# ESS-DIVE Data Repository:Updates

April 29, 2019

ESS Cyberinfrastructure Working Group Meeting

ESS-DIVE

Deep Insight for Earth Science Data

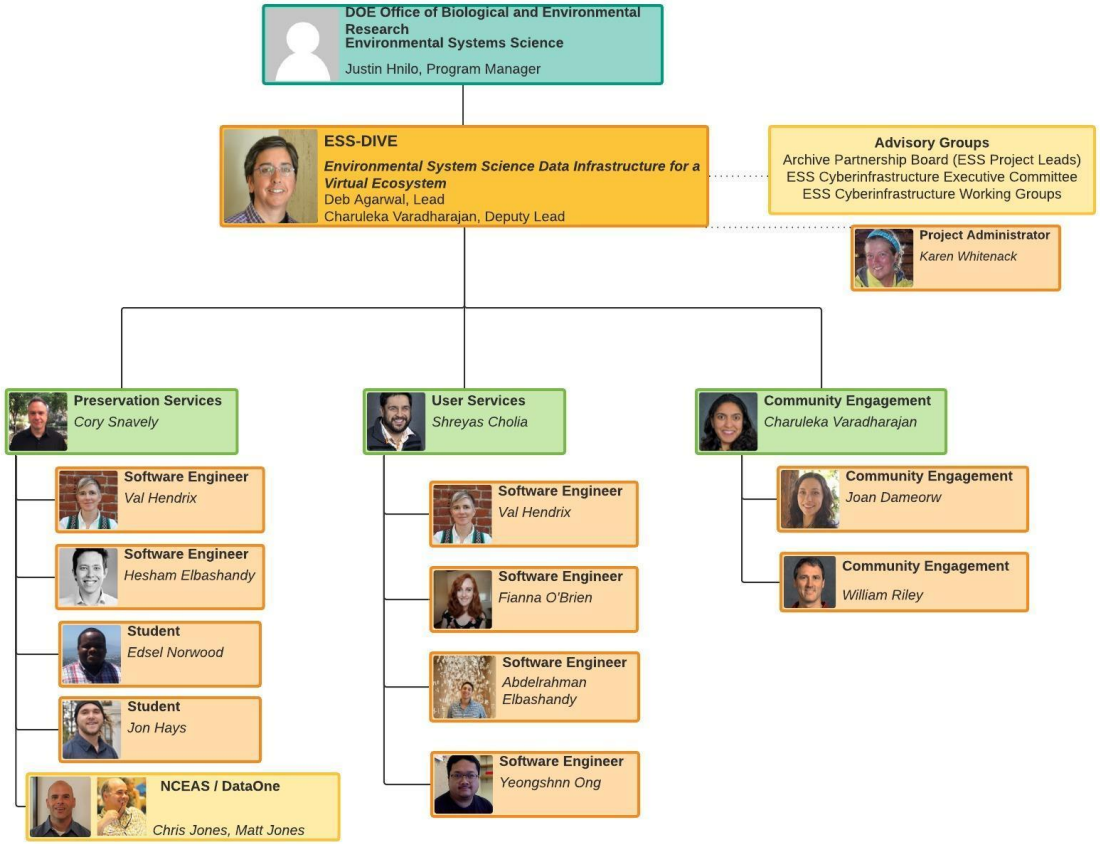
U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB
Lawrence Berkeley National Laboratory

NERSC

NCEAS

# ESS-DIVE team has expanded



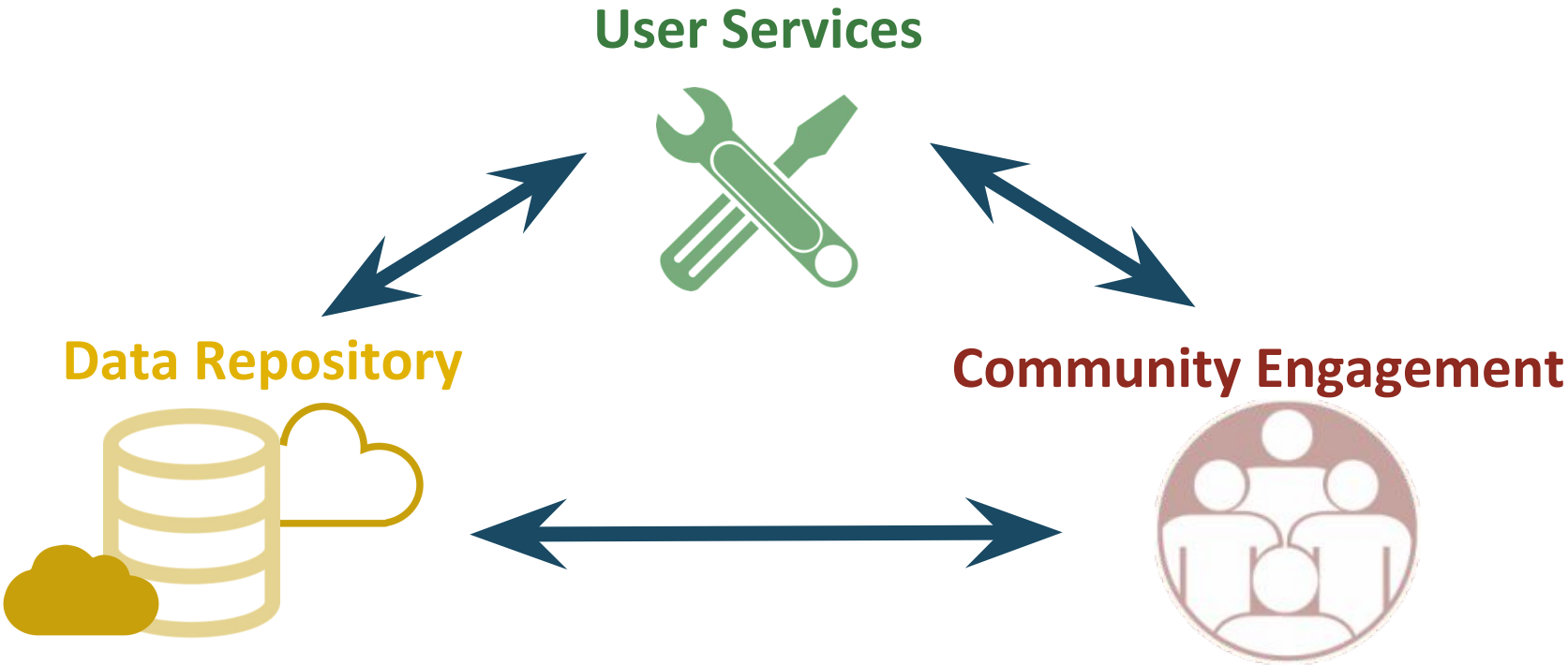

**Environmental Scientists**

**Data Scientists**

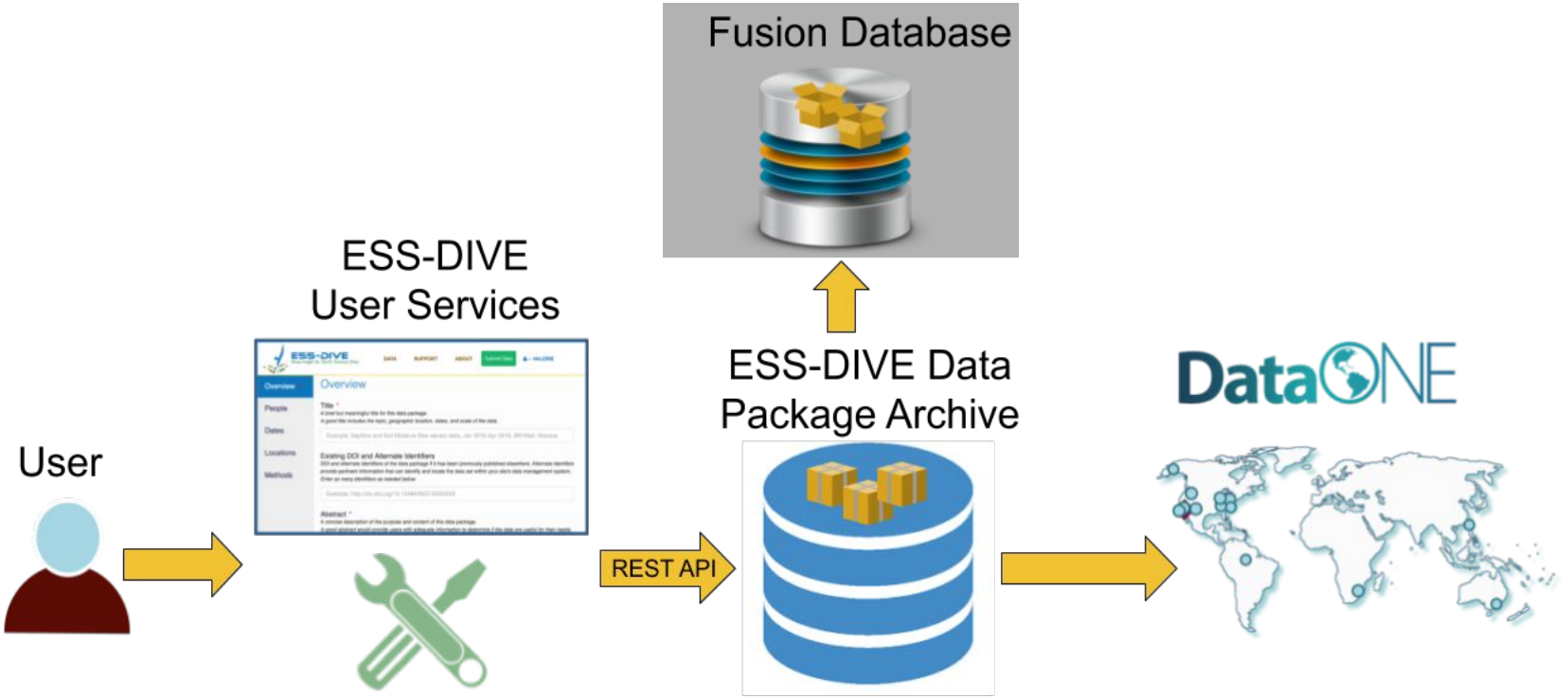
**Software Engineers**

**Digital Librarians**

# Three Pronged Approach to Repository



# ESS-DIVE Archiving Features



# ESS-DIVE Implementation Timeline

**July 2017**
Project Start

**April 2018**
ESS-DIVE Accepting Data

**Sept 2018**
Completed Translation of Previous Repo

**March 2019**
Package Service 1.0.0 API Feature Complete

**Sept 2017**
Previous Repo Transferred

**Aug 2018**
Joined DataONE


**Dec 2018**
Prototype Package Service Available

**April 2019**
Initial data integrity audit & report

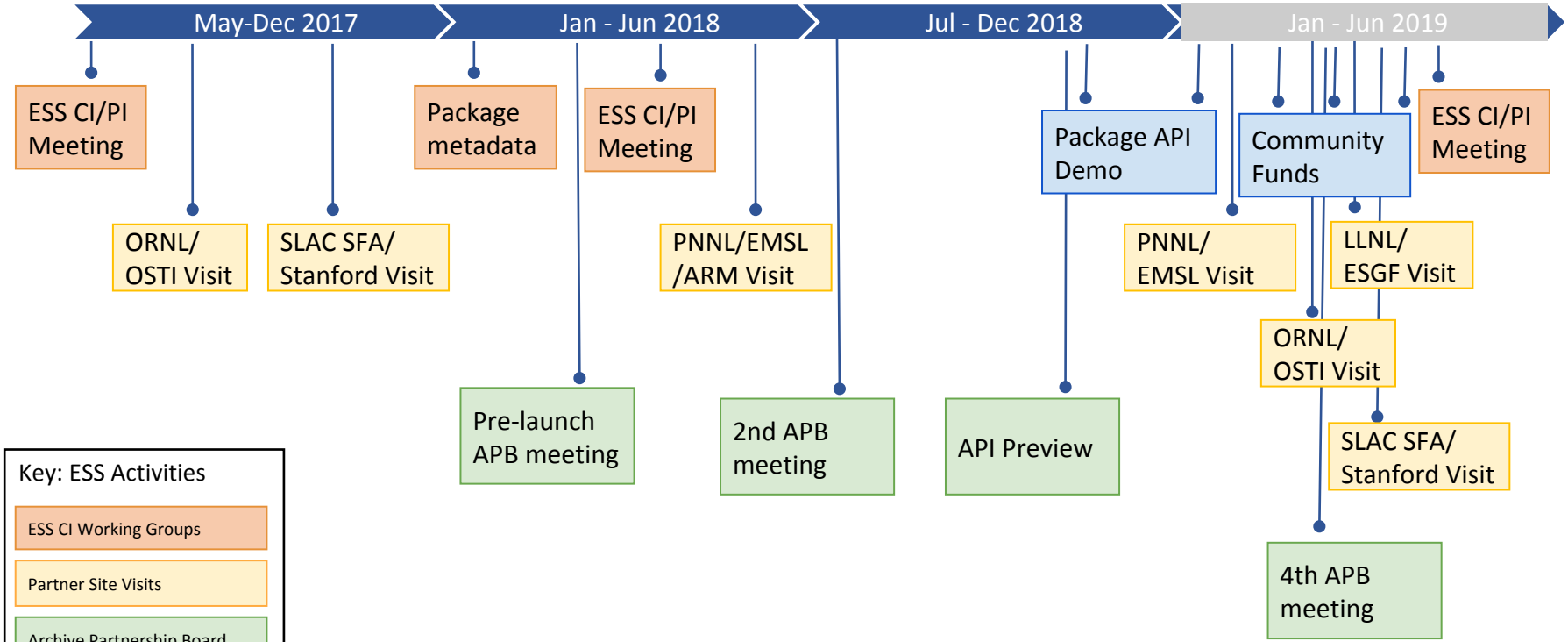
# Community Engagement

How we interact with the community

Gather requirements in several ways

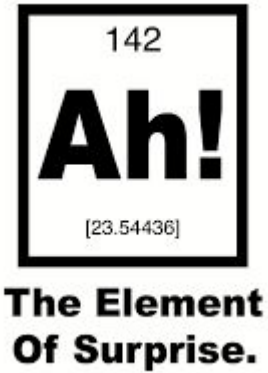
- Partner Site visits
- Webinars
- ESS CI Working Groups
- Archive Partnership Board
- Other Meetings/Conferences
- Surveys

# Timeline of Community Engagement Activities




*Visits to conferences, meetings, workshops not shown*

# Lessons Learned and ESS-DIVE Impacts




| Assumptions | Reality |
|---|---|
| DataONE API sufficient | User-oriented API |
| 20TB cap with small file uploads | Large data files needed in this phase |
| People upload metadata to us through webform or API | Metadata harvesting from OSTI |
| Project Spaces could wait | Need ASAP! |
| File-level metadata standards | Sample ID standards |

# Standards Development and Data Quality Review

- Support needs of contributing scientists
- Maximize the value of data into the future
- Incorporate into data quality review

# Complete: Package-Level Metadata

Gathered information from standards group and the community

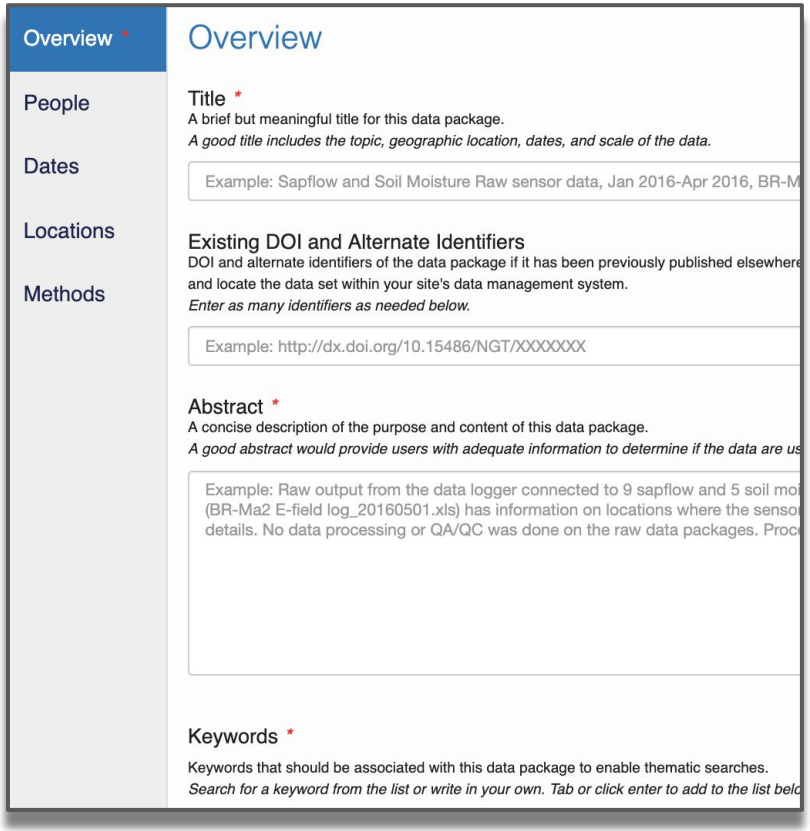
Cyberinfrastructure Working Group

DataCite, OSTI, ORNL, EMSL, PNNL

Cross-walk comparison - everyone sees how to translate their standard to ours

Finalized in April 2018

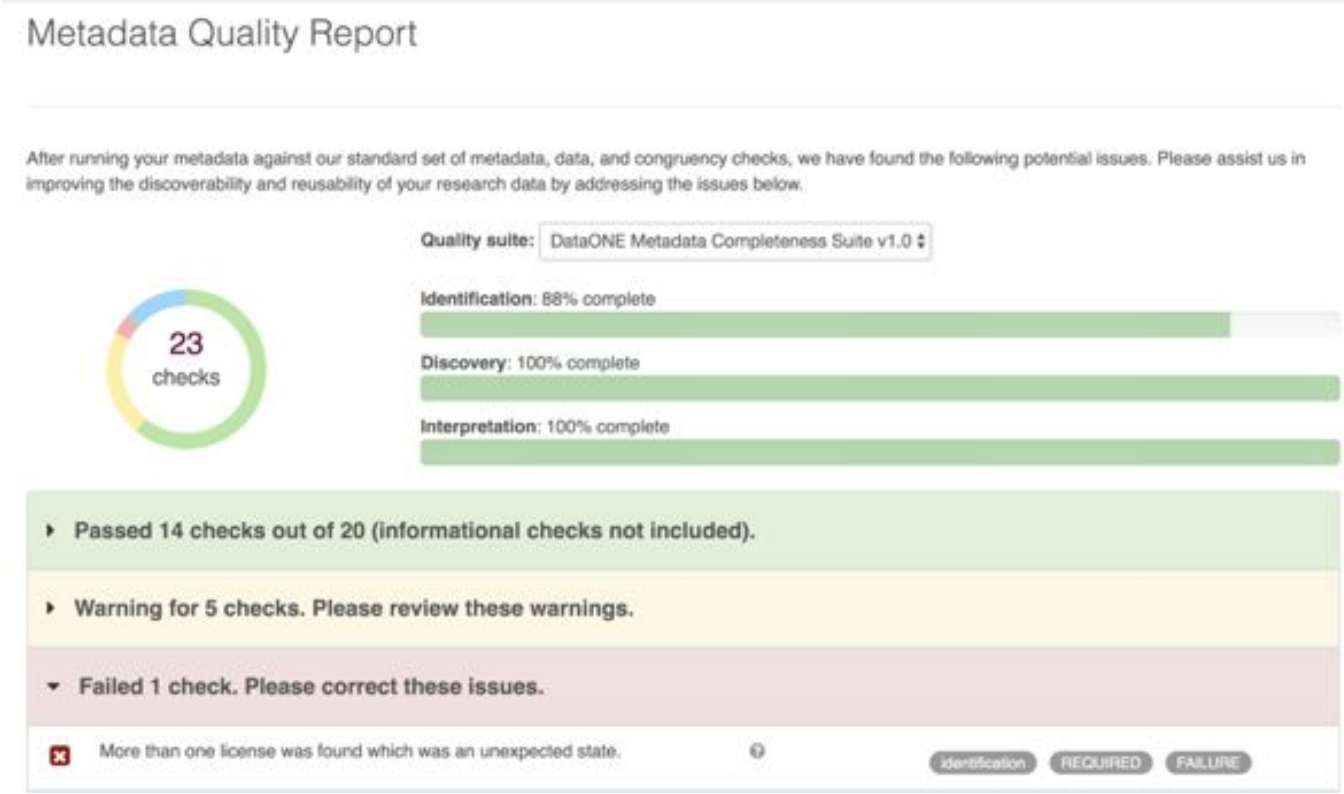
| ESS-DIVE Field | JSON-LD | DataCite 4.1 |
|---|---|---|
| Title | name | title |
| Alternative Identifiers | alternateName | alternateIdentifiers |
| Abstract | description | description[@type=Abstract] |
| Keywords | keywords | subjects |
| Data Variables | variablesMeasured | subjects |
| Publication Date | datePublished | publicationYear |



# Package-Level Metadata and Data Review

Manual review process

Automated metadata checks in the web form and API



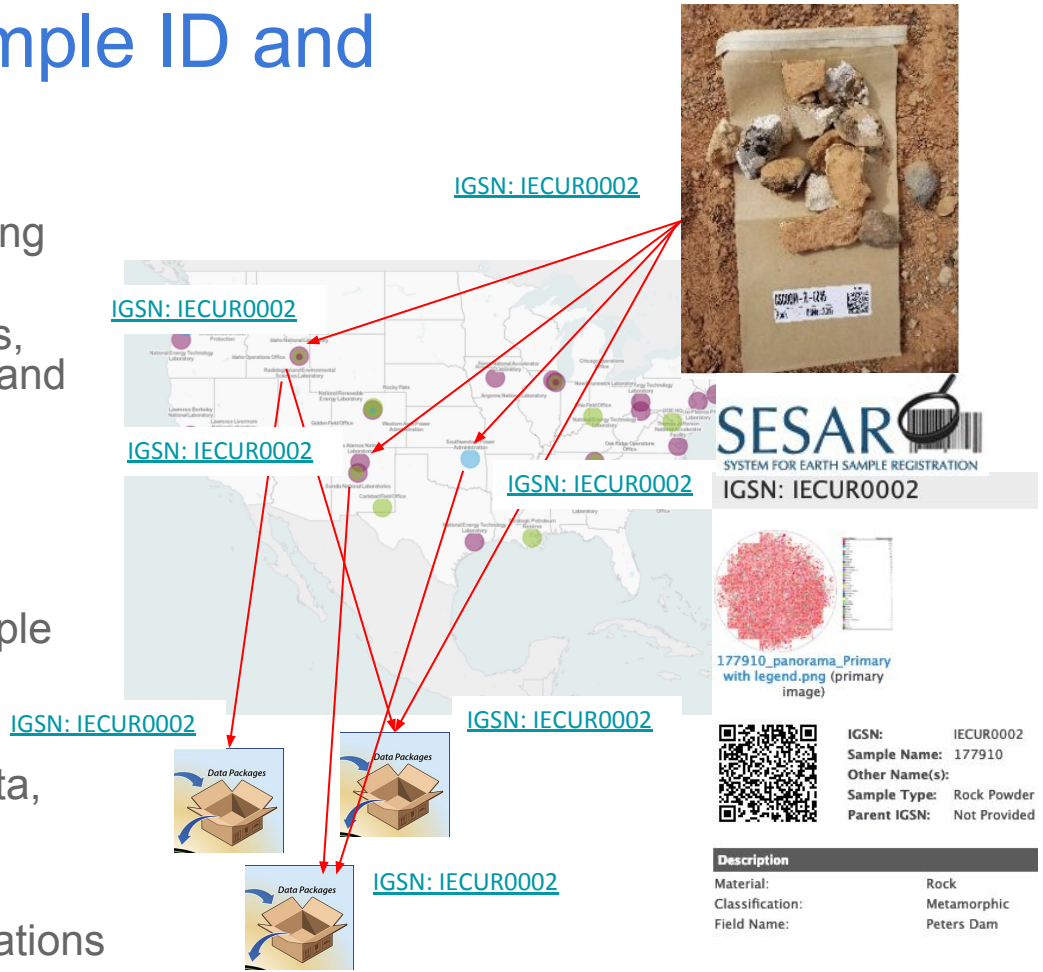
# Community Need: Sample ID and Tracking

Kate Maher - Sample naming and tracking from field to dataset publication

**Research:** Lit Review, other repositories, user facilities (JGI, EMSL, KBase), PID and metadata specialists (Kerstin Lehnert), RDA

**Draft Proposal:** International Geo Sample Numbers (IGSNs) for ESS samples

- Standardized core sample metadata, templates
- Linking to other samples, online metadata profiles, datasets, publications

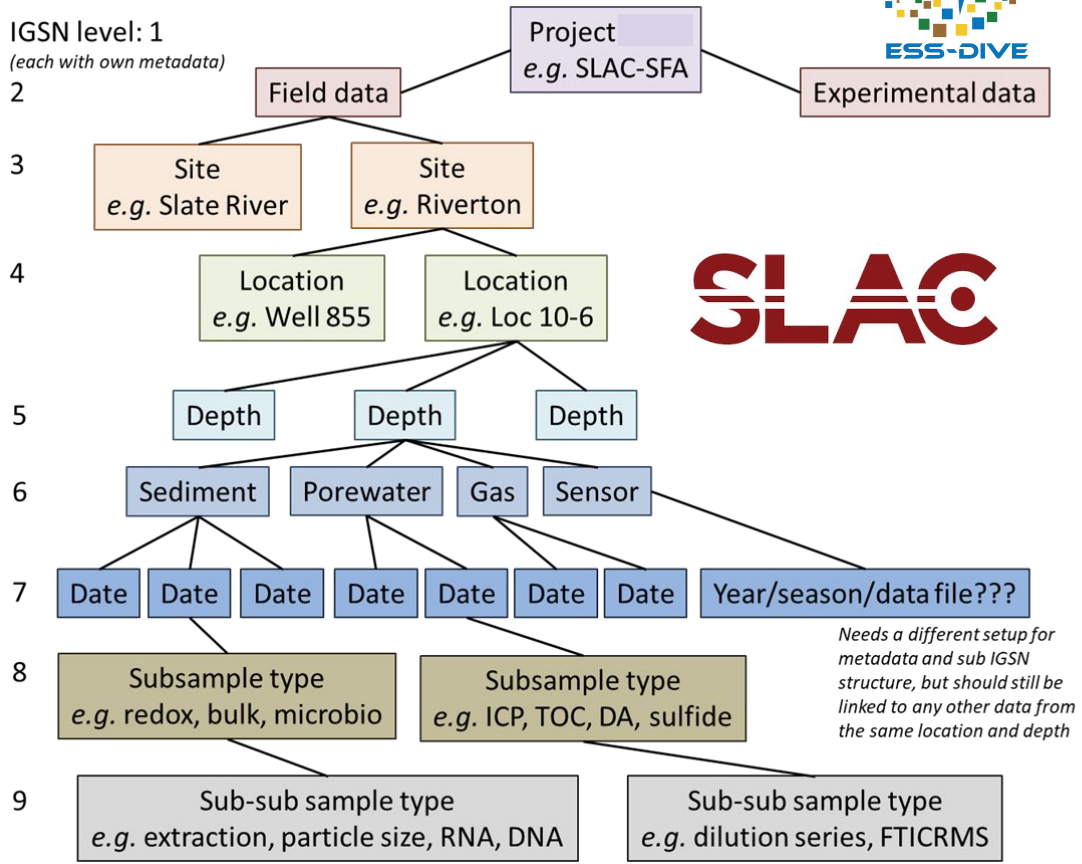


# Pilot IGSN and Sample Tracking

SLAC-SFA, SBR SFAs, WHONDRS

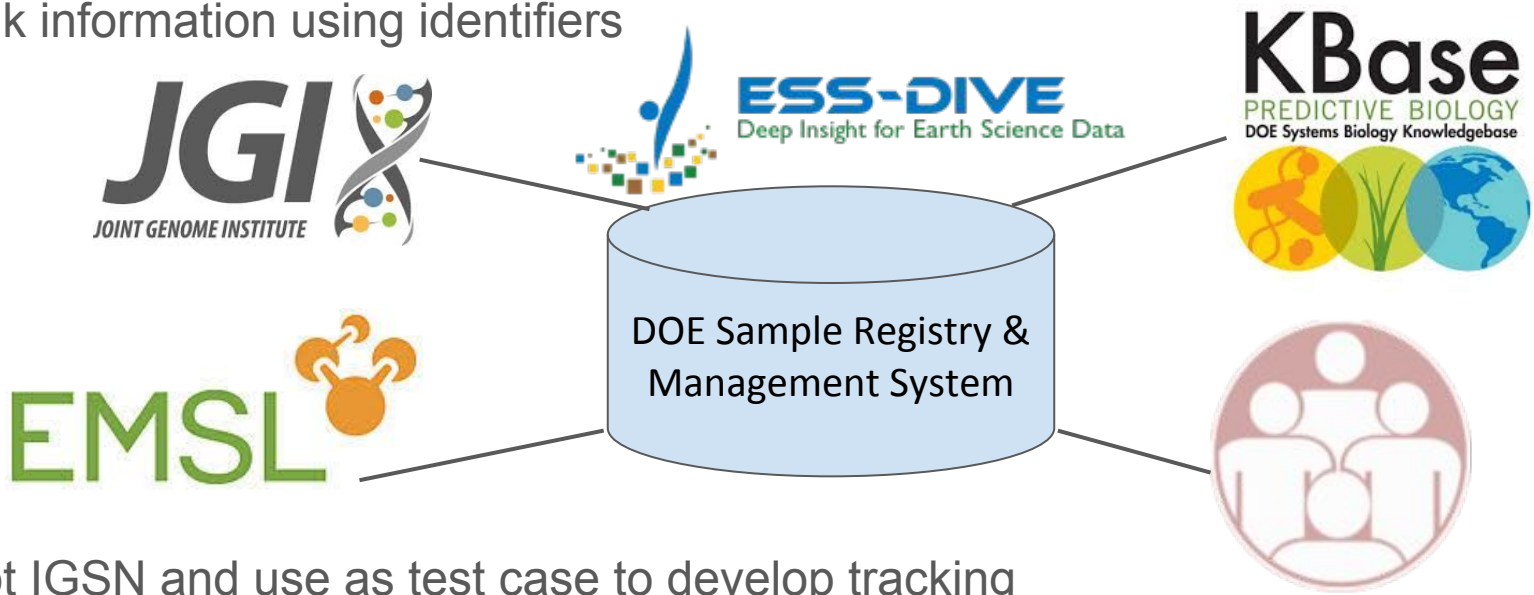
**Register IGSNs**: Decide what gets IGSN, sample relationships, and metadata needed

**Develop workflows**: Fit IGSN and metadata collection into planned field and lab workflows



# Sample Identification - Across User Facilities

Need central DOE system to register samples, obtain PIDs, add metadata, and link information using identifiers




Pilot IGSN and use as test case to develop tracking system across facilities

# Future Plans for Standardization and Review

**File-level metadata** is a top priority to make files machine-readable


- File format and software used
- Variable names, descriptions and units
- Date/Time (e.g. YYYY-MM-DD, ISO 8601)
- Latitude/Longitude (WGS 84)
- Values for no data
- Links to other files
- Any manipulations of the file

Standards needed for fusion database

practical data curation and submission process

Necessary for **Fusion database** → advanced search within and across datasets

*Gathering information file-metadata captured by projects and repositories*

Become part of **data quality reviews**


Community need for ***usable APIs*** that could be used without software engineers.
-- Community Engagement

Community Driven

# User Services

Provide key front-end capabilities for ESS users.

Four major interfaces and support services

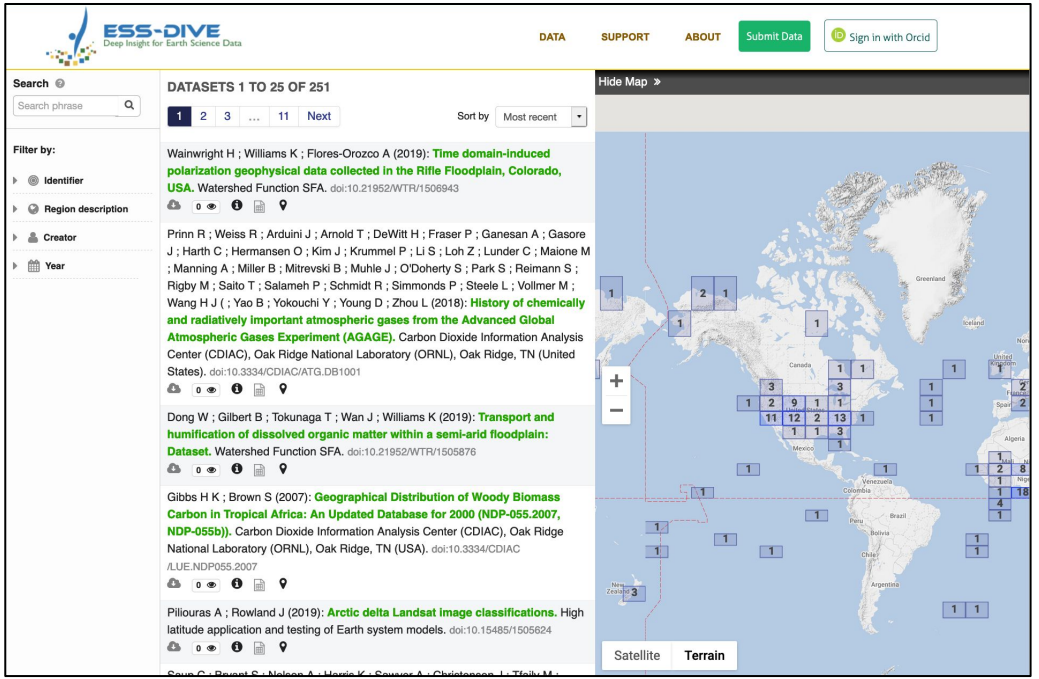
- Web Portal
- Package Service API
- API Documentation Portal
- API Tutorials & Code

# ESS-DIVE Data Portal Launched on April 1, 2018

## Web Portal

*http://data.ess-dive.lbl.gov*

- Multiple ways to **find data** through keyword search and filters
- **Public download** of data and metadata
- **Tracking of downloads** for data contributors and programs
- **ORCiD logins** provide federated access



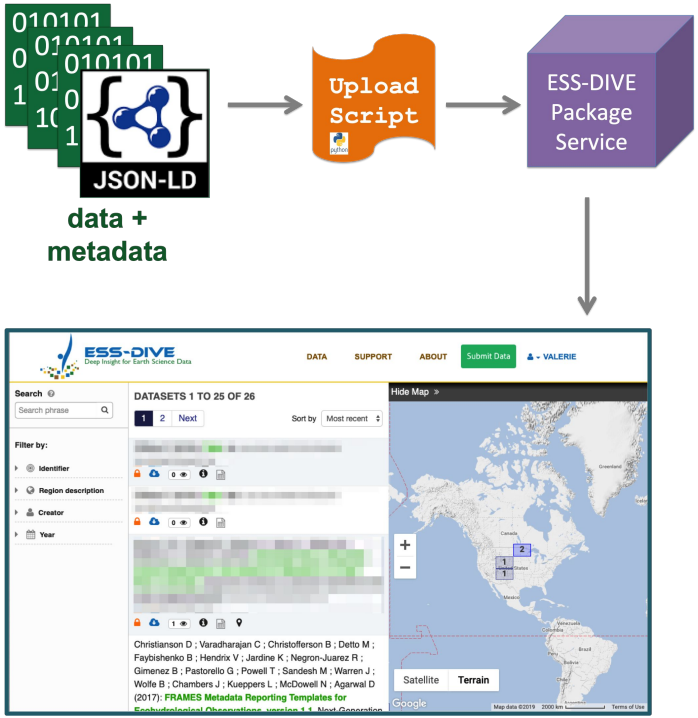
# User Services

Community Driven

**Package Service API**

*http://api.ess-dive.lbl.gov/packages*

- Released 1.0.0 March 2019
- User friendly web service for programmatic submission of ***new*** data packages.
- JSON-LD metadata standard for scripting ease.
- Users can upload metadata+data to ESS-DIVE via scripts
- ***Next:*** updates to existing data packages.

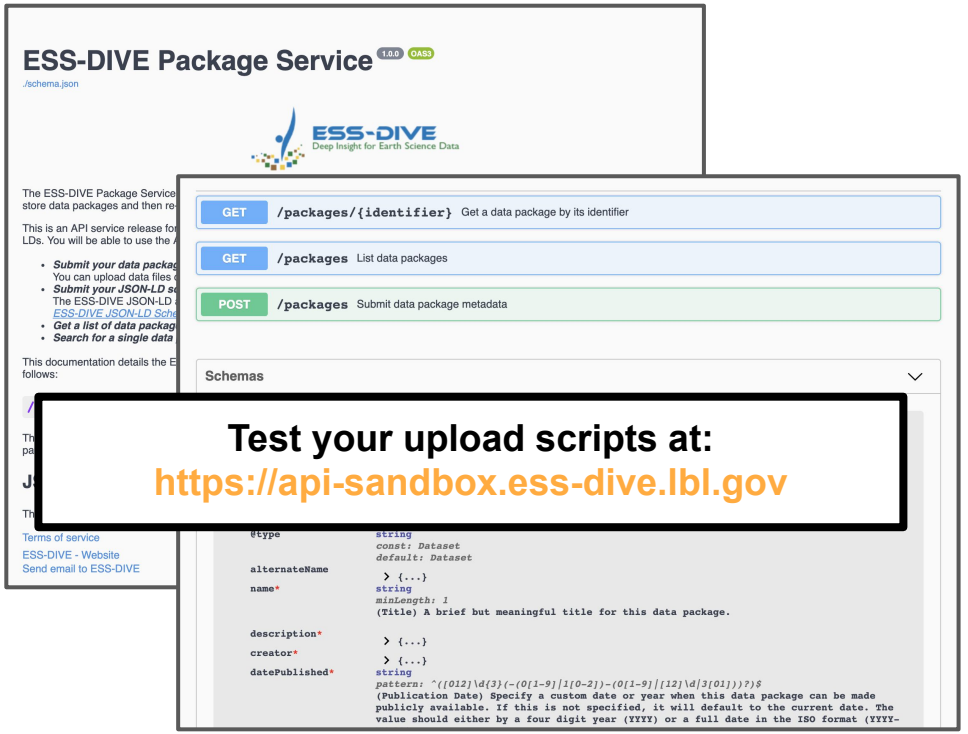


# User Services

**API Documentation Portal**

*http://api.ess-dive.lbl.gov*

- Detailed technical documentation for ESS project developers.
- ESS users learn how to create, list and view a single data package
- Access to a sandbox system for testing API usage by ESS project developers.



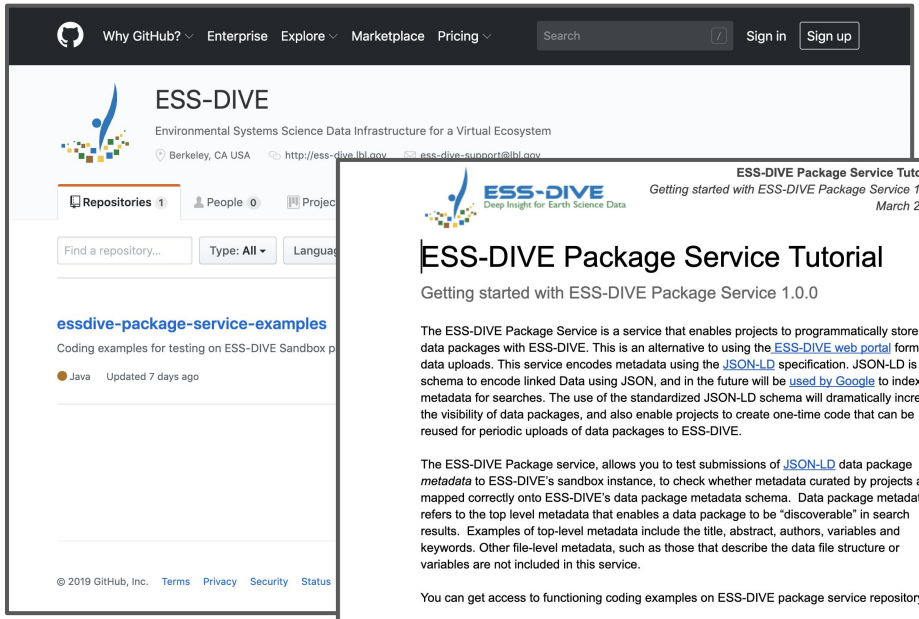
# User Services

**ESS-DIVE Tutorial - April 30th 12-1, Rm 15/16**

## API Tutorials & Code

- API Tutorials provided in three languages (Python, R and Java)
- Example code published on Github
- Metadata crosswalk provided to help projects get started.

| ESS-DIVE Field | JSON-LD | DataCite 4.1 |
|---|---|---|
| Title | name | title |
| Alternative Identifiers | alternateName | alternateIdentifiers |
| Abstract | description | description[@type=Abstract] |
| Keywords | keywords | subjects |
| Data Variables | variablesMeasured | subjects |
| Publication Date | datePublished | publicationYear |



# User Interfaces

Community Driven

## Web Portal

*http://data.ess-dive.lbl.gov*

### Project Spaces

- A ***beta*** version has been release in MetacatUI.
- An ***ESS project view*** can be defined in ESS-DIVE via XML.
- ***Next steps*** are to work with Community Engagement and NCEAS to define the next iteration which will allow projects to manage their space.

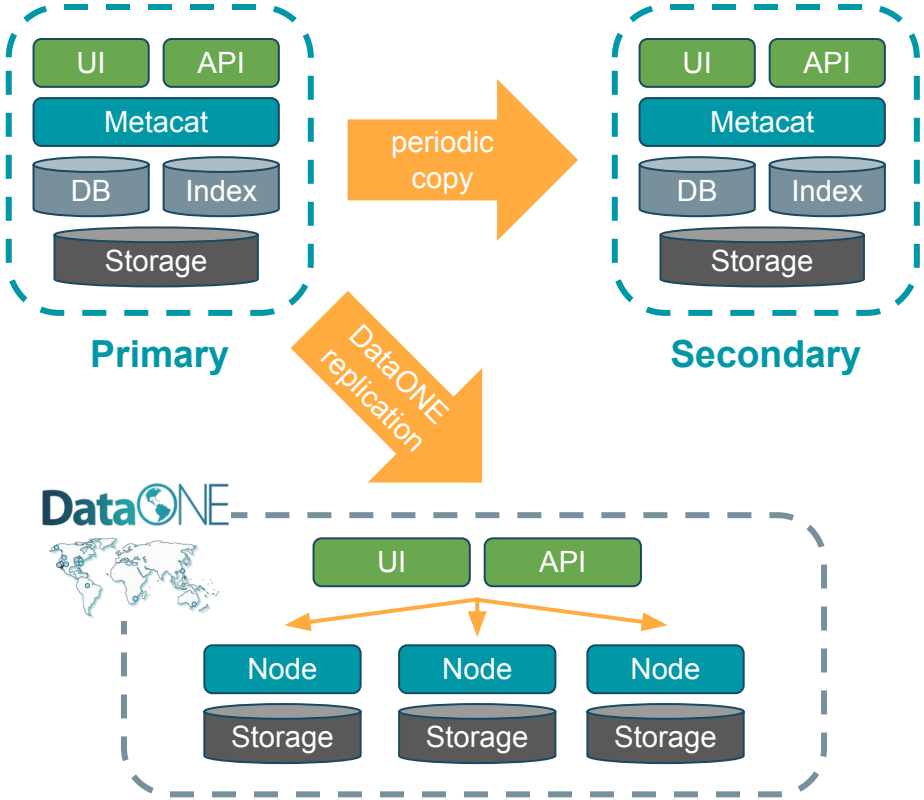
ESS-DIVE

# Preservation Services

Manages long-term data package preservation and availability.

- ✓ **Durable identifiers** *for citation accuracy*
- ✓ **Quality metadata** *for discoverability and provenance*
- ✓ **Change management rigor** *for predictability and reliability*
- ✓ **Redundancy** *to prevent loss of data or service*
- ✓ **Data auditing and reporting** *to detect data loss/corruption*

# ESS-DIVE Redundancy Model




**ESS-DIVE is *highly redundant*:**

- ✓ Two instances at Berkeley
- ✓ Replication to three nodes in the DataONE network

**The architecture spans**

- ✓ Five sites
- ✓ Multiple organizations
- ✓ Multiple geographies

***18-month uptime: 99.9778%***

# Updates and Next Steps in Preservation Services

- **Transfer of CDIAC Data sets**
- **DataONE Federation**
- ***Next:*** **Alternative Data Upload Tools**
  - needs surfaced during Community Engagement
  - will handle data files from 2GB to 100GB
  - will leverage Globus transfers
  - designed for automated processes

# Discussion on Community Funds