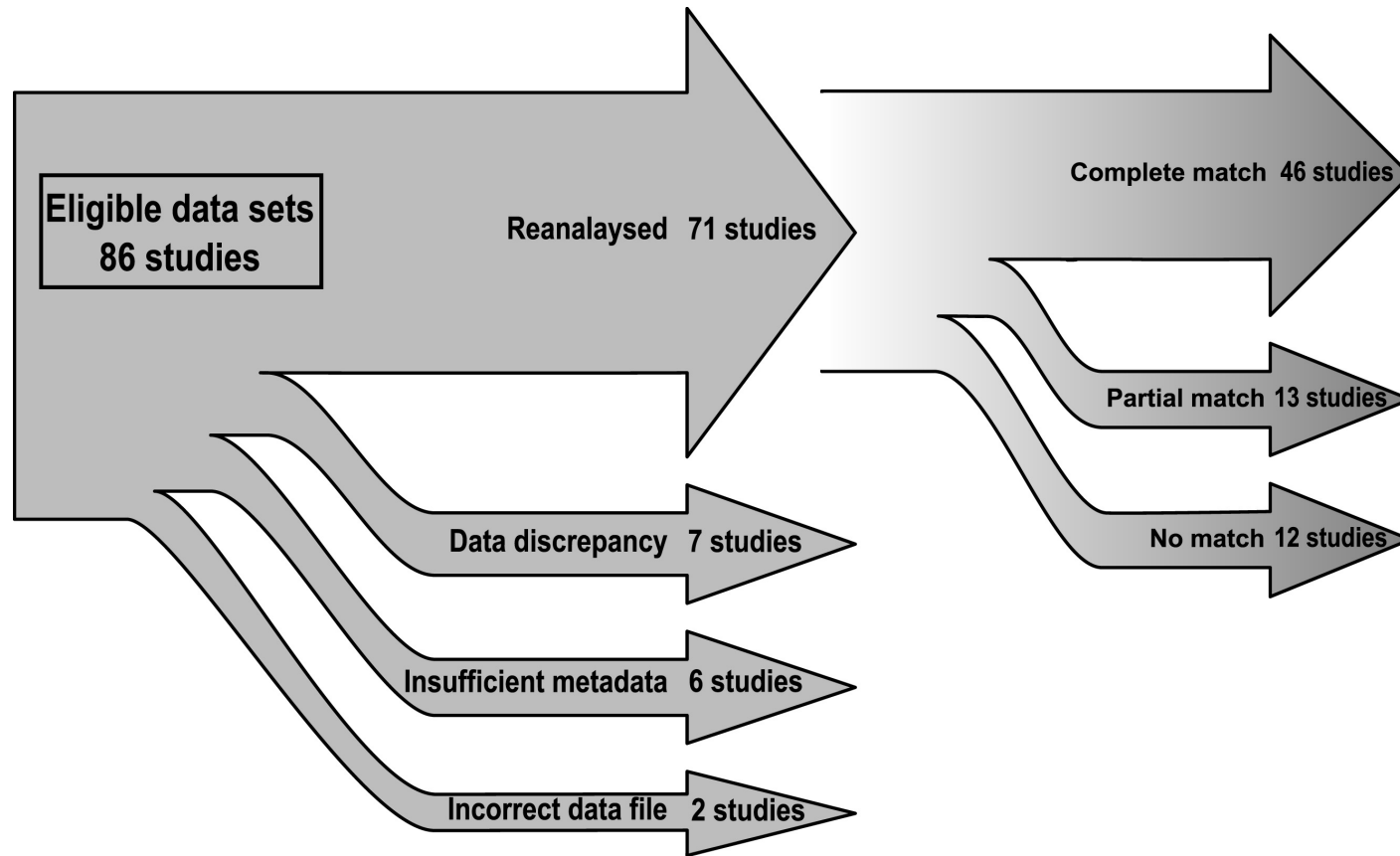
Shen, Y. Strategic Planning for a Data-Driven, Shared-Access Research Enterprise: Virginia Tech Research Data Assessment and Landscape Study. *Coll. Res. Libr.* **77**, 500–519 (2016). doi: [10.5860/crl.77.4.500](https://doi.org/10.5860/crl.77.4.500)

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



Data Quality

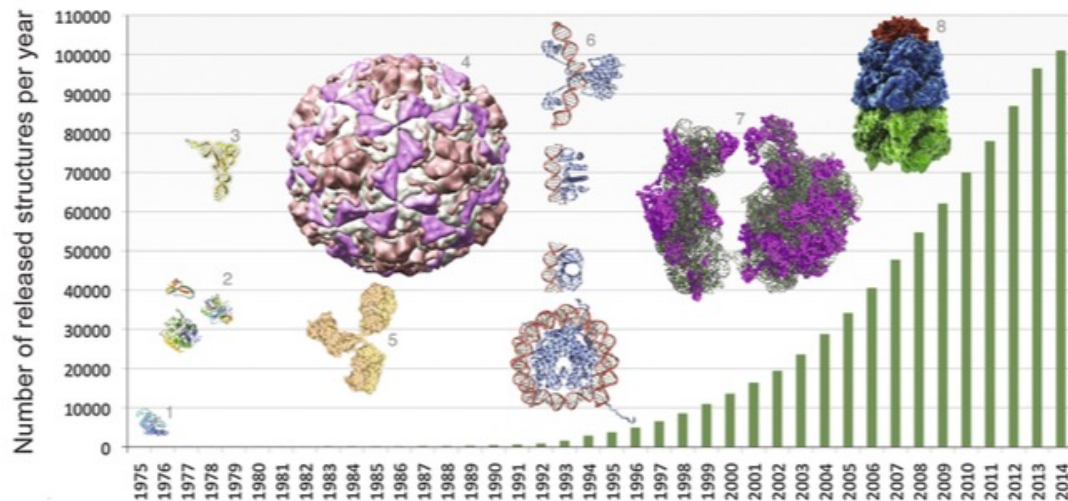


On Average, 13%
of Papers Had
Usable Data

Why is it better to have data available?



www.rcsb.org



Biological assembly 1 assigned by authors

Select a Viewer

JSmol (JavaScript)

Stoichiometry: A

Select Orientation: Front

Select Display Mode: Secondary Structure, Subunit, Symmetry

Display Options:

- Style: Cartoon
- Color: Secondary Structure
- Surface: None
- H-Bonds
- SS Bonds
- Rotation
- Black Background
- Polyhedron
- Axes

wwPDB Validation

Metric	Percentile Ranks	Value
Rfree		0.202
Clashscore		4
Ramachandran outliers		0
Sidechain outliers		0.6%
RSRZ outliers		1.7%

Worse | Percentile relative to all X-ray structures | Percentile relative to X-ray structures of similar resolution | Better

3D Report Full Report

Scripting Options

Ligands Domain Modification

Ligand ID	View Interactions	Ligand Electron Density	Image	Name / Formula / Weight
MTA	View SHJM - MTA Pocket Interaction	A:401		5'-DEOXY-5-METHYLTHIOADENOSINE C11 H15 N5 O3 S 297.334

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

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

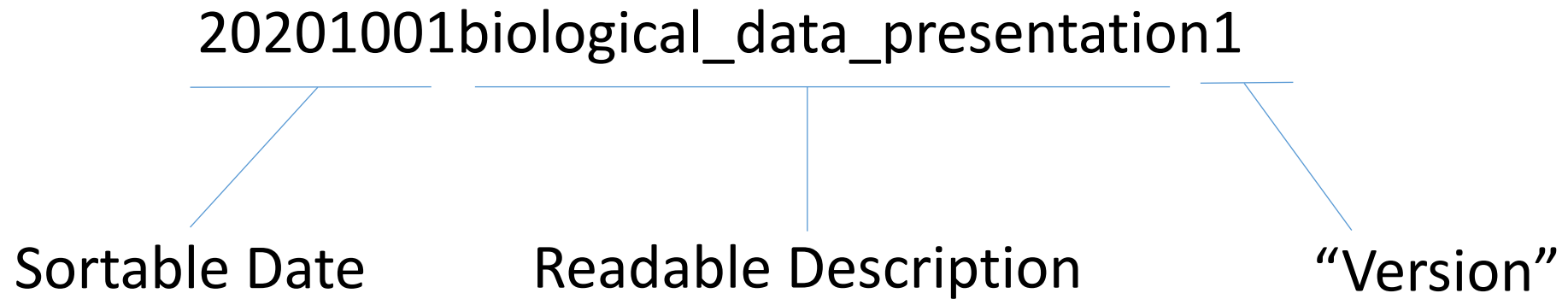


Simple Solutions

- 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

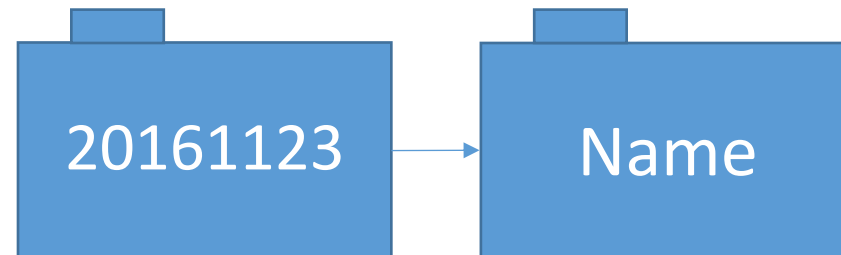
Simple



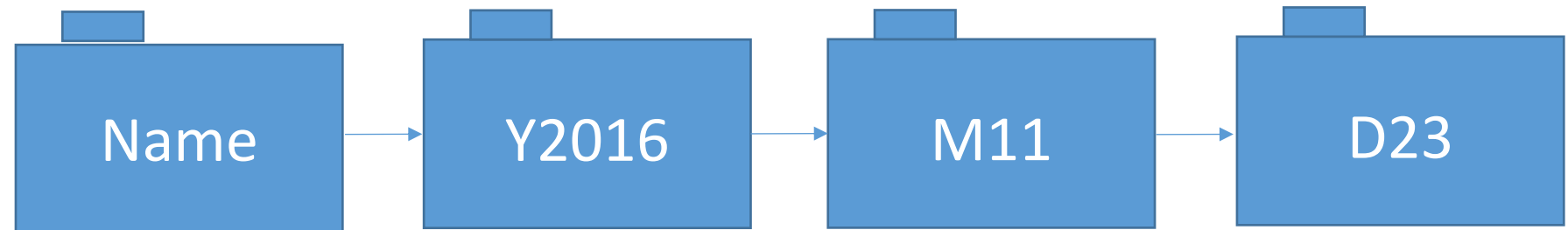
Full worksheet at:

<https://resolver.caltech.edu/CaltechAUTHORS:20200601-161923247>

Date Based



Complex



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)



Data Sharing

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

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

The Zenodo logo consists of the word "zenodo" in a white, lowercase, sans-serif font, centered within a solid blue rectangular background.

- 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



GitHub

Discoverability

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

The image shows two search results for the same research paper. The top screenshot is from DataCite, showing 2 results for 'Titan'. The bottom screenshot is from Google, showing about 153,000 results for 'Titan VIMS and ISS'.

DataCite Search Results:

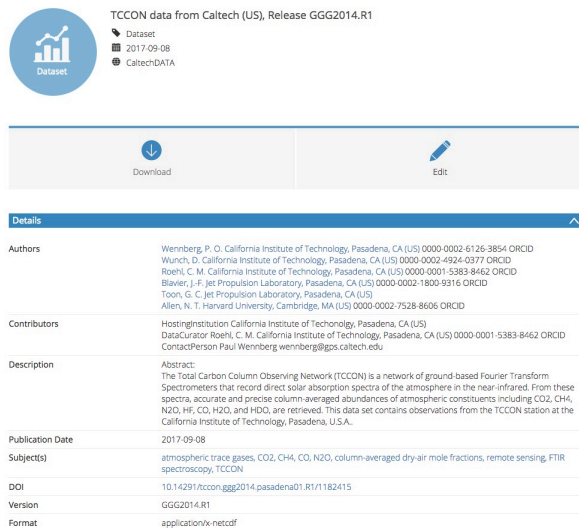
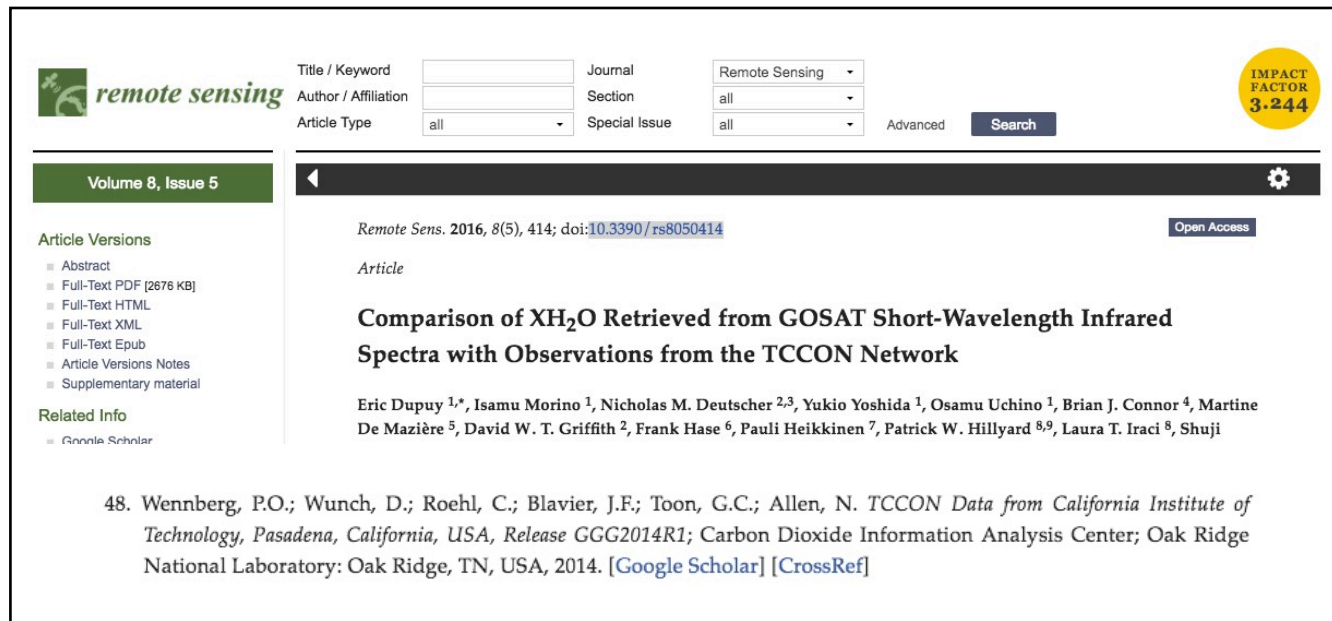
- Search term: Titan
- Sorted by: Relevance
- 2 results
- Resource Type: Image
- Author: Brown, Robert H.
- Registration Year: 2019 (316)

Google Search Results:

- Search term: Titan VIMS and ISS
- About 153,000 results (0.26 seconds)
- PDF: TITAN'S GLOBAL MAP COMBINING VIMS AND ISS MOSAICS. B. Seignovert et al. at next 2019 LPSC (cf. related ...)
- PDF: This poster VIMS-ISS map Titan VIMS rotating globe
- Visible and Infrared Mapping Spectrometer (VIMS) | Cassini Orbiter ...
- Titan's global map combining VIMS and ISS mosaics - CaltechDATA

Unique Views: 316
Unique Downloads: 79
between February 14, 2019 and June 12, 2019
[More info on how stats are collected](#)

Citations


<https://doi.org/10.14291/tcon.ggg2014.pasadena01.R1/1182415>

Update Record

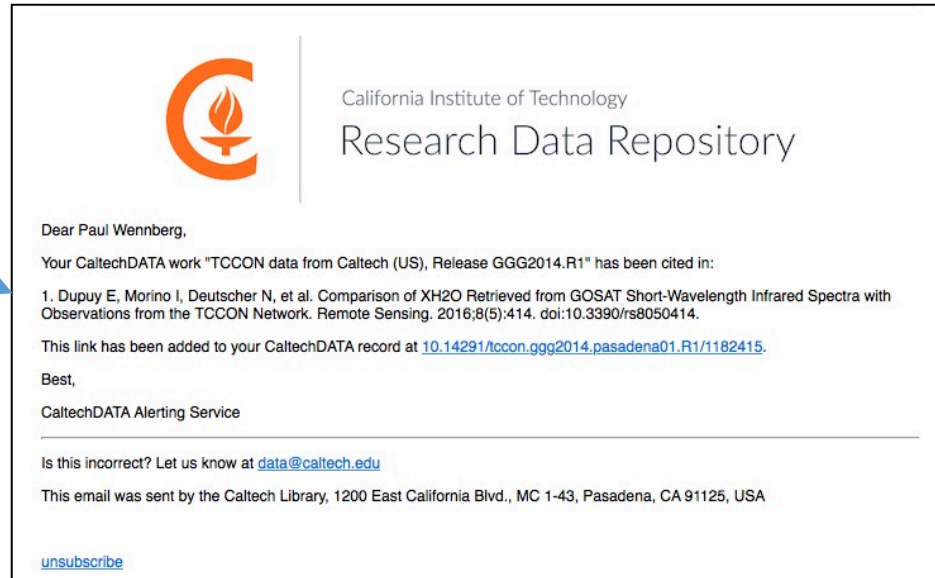
Related Identifier(s)

- IsDocumentedBy (URL): https://tcon-wiki.caltech.edu/Network_Policy/Data_Use_Policy/Data_Description
- IsDocumentedBy (URL): <https://tcon-wiki.caltech.edu/Sites>
- IsPartOf (URL): <http://tcondata.org>
- IsDocumentedBy (DOI): [10.14291/tcon.ggg2014.documentation.R0/1221662](https://doi.org/10.14291/tcon.ggg2014.documentation.R0/1221662)
- IsCitedBy (DOI): [10.5194/amt-9-683-2016](https://doi.org/10.5194/amt-9-683-2016)
- IsCitedBy (DOI): [10.5194/amt-9-227-2016](https://doi.org/10.5194/amt-9-227-2016)
- IsCitedBy (DOI): [10.5194/amt-9-3491-2016](https://doi.org/10.5194/amt-9-3491-2016)
- IsCitedBy (DOI): [10.5194/amt-9-3527-2016](https://doi.org/10.5194/amt-9-3527-2016)
- IsNewVersionOf (DOI): [10.14291/tcon.ggg2014.pasadena01.R0/1149162](https://doi.org/10.14291/tcon.ggg2014.pasadena01.R0/1149162)
- IsPartOf (DOI): [10.14291/TCCON_GGG2014](https://doi.org/10.14291/TCCON_GGG2014)
- IsCitedBy (DOI): [10.3390/rs8050414](https://doi.org/10.3390/rs8050414)

<https://doi.org/10.3390/rs8050414>

Citation

Email Alert





California Institute of Technology

Research Data Repository

Demo

Use Cases - Theses

Upload files while writing



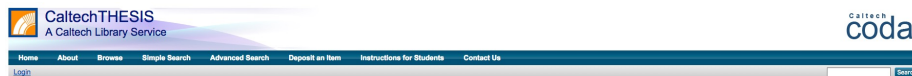
California Institute of Technology
Research Data Repository

<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>

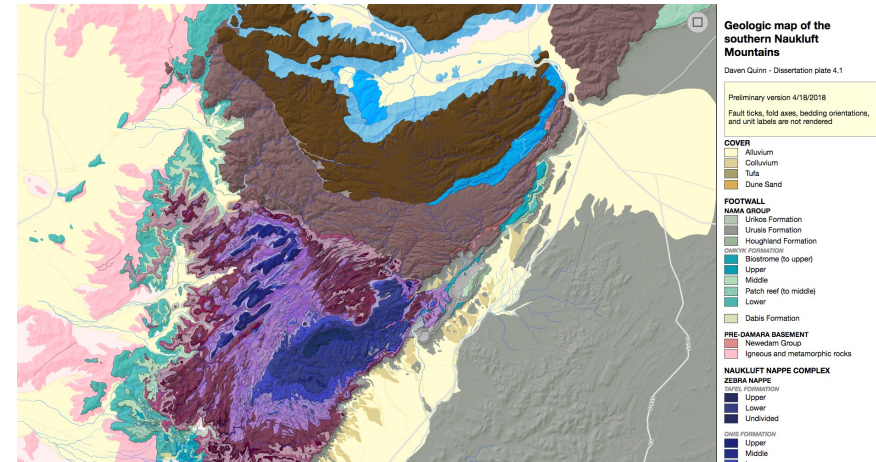


Engineered Viral Vectors and Developed Tissue Clearing Methods for Single-cell Phenotyping in Whole Organs

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. <https://resolver.caltech.edu/CaltechTHESIS:10.22002/D1.234>

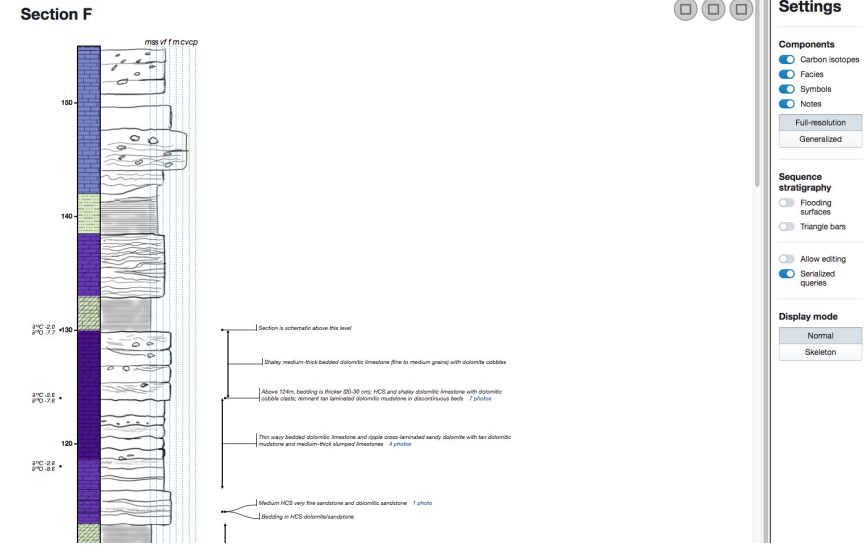
Abstract
A central question in biology is how different cell types interact with each other and their native environment to form complex functional systems and networks. Although our ability to investigate this question has considerably expanded from the development of genetically encoded tools, some 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 transgene 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 sensitively 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 optiques. These earlier methods eventually became inapplicable 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 viscoelasticity to perfuse clearing reagents. This technique, called perfusion assisted agent release in situ (PARIS) 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>

Use Cases



bioRxiv
beta
THE PREPRINT SERVER FOR BIOLOGY

HOME | ABO

Search



California Institute of Technology

Research Data Repository

New Results

An allosteric theory of transcription factor induction

Manuel Razo-Mejia, Stephanie L. Barnes, Nathan M. Belliveau, Griffin Chure, Tal Einav, Rob Phillips

doi: <https://doi.org/10.1101/111013>

This article is a preprint and has not been peer-reviewed [what does this mean?].

Abstract Info/History Metrics Supplementary material Preview PDF

Abstract

Allosteric molecules serve as regulators of cellular activity across all domains of life. We present a general theory of allosteric transcriptional regulation that permits quantitative predictions for how physiological responses are tuned to environmental stimuli. To test the model's predictive power, we apply it to the specific case of the ubiquitous simple repression motif in bacteria. We measure the fold-change in gene expression at different inducer concentrations in a collection of strains that span a range of repressor copy numbers and operator binding strengths. After inferring the inducer dissociation constants using data from one of these strains, we show the broad reach of the model by predicting the induction profiles of all other strains. Finally, we derive an expression for the free energy of allosteric transcription factors which enables us to collapse the data from all of our experiments onto a single master curve, capturing the diverse phenomenology of the induction profiles.

<https://doi.org/10.1101/111013>

Paper Website
on GitHub



The screenshot shows a GitHub repository page. On the left is a navigation menu with links for ABOUT, ANALYSIS, DATA, PEOPLE, and ACKNOWLEDGEMENTS. Below the menu is the Caltech logo and a link to the Phillips Lab GitHub Repo. The main content area features a large green oval containing a handwritten mathematical equation:
$$\text{Fold-Change} \approx \left(1 + \frac{P}{K} \frac{R}{N_{NS}} e^{-\frac{R \Delta G_{RA}}{RT}}\right)^{-1}$$
 Below the equation is the title "An Allosteric Theory of Transcription Factor Induction" and a paragraph of text describing the website's purpose. At the bottom, there are links for "Main Text" and "Supplementary Information".

Data Files



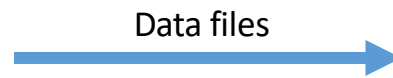
<https://doi.org/10.22002/D1.224>
<https://doi.org/10.22002/D1.227>
<https://doi.org/10.22002/D1.228>
<https://doi.org/10.22002/D1.229>

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

Use Case - TCCON



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



Data Curation and Processing

Use Case - TCCON



tcon.ornl.gov

TCCON Data Archive HOME GGG2014 GGG2012 GGG2009

Total Carbon Column Observing Network (TCCON)
The TCCON Data Archive

TCCON is a network of ground-based Fourier Transform Spectrometers recording direct solar spectra in the near infrared spectral region. From these spectra, accurate and precise column-averaged abundances of CO₂, CH₄, H₂O, HF, CO, H₂O, and HDO are retrieved. The HF and HDO retrievals are uncalibrated and hence preliminary. Data are updated monthly on the first of the month. The data become publicly available no later than one year after the measurements are recorded, and many sites choose to release their data much sooner.

For the latest TCCON information, please visit the [TCCON Wiki](#). For citation information and our data policy, please see our [Data Use Policy](#). For site-specific information and data analysis descriptions, please read the [Data Description](#). Auxiliary data (column averaging kernels, a priori profiles) are included in the netCDF files provided below. Information on how to use our column averaging kernels and a priori profiles can be found on our [Auxiliary Data](#) page.

A technical report describing the GGG2014 TCCON data version can be found on the [documentation](#) page. Our telluric line list can be downloaded from the [slm](#) page. Our solar line list can be downloaded from the [solar](#) page. A program to generate our a priori profiles can be downloaded from the [a priori](#) page. Please note that the a priori profiles used in the TCCON retrievals are included in the data files below. If you need to produce TCCON a priori profiles for locations and times where there are no TCCON measurements, please use the program linked above.

The TCCON is closely affiliated with the Network for the Detection of Atmospheric Composition Change Infrared Working Group (NDACC-IRWG). In contrast with TCCON, which produces column-averaged dry-air mole fractions, the NDACC produces vertical profiles of the concentrations of many of the same gases and several others. The NDACC website and links to their database can be found at www.acd.ucar.edu/irwg.

Sign up to the TCCON Users email list to get email updates on TCCON data releases. Note that the website is self-signed; you can safely add an exception.

[Login for TCCON Partners](#)

Private data files

- Sites
- Ascension Island
- [0ae20120522_20120831.nc](#)
 - [0ae20130317_20130618.nc](#)
 - [0ae20130911_20131229.nc](#)
 - [0ae20140108_20140716.nc](#)
 - [0ae20140717_20141019.nc](#)
 - [0ae20141021_20141231.nc](#)
 - [0ae20150101_20150310.nc](#)
 - [0ae20150311_20150409.nc](#)
 - [0ae20150410_20150630.nc](#)
 - [0ae20150701_20150926.nc](#)
 - [0ae20151005_20151218.nc](#)

Public data files

Index of /2014Public/ascension01

Name	Size	Date Modified
[parent directory]		
README.txt	11.8 kB	10/20/14, 5:00:00 PM
ae20120522_20161221.public.nc	10.1 MB	5/31/17, 5:25:00 PM

Automatically released 1x/month

Departmental Server at Caltech

Login for TCCON Partners TCCON Data Archive GGG2014 GGG2012 GGG2009

Total Carbon Column Observing Network (TCCON)



TCCON is a network of ground-based Fourier Transform Spectrometers recording direct solar spectra in the near infrared spectral region. From these spectra, accurate and precise column-averaged abundances of CO₂, CH₄, N₂O, HF, CO, H₂O, and HDO are retrieved and reported here. A technical report describing the retrievals is found here; solar and telluric spectral line lists used in the retrievals are publicly available.

Data in netCDF format are publicly available no later than one year after the spectra are recorded; many sites release their data earlier. Citation and data use requirements are included in the license associated with each record. Column averaging kernels and a priori profiles are included in the files. Information on how to use these can be found here. To produce TCCON a priori profiles for locations and times where there are no TCCON measurements, a stand-alone program can be downloaded.

[Sign up to the TCCON Users email list to get email updates on TCCON data releases.](#)

tcondata.org

CaltechDATA

Migration

TCCON data from Park Falls (US), Release GGG2014.R1

Dataset

2017-09-27

CaltechDATA

Download Edit

Details

Authors
Wernberg, P. D. California Institute of Technology, Pasadena, CA (US) 0000-0002-6126-3854 ORCID
Roehl, C. M. California Institute of Technology, Pasadena, CA (US) 0000-0001-5383-8462 ORCID
Wunch, D. California Institute of Technology, Pasadena, CA (US) 0000-0002-4024-0377 ORCID
Toon, G. C. Jet Propulsion Laboratory, Pasadena, CA (US)
Blavier, J. F. Jet Propulsion Laboratory, Pasadena, CA (US) 0000-0002-1808-8316 ORCID
Wagner, R. University of Colorado, NOAA, Boulder, CO (US) 0000-0002-8106-3702 ORCID
Koppel-Aleks, G. University of Michigan, Ann Arbor, MI (US) 0000-0003-2119-0044 ORCID
Allen, N. T. Harvard University, Cambridge, MA (US) 0000-0002-7528-8605 ORCID
Ayers, J. Wisconsin Educational Communications Board, Park Falls, WI (US)

Contributors
HostInstitution California Institute of Technology, Pasadena, CA (US)
DataCurator Roehl, C. M. California Institute of Technology, Pasadena, CA (US) 0000-0001-5383-8462 ORCID
ContactPerson Paul Wernberg wernberg@jplgs.caltech.edu

Description
Abstract:
The Total Carbon Column Observing Network (TCCON) is a network of ground-based Fourier Transform Spectrometers that record direct solar absorption spectra of the atmosphere in the near-infrared. From these spectra, accurate and precise column-averaged abundances of atmospheric constituents including CO₂, CH₄, N₂O, HF, CO, H₂O, and HDO, are retrieved. This data set contains observations from the TCCON station at Park Falls, U.S.A..

Publication Date
2017-09-27

Subjects
atmospheric trace gases, CO₂, CH₄, CO, N₂O, column-averaged dry-air mole fractions, remote sensing, FTIR spectroscopy, TCCON

DOI
10.14291/tcon-ggg2014-parkfalls01.R1

Version
GGG2014.R1

Format
application/x-netcdf

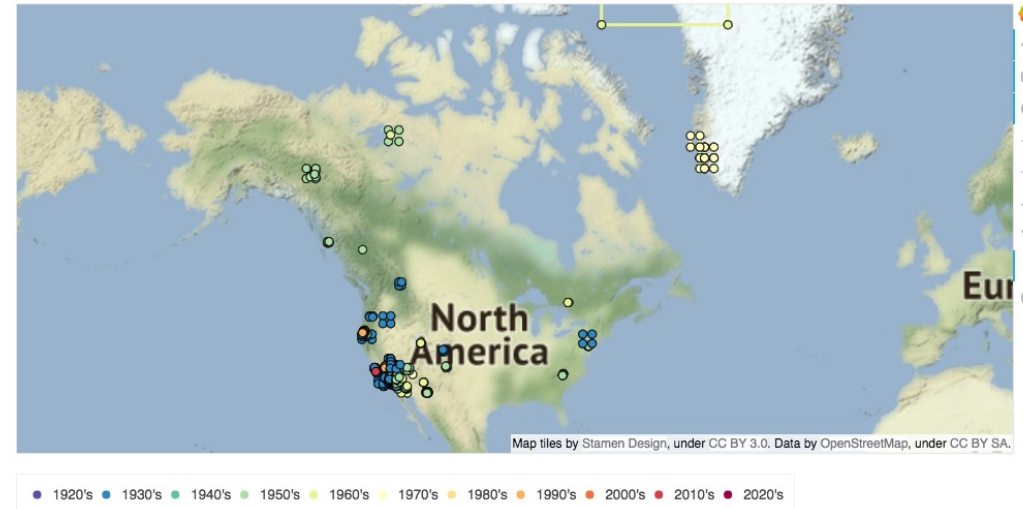
<https://doi.org/10.14291/tcon.ggg2014.parkfalls01.R1>



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 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