Caltech Library

Deploying InvenioRDM as an institutional repository platform for data, software, and publications

Tom Morrell

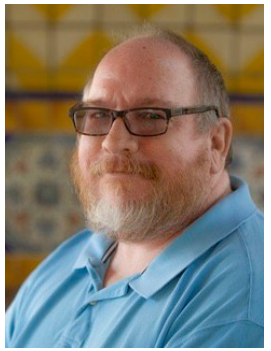
Coalition for Networked Information (CNI) Spring Meeting

April 4, 2023

https://doi.org/10.5281/zenodo.7799359

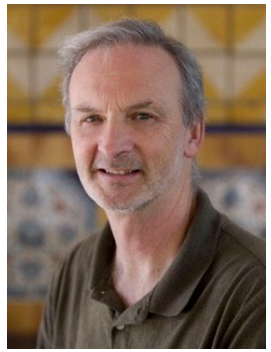
Caltech

- Big impact – 46 Nobel prizes
- Manages the Jet Propulsion Laboratory (JPL), Palomar and W. M. Keck Observatories, and co-manages LIGO
- Small - 300 faculty, 1,000 undergraduates, 1,400 graduate students
- Library has run institutional repositories since 2001; over 100,000 items


Robert Doiel


Tommy Keswick


Mike Hucka


Stephen Davison


Kathy Johnson


George Porter

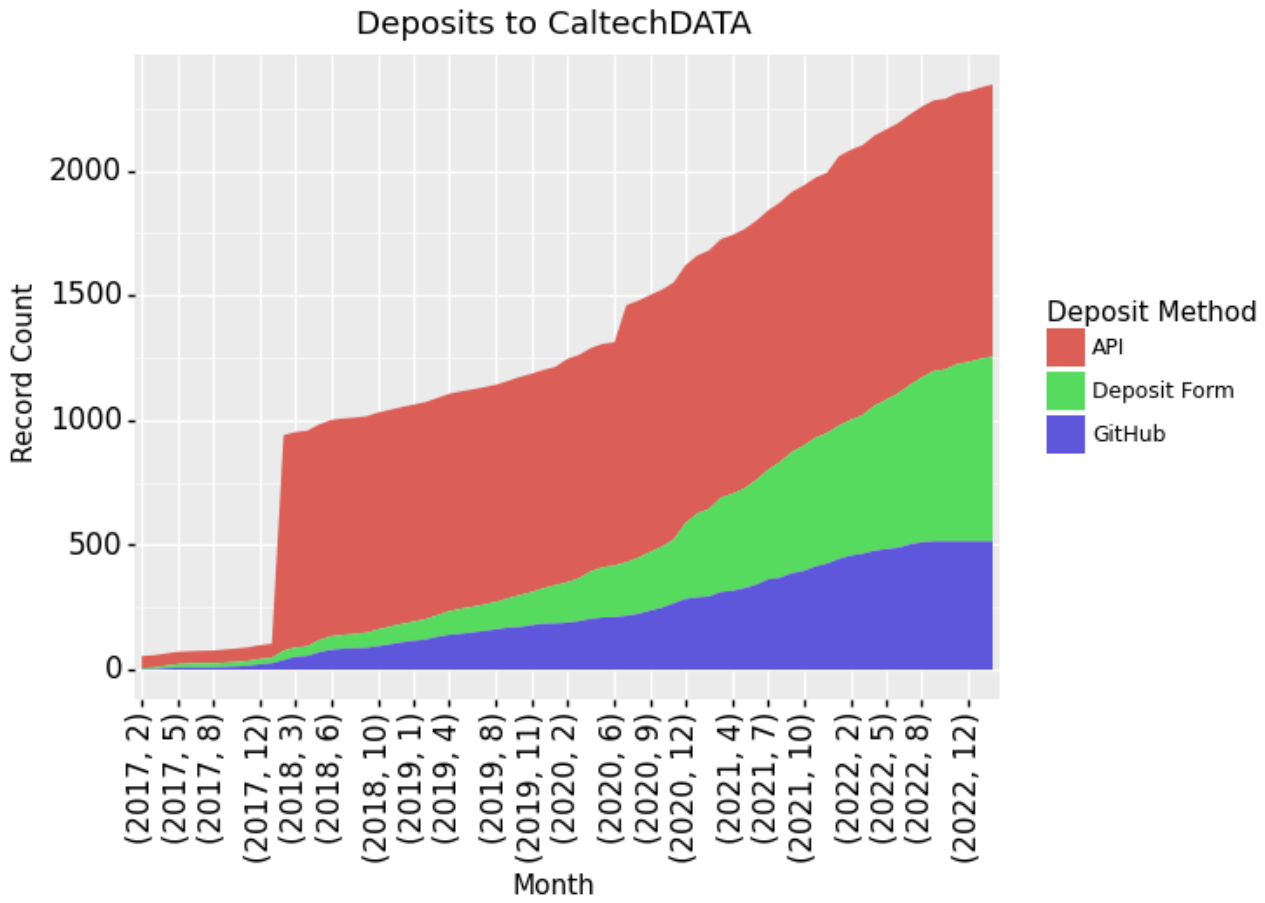

Tony Diaz

Caltech Library

https://www.caltech.edu/about
www.keckobservatory.org

CaltechDATA

- Institutional repository for Caltech researchers (https://data.caltech.edu)
- Started in 2017
- Grown to over 26,000 records; over 10 TB of storage
- Submissions from over 6% of campus



Note: Excludes ~23,000 API-migrated records from MEAD

Caltech Library

What's in CaltechDATA?

Table 1. Compilation of dip angles and subduction parameters along the 153 transects.

Index	[lat lon]	Trench	φ_{azi} (°)	V_{conv} (cm/yr)	V_{up} (cm/yr)	V_{sub} (cm/yr)	T_{conv} (km)	A_{slab} (Ma)	A_{re} (Ma)	A_{long} (Ma)	OPN	D_{Moho} (km)
1	[60.09 -145.46]	AL	323.36	4.85	-0.11	4.75	2441.60	42.49	153-120	230	c	40.90
2	[58.59 -149.20]	AL	320.16	5.12	-0.13	4.99	2544.49	45.70	153-120	230	c	45.27
3	[56.98 -152.45]	AL	315.32	5.34	-0.23	5.12	2640.72	48.47	153-120	230	c	34.47
4	[55.44 -155.22]	AL	314.56	5.60	-0.19	5.41	2726.20	49.82	153-120	230	c	32.06
5	[54.31 -158.11]	AL	326.59	5.88	0.31	6.20	2805.70	51.45	153-120	230	c	32.35
6	[53.33 -161.60]	AL	337.29	5.85	0.77	6.62	2891.46	53.01	153-120	230	c	30.37
7	[52.47 -165.32]	AL	340.66	5.84	0.95	6.79	2975.09	54.83	153-120	230	c	28.68
8	[51.64 -168.96]	ALU	343.50	5.78	1.11	6.89	3052.13	57.64	69	NA	c	23.60
9	[50.90 -172.64]	ALU	340.64	5.98	1.08	7.06	3123.34	59.86	69	NA	c	23.71
10	[50.60 -176.38]	ALU	352.91	5.18	1.47	6.64	3181.51	53.98	69	NA	c	20.94
11	[50.46 179.93]	ALU	6.93	3.83	1.78	5.61	3231.32	52.14	69	NA	c	22.16
12	[50.85 176.23]	ALU	10.74	3.30	1.84	5.14	3266.61	49.08	69	NA	c	17.41
13	[51.82 172.64]	ALU	25.23	1.47	1.92	3.40	3286.43	42.66	69	NA	c	18.16
14	[9.23 -85.37]	MAM	39.04	8.47	-0.63	7.84	1843.77	18.69	75-73	NA	c	39.25
15	[15.17 -96.80]	MEX	9.77	6.04	0.13	6.17	2566.01	16.48	NA	220	c	39.53
16	[7.85 -83.60]	MAM	37.13	8.93	-0.80	8.13	1822.99	14.74	75-73	NA	c	29.59
17	[10.90 -87.39]	MAM	27.60	8.14	-0.57	7.57	1864.96	61.93	75-73	NA	c	33.69
18	[15.8 -98.8]	MEX	22.50	6.16	0.70	6.87	2501.70	9.67	NA	220	c	40.20
19	[11.85 -89.30]	MAM	27.82	7.74	-0.43	7.31	1881.52	23.27	75-73	NA	c	39.65
20	[16.6 -100.7]	MEX	21.24	5.59	0.67	6.26	2431.92	7.67	NA	220	c	40.57
21	[12.67 -90.96]	MAM	32.29	7.35	-0.27	7.08	1893.90	24.86	75-73	NA	c	41.55
22	[17.6 -102.9]	MEX	27.54	5.10	0.92	6.03	2347.35	4.91	NA	220	c	36.43
23	[13.70 -92.91]	MAM	26.87	6.95	-0.17	6.78	1906.25	25.42	75-73	NA	c	35.04
24	[18.50 -104.40]	MEX	46.60	4.65	1.34	5.98	570.27	7.65	NA	220	c	35.15

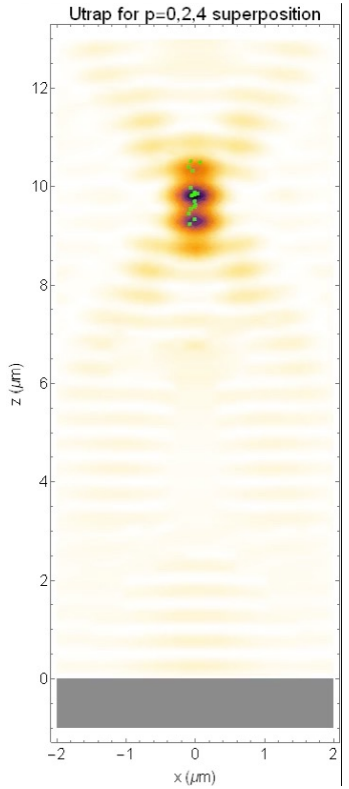
Hu, Jiashun and Gurnis, Michael; Supplementary Data for manuscript entitled "Subduction Duration and Slab Dip" https://10.22002/D1.1380

```
# csv file with metadata information for each dataset
# must have the species information, subsampling levels for reads, total reads,
# technology used and path to one fastq file for R1 and one for R2
# if you have multiple files for each read tehy must be concatenated first
metadatas=pd.read_csv('./cell-depth-tradeoff-metadata.tsv', sep='\t')

def make_t2g_path(wildcards):
    DATASET_SAMPLE_ID = wildcards.dataset_sample_id
    species = metadatas[metadatas['dataset_sample_id']==DATASET_SAMPLE_ID]['species'].values[0]
    if species=='mouse':
        T2G = os.path.join(REF_PATH,'mus_musculus-ensembl-96/transcripts_to_genes.txt')
    if species=='human':
        T2G = os.path.join(REF_PATH,'homo_sapiens-ensembl-96/transcripts_to_genes.txt')
    return T2G
```

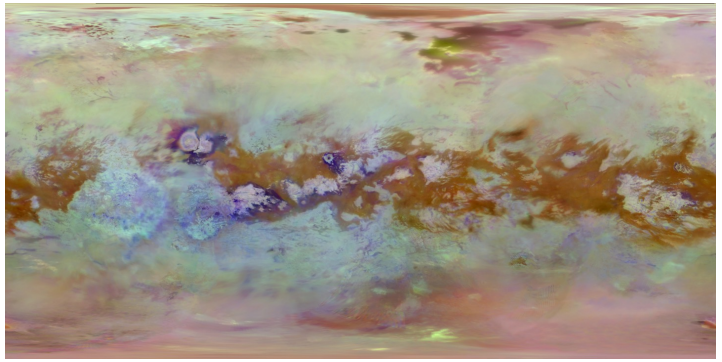
Valentine Svensson, Eduardo da Veiga Beltrame, and Lior Pachter. Code for producing the analysis in the "Quantifying the tradeoff between sequencing depth and cell number in single-cell RNA-seq" https://10.22002/D1.1276

J.-B. Béguin et al. Reduced volume and reflection for optical tweezers with radial Laguerre-Gauss beams https://10.22002/D1.1346

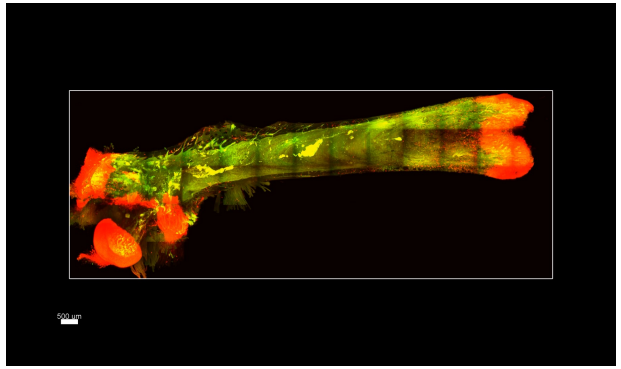


Iwashita, Yumi et al.; JPL Mars Yard Database https://10.22002/D1.1332

Witkosky, Rebecca; Mako thermal infrared hyperspectral airborne emissivity image, field photographs, and ground-based spectra of the San Andreas fault and Thousand Palms Oasis in the Coachella Valley, California https://10.22002/D1.1236



Seignovert, Benoît et al. Titan's global map combining VIMS and ISS mosaics https://10.22002/D1.1173



Chan, Ken , Greenbaum, Alon, Gradinaru, Viviana; Visualizing endogenous fluorescence throughout a cleared mouse femur https://10.22002/D1.1234

Caltech Library

Repository details

- The initial version of CaltechDATA was inspired by Zenodo
  - Easy to describe and upload files
  - Researchers control their records
  - All records get a DOI
  - Integration with GitHub
  - API for accessing data
- Lots of other institutions had the same idea

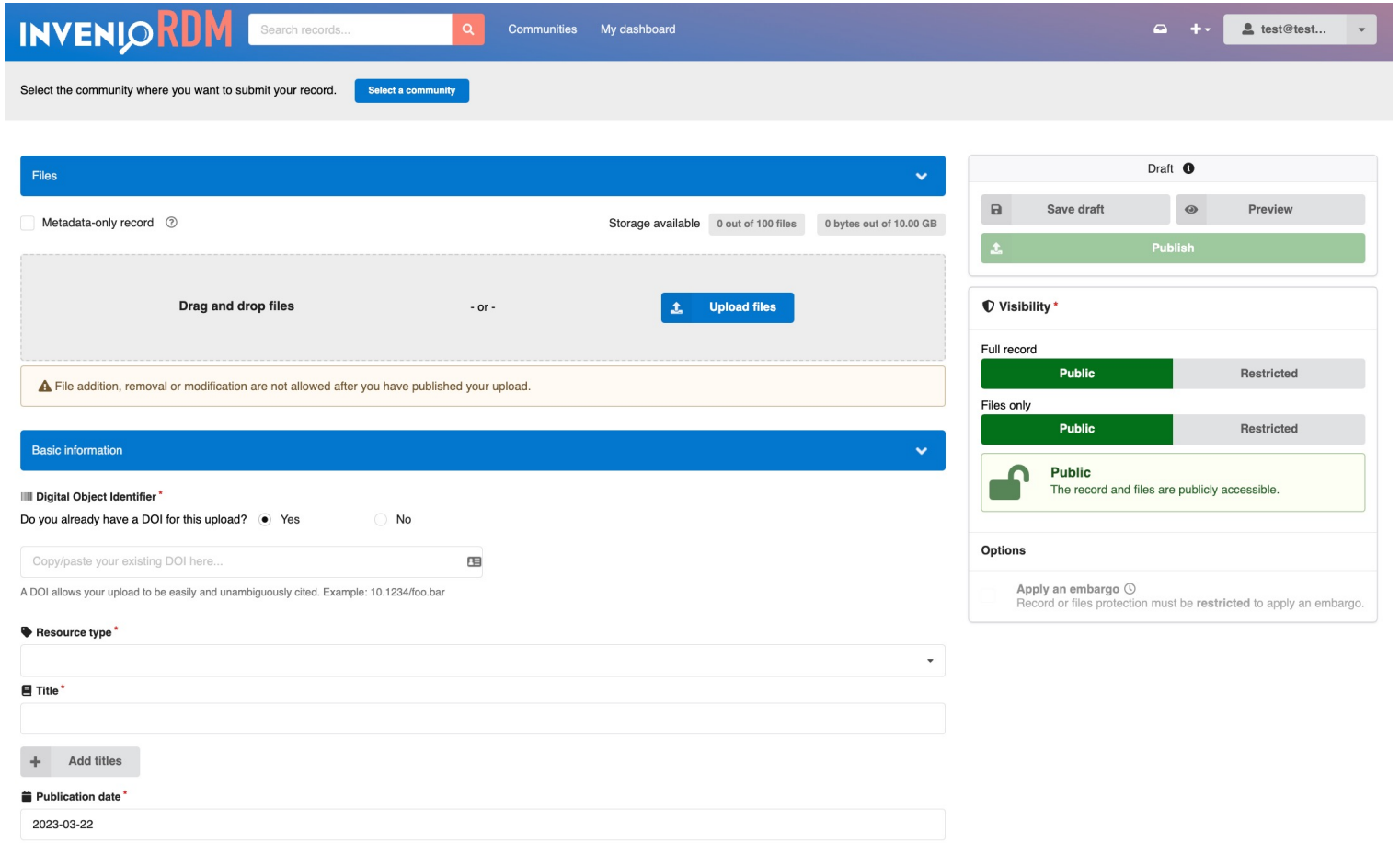
InvenioRDM Partners

Introducing InvenioRDM

- InvenioRDM is build on the Invenio repository platform (Python!)
- Inspired by Zenodo, but customizable by institutions
- Designed around data and software, but supports all item types
- CaltechDATA was an early migration; Zenodo itself is migrating this fall

Built-in features

- User-friendly deposit form
  - Auto-complete
    - Creators and contributors with ORCIDS
    - Affiliation with RORs
    - Subjects
    - Awards
    - Funders with RORs
  - Drag and drop file upload
  - Automatic DOI registration
  - Draft records
- Community record curation



Want to try it out? https://inveniordm.web.cern.ch/

Caltech Library

Migration requirements

- Move all ~20,000 records and files
- Customize the repository for Caltech
  - Theming
  - ORCIDs
- Ensure API integrations continued to work

Caltech DATA
by Caltech Library

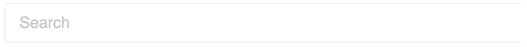
Migration strategy

- Relied on standard DataCite metadata
  - We used and validated our exporter as part of our backup and API work
    - https://github.com/caltechlibrary/caltechdata_api
- Exported all metadata and files
- Imported records using the InvenioRDM API
- Switched from old to new repository once all records were in place

Metadata enhancements

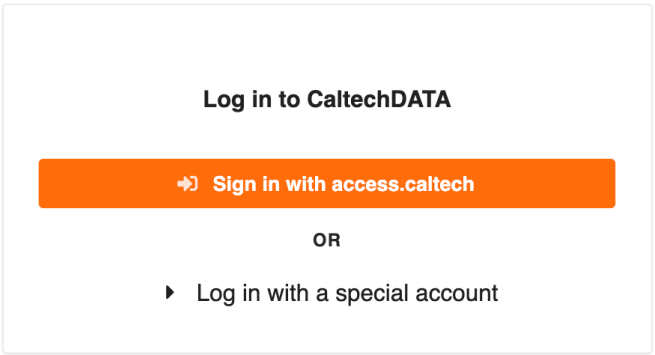
- ROR didn’t exist when CaltechDATA started
  - Started with automatic mapping from https://github.com/Metadata-Game-Changers/RORRetriever, followed by manual verification
  - Mapped and split free-text affiliations
  - Mapped funders as well
- Other minor cleanup, like splitting subjects
- https://github.com/caltechlibrary/inveniordm-migrate/blob/master/migrate_caltechdata.py

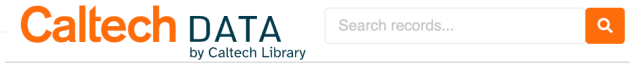
Theming



Data Sets

Software





Published March 20, 2023 | Version v1

The Materials Experiment Knowledge Graph

Statt, Michael[1]; Rohr, Brian[1]; Guevarra, Dan[2]; Breeden, Ja'Nya[2]; Suram, Santosh[3]; Gregoire, John[2]

Show affiliations

Citation

Style APA

Statt, M., Rohr, B., Guevarra, D., Breeden, J., Suram, S., & Gregoire, J. (2023). The Materials Experiment Knowledge Graph [Data set]. CaltechDATA. https://doi.org/10.22002/h88fq-dk449

Description

A Neo4j dump of the Materials Experiment Knowledge Graph (MEKG), a graph database containing metadata and experimental provenance from synthesis to measurement of primarily metal oxide solid state materials resulting from high throughput experimentation. The MEKG is a practical graph restructuring of the Materials Provenance Store, designed for greater querying efficiency through explicit entity relationships which would otherwise require slow table joins in the SQL paradigm. The higher search efficiency of long paths (equivalently multi-table joins) comes at the cost of slower aggregation operations.

Files

Files (5.5 GB)

Name	Size	
public-release-neo4j-20230228.dump md5:340d15e83ffea78c09e36fe19b10e594	5.5 GB	Download

Versions

Version v1 Mar 20, 2023
10.22002/h88fq-dk449

Details

DOI

DOI 10.22002/h88fq-dk449

Resource type

Dataset

Publisher

CaltechDATA

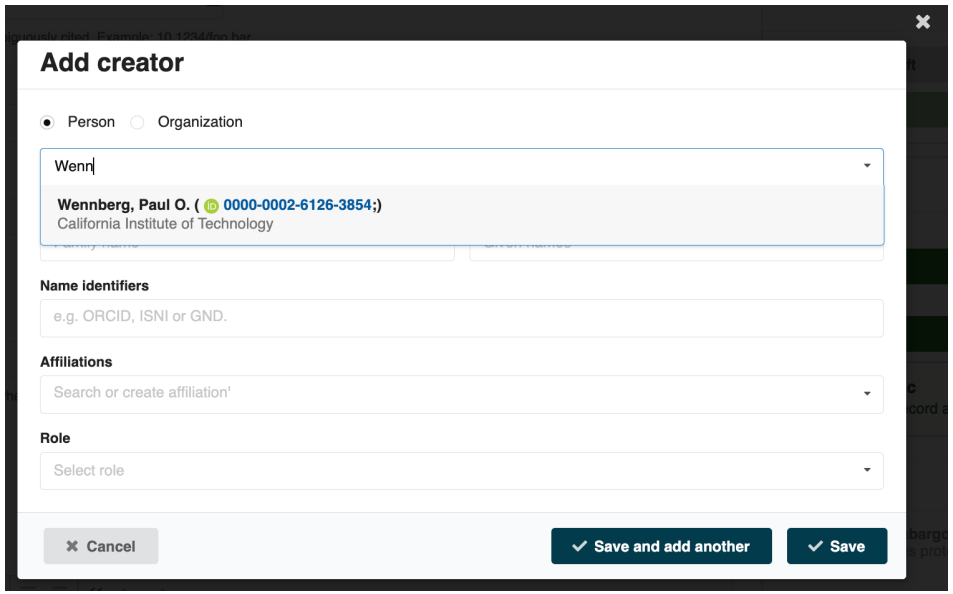
Rights

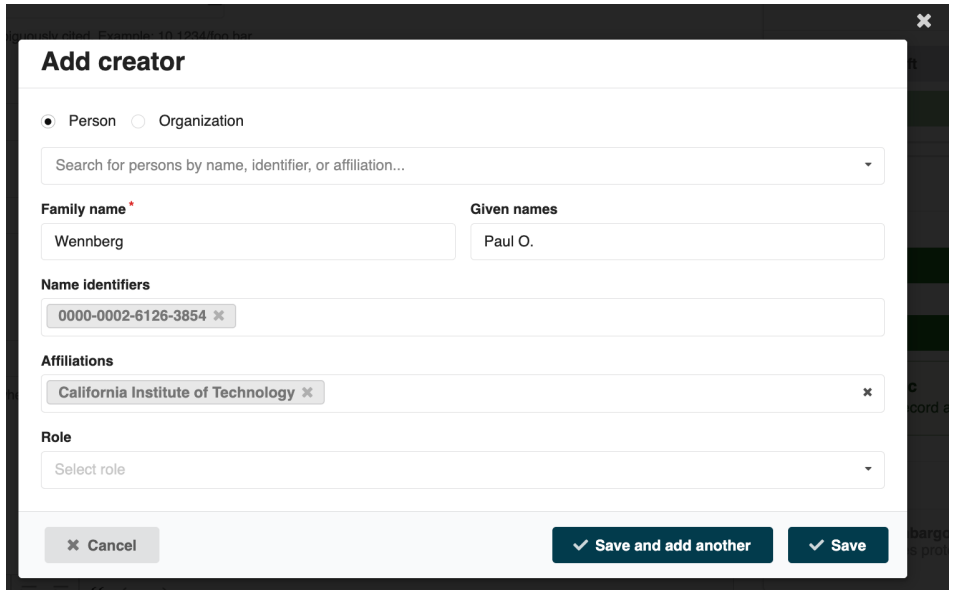
Creative Commons Attribution 4.0 International

Caltech Library

CaltechPEOPLE

- Library-wide effort to identify people associated with Caltech and their ORCIDS
- Powers our metadata service https://feeds.library.caltech.edu/
- Added to InvenioRDM as a name vocabulary





Caltech Library

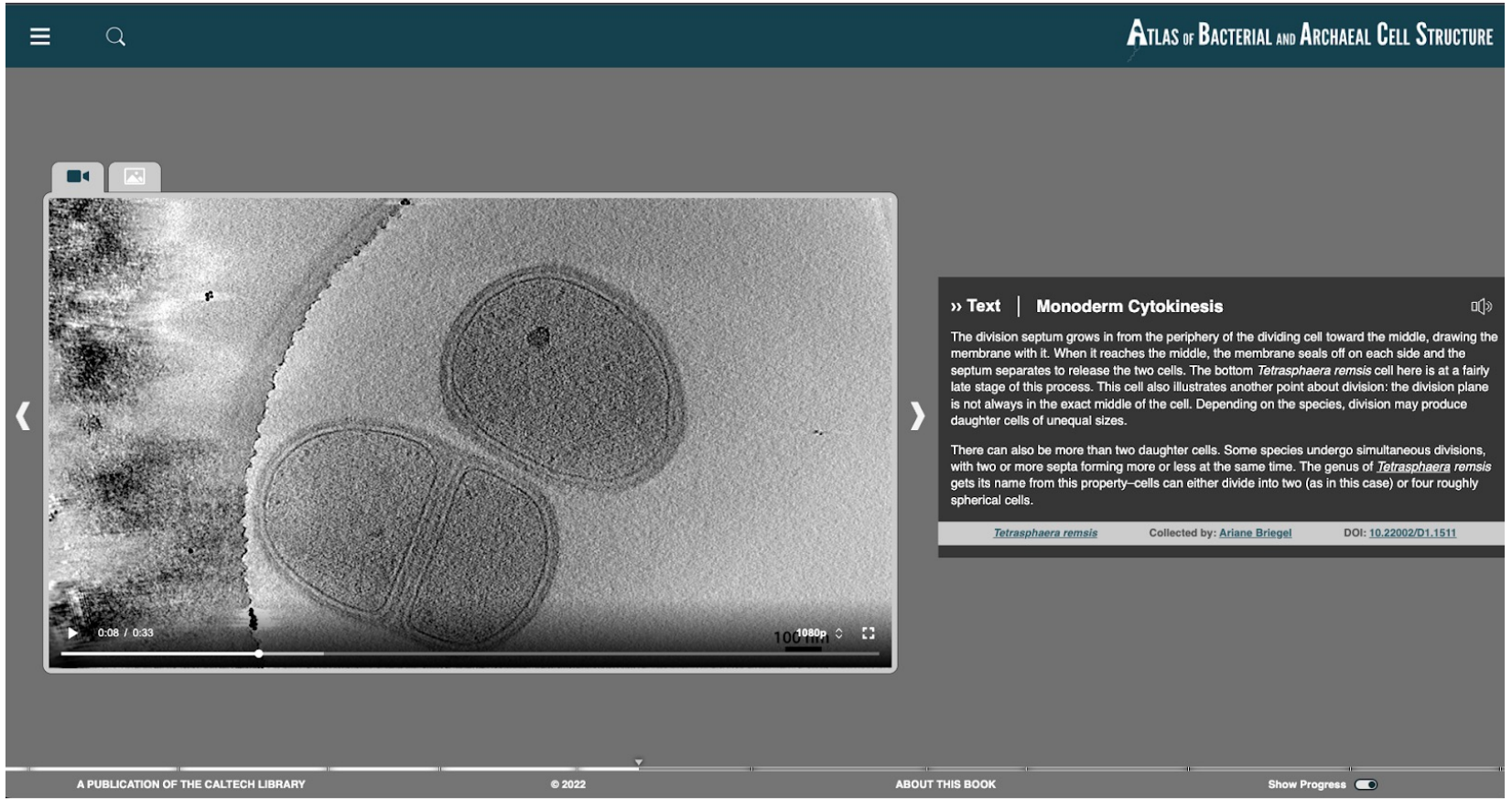
Automation with APIs

- Cell Atlas
- TCCON
- Micropublication

More details:
https://www.cni.org/topics/digital-curation/a-sustainable-lightweight-approach-to-digital-content-management-and-publication

CaltechDATA and the Cell Atlas

https://cellstructureatlas.org

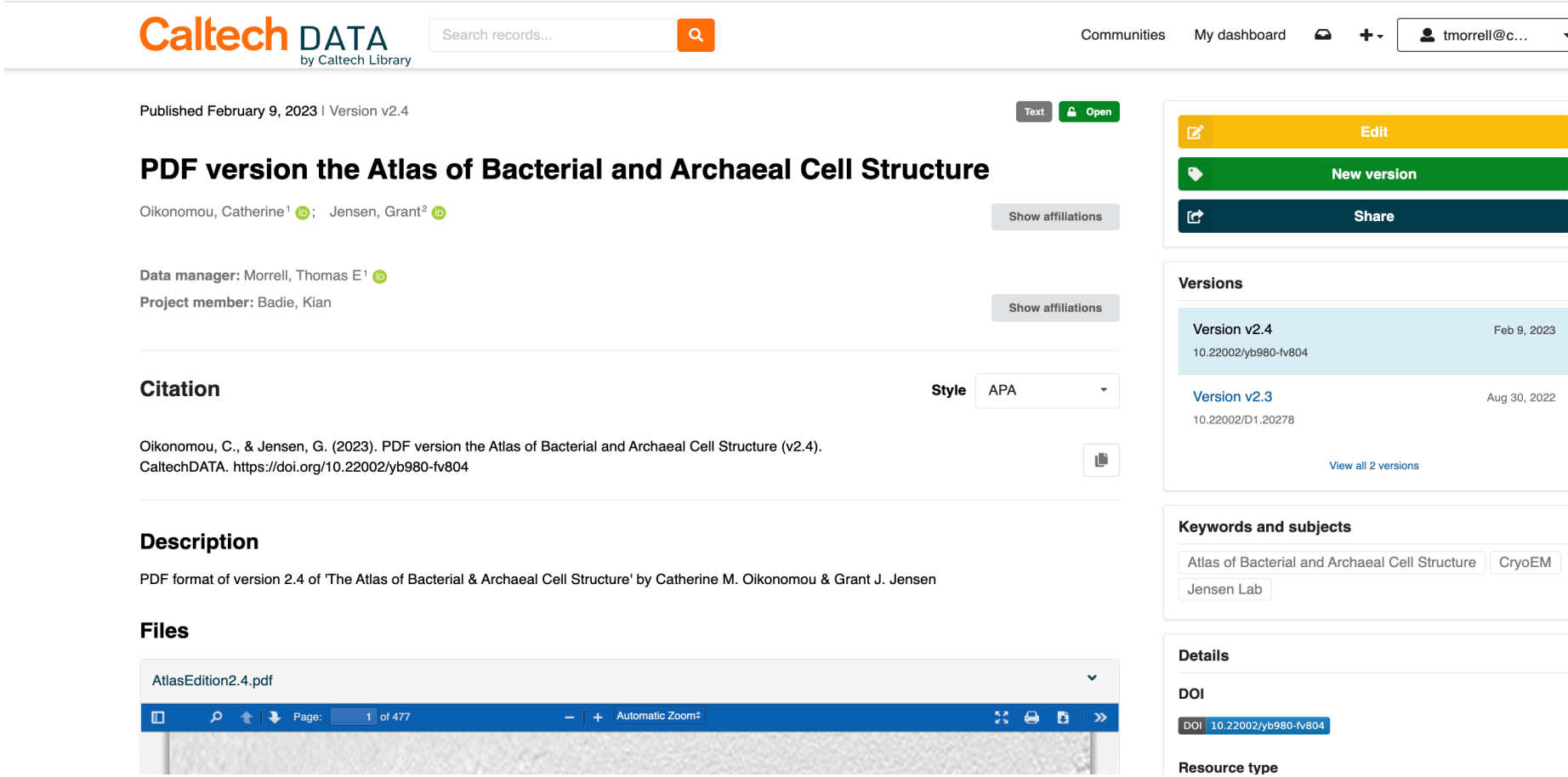


- Open-access textbook on microbial cells
- Over 150 videos with text and narration
- Videos and other media stored in CaltechDATA; automatically created with CaltechDATA API

Caltech Library

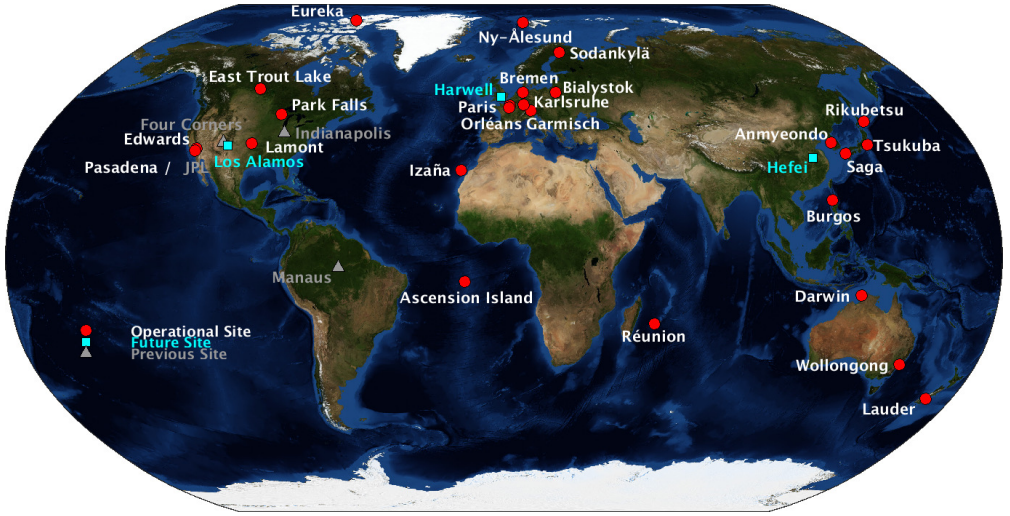
CaltechDATA and the Cell Atlas

- The v2.4 release was automatically uploaded to InvenioRDM version of CaltechDATA
- Only minor changes required (like addition of ROR identifiers)
- Now has built-in versioning



Caltech Library

TCCON



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

Data files

Caltech DATA
by Caltech Library

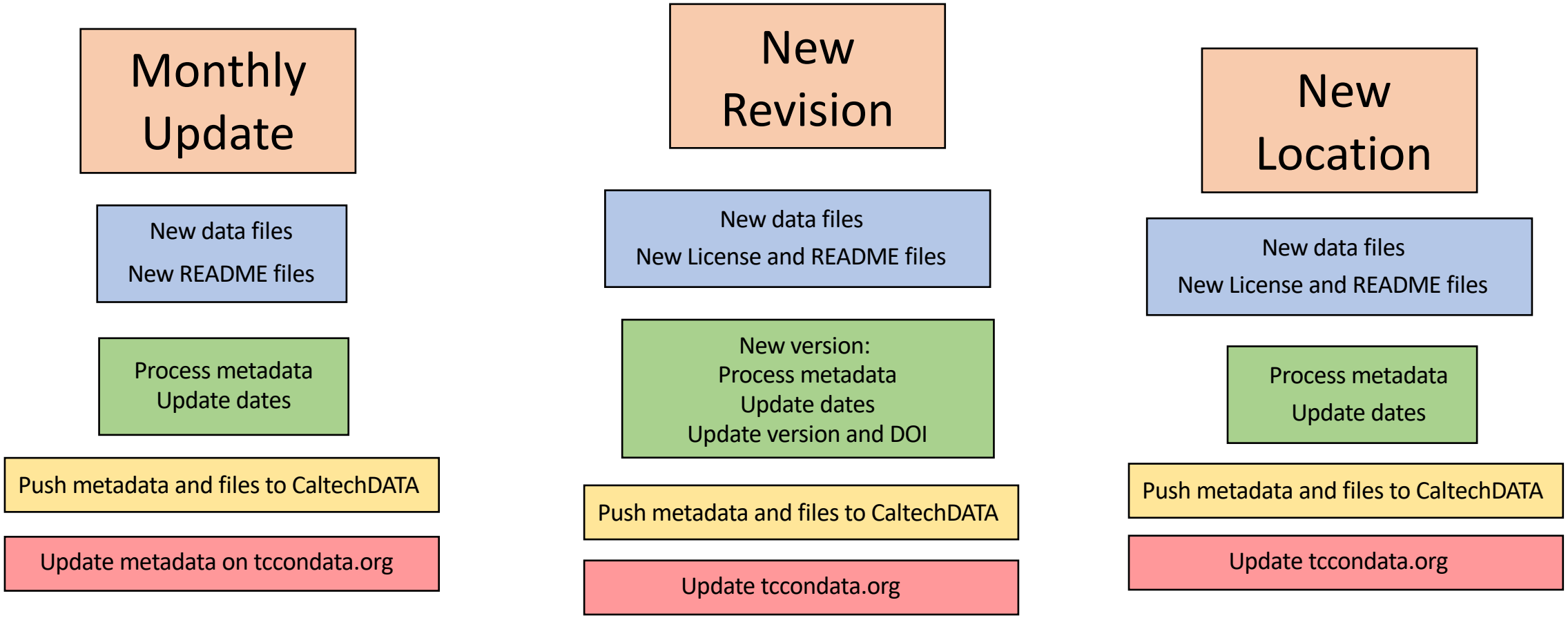
Processed Data

Data Curation and Processing

Caltech Library

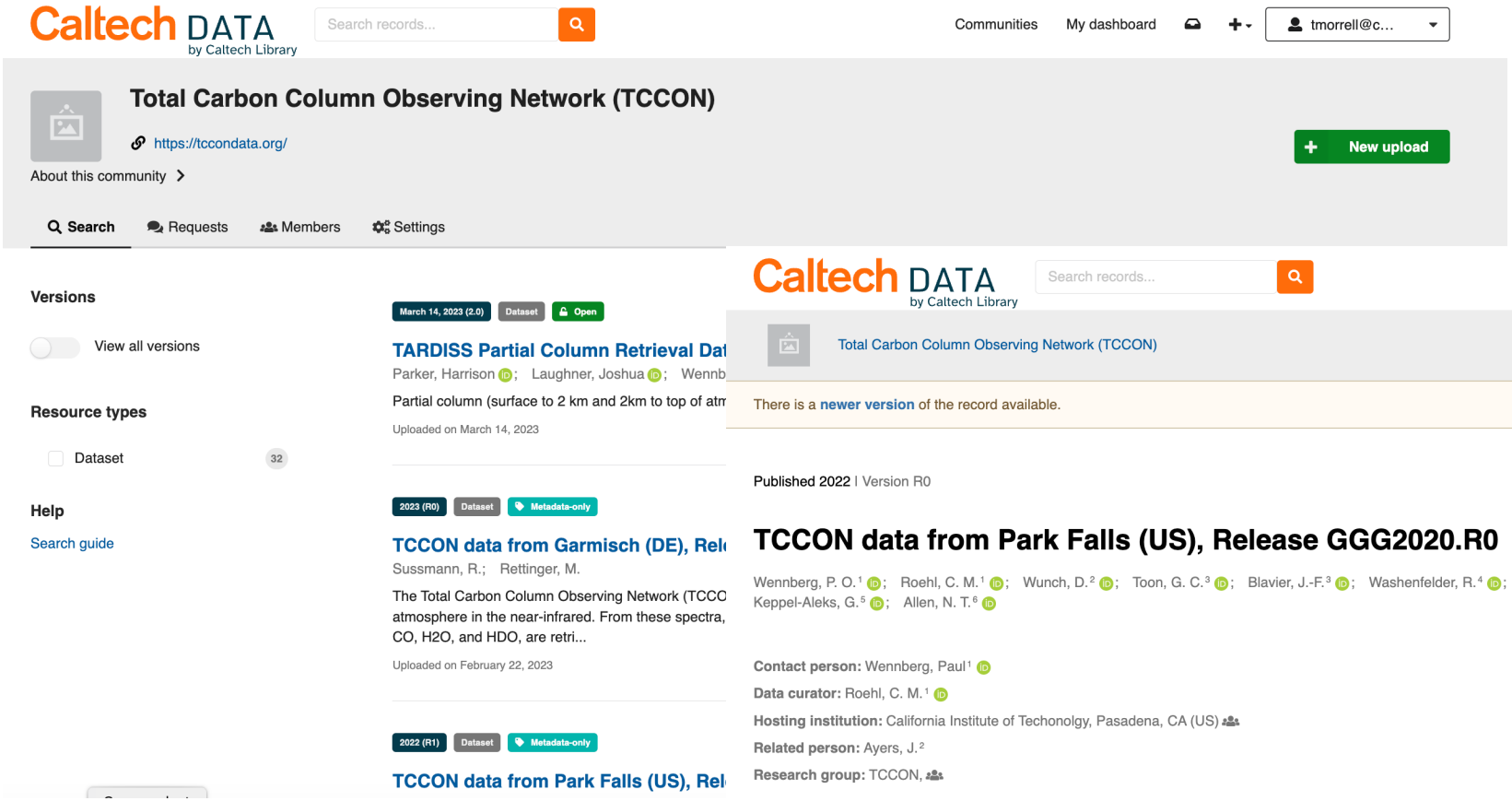
https://tccon-wiki.caltech.edu/Sites/Park_Falls
http://tccondata.org/

TCCON Automation



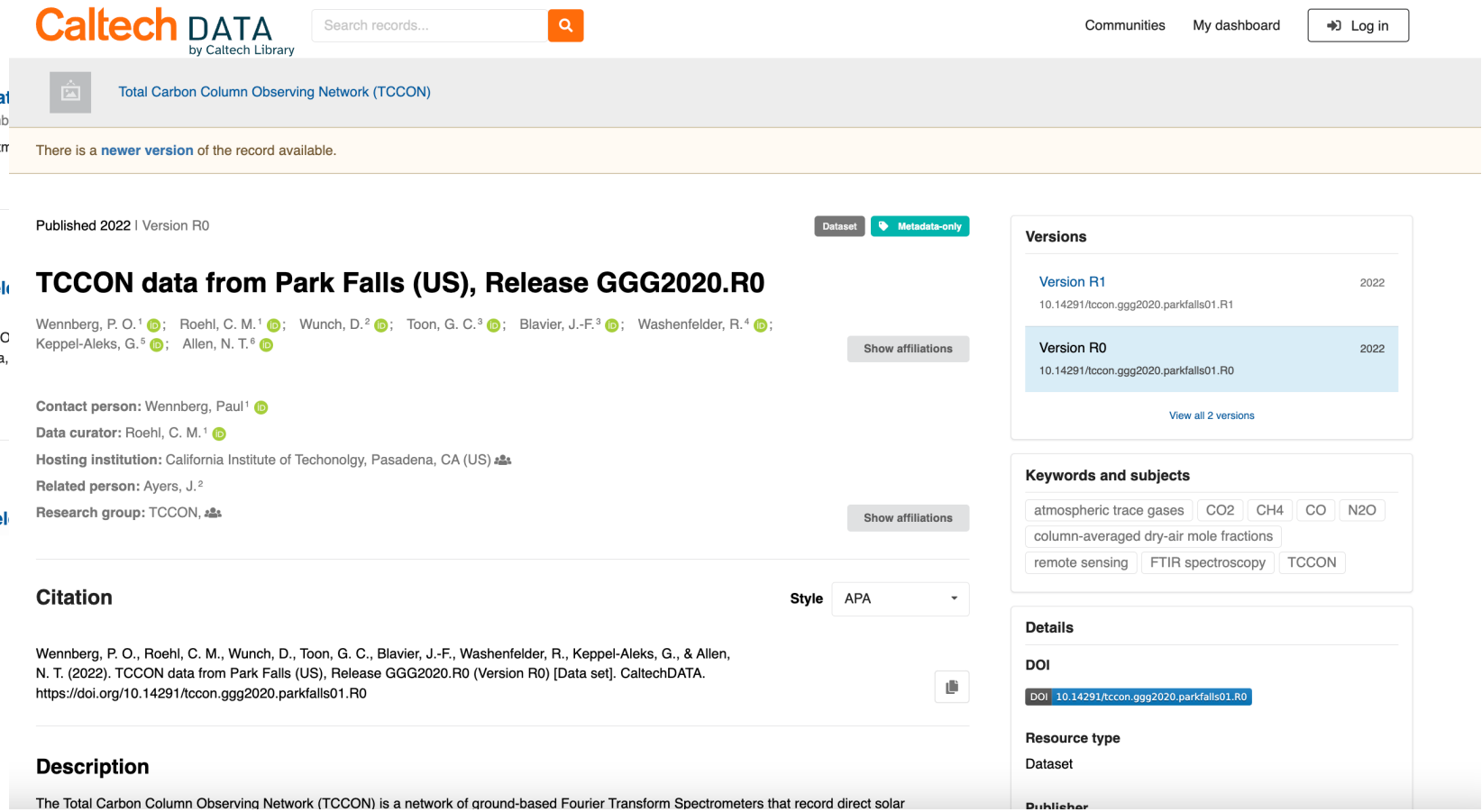
Caltech Library

Improvements in InvenioRDM



Community Landing Page

Record With Automatic Versioning



Caltech Library

microPublication Biology

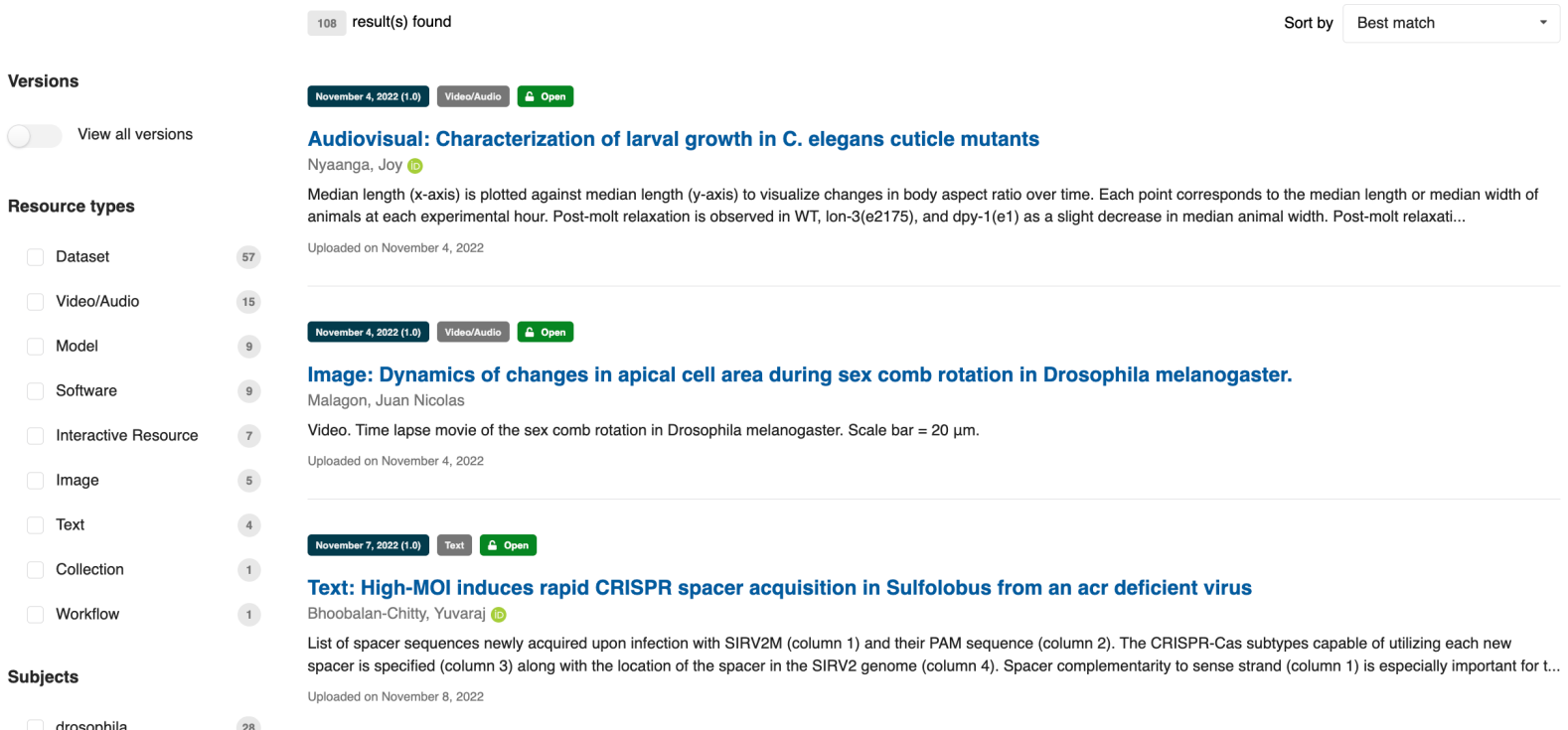
- Innovative journal for single findings
  - May be novel, negative, or reproduced
  - May lack an overall narrative
- Peer-reviewed
- Data files automatically uploaded to appropriate partner repositories

https://www.micropublication.org/

Caltech Library

microPublication Biology and CaltechDATA

- micropublication uses CaltechDATA for supplementary files that don’t fit in a domain repository
- Part of our library publishing services
- Automated using the CaltechDATA API
- The micropublication team implemented this independently, and migrated independently



Caltech Library

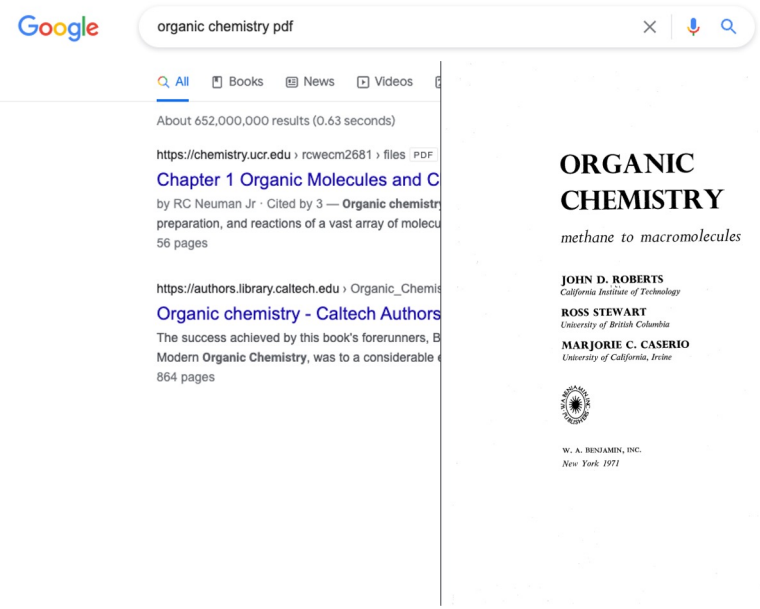
Migration Completed!

- We successfully migrated all content by our contract deadline
- API integrations continue to work
- Significant improvements to landing pages and versioning
- GitHub support coming soon

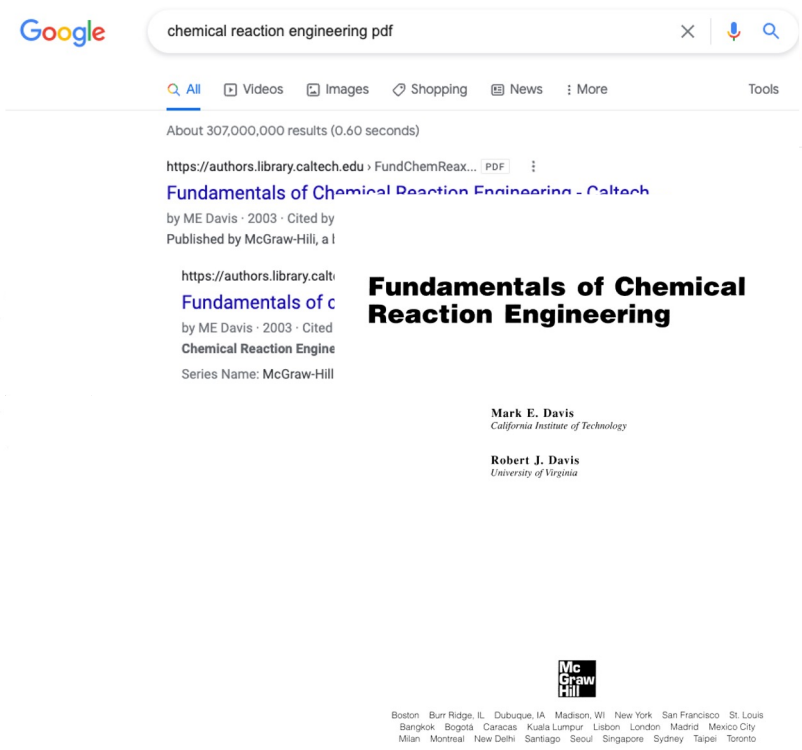
Caltech DATA
by Caltech Library

CaltechAUTHORS

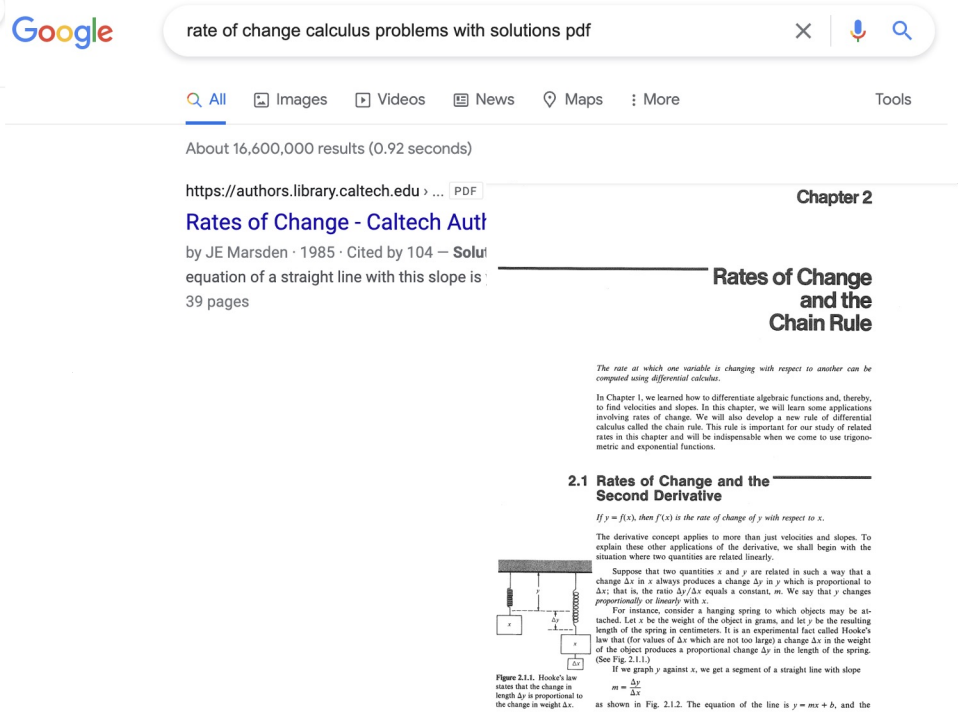
- Over 100,000 records of work by Caltech authors
- Hosted in Eprints since 2004



75,000 clicks/year
Over 1 million downloads



47,000 clicks/year
Over 500,000 downloads

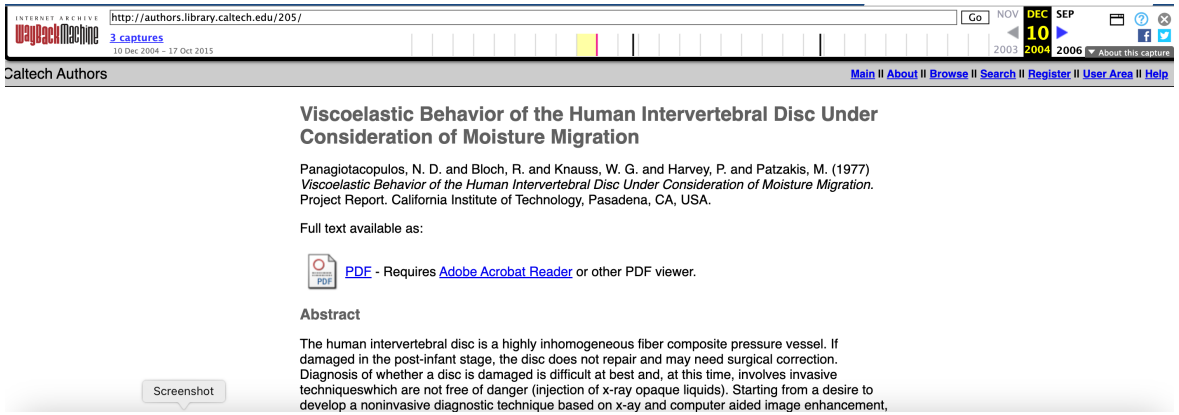


26,000 clicks/year
Over 180,000 downloads

Caltech Library

CaltechAUTHORS

- Over 100,000 records of work by Caltech authors
- Hosted in Eprints since 2004



2004 Landing Page

Current Landing Page

Caltech Library

We’re migrating CaltechAUTHORS to InvenioRDM

- Move all ~100,000 records
- Capture all customized metadata
- Fully redirect all old URLs
- Build more automation for record creation utilizing APIs

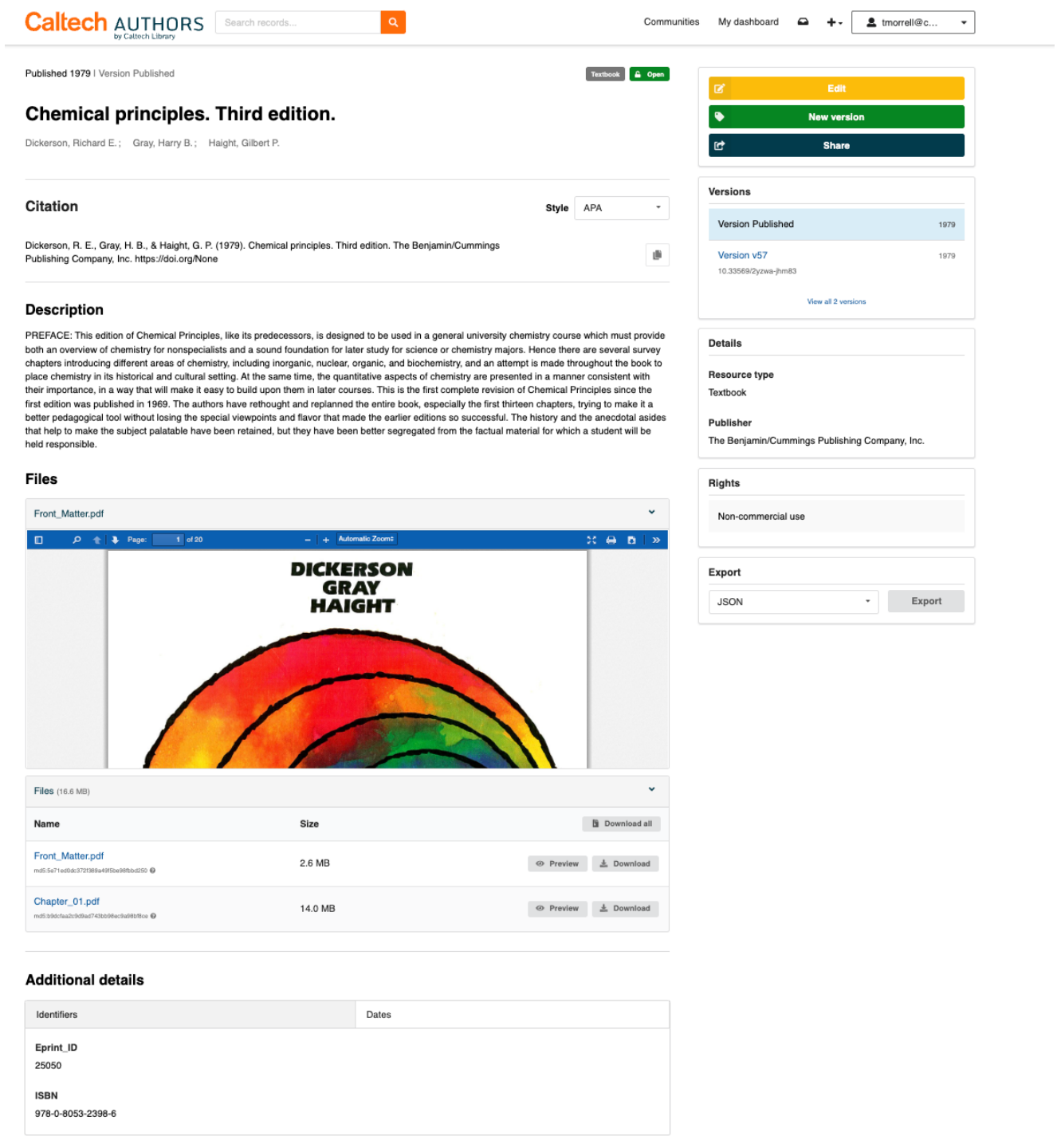
Caltech AUTHORS
by Caltech Library

Example Customization: Resource Types

- Conference or Workshop / Conference or Workshop Paper
- Conference or Workshop / Conference or Workshop Poster
- Conference or Workshop / Conference or Workshop Presentation
- Dataset
- Image
- Image / Map
- Lab Notebook
- Other
- Presentation or Speech
- Publication / Annotation Collection
- Publication / Atlas
- Publication / Book
- Publication / Book Section - Chapter
- Publication / Data Management Plan
- Publication / Discussion Paper
- Publication / Documentation or Manual
- Publication / Erratum
- Publication / Journal Article
- Publication / Journal Issue
- Publication / Map
- Publication / Newspaper Issue
- Publication / Oral History
- Publication / Other
- Publication / Patent
- Publication / Project Report
- Publication / Report
- Publication / Software Documentation
- Publication / Technical Report
- Publication / Thesis
- Publication / White Paper
- Publication / Working Paper
- Software
- Teaching Resource
- Teaching Resource / Lecture Notes
- Teaching Resource / Textbook
- Video/Audio

Caltech Library

Current Status



Caltech Library

Conclusion

- InvenioRDM is a powerful, open-source platform for institutional repositories
- We successfully migrated CaltechDATA by focusing on standardized metadata
- Customized resources can utilize API integrations
- We’re in the process of migrating all our library repositories to InvenioRDM

tmorrell@caltech.edu