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

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

Robert Doiel
Tommy Keswick
Mike Hucka
Stephen Davison

Kathy Johnson
George Porter
Tony Diaz
CaltechDATA

- Institutional repository for Caltech researchers ([https://data.caltech.edu](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
What’s in CaltechDATA?

### Table 1. Compilation of dip angles and subduction parameters along the 155 transects.

<table>
<thead>
<tr>
<th>Transect</th>
<th>$\varphi_0$ (°)</th>
<th>$\varphi_{20}$ (°)</th>
<th>$\varphi_{40}$ (°)</th>
<th>$\varphi_{60}$ (°)</th>
<th>$\varphi_{80}$ (°)</th>
<th>$\varphi_{100}$ (°)</th>
<th>$\varphi_{120}$ (°)</th>
<th>$\varphi_{140}$ (°)</th>
<th>$\varphi_{160}$ (°)</th>
<th>$\varphi_{180}$ (°)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AL</td>
<td>325.36</td>
<td>8.65</td>
<td>0.11</td>
<td>0.95</td>
<td>4.75</td>
<td>2440.60</td>
<td>43.49</td>
<td>151.12</td>
<td>120.00</td>
<td>40.90</td>
</tr>
<tr>
<td>AL</td>
<td>320.16</td>
<td>5.12</td>
<td>0.13</td>
<td>4.99</td>
<td>2544.49</td>
<td>41.47</td>
<td>151.12</td>
<td>120.00</td>
<td>45.27</td>
<td></td>
</tr>
<tr>
<td>AL</td>
<td>315.32</td>
<td>5.34</td>
<td>0.19</td>
<td>5.12</td>
<td>2690.72</td>
<td>48.47</td>
<td>151.12</td>
<td>120.00</td>
<td>34.47</td>
<td></td>
</tr>
<tr>
<td>AL</td>
<td>314.56</td>
<td>5.60</td>
<td>0.19</td>
<td>5.41</td>
<td>2720.20</td>
<td>48.47</td>
<td>151.12</td>
<td>120.00</td>
<td>32.06</td>
<td></td>
</tr>
<tr>
<td>AL</td>
<td>326.59</td>
<td>5.88</td>
<td>0.31</td>
<td>6.20</td>
<td>2890.70</td>
<td>51.45</td>
<td>151.12</td>
<td>120.00</td>
<td>32.35</td>
<td></td>
</tr>
<tr>
<td>AL</td>
<td>337.25</td>
<td>5.85</td>
<td>0.77</td>
<td>6.62</td>
<td>2985.46</td>
<td>52.47</td>
<td>151.12</td>
<td>120.00</td>
<td>30.37</td>
<td></td>
</tr>
<tr>
<td>AL</td>
<td>340.46</td>
<td>5.84</td>
<td>0.93</td>
<td>6.79</td>
<td>2979.07</td>
<td>54.43</td>
<td>151.12</td>
<td>120.00</td>
<td>28.68</td>
<td></td>
</tr>
<tr>
<td>AL</td>
<td>343.50</td>
<td>5.78</td>
<td>1.11</td>
<td>6.89</td>
<td>3002.53</td>
<td>57.04</td>
<td>60.00</td>
<td>120.00</td>
<td>23.60</td>
<td></td>
</tr>
<tr>
<td>AL</td>
<td>344.60</td>
<td>5.78</td>
<td>1.08</td>
<td>7.06</td>
<td>3127.34</td>
<td>59.85</td>
<td>60.00</td>
<td>120.00</td>
<td>23.71</td>
<td></td>
</tr>
<tr>
<td>AL</td>
<td>352.33</td>
<td>5.38</td>
<td>1.47</td>
<td>6.64</td>
<td>3181.51</td>
<td>53.08</td>
<td>60.00</td>
<td>120.00</td>
<td>20.94</td>
<td></td>
</tr>
<tr>
<td>AL</td>
<td>354.36</td>
<td>5.83</td>
<td>1.78</td>
<td>5.65</td>
<td>3291.32</td>
<td>52.14</td>
<td>60.00</td>
<td>120.00</td>
<td>22.16</td>
<td></td>
</tr>
<tr>
<td>AL</td>
<td>354.85</td>
<td>6.00</td>
<td>1.84</td>
<td>5.14</td>
<td>3296.61</td>
<td>49.00</td>
<td>60.00</td>
<td>120.00</td>
<td>17.41</td>
<td></td>
</tr>
<tr>
<td>AL</td>
<td>352.30</td>
<td>2.73</td>
<td>1.52</td>
<td>4.40</td>
<td>3289.40</td>
<td>61.04</td>
<td>60.00</td>
<td>120.00</td>
<td>18.17</td>
<td></td>
</tr>
<tr>
<td>AL</td>
<td>343.83</td>
<td>8.67</td>
<td>0.63</td>
<td>7.84</td>
<td>1843.77</td>
<td>18.69</td>
<td>75.73</td>
<td>120.00</td>
<td>38.35</td>
<td></td>
</tr>
<tr>
<td>AL</td>
<td>353.77</td>
<td>6.04</td>
<td>0.13</td>
<td>6.17</td>
<td>2586.52</td>
<td>16.49</td>
<td>60.00</td>
<td>120.00</td>
<td>39.13</td>
<td></td>
</tr>
<tr>
<td>AL</td>
<td>377.81</td>
<td>8.93</td>
<td>0.80</td>
<td>8.13</td>
<td>1822.09</td>
<td>17.44</td>
<td>75.73</td>
<td>120.00</td>
<td>29.19</td>
<td></td>
</tr>
<tr>
<td>AL</td>
<td>399.97</td>
<td>8.54</td>
<td>0.57</td>
<td>7.57</td>
<td>1894.06</td>
<td>19.54</td>
<td>75.73</td>
<td>120.00</td>
<td>33.69</td>
<td></td>
</tr>
<tr>
<td>AL</td>
<td>391.58</td>
<td>8.61</td>
<td>0.70</td>
<td>6.87</td>
<td>2015.70</td>
<td>9.46</td>
<td>60.00</td>
<td>120.00</td>
<td>40.10</td>
<td></td>
</tr>
<tr>
<td>AL</td>
<td>410.05</td>
<td>7.74</td>
<td>0.43</td>
<td>7.31</td>
<td>1880.52</td>
<td>23.57</td>
<td>75.73</td>
<td>120.00</td>
<td>39.65</td>
<td></td>
</tr>
<tr>
<td>AL</td>
<td>416.16</td>
<td>5.78</td>
<td>0.87</td>
<td>7.86</td>
<td>2171.09</td>
<td>7.86</td>
<td>60.00</td>
<td>120.00</td>
<td>40.17</td>
<td></td>
</tr>
<tr>
<td>AL</td>
<td>415.67</td>
<td>7.39</td>
<td>0.27</td>
<td>7.08</td>
<td>1899.90</td>
<td>24.89</td>
<td>75.73</td>
<td>120.00</td>
<td>41.30</td>
<td></td>
</tr>
<tr>
<td>AL</td>
<td>417.16</td>
<td>6.10</td>
<td>0.32</td>
<td>6.05</td>
<td>2362.25</td>
<td>4.91</td>
<td>60.00</td>
<td>120.00</td>
<td>36.43</td>
<td></td>
</tr>
<tr>
<td>AL</td>
<td>415.70</td>
<td>6.72</td>
<td>0.17</td>
<td>6.08</td>
<td>1906.25</td>
<td>21.41</td>
<td>75.73</td>
<td>120.00</td>
<td>35.04</td>
<td></td>
</tr>
<tr>
<td>AL</td>
<td>18.00-20.40</td>
<td>4.65</td>
<td>1.24</td>
<td>5.96</td>
<td>570.27</td>
<td>7.86</td>
<td>60.00</td>
<td>120.00</td>
<td>35.15</td>
<td></td>
</tr>
</tbody>
</table>

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

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
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/](https://inveniordm.web.cern.ch/)
Migration requirements

• Move all ~20,000 records and files
• Customize the repository for Caltech
  • Theming
  • ORCIDs
• Ensure API integrations continued to work
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
The Materials Experiment Knowledge Graph

Statt, Michael; Rohr, Brian; Guervara, Dan; Breeden, Ja’Nya; Suram, Santosh; Gregoire, John

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 MERG 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

<table>
<thead>
<tr>
<th>Name</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>public-release-neo4j-20230328.dump</td>
<td>5.5 GB</td>
</tr>
</tbody>
</table>

Citation

CaltechPEOPLE

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

- Cell Atlas
- TCCON
- Micropublication

More details:
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
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
TCCON

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

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

Caltech DATA by Caltech Library

Data Curation and Processing
TCCON Automation

**Monthly Update**
- New data files
- New README files
- Process metadata
- Update dates

- Push metadata and files to CaltechDATA
- Update metadata on tccondata.org

**New Revision**
- New data files
- New License and README files
- New version:
  - Process metadata
  - Update dates
  - Update version and DOI

- Push metadata and files to CaltechDATA

**New Location**
- New data files
- New License and README files
- Process metadata
- Update dates

- Push metadata and files to CaltechDATA
- Update tccondata.org
Improvements in InvenioRDM

Record With Automatic Versioning

Total Carbon Column Observing Network (TCON)

Published 2022 | Version R0

TCON data from Park Falls (US), Release GGG2020.R0

Citation


Details

DOI

10.14299/tcon.ggg2020.parkfalls.r0

Resource type

Dataset

Description

The Total Carbon Column Observing Network (TCON) is a network of ground-based Fourier Transform Spectrometers that record direct solar
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/
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
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
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
CaltechAUTHORS

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

2004 Landing Page

Current Landing Page
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
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
- Teaching Resource
- Teaching Resource / Lecture Notes
- Teaching Resource / Textbook
- Video/Audio
Current Status
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