



Scientific Data Management for DOE

March 29, 2023

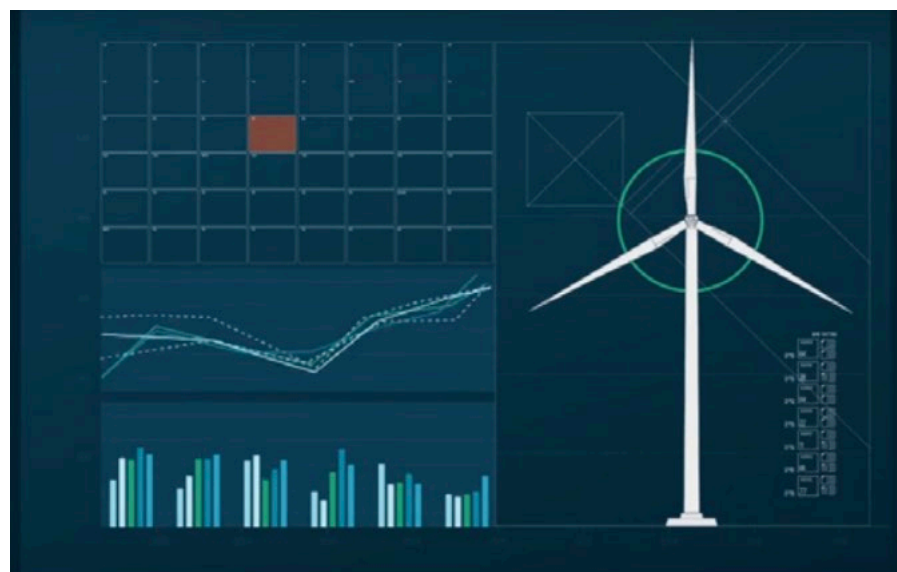
Matt Macduff



PNNL is operated by Battelle for the U.S. Department of Energy



Atmosphere to Electrons (A2e) Initiative: Wind Data Hub



Challenge



Result

The Wind Data Hub (WDH) will **collect, catalog, process, store, preserve, and disseminate** all laboratory, field, and benchmark model data produced by the land-based and offshore demonstration projects in support of mission of the Wind Energy Technology Office



Livewire Data Platform

- 2018 Multi-laboratory collaboration
- Transportation and mobility-related data
- API/External Download Options
- Granular membership management



ENERGY.GOV
Office of
ENERGY EFFICIENCY &
RENEWABLE ENERGY

LIVEWIRE
DATA PLATFORM

Home About

WHAT IS THE LIVEWIRE DATA PLATFORM?

The Livewire Data Platform makes it easy to search and share transportation and mobility-related data. The Livewire Data Platform supports the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy's Energy Efficient Mobility Systems (EEMS) Program goal of providing an affordable, efficient, safe, and accessible transportation future where mobility is decoupled from energy consumption.

WHAT TYPE OF DATA ARE THERE?

The Livewire Data Platform collects data to support EEMS research. analytical, and raw data Data Platform data aim

Datasets / DOE-NREL Fleet DNA

[Summary](#) [API Details](#) [Help](#)

US DEPARTMENT OF ENERGY/NATIONAL RENEWABLE ENERGY LABORATORY'S FLEET DNA: COMMERCIAL FLEET VEHICLE OPERATING DATA

The Fleet DNA offers access to vehicle fleet data summaries similar to real-world "genetics" for medium- and heavy-duty commercial fleet vehicles operating within a variety of vocations. The data available are grouped by vehicle day, which consists of a 24-hour period of operation. The data in Fleet DNA are organized by provider, deployment, and vehicle. Each provider has multiple deployments that consist of a series of vehicles with the same configuration operating in the same location. While the provider is not identified by name, this method of organization allows the drive-cycle metric data to be organized and assessed using any arrangement of the classification system: city of vehicle depot, state of vehicle depot, class (vehicle weight), type (shape of the vehicle), vocation (operation of the vehicle), drivetrain (hybrid/conventional/electric/etc.), and fuel type. For each vehicle day there are over 350 unique results ranging from statistics indicating the type of roads used during travel to drive-cycle metrics that characterize vehicle operating behavior. A data dictionary of these metrics is available at <https://www.nrel.gov/docs/fy14osti/62572.pdf>.

Regina
Winnipeg

+

-

Data Access

[API](#)

Contact

[Eric Miller](#)

National Renewable Energy Laboratory

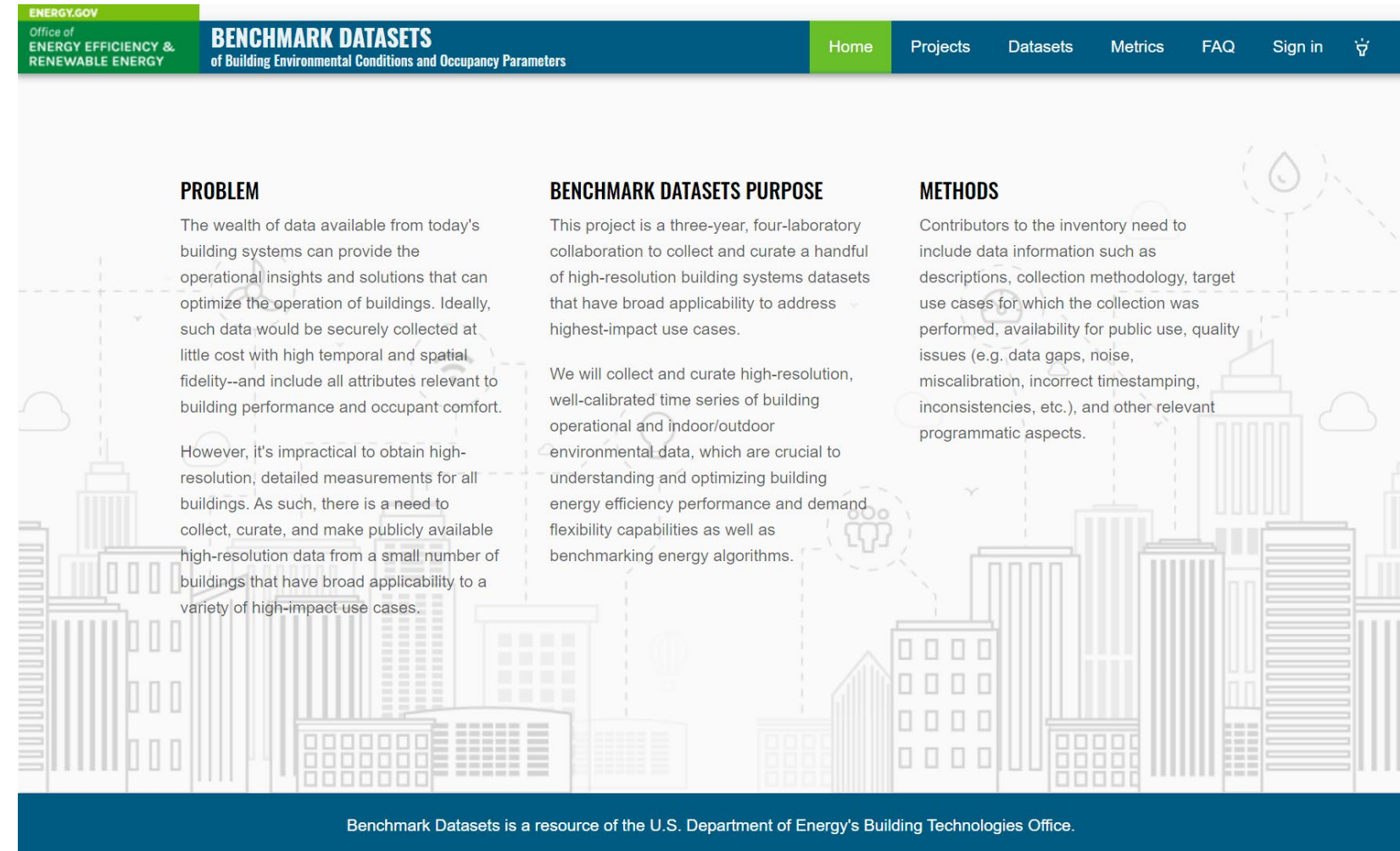
Timeline

[Starts 2015-10-01](#)



Building Benchmark Datasets and Solid Phase Processing Portal

- 2019 Building Benchmark Datasets – building energy
- 2021 Solid Phase Processing Portal – materials
- New domains with unique workflows



The screenshot shows the 'BENCHMARK DATASETS' website. The header includes the 'ENERGY.GOV' logo, 'Office of ENERGY EFFICIENCY & RENEWABLE ENERGY', and the title 'BENCHMARK DATASETS of Building Environmental Conditions and Occupancy Parameters'. Navigation links include 'Home', 'Projects', 'Datasets', 'Metrics', 'FAQ', and 'Sign in'. The main content area is divided into three columns: 'PROBLEM', 'BENCHMARK DATASETS PURPOSE', and 'METHODS'. The 'PROBLEM' section discusses the challenges of obtaining high-resolution data from many buildings. The 'PURPOSE' section describes a three-year project to collect and curate high-resolution data. The 'METHODS' section lists the types of data and information contributors need to provide. The background features a stylized city skyline illustration.

PROBLEM

The wealth of data available from today's building systems can provide the operational insights and solutions that can optimize the operation of buildings. Ideally, such data would be securely collected at little cost with high temporal and spatial fidelity--and include all attributes relevant to building performance and occupant comfort.

However, it's impractical to obtain high-resolution, detailed measurements for all buildings. As such, there is a need to collect, curate, and make publicly available high-resolution data from a small number of buildings that have broad applicability to a variety of high-impact use cases.

BENCHMARK DATASETS PURPOSE

This project is a three-year, four-laboratory collaboration to collect and curate a handful of high-resolution building systems datasets that have broad applicability to address highest-impact use cases.

We will collect and curate high-resolution, well-calibrated time series of building operational and indoor/outdoor environmental data, which are crucial to understanding and optimizing building energy efficiency performance and demand flexibility capabilities as well as benchmarking energy algorithms.

METHODS

Contributors to the inventory need to include data information such as descriptions, collection methodology, target use cases for which the collection was performed, availability for public use, quality issues (e.g. data gaps, noise, miscalibration, incorrect timestamping, inconsistencies, etc.), and other relevant programmatic aspects.

Benchmark Datasets is a resource of the U.S. Department of Energy's Building Technologies Office.



Specific Platform Requirements

The diverse nature of these projects added unique platform requirements:

- Two-factor authentication for proprietary datasets with nondisclosure agreements (NDAs)
- High granularity of access controls to support many small groups with different access requirements for different data
- Real-time streaming data
- “Big Data” model output datasets
- Data transformation pipelines
- Programmatic access via application programming interfaces (APIs)

To implement and maintain this platform as cost effectively as possible, we leveraged the Amazon Web Services (AWS) cloud

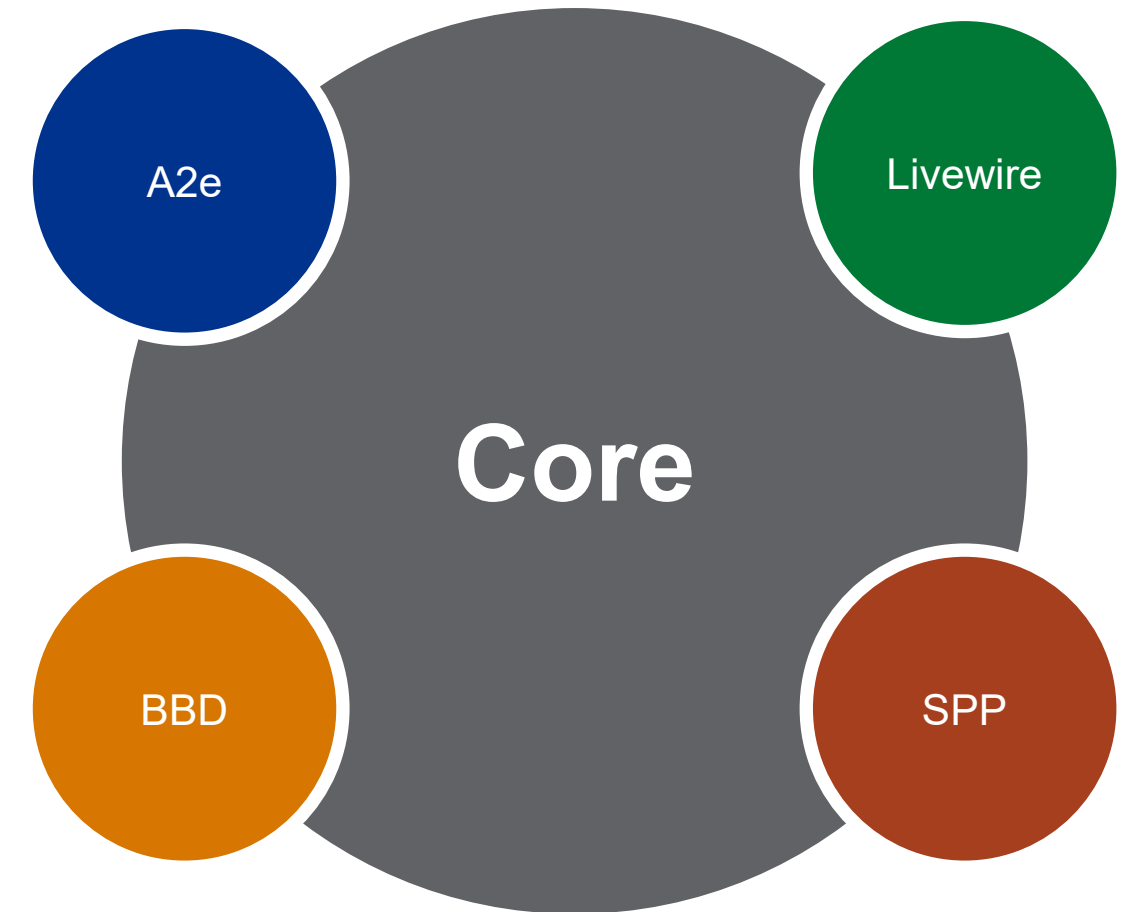


Data Archive and Portal (DAP) Platform

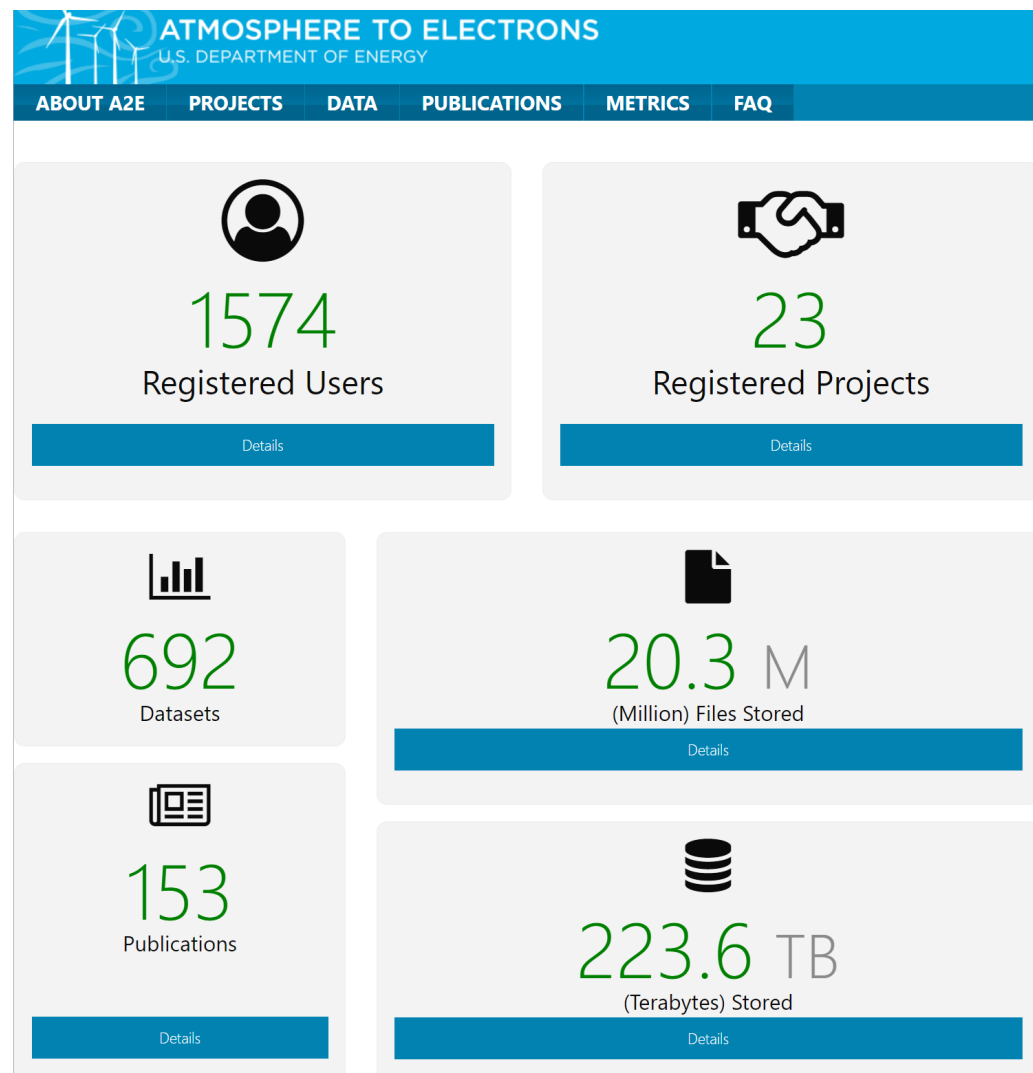
Generic data platform framework

Currently leveraged by:

- *Livewire (transportation)*
<https://livewire.energy.gov>
- *BBD (buildings)*
<https://bbd.labworks.gov>
- *WDH (wind energy)*
<https://a2e.energy.gov>
- *SPP (materials)*



DAP Capabilities



- Data archive and storage
 - Hybrid, configurable
- Data collection
 - HTTPS-REST, portal drag and drop
- Secure data access
 - Tiered access with two-factor authentication
- Metadata standards
 - Project Open Data, customizable
- Metadata curation
- Data publishing
 - Data.gov, digital object identifier (DOI)
 - Data metrics
- Metadata search and discovery
 - ElasticSearch
- Data downloads
 - Portal, HTTPS-REST
- Operational support, email reports



ⓘ You have signed in without your second-factor token. You will NOT see any proprietary datasets.

🔍 Search

Project: buoy X Preferred: Yes X

☰ Hide Map

Project

> wfip2	75
> tap	62
> uae6	30
> wfip1	30
> xpia	13
> mmc	12
> buoy	8
> awaken	6
> test	4
> oc5	3
> wake	3
> impowr	2
> sumr-d	2
> aawpc	1
> lees	1



19 datasets found (page 1 of 1)

⌵ Timelines

Last Updated ⌵

Date Range

Start Date → End Date

2020 10 03

View > Basic Daily Availability

2021 12 11

Preferred for General Use

<input type="checkbox"/> No	14
<input checked="" type="checkbox"/> Yes	8

Data Level

<input type="checkbox"/> Processed Data	5
<input type="checkbox"/> Derived Data	3

Instrument

<input type="checkbox"/> Buoy	3
<input type="checkbox"/> Auxiliary	2
<input type="checkbox"/> Lidar	2
<input type="checkbox"/> Reanalysis	1

buoy / buoy.z05.a0

13,245 files, 35.1 GB, Updated an hour ago 🛒

*.10m.a2e.nc				
*.imu.a2e.nc				
*.waves.a2e.nc				

buoy / buoy.z05.00

31,690 files, 25.0 GB, Updated an hour ago 🛒

*.conductivity.csv				
*.currents.csv				
*.gill.csv				
*.gnss.bin				
*.gps.csv				
*.imu.bin				
*.pressure.csv				
*.pyranometer.csv				
*.rh.csv				



ABOUT A2E

PROJECTS

DATA

PUBLICATIONS

METRICS

FAQ



You have signed in without your second-factor token. You will NOT see any proprietary datasets.

Search

Project: buoy X Preferred: Yes X

Hide Map

Project

> wfip2	75
> tap	62
> uae6	30
> wfip1	30
> xpia	13
> mmc	12
> buoy	8
> awaken	6
> test	4
> oc5	3
> wake	3
> impowr	2
> sumr-d	2
> aawpc	1
> lees	1

Date Range

Start Date → End Date

Preferred for General Use

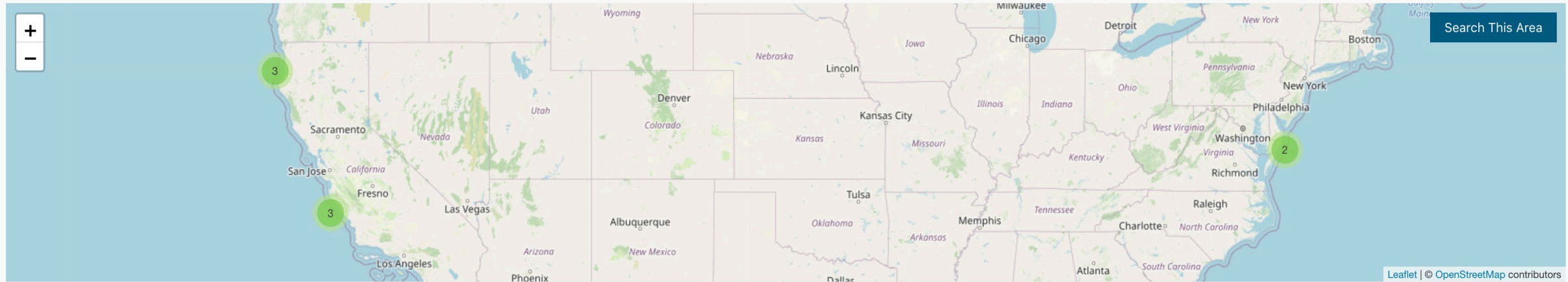
- ☐ No 14
- ☒ Yes 8

Data Level

- ☐ Processed Data 5
- ☐ Derived Data 3

Instrument

- ☐ Buoy 3
- ☐ Auxiliary 2
- ☐ Lidar 2
- ☐ Reanalysis 1



8 datasets found (page 1 of 1)

Visualization

Last Updated

2020 10 03

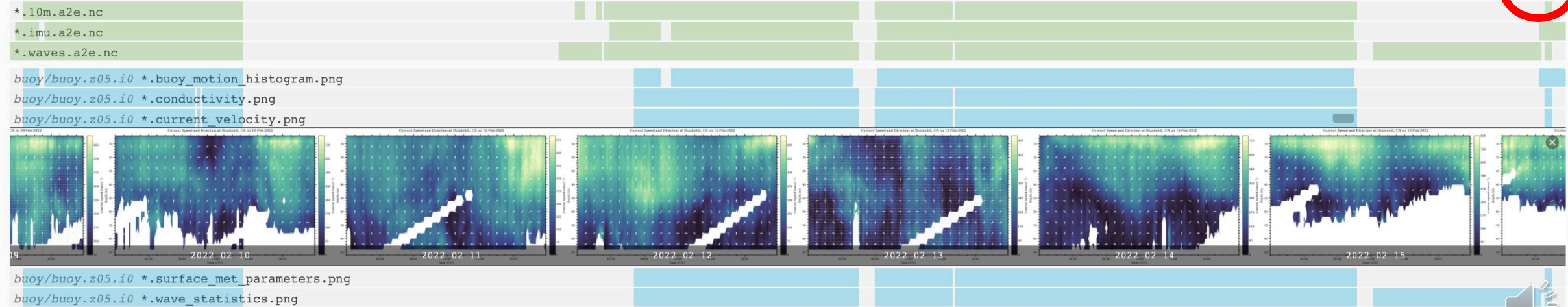
View > Basic Daily Availability

Data Images

2022 05 07

buoy / buoy.z05.a0

17,212 files, 45.3 GB, Updated 12 hours ago

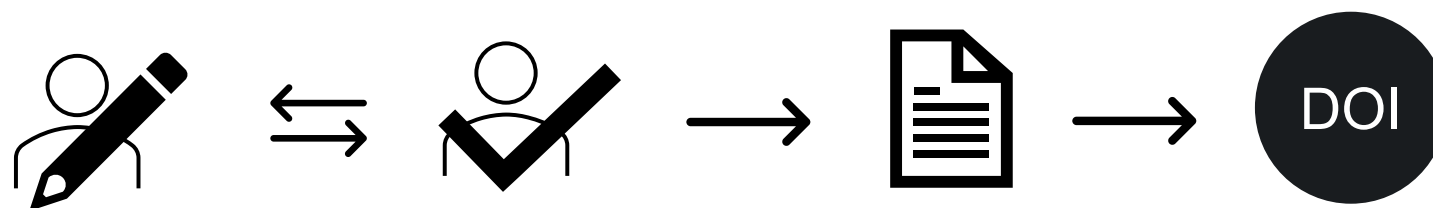


buoy / lidar.z05.a0

382 files, 28.4 MB, Updated 14 hours ago

Metadata Management

- Project (i.e. Catalog) metadata managed internally
- Dataset / file naming convention done collaboratively
- Dataset metadata managed by data owners
- All metadata internally reviewed before published
- DOI assigned post approval



- Extends DCAT-US Schema v1.1 (Project Open Data)
- Form generation / validation using JSON Schema

< Metadata Submissions

Edit Metadata

Steps

- ✓ SELECT DATASET
test.z03/lidar.z10.a0
- ✓ IMPORT METADATA
Optional
- 3 EDIT METADATA
- 4 REVIEW

Edit Metadata for test.z03/lidar.z10.a0

DAP Project Time-Series / Sensor Dataset

Extended dataset metadata for describing sensor-based sources

Title*
Lidar / Processed Data
Human-readable name of the dataset; should be in plain English and include sufficient detail to facilitate search and discovery.

Short Dataset Name
Concise (i.e. less than 30 characters) name for this dataset; should be human-readable; may be revised by curators for standardization across the project catalog.

Description*
Human-readable description (e.g., an abstract) with sufficient detail to enable a user to quickly understand whether the dataset is of interest.

Timezone Offset
0
The number of timezones from UTC

☒ Recommended for General Use

☐ Data Quality

Explanation of Data Quality

Supports markdown

Explanation of Uncertainty

Supports markdown

Explanation of Constraints

Supports markdown

Data Details

Additional information about the data that would be relevant to end users

References

Lists and describes relevant external references (blank line between each)

Dataset Point(s) of Contact
Contact person(s)'s contact information for the dataset

+ ADD ITEM

Geographic Location(s)
Coordinates where data samples were collected

+ ADD ITEM

Measurements
List of measurements

+ ADD ITEM

Dimensions List
Relevant dimensions list

+ ADD ITEM

Data Relevant Events
Time-bound events with details relevant to the data and potentially the data quality

+ ADD ITEM

Related Documents
Related documents such as technical information about a dataset, developer documentation, etc.

Step 1 (optional): Upload an Attachment

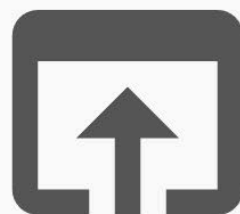
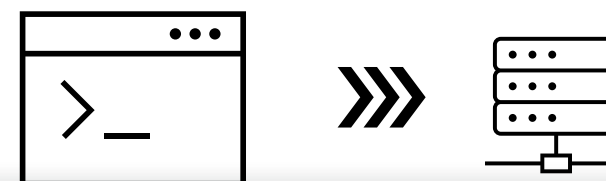
To upload, drop files here or click to browse

Step 2: Select Uploaded Files to Attach

BACK CANCEL SAVE DRAFT SAVE & REVIEW



Data Upload



Drag and Drop

Ideal for files that are:

- small (in size *and* quantity),
- on this computer, *and*
- a static set.

Select Dataset ... ▼

> GET STARTED

OR



Uploader Client

Ideal for files that are:

- large (in size *or* quantity),
- remote (reside on another computer), *or*
- routine / ongoing.

Select Project ... ▼

> GET STARTED



Access Control

PROJECT

DATASET 1

- Open / Public
- Requires login to access data
- Includes automatically minted DOI
- Harvested by OpenEI

DATASET 2

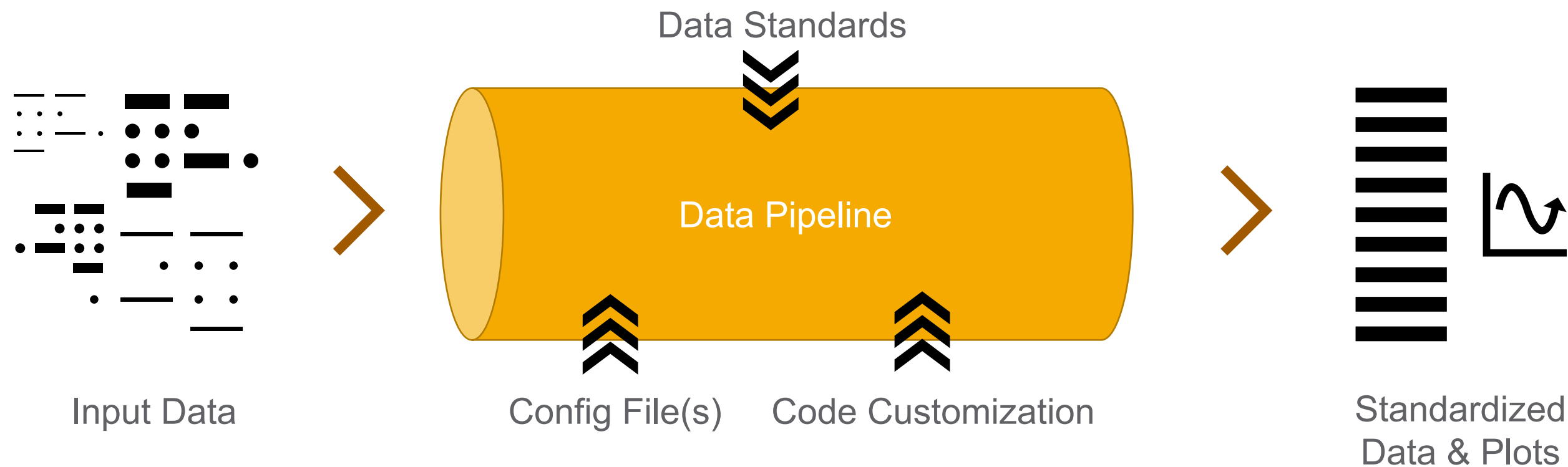
- Restricted
- Requires approval to access data
- No DOI until/unless made public

DATASET 3

- Proprietary
- Requires approval and MFA to access data
- No DOI

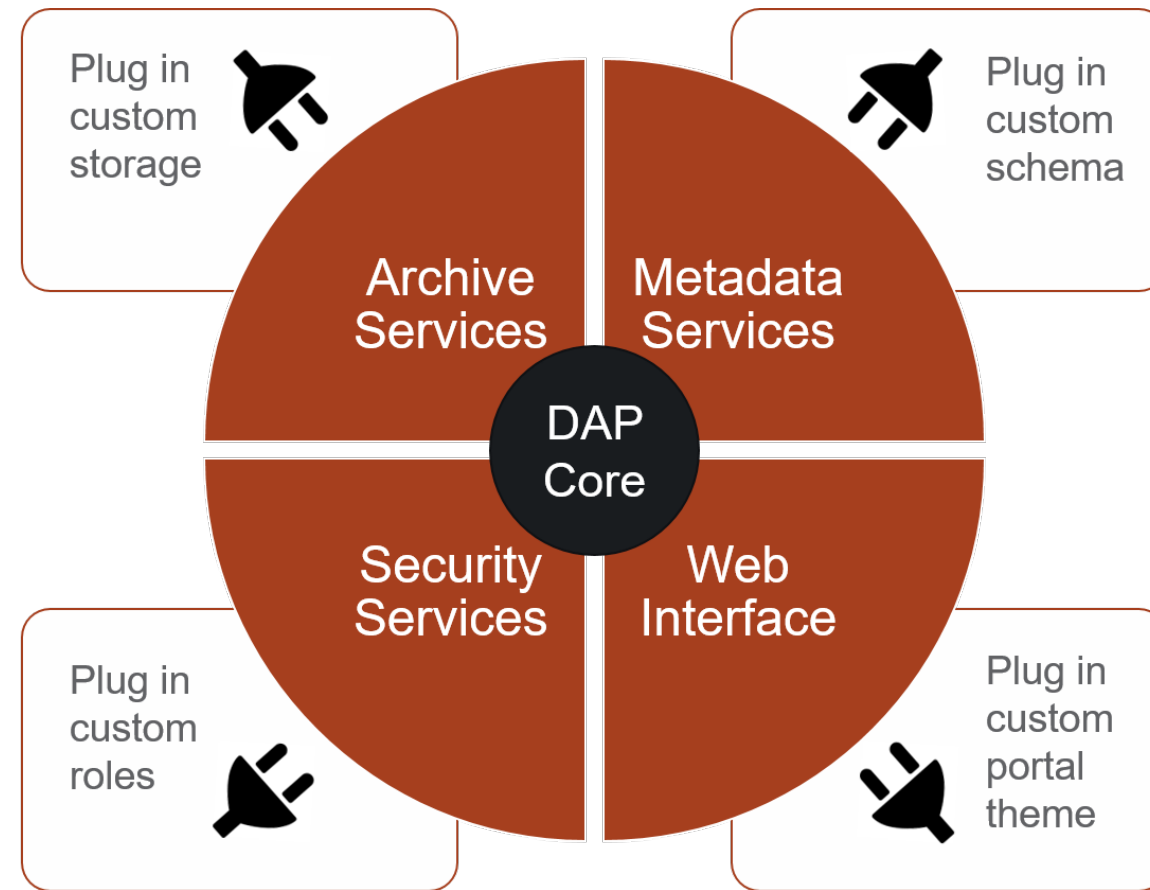


Standardized Data Pipeline



Extensible DAP Platform

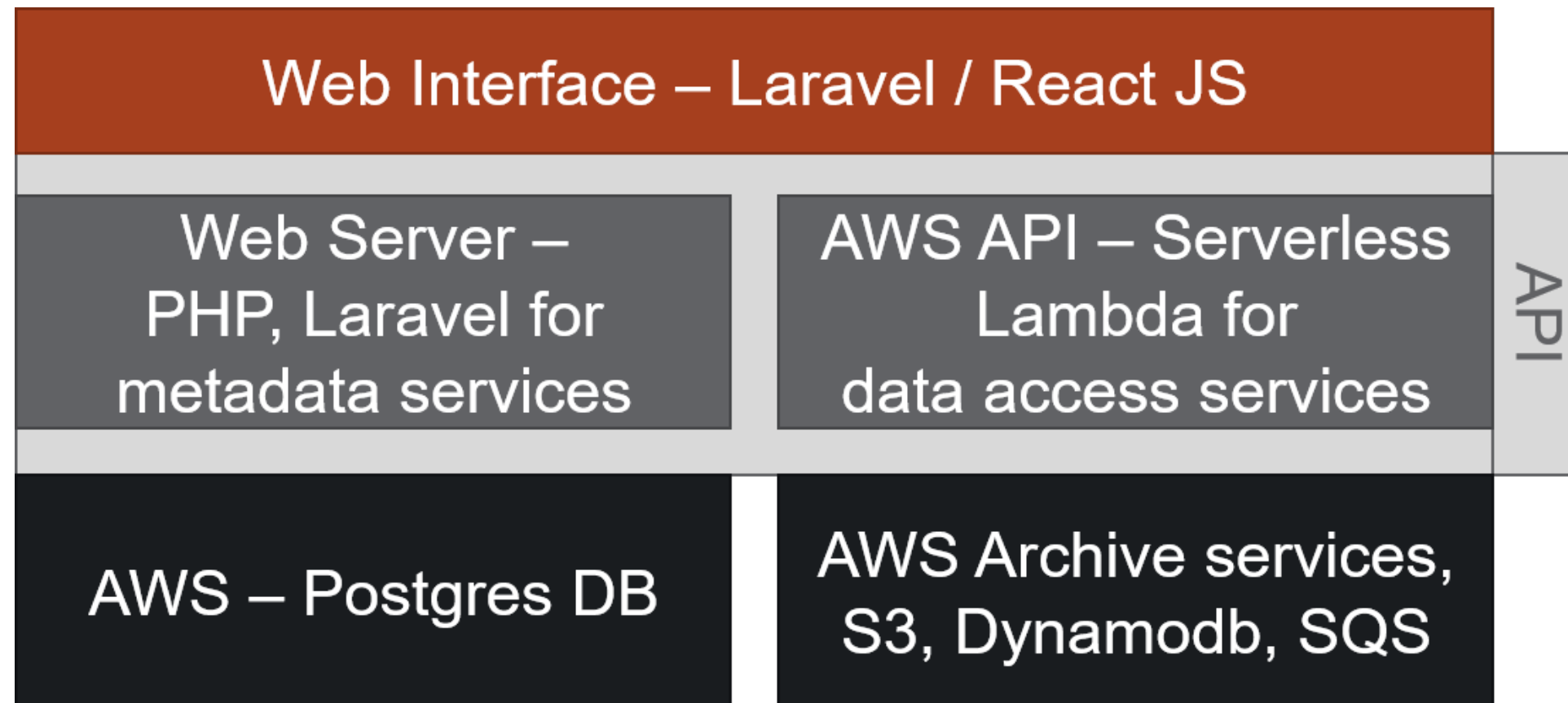
- Leverages the core DAP architecture and automated build system to be easily reused by other projects
- Extensibility hooks to customize key services
- Adding scripts to rebuild entire stack for any environment



- Rapid deployment of new projects
- Leverages new features and bug fixes across projects
- Accelerated development
- Improved stability
- Reduced cost and risk

DAP Software Architecture

- User-friendly, responsive portal based upon best-of-class technology

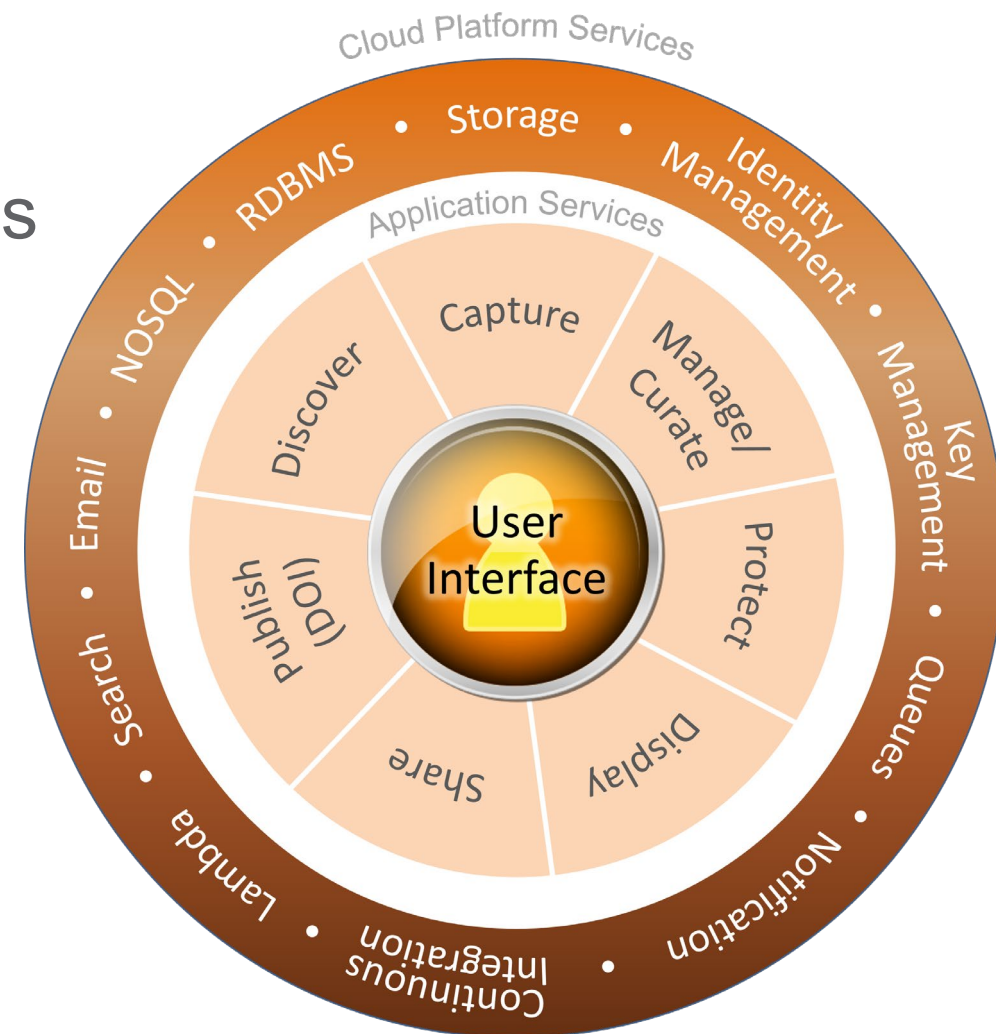


Cloud-based Architecture

- Built on the AWS cloud platform
- Leverages many reliable, best-of-class AWS services
- Well-documented REST APIs

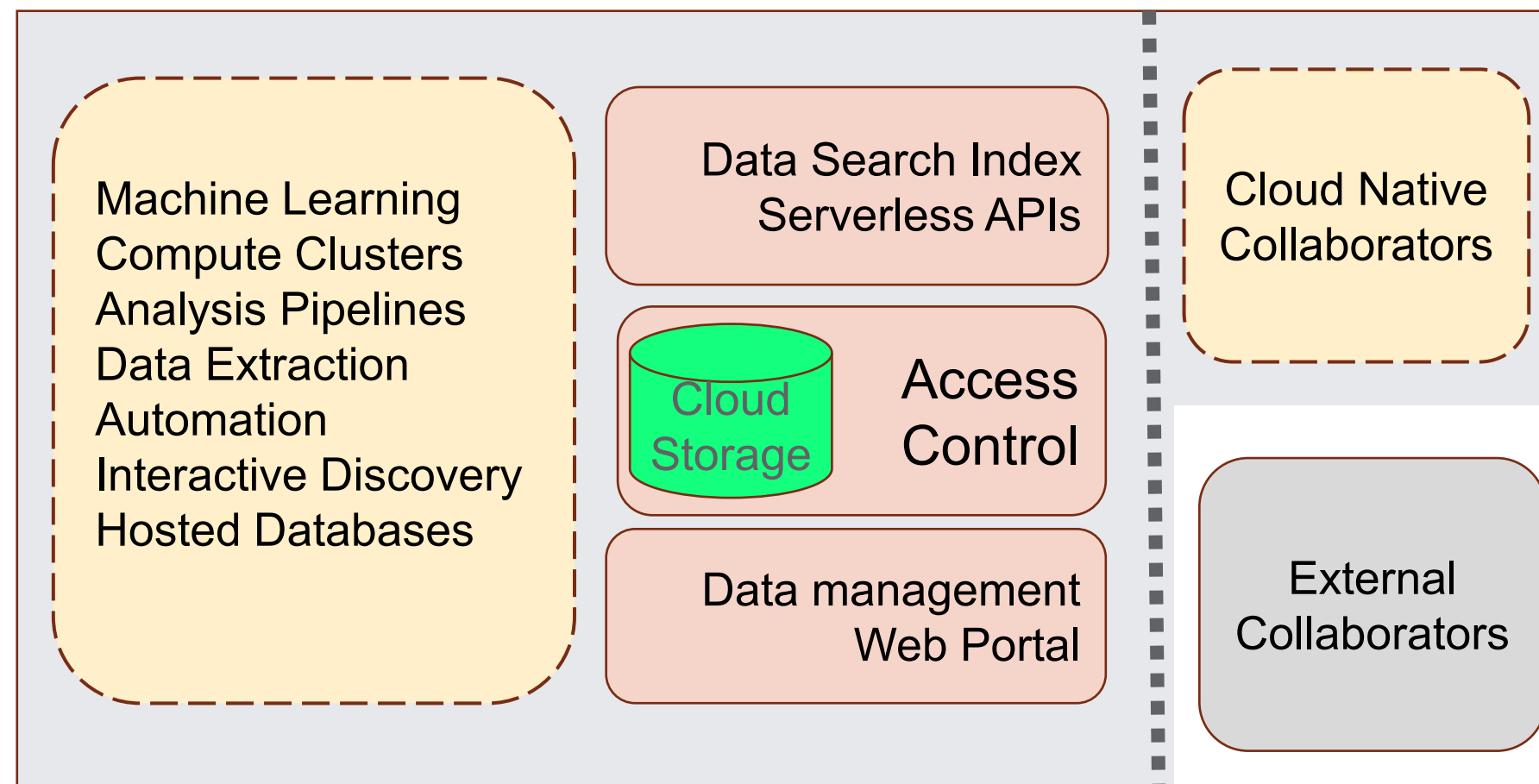
Advantages of AWS

- Simplified development and deployment
- Scale on demand
- Reduced cost and risk
- Accelerated analytics
- Reliable managed services
- Uptime > 99.5%



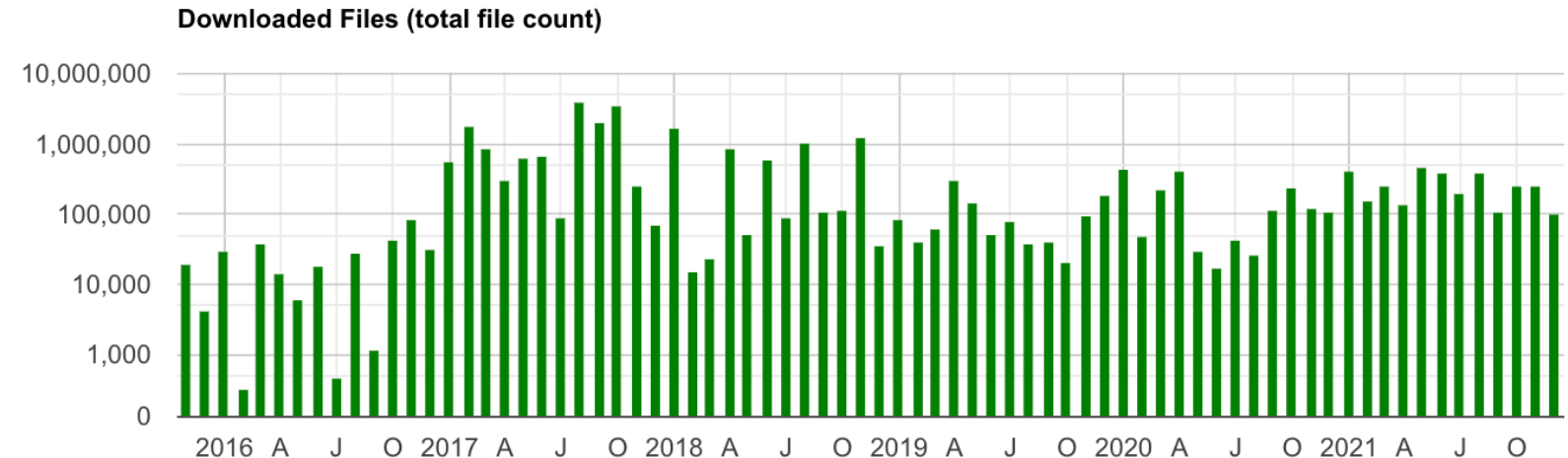
Benefits of Cloud

- Flexible
 - Huge variety of ready-to-use PAAS, SAAS, Compute and Storage services
- Scalable
 - Effectively infinite
- Manage cost
 - Right-size, pay just for what you use
- Collaboration
 - Easy access for everyone
- Security
 - FISMA/FedRAMP
 - Layers of additional controls



S3 – Simple Storage Service

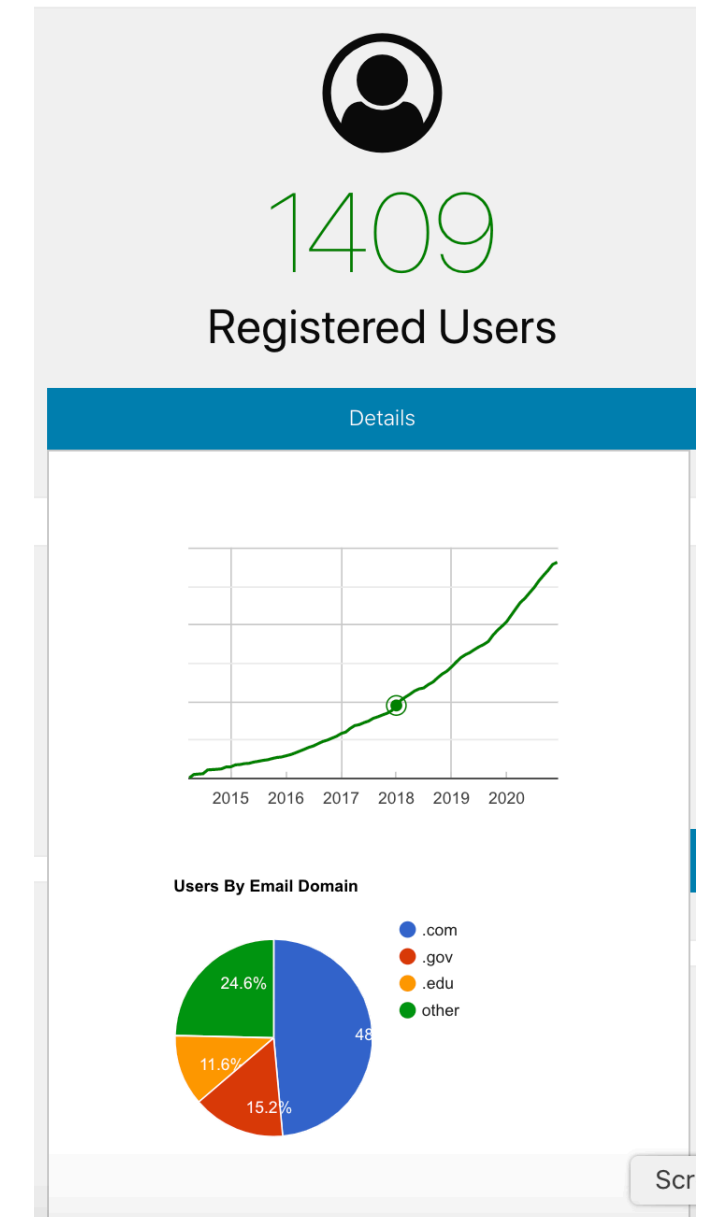
- Industry standard object store
- Robust reliability
- Triggers for workflows
- Granular access control
- What about download cost?



Opportunities and Challenges

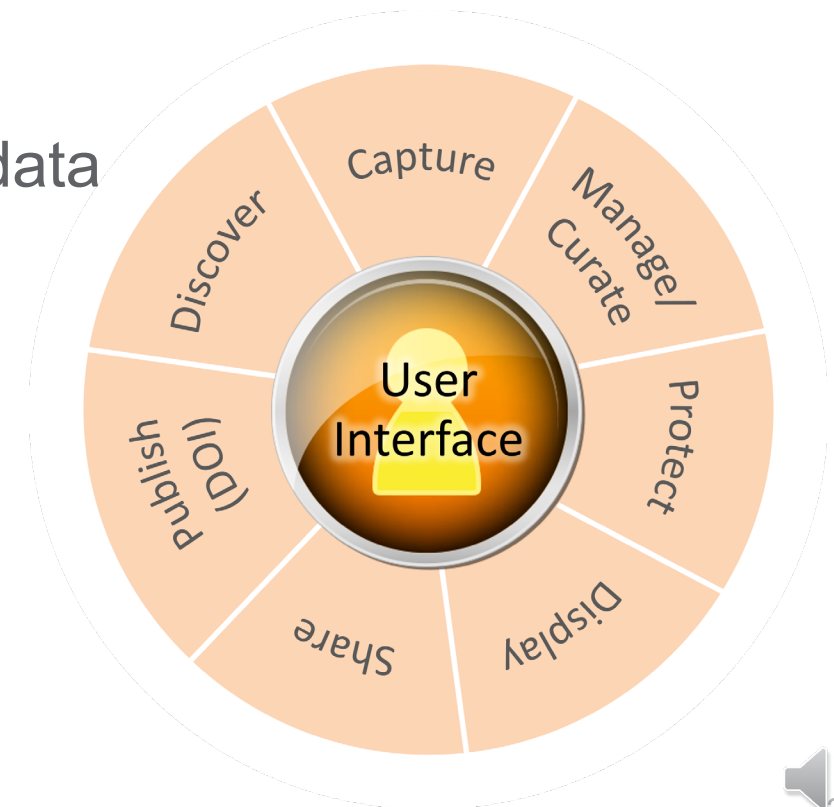
Develop and build or adapt data management efforts to accommodate:

- Growth of data volume and users
- Coordinate and monitor real-time processing
- Metadata coordination is work
- User Support
- Sustaining M&O
- Analysis near data - Jupyter et al
- Data Analysis Center
- Unique process of science for each domain



DAP Conclusions

- Leveraging cloud services and dev-ops significantly reduced software implementation costs and time to market
- Platform stability and scalability increased over on-site managed servers
- New deployments can focus on domain-specific tasks instead of implementing core software capabilities:
 - Identification of standard data formats and metadata
 - Identification of access roles and security policy
 - Custom branding
 - Hooking in custom back-end storage



Thank you

