

Launching a Research-Focused Data Repository

Thomas Morrell

California Institute of Technology Library, Pasadena, CA, USA

Open Repositories 2017

June 28, 2017

Current Research Data Practices



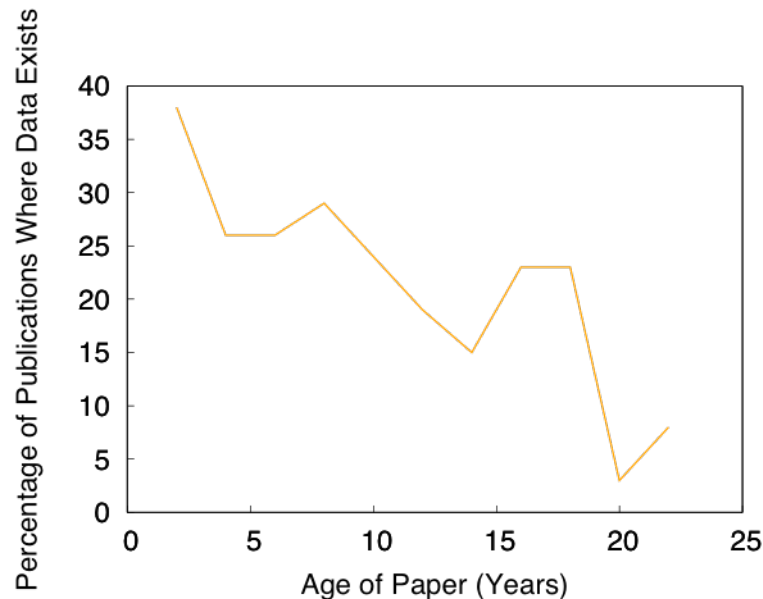
Most researchers store data on local computer hard drives

Researchers report that finding data is their biggest challenge

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

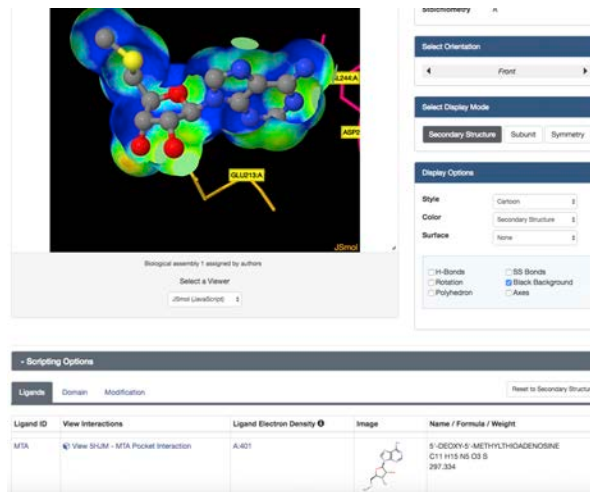
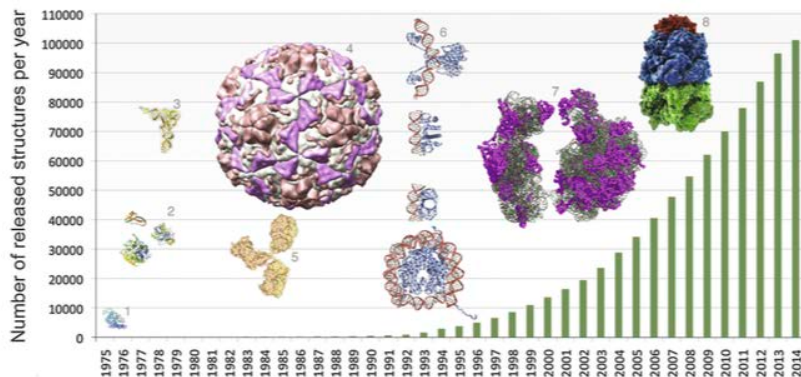
Who Maintains the Files?



Disciplinary Repositories



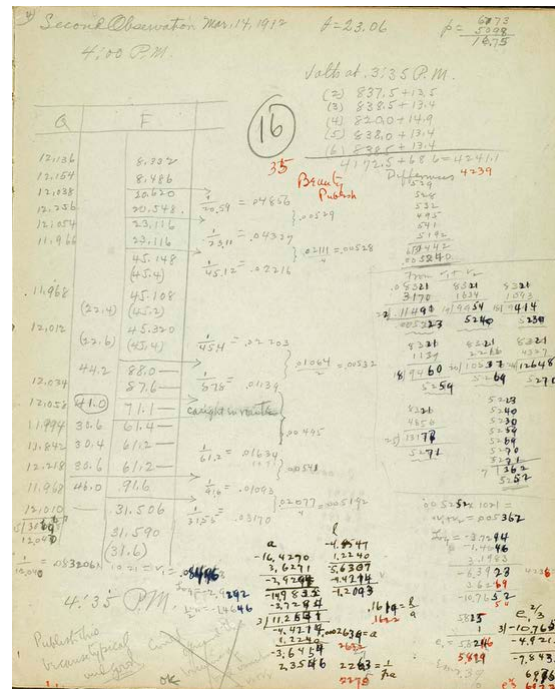
www.rcsb.org



Institutional repository?

Libraries have Institutional Backing

Libraries have a History of Managing Objects

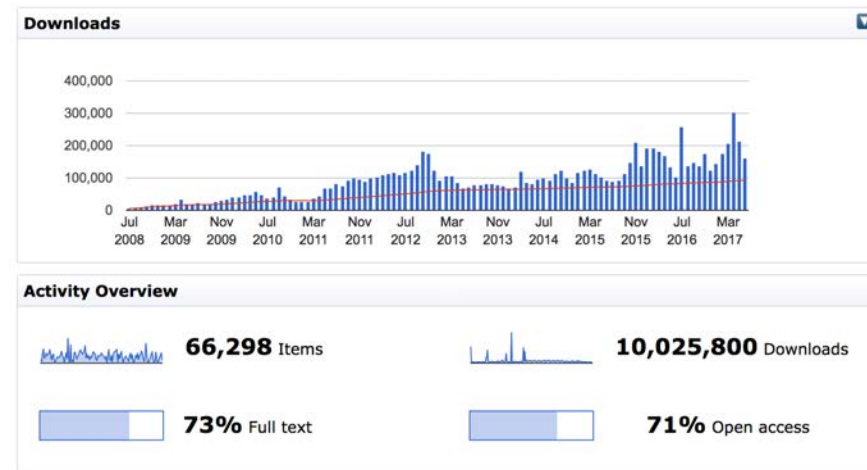


Caltech

- Big impact – 34 Nobel prizes
- Manages the Jet Propulsion Laboratory (JPL), Seismological Laboratory, Palomar and W. M. Keck Observatories, and co-manages LIGO
- Small - 300 faculty, 1,000 undergraduates, 1,250 graduate students
- Library has run an institutional repository since 2001.
- 3.5 library developers



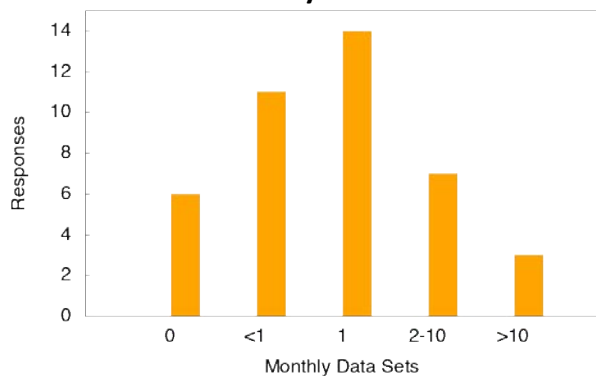
CaltechAUTHORS Downloads



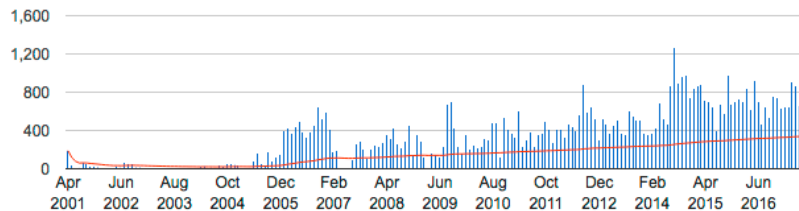
Data Repository Models

- Goal: Capture research data

2017 Survey of ARL Libraries
Monthly Data Sets



Monthly CaltechAUTHORS Deposits



General Repository Sizes:

Zenodo – 185,101 Files

Dataverse – 341,576 Files

Dryad – 50,441 Files

Invenio and TIND

- Developed by CERN
- Invenio 1 – MARC based document management
 - CERN Document Server (cds.cern.ch)
- Invenio 3 – JSON based repository software
 - Zenodo Data Repository (zenodo.org)
 - HEPdata (hepdata.net)
- TIND is a CERN startup that does custom deployments of Invenio systems

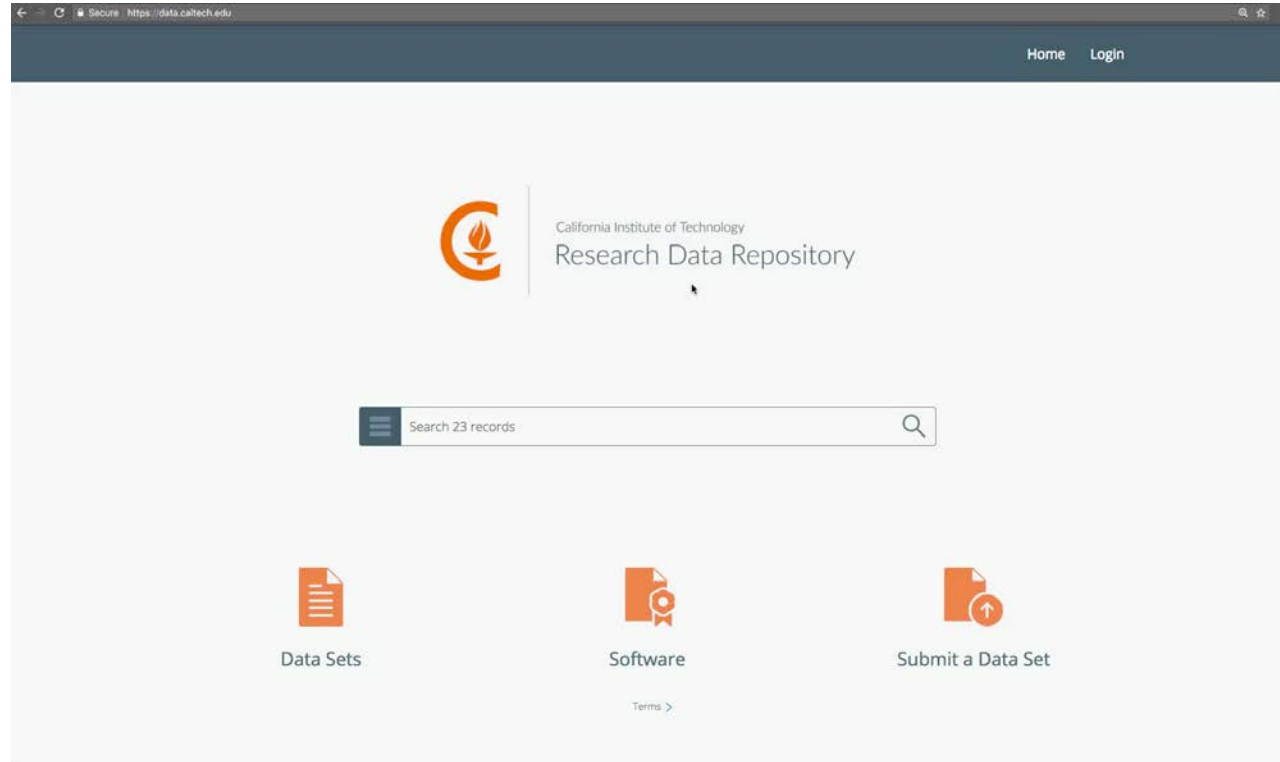


Researcher-Focused Design Principles

- Simple deposit process
- Library manages data availability
- Support re-use of data
- Automate where possible

Single Deposit Form

- Institutional (Shibboleth) login
- One page
- Minimal mandatory fields
- Collapsed sections
- Drag and drop file upload



Data Availability



- All records assigned permanent identifiers (DOIs)
- DataCite4.0 metadata
 - Added explicit Related Publication fields
 - Added Contributor Email
- AWS S3 Data storage + Backups

Related Publication

Identifier - DOI

Title

Author Information

Author Name

Author Affiliation

Journal

Publication Date

Language

Contributors

Contributor

Contributor Name

Contributor Affiliation

Contributor Identifier

Identifier Type

Contributor Role

Contributor Email

Discoverability

- CaltechDATA site search
- DOIs appear in DataCite search
- and Google

The image displays three overlapping search results for the paper "Identifying and Quantifying Mineral Abundance through VSWIR Microimaging Spectroscopy: A Comparison to XRD and SEM" by Leask, Ellen K.; Ehlmann, Bethany L.

Top Panel (DataCite Search): Shows the search results for "VSWIR microimaging". The result is listed as a "Dataset" with the title "Identifying and Quantifying Mineral Abundance through VSWIR Microimaging Spectroscopy: A Comparison to XRD and SEM" by Leask, Ellen K.; Ehlmann, Bethany L. The right sidebar shows filters for "Resource Type" (Dataset) and "Author" (Ehlmann, Bethany L., Leask, Ellen K.).

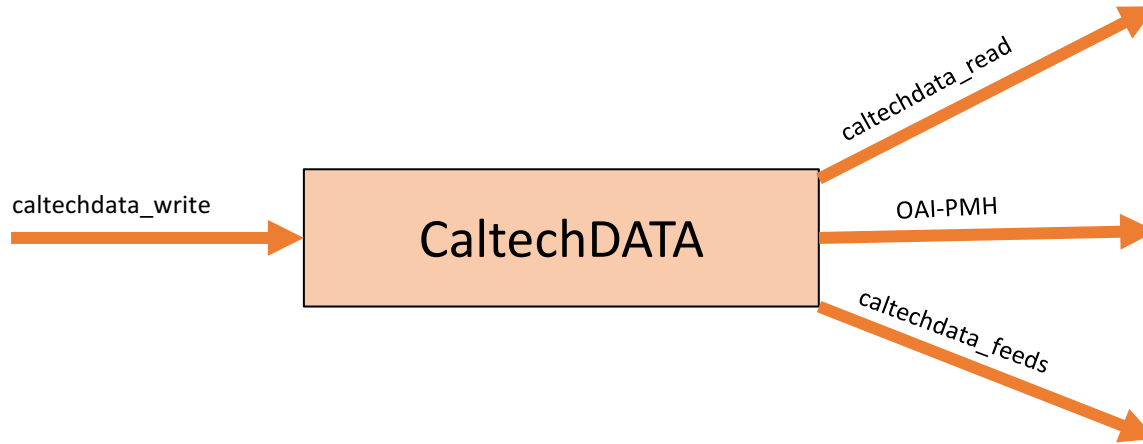
Middle Panel (DataCite Search): A zoomed-in view of the same search result, showing the title, authors, and a "Search" button.

Bottom Panel (Google Search): Shows the search results for "VSWIR microimaging". The first result is the paper by Leask and Ehlmann, with the DOI <https://doi.org/10.22002/D1.222>. The snippet indicates it is a "Dataset" and mentions "Mar 13, 2017".

Right Panel (CaltechDATA Search): Shows the search results for "VSWIR microimaging". The result is listed as a "Dataset" with the title "Identifying and Quantifying Mineral Abundance through VSWIR Microimaging Spectroscopy: A Comparison to XRD and SEM" by Leask, Ellen K.; Ehlmann, Bethany L. The right sidebar shows filters for "Resource Types" (Dataset) and "Publication Year" (2017).

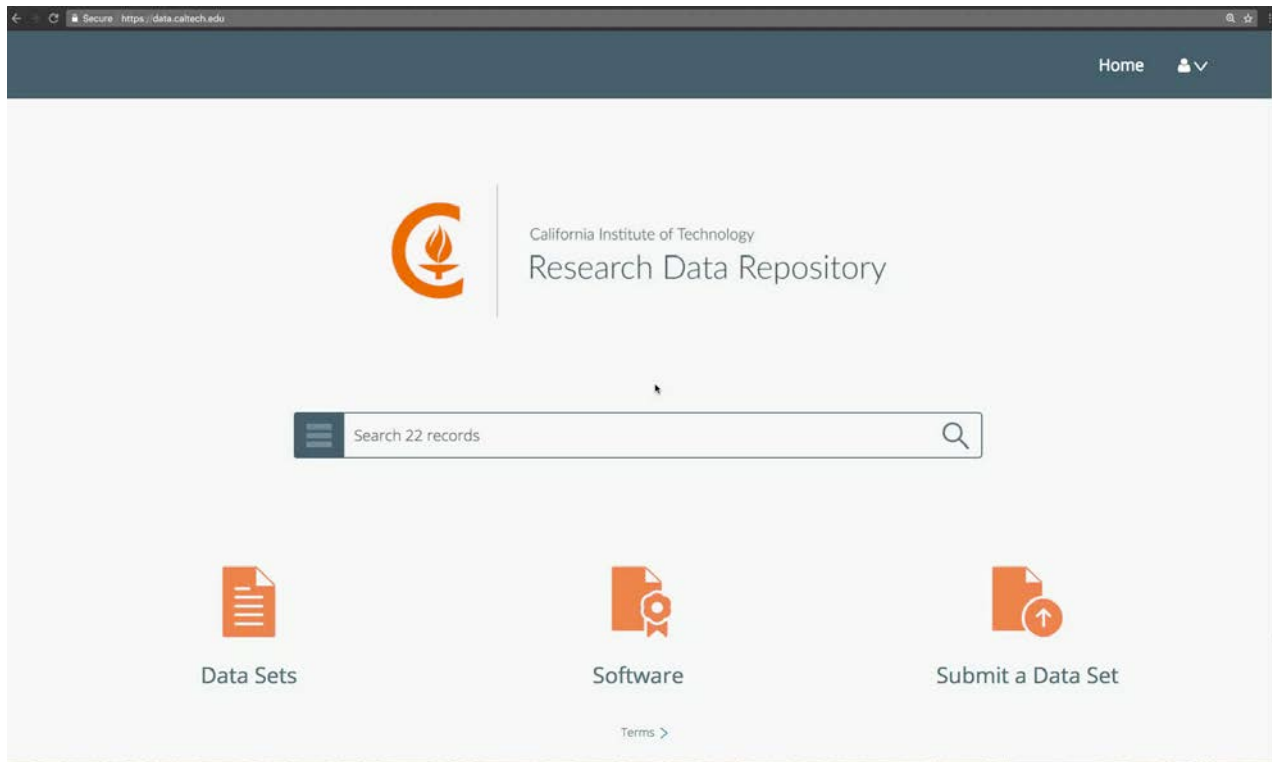
Bottom Right Panel (DOI Link): Shows the DOI link <https://doi.org/10.22002/D1.222> and the title "Identifying and Quantifying Mineral Abundance through VSWIR ... - DOIs".

Encourage re-use of data



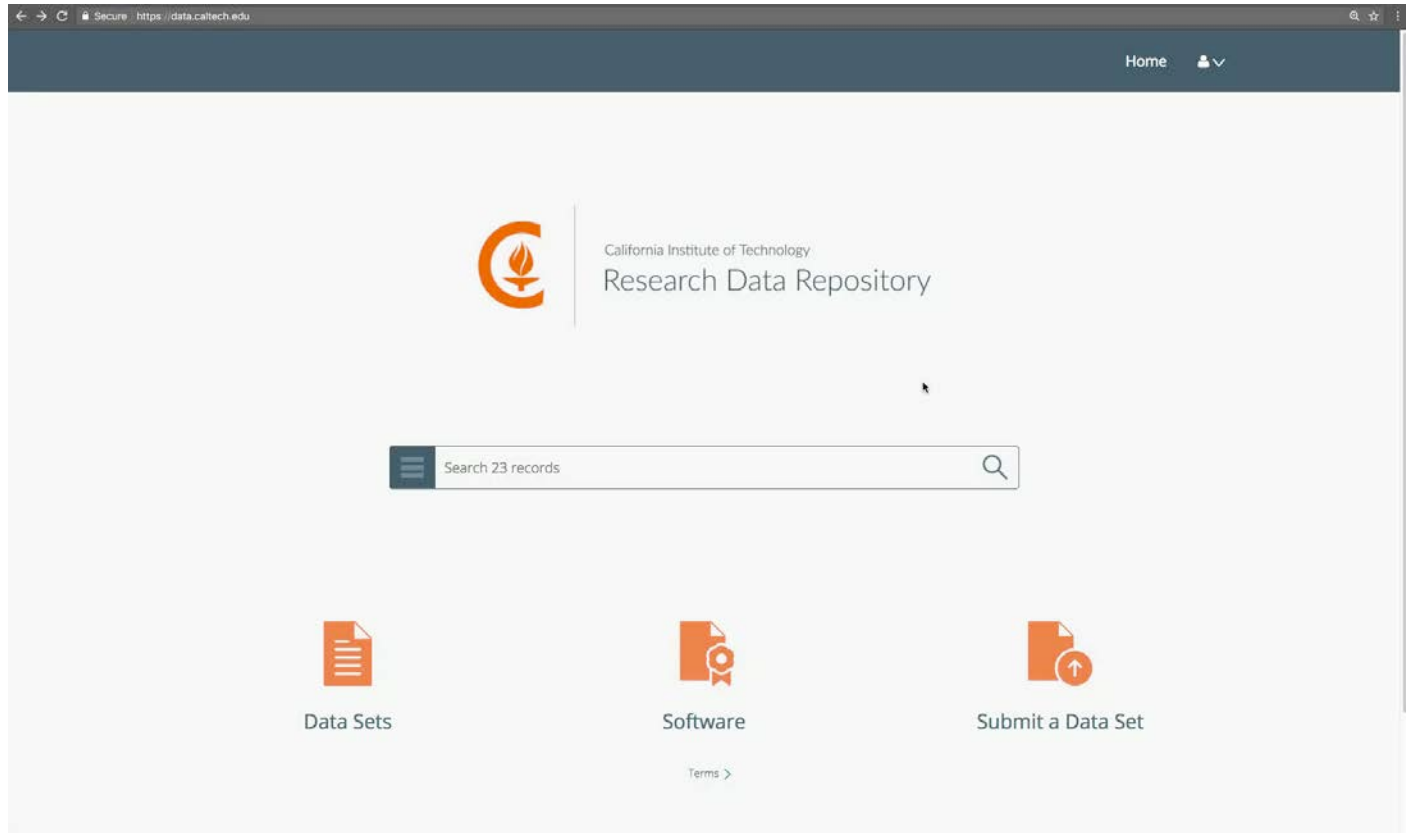
- Enable open science
- Support library development (11:00 AM Thursday, Ballroom A)

Read API Example: Interactive Plotting



plots.caltechlibrary.org
doi.org/10.22002/D1.240

Automation



Current Status

- Soft launch March 2017
- Campus launch June 2017
- 23 Records from 9 groups



Dataset

TCCON data from Park Falls, Wisconsin, USA, Release GGG2014R0
Wennberg, P. O.; C. Roehl; D. Wunch; G. C. Toon; (5 more)

✓ | Dataset | 2017-01-25 | CaltechDATA



Dataset

Visualizing endogenous fluorescence throughout a cleared mouse tibia
Chan, Ken; Greenbaum, Alon; Gradinaru, Viviana

✓ | Dataset | 2017-05-04 | CaltechDATA



Dataset

Microscopy image files for "An allosteric theory of transcription factor induction."
Manuel Razo-Mejia; Stephanie L. Barnes; Nathan M. Belliveau; (3 more)

✓ | Dataset | 2017-04-04 | CaltechDATA



Dataset

Flow Cytometry files for "An allosteric theory of transcription factor induction"
Manuel Razo-Mejia; Stephanie L. Barnes; Nathan M. Belliveau; (3 more)

✓ | Dataset | 2017-03-27 | CaltechDATA



Dataset

AAV-PHP.S transduction across multiple layers of the small intestine
Chan, Ken; Gradinaru, Viviana

✓ | Dataset | 2017-05-04 | CaltechDATA

Use Cases



California Institute of Technology

Research Data Repository



bioRxiv
beta
THE PREPRINT SERVER FOR BIOLOGY

HOME | ABO

Search

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



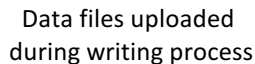
Data Files

Caltech Library

http://rpgroup-pboc.github.io/mwc_induction

Thesis Preparation

<https://doi.org/10.7907/Z9NC5Z7H>



<http://dx.doi.org/10.22002/D1.234>

<http://dx.doi.org/10.22002/D1.235>

<http://dx.doi.org/10.22002/D1.236>

<http://dx.doi.org/10.22002/D1.237>

Software in CaltechDATA

<http://dx.doi.org/10.22002/D1.218>



Software
2017-02-27
CaltechDATA



Details

Authors

Holzmann, Gerard Department of Computing & Mathematical Sciences, Caltech

Description

Abstract

Abstract: Tau is a small Tc/Tr application that can be used to quickly specify and formally verify small automata models (the name 'tau' is short for 'tiny automata'). It is used as a teaching aid in CS118, a course on the formal verification of asynchronous software systems using logic model checking. Tau requires the availability of a standard C compiler (e.g., gcc) and a recent version of the Spin model checker (e.g., Version 6.4.3 or later) as background tools.

<http://dx.doi.org/10.22002/D1.240>



Software
2017-05-24
CatechDATA



Details

Authors

Tom Morrell Caltech

Description

Abstract

Abstract:
Includes new documentation with improved AWS setup.
Other:
Demo interactive plotting tool that uses Bokeh server to produce an interactive plot by calling the caltechDATA (invenio 3) API

Use Case - TCCON



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

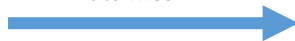


↑
Processed Data



Data Curation and Processing

Data files



Use Case - TCCON



tcccon.ornl.gov

TCCON Data Archive

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, N₂O, 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](#). Available data include 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 full-time staff can be downloaded from the [data](#) 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 [prior](#) page. Please note that the a priori profiles used in the TCCON networks 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 (NDACC) and the Working Group (WG) for the Detection of Atmospheric Composition Change (WDACC), which produces column-averaged dry-air mole fractions, the NDACC website and links to their database can be found at [www.ndacc.org/ndacc/](#).

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

[Login to TCCON Archives](#)

Private data files

Sites	
Ascension Island	
asc20120522_20120831.nc	
asc20130317_20130618.nc	
asc20130911_20131229.nc	
asc20140108_20140716.nc	
asc20140717_20141019.nc	
asc20141021_20141231.nc	
asc20150101_20150310.nc	
asc20150311_20150409.nc	
asc20150410_20150630.nc	
asc20150701_20150926.nc	
asc20151005_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
asc20120522_20161221_public.nc	10.1 MB	5/31/17, 5:25:00 PM

Automatically released 1x/month



Departmental Server at Caltech

Migration Plan

Library-hosted CaltechDATA Repository



TCCON data from Ascension Island (SH), Release GGG2014R0

Dataset
2017-02-21
CaltechDATA



Download



Edit

Details

Authors

Feist, D. G. Max Planck Institute for Biogeochemistry, Jena (DE) ORCID 0000-0002-5890-6687
Arnold, S. G. Max Planck Institute for Biogeochemistry, Jena (DE)
John, N. Ariane Tracking Station, Ascension Island (SH)
Geibel, M. C. Stockholm University, Stockholm (SE) ORCID 0000-0002-7369-0781

Contributors

DataCuration Roehl, C. M. California Institute of Technology, Pasadena, CA, U.S.A. ORCID 0000-0001-5383-8462

Description

Abstract:
The Total Carbon Column Observing Network (TCCON) is a network of ground-based Fourier Transform

<https://data.caltech.edu/records/210>

Launching a Researcher-focused Data Repository

- Invenio 3-powered data repository
- Simplified, general deposit form
- DOIs and standard metadata
- API access
- Automation



California Institute of Technology

Research Data Repository

tmorrell@caltech.edu

data.caltech.edu

<https://www.library.caltech.edu/resources/caltechdata>

Deposit Form 1

Resource Type

Resource Type

Dataset

Basic Information

Title *

Briefly describe your dataset

Author(s)

Author Name *

Family name, First name

Author Affiliation *

Caltech

Author Identifier

Author Identifier

Author Identifier Type

ORCID

Keywords *

Keywords

Language

English

Deposit Form 2

Description(s)

Description *

Provide sufficient technical information so that others will be able to use the data set.

Description Type

Other

Format

Funding

Funding Agency

Grant Number

Related Identifier(s)

Related Publication

Contributors

Relevant Dates

Geolocation

Deposit Form 3

Related Identifier(s) 

Related Identifier(s)

This CaltechDATA Record

Is Part Of

Related Identifier

Identifiers of related resources

Identifier Type

DOI

+

Related Publication 

Contributors 

Relevant Dates 

Geolocation 

Location Place

Name of a place

Location Point

Longitude

Longitude of a point

Latitude

Latitude of a point

Location Area

East

Eastern longitudinal dimension of area

West

Western longitudinal dimension of area

North

Northern latitudinal dimension of area

Deposit Form 4

Embargo

Embargoed Status

☐ Embargoed Access ☒ Open Access

Embargo Date

YYYY-MM-DD

License/Waiver

License/Waiver Name *

Creative Commons CCZero

Version

DOI

Upload

Drop files here to upload (accepts: any type)

Read API Example: Interactive Plotting



Orange-brown jasper spectrum

Dataset
2017-02-16
CaltechDATA



Download



Edit



Red jasper spectrum

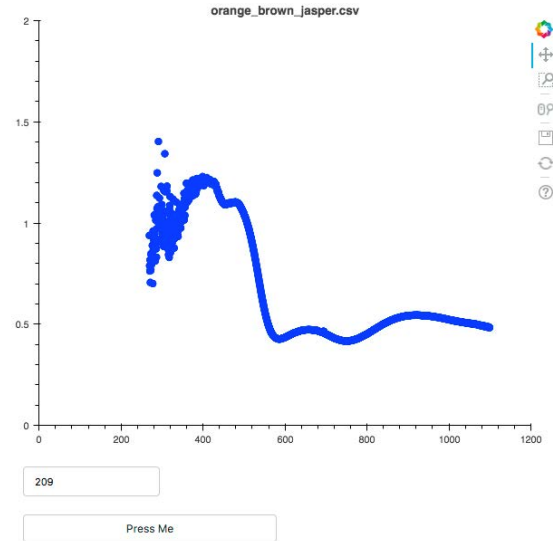
Dataset
2017-02-14
CaltechDATA



Download



Edit



plots.caltechlibrary.org
doi.org/10.22002/D1.240

GitHub Repositories

(updated now) Sync now ...

Get started

1 Flip the switch

Select the repository you want to preserve, and toggle the switch below to turn on automatic preservation of your software.

☒

2 Create a release

Go to GitHub and [create a release](#). CaltechDATA will automatically download a .zip-ball of each new release and register a DOI.

3 Get the badge

After your first release, a DOI badge that you can include in GitHub README will appear next to your repository below.

DOI 10.5281/zenodo.8475

(example)

Enabled Repositories

 caltechlibrary/dataset

DOI 10.22002/D1.239

☒

 tmorrell/caltechdata_plot

DOI 10.22002/D1.240

☒

Repositories

If your organization's repositories do not show up in the list, please ensure you have enabled [third-party access](#) to the CaltechDATA application. Private repositories are not supported.

 AuthorCarpentry/CODATA-RDA-Trieste-2016

☐