Launching a Researcher-Focused Data Repository

Thomas Morrell
California Institute of Technology Library, Pasadena, CA, USA

Open Repositories 2017
June 28, 2017

Caltech Library

Current Research Data Practices

Most researchers store data on local computer hard drives

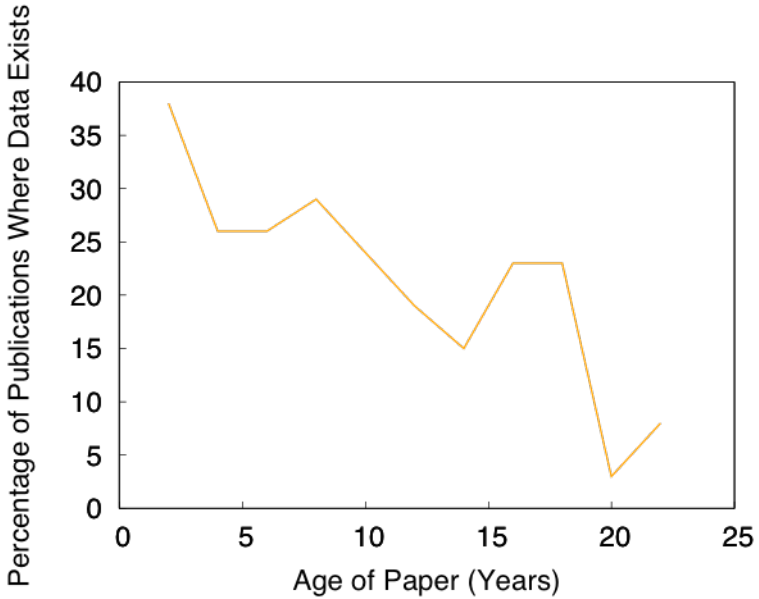
Researchers report that finding data is their biggest challenge

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

Who Maintains the Files?

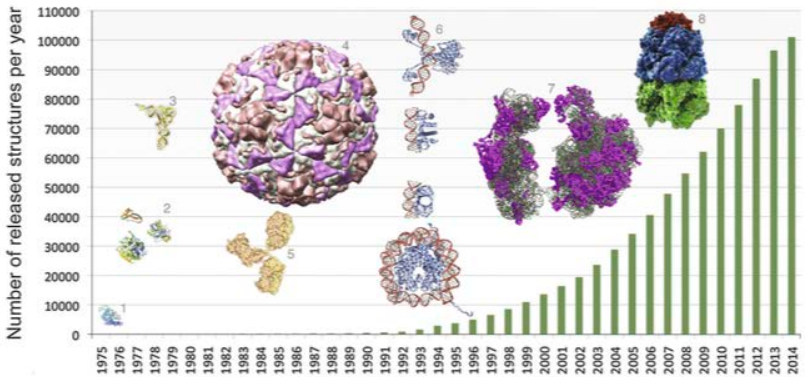


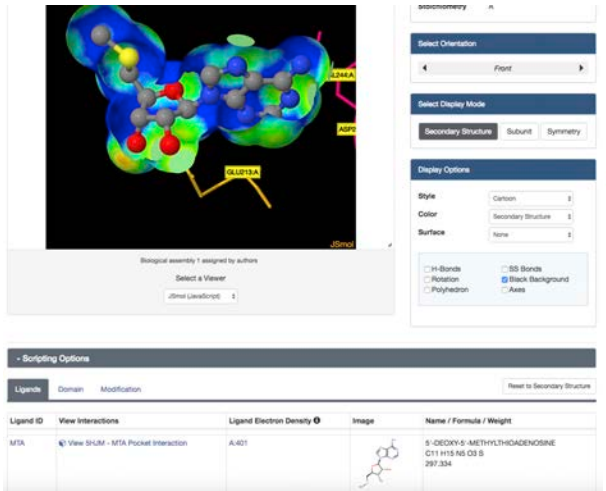
Vines, T. H. *et al.* The availability of research data declines rapidly with article age. *Curr. Biol.* **24,** 94–97 (2014).

Disciplinary Repositories

RCSB PDB
PROTEIN DATA BANK

www.rcsb.org





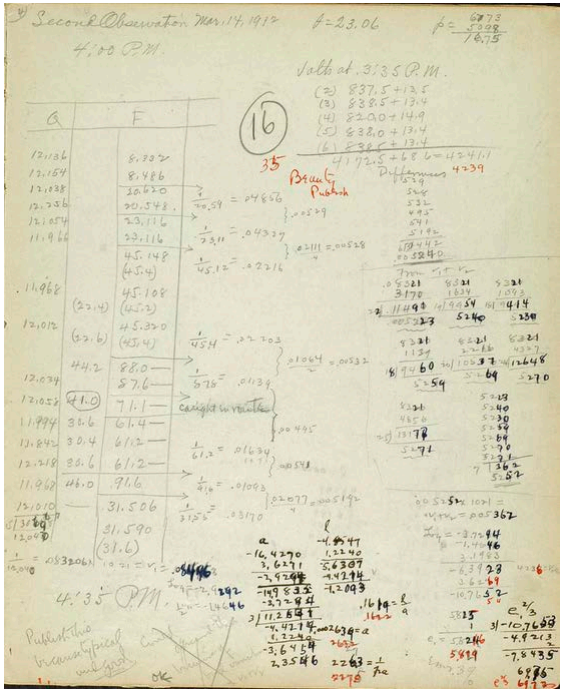
Caltech Library

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

Institutional repository?

Libraries have Institutional Backing

Libraries have a History of Managing Objects



Caltech Library

Robert A. Millikan Oil Drop Experiment Notebooks. Lab Notes Online. California Institute of Technology Archives. Retrieved 10/16 from the World Wide Web: http://resolver.caltech.edu/CaltechLN:LN_Millikan_R_2

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

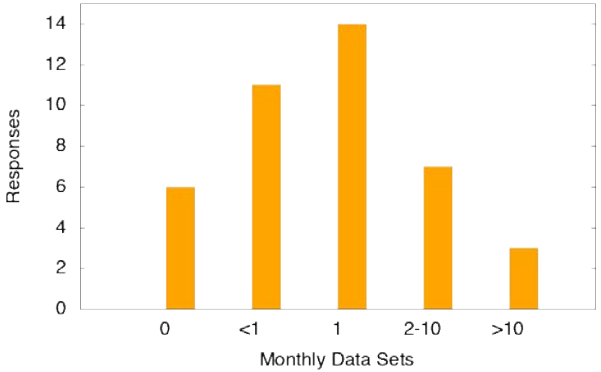
Caltech Library

http://www.caltech.edu/content-historic-awards-honors
www.keckobservatory.org

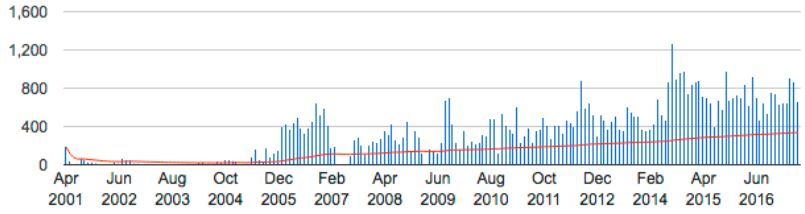
Data Repository Models

- 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

Caltech Library

SPEC Kit 354, Data Curation, May 2017, Association of Research Libraries (http://publications.arl.org/Data-Curation-SPEC-Kit-354/)
Repository sizes as of May 30, 2017

Invenio and TIND

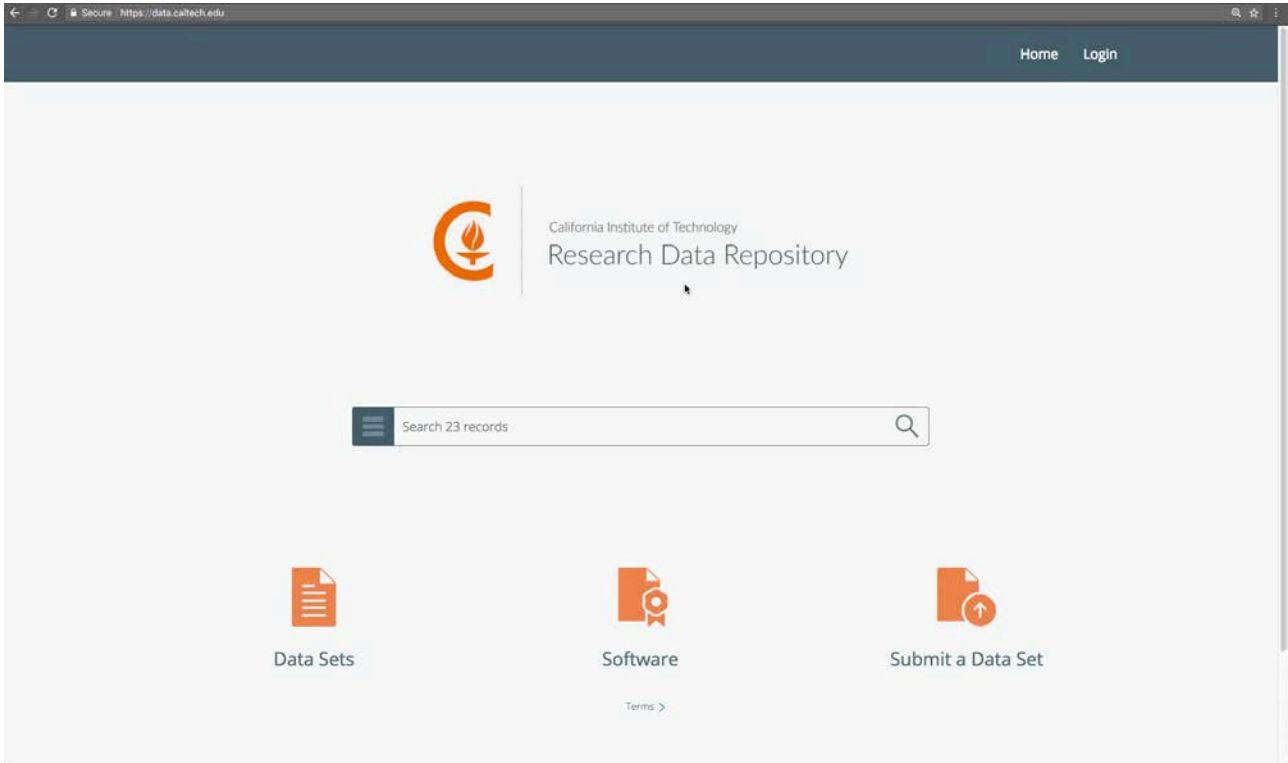
- Developed by CERN
- Invenio 1 – MARC based document management
  - CERN Document Server (cds.cern.ch)
- Invenio 3 – JSON based repository software
  - Zenodo Data Repository (zenodo.org)
  - HEPdata (hepdata.net)
- TIND is a CERN startup that does custom deployments of Invenio systems

Researcher-Focused Design Principles

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

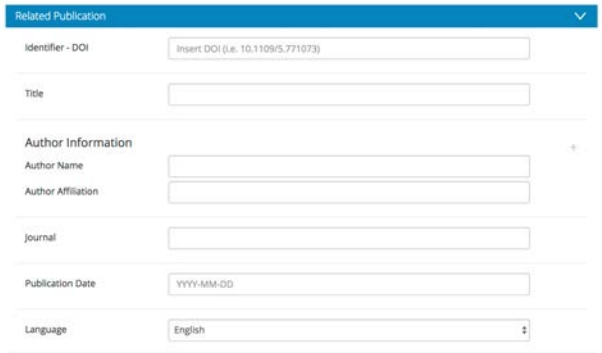
Single Deposit Form

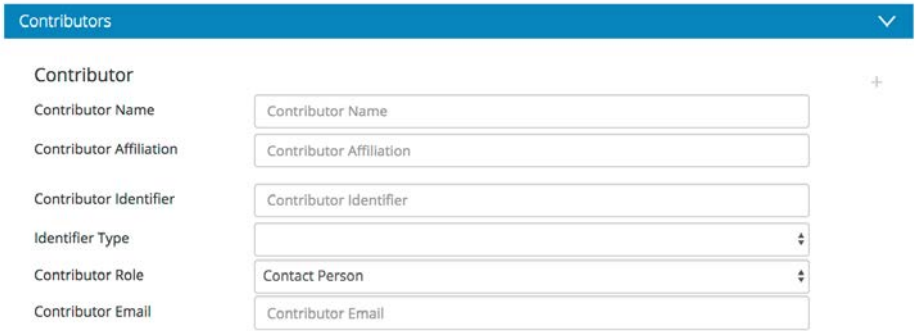
- Institutional (Shibboleth) login
- One page
- Minimal mandatory fields
- Collapsed sections
- Drag and drop file upload



Data Availability

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

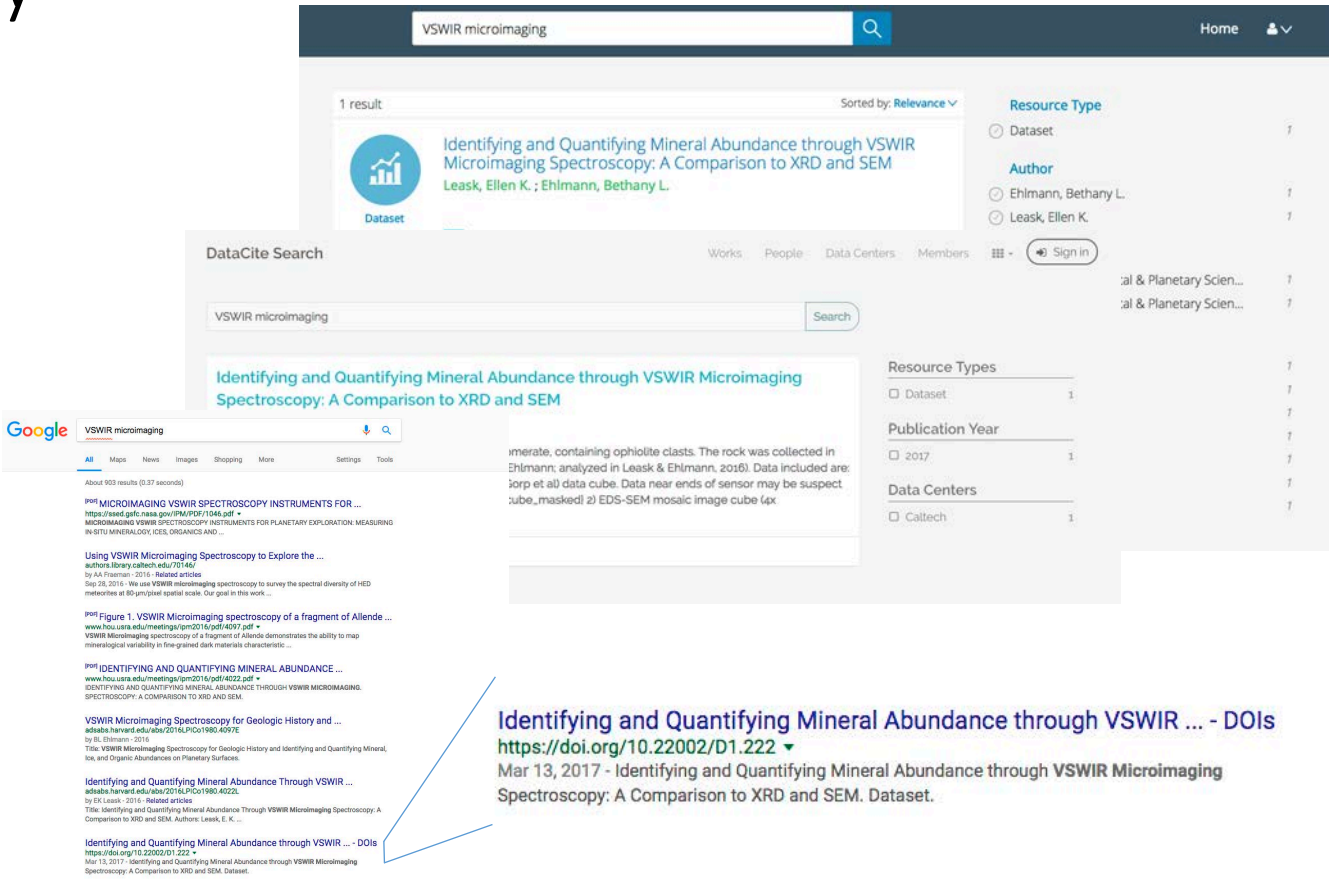




Caltech Library

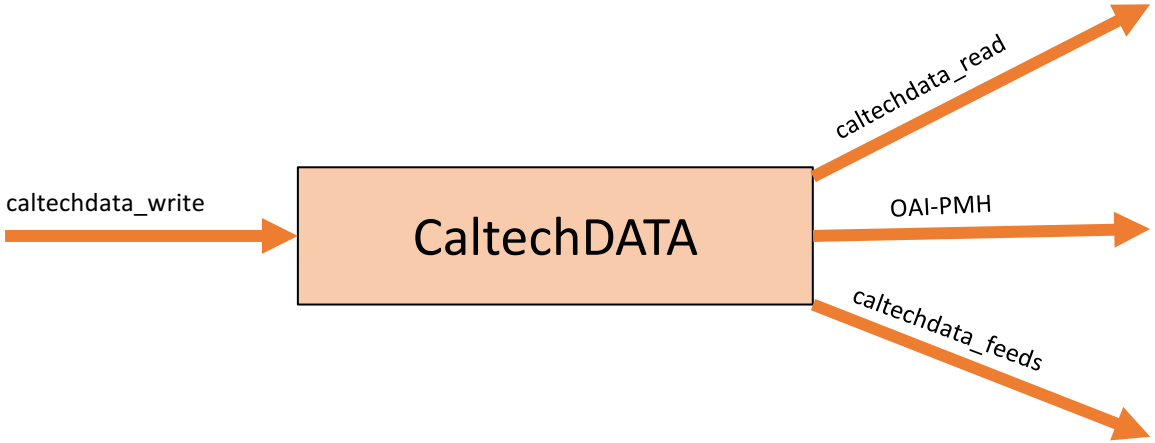
Discoverability

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



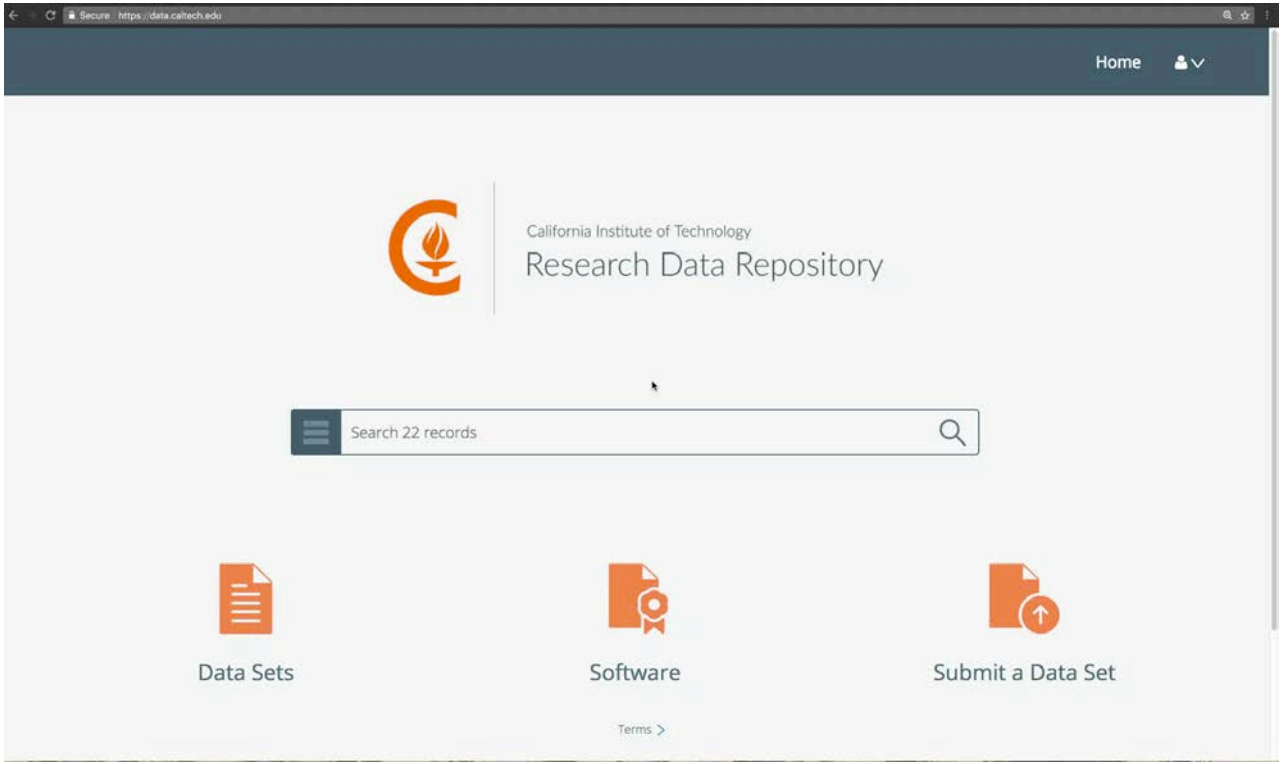
Caltech Library

Encourage re-use of data



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

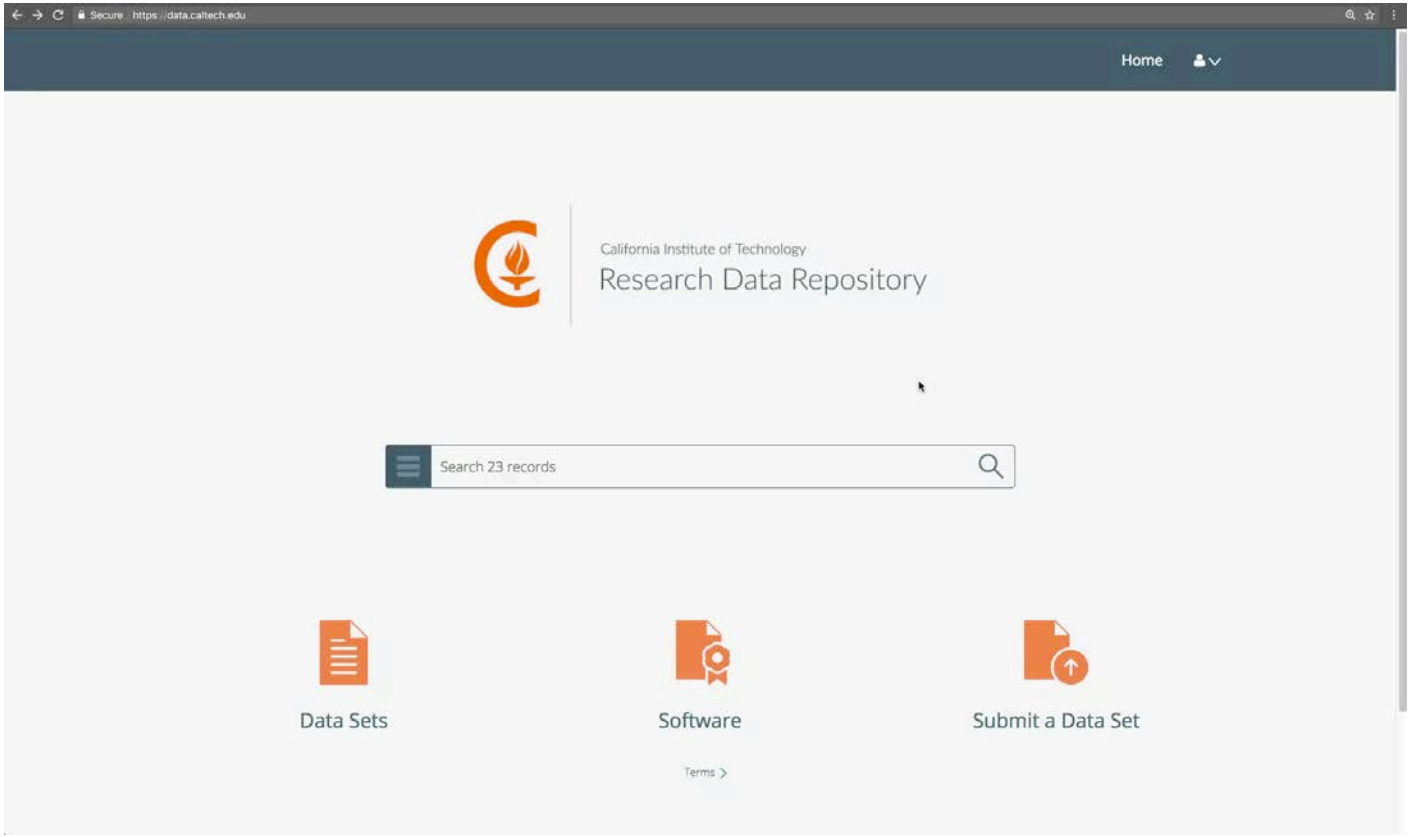
Read API Example: Interactive Plotting



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

Caltech Library

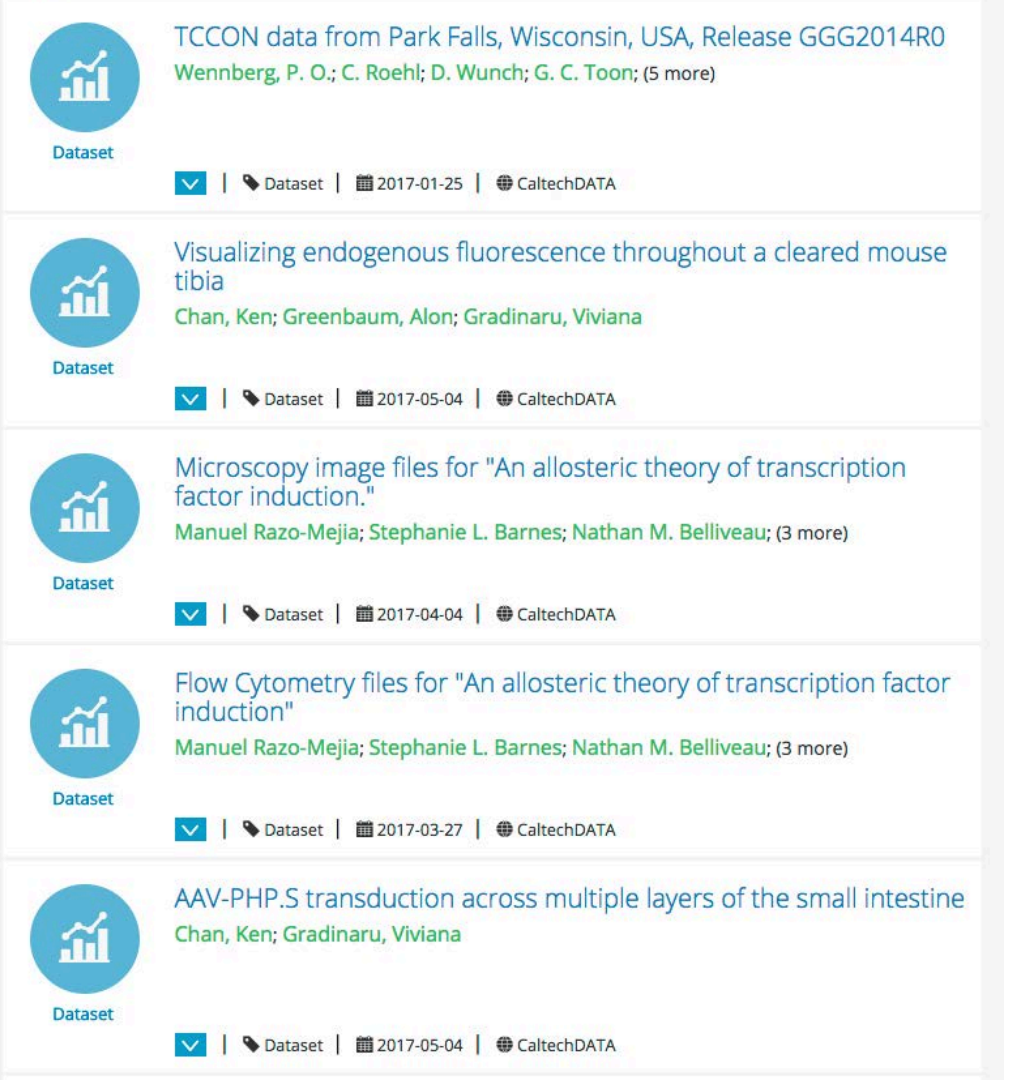
Automation



Caltech Library

Current Status

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

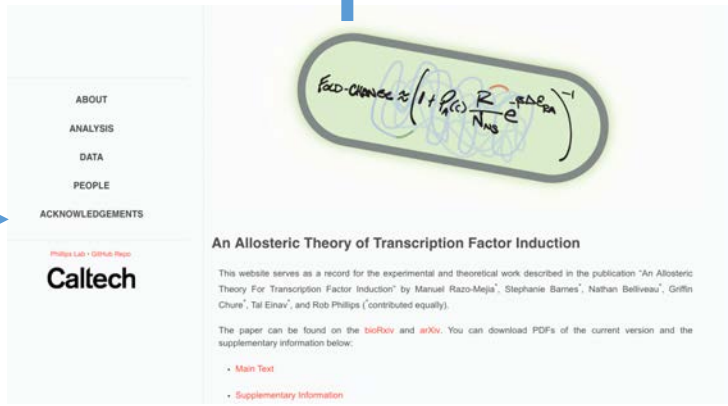


Caltech Library

Use Cases

https://doi.org/10.1101/111013

Paper Website on GitHub



Data Files

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

http://dx.doi.org/10.22002/D1.224
http://dx.doi.org/10.22002/D1.227
http://dx.doi.org/10.22002/D1.228
http://dx.doi.org/10.22002/D1.229

Caltech Library

Use Cases

Thesis Preparation

https://doi.org/10.7907/Z9NC5Z7H

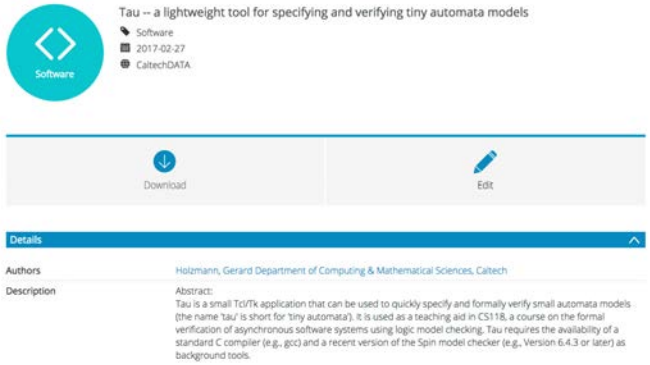
Data files uploaded during writing process

California Institute of Technology
Research Data Repository

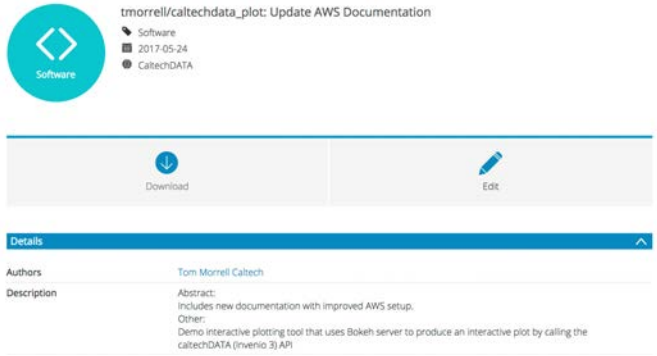
http://dx.doi.org/10.22002/D1.234
http://dx.doi.org/10.22002/D1.235
http://dx.doi.org/10.22002/D1.236
http://dx.doi.org/10.22002/D1.237

Software in CaltechDATA

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

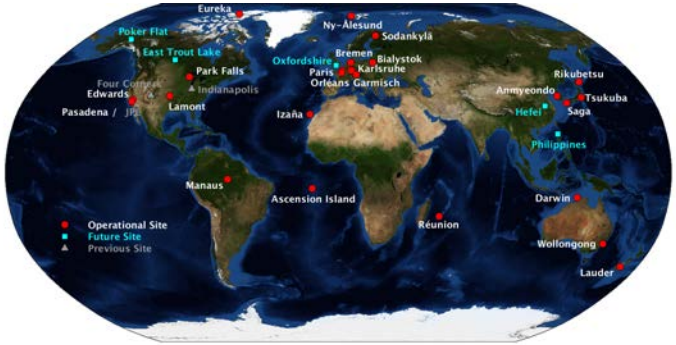


http://dx.doi.org/10.22002/D1.240



Caltech Library

Use Case - TCCON



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

Data files

CDIAC
Carbon Dioxide Information Analysis Center

OAK RIDGE
National Laboratory

Processed Data

Data Curation and Processing

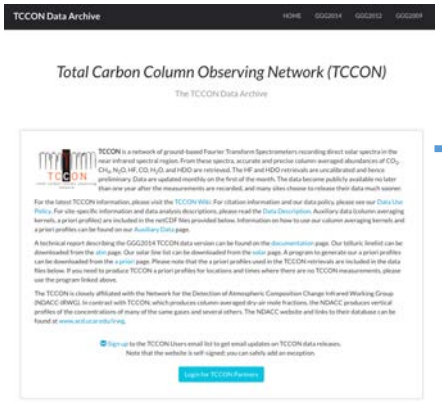
Caltech Library

https://tccon-wiki.caltech.edu/Sites/Park_Falls
http://tccon.ornl.gov/

Use Case - TCCON

OAK RIDGE National Laboratory

tccon.ornl.gov



Private data files

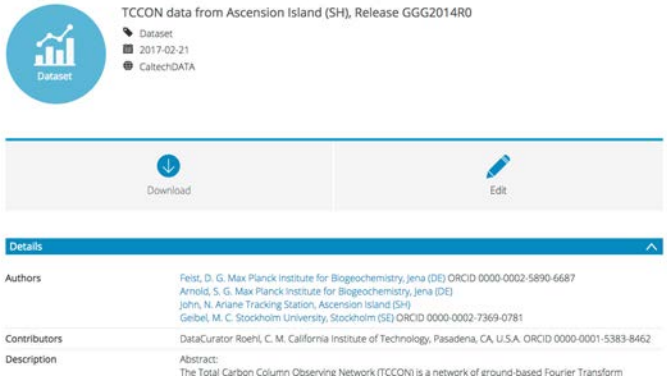
Public data files

Automatically released 1x/month

Migration Plan

Departmental Server at Caltech

Library-hosted CaltechDATA Repository



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

Caltech Library

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

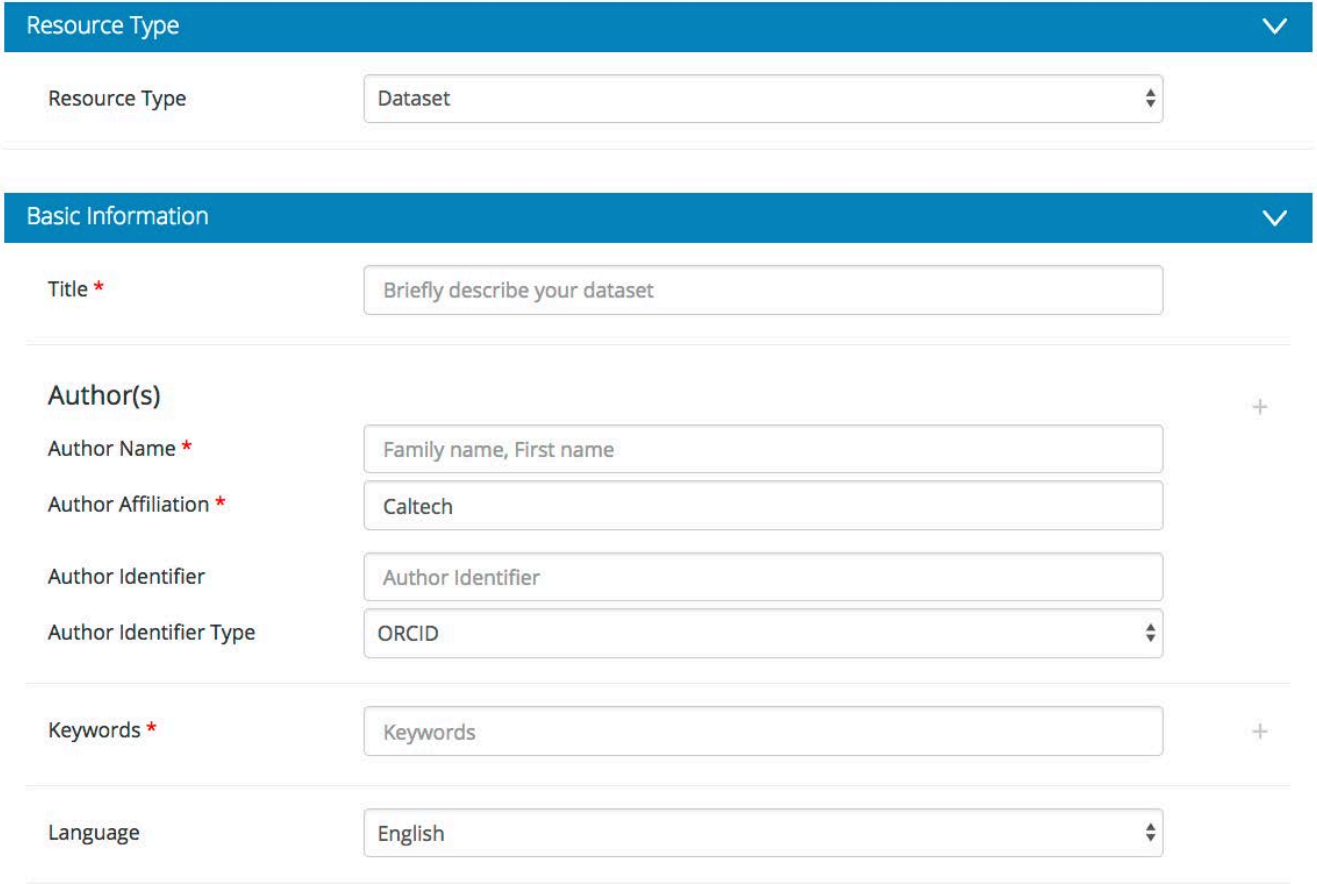
tmorrell@caltech.edu

data.caltech.edu

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

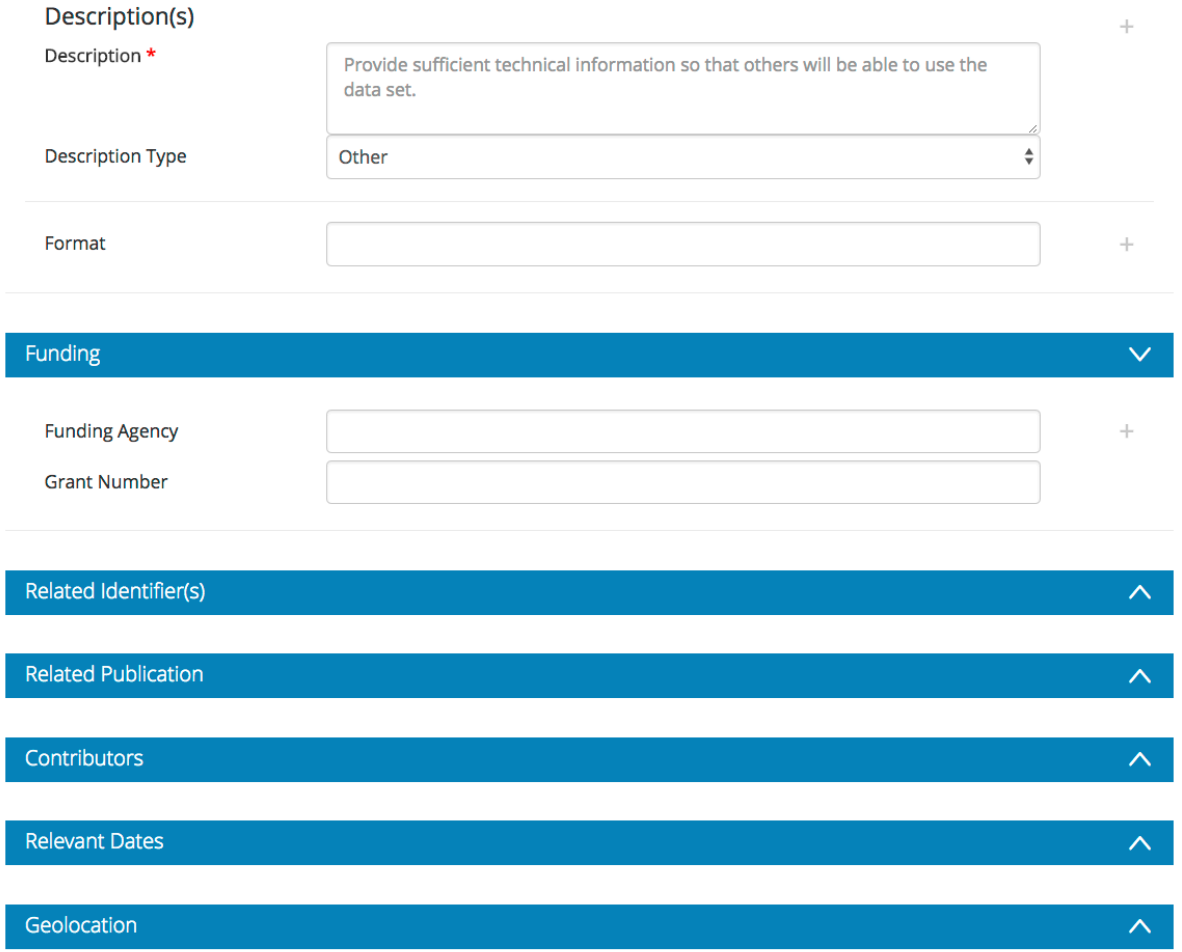
Caltech Library

Deposit Form 1



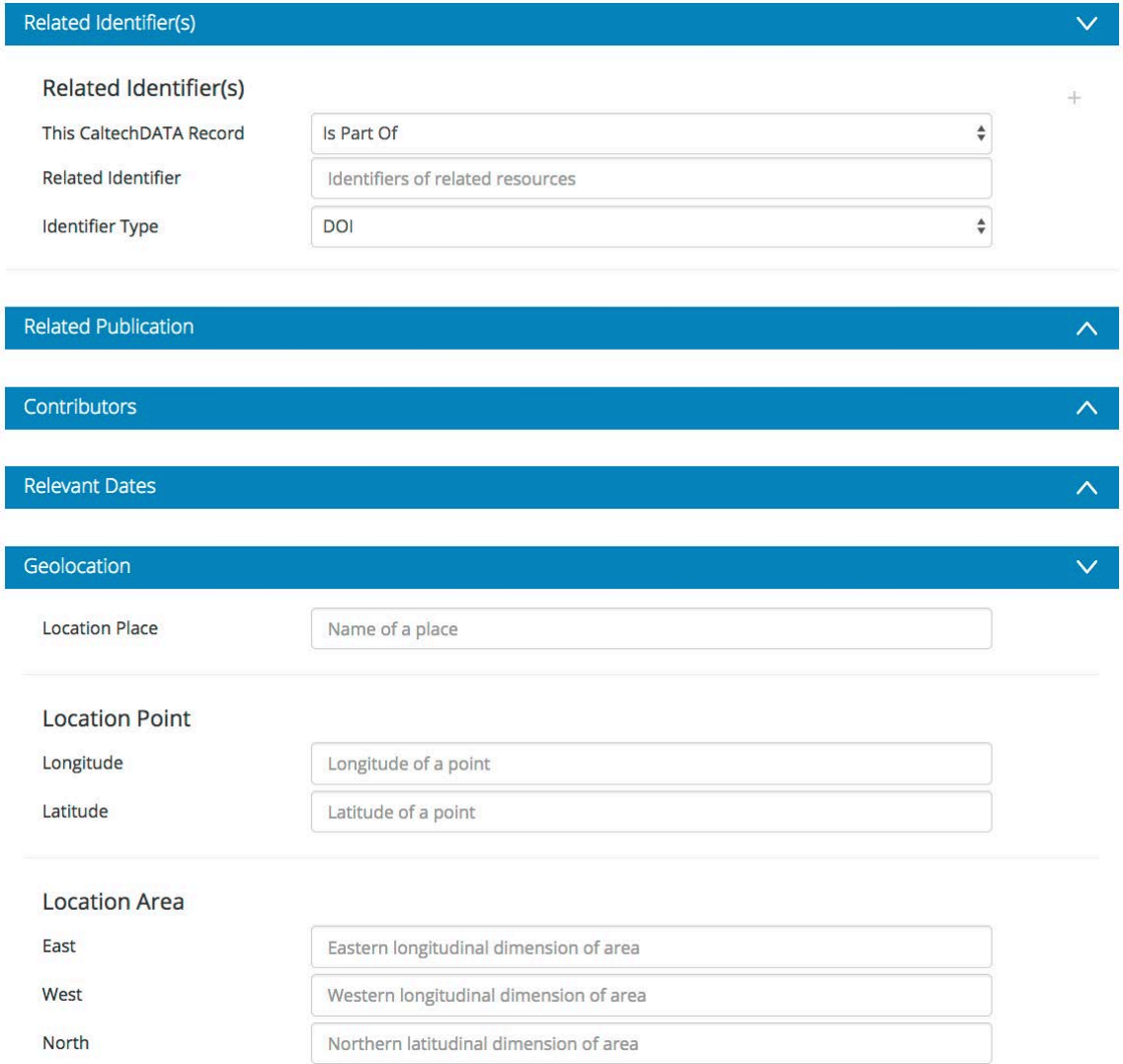
Caltech Library

Deposit Form 2



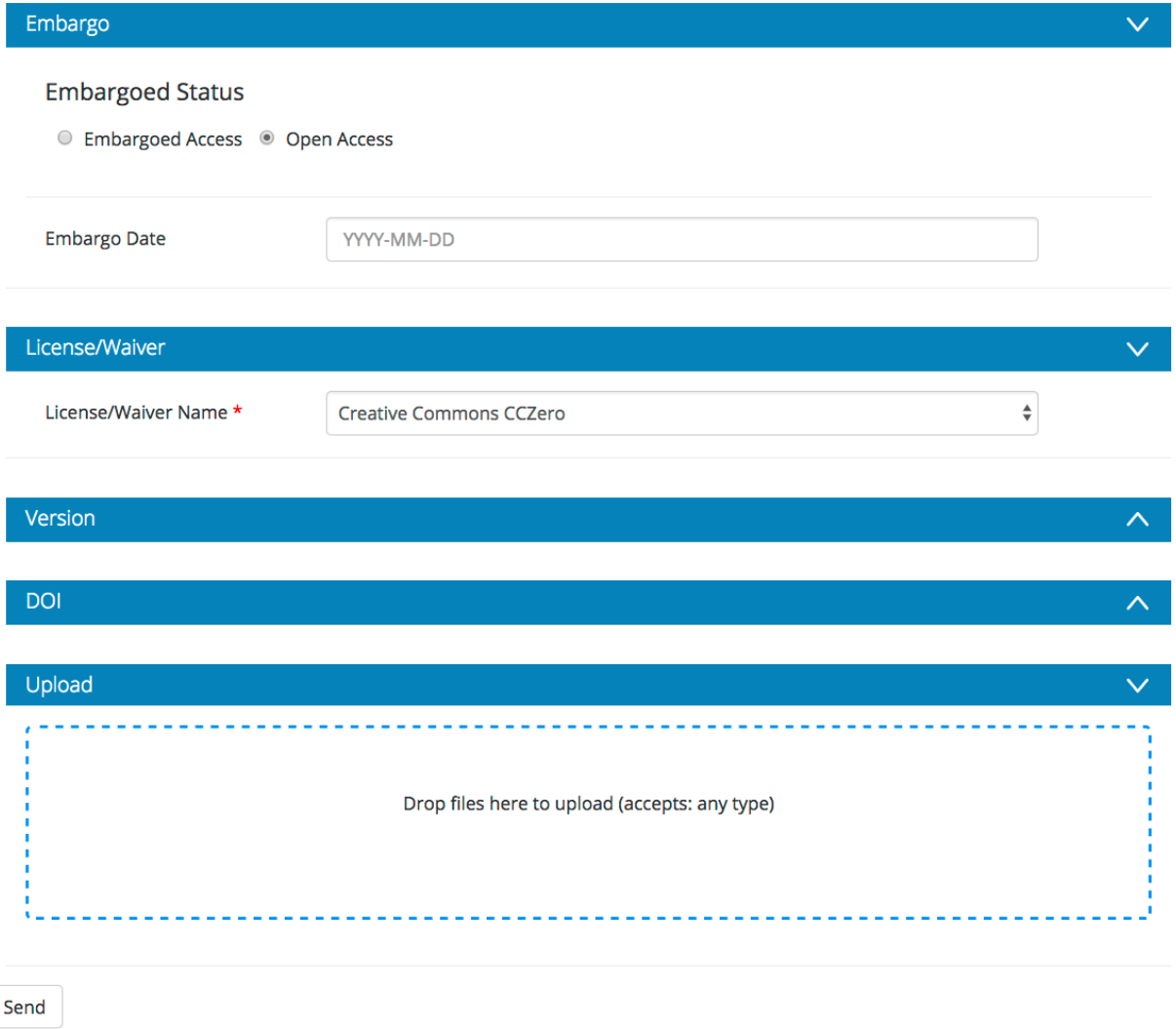
Caltech Library

Deposit Form 3



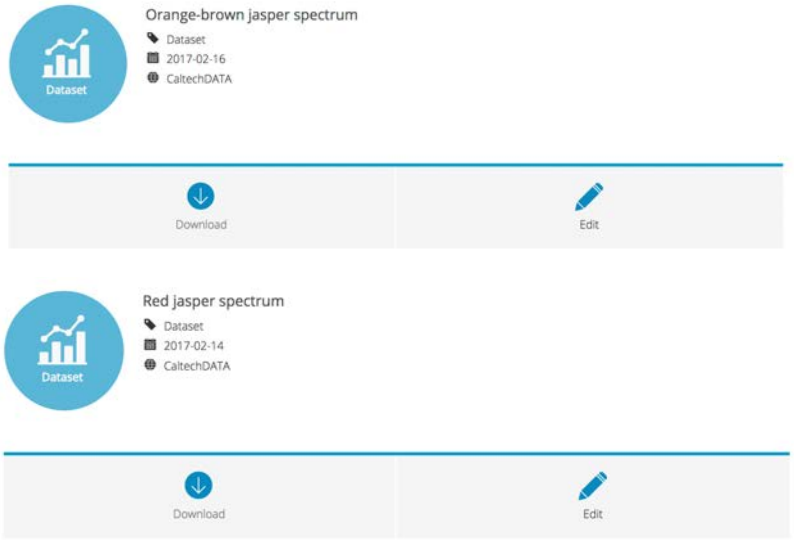
Caltech Library

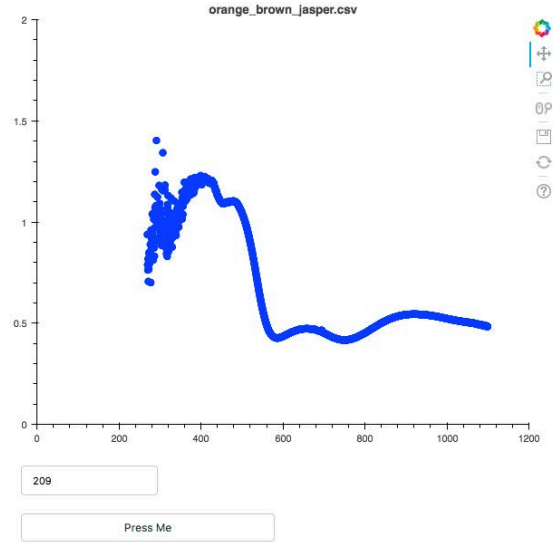
Deposit Form 4



Caltech Library

Read API Example: Interactive Plotting

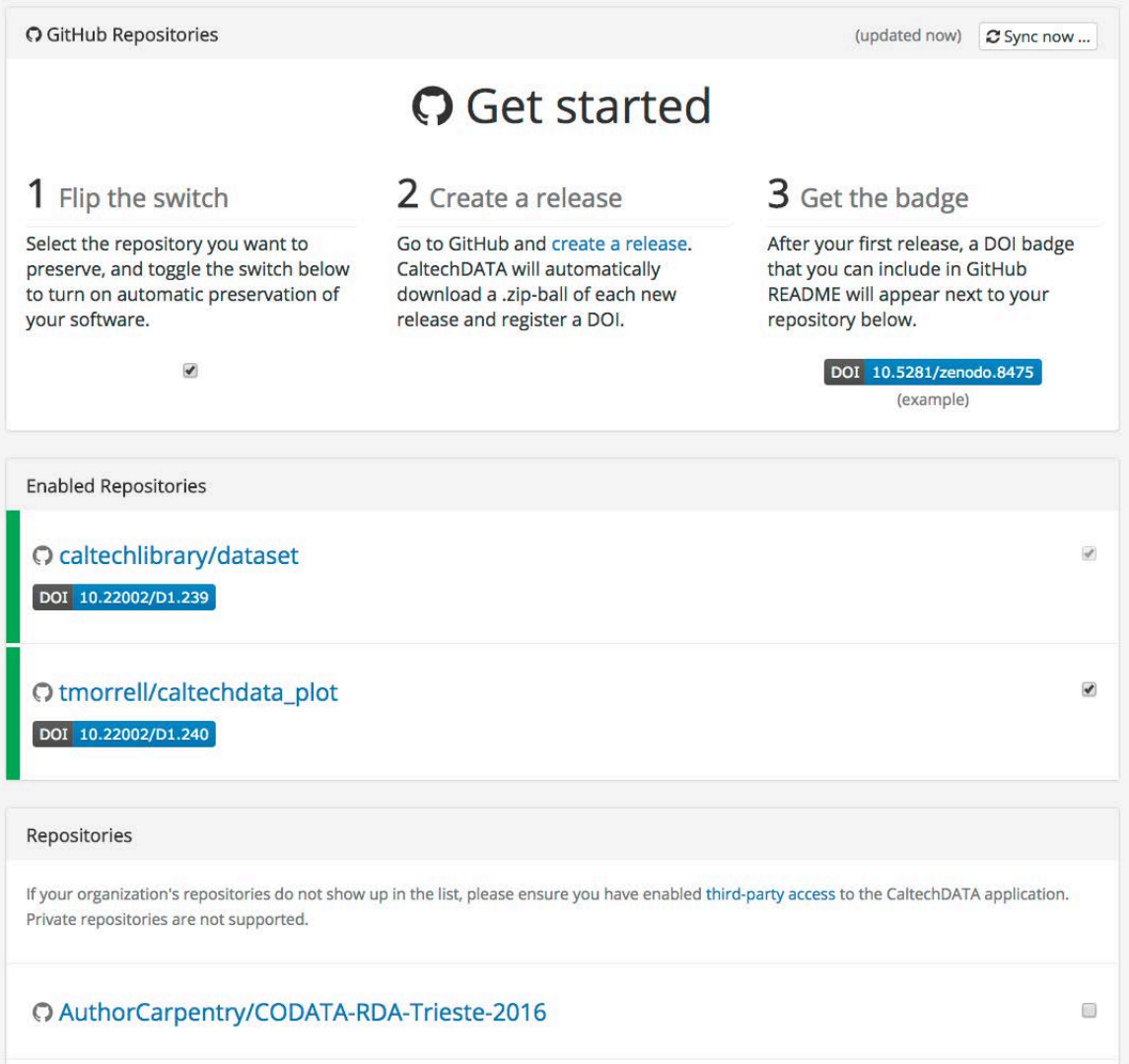




plots.caltechlibrary.org
doi.org/10.22002/D1.240

Caltech Library

Github



Caltech Library