



# 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**

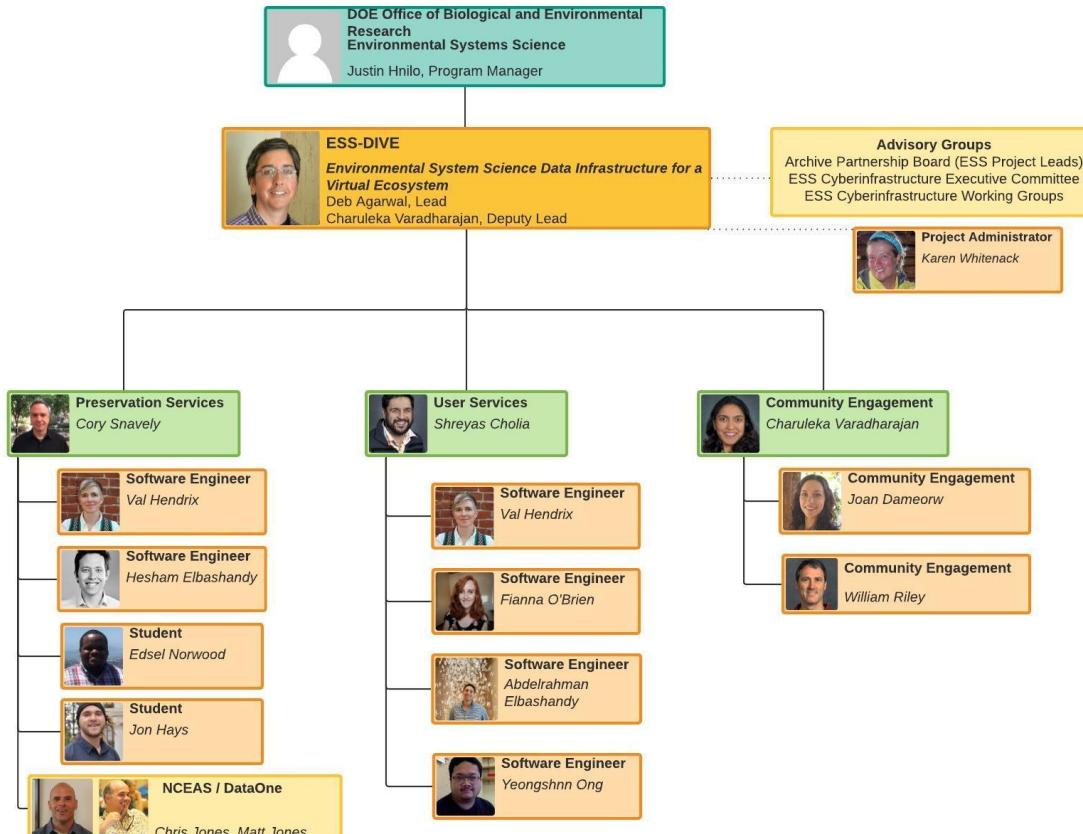
Office of  
Science



BERKELEY LAB  
Lawrence Berkeley National Laboratory



# ESS-DIVE team has expanded



**Environmental  
Scientists**

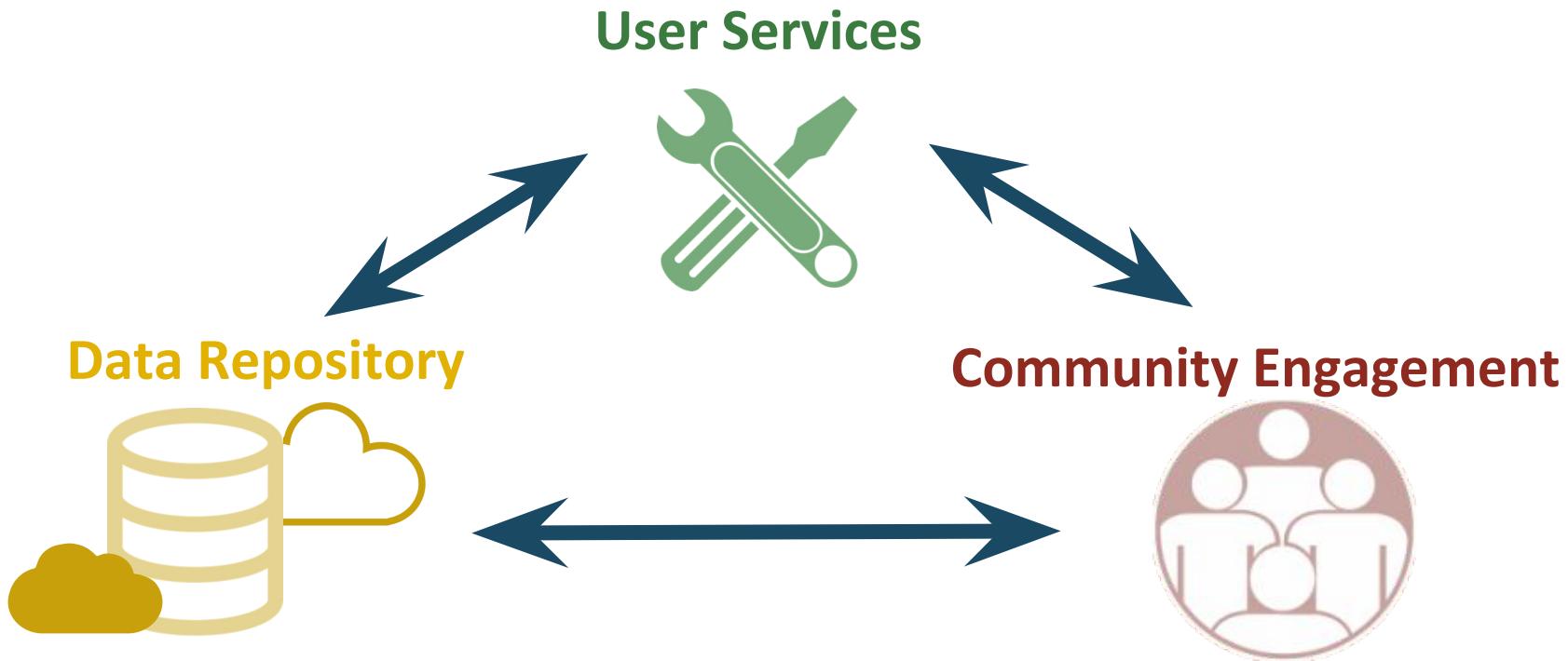
**Data Scientists**

**Software Engineers**

**Digital Librarians**

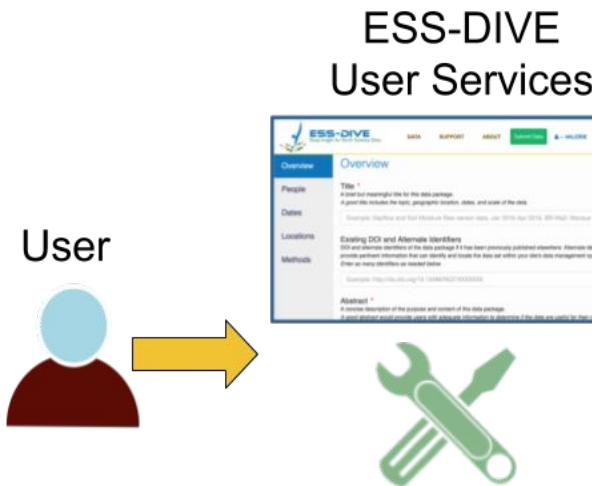


# Three Pronged Approach to Repository





# ESS-DIVE Archiving Features



DataONE





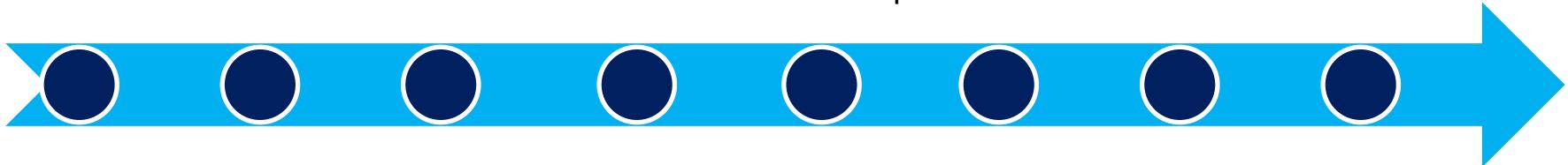
# ESS-DIVE Implementation Timeline

**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



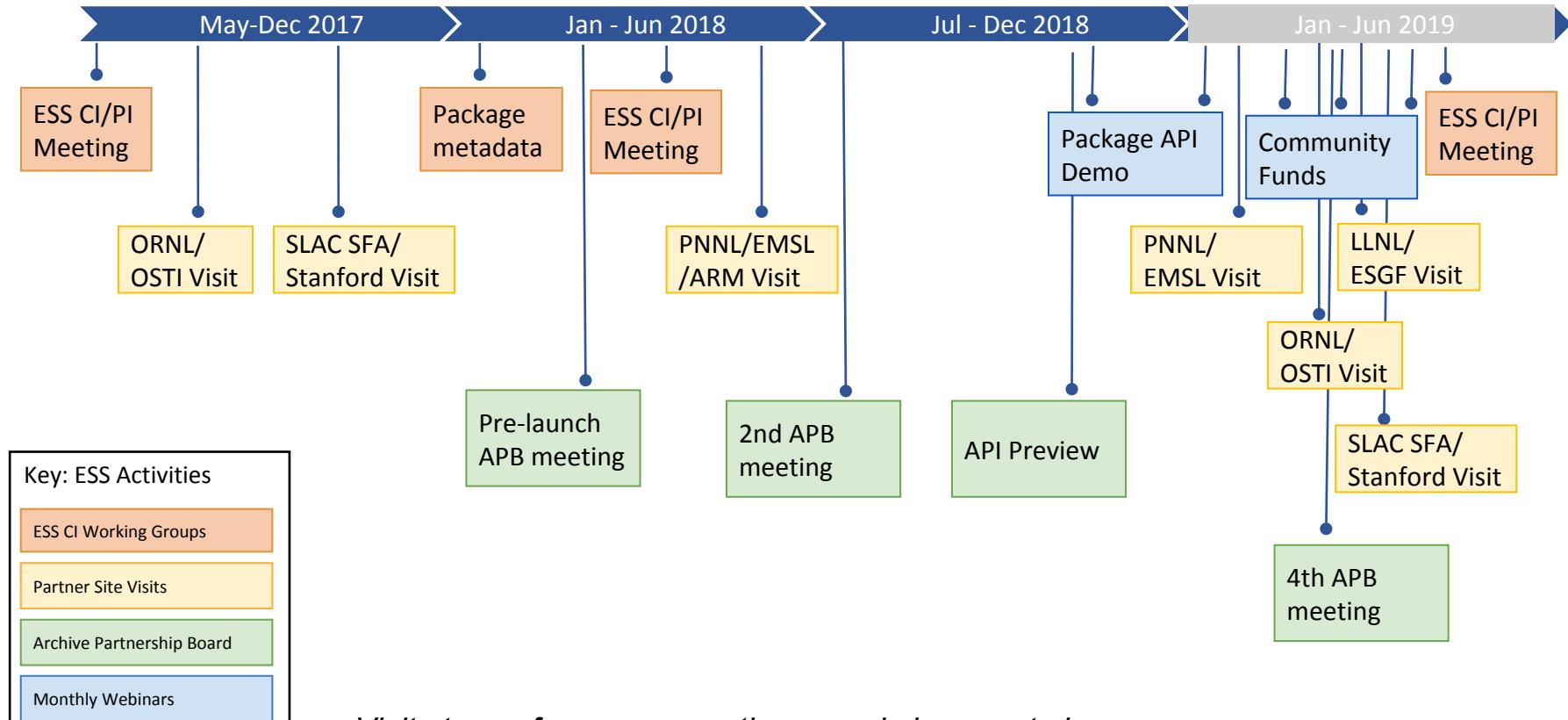
# Community Engagement

How we interact with the community

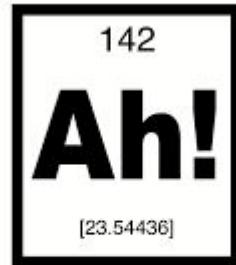
Gather requirements in several ways

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

# Timeline of Community Engagement Activities



# Lessons Learned and ESS-DIVE Impacts



**The Element  
Of Surprise.**

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	alternatelIdentifiers
Abstract	description	description[@type=Abstract]
Keywords	keywords	subjects
Data Variables	variablesMeasured	subjects
Publication Date	datePublished	publicationYear

Overview

People

Dates

Locations

Methods

Overview

**Title \***  
A brief but meaningful title for this data package.  
*A good title includes the topic, geographic location, dates, and scale of the data.*

Example: Sapflow and Soil Moisture Raw sensor data, Jan 2016-Apr 2016, BR-M

Existing DOI and Alternate Identifiers  
DOI and alternate identifiers of the data package if it has been previously published elsewhere and locate the data set within your site's data management system.  
*Enter as many identifiers as needed below.*

Example: <http://dx.doi.org/10.15486/NGT/XXXXXX>

**Abstract \***  
A concise description of the purpose and content of this data package.  
*A good abstract would provide users with adequate information to determine if the data are useful.*

Example: Raw output from the data logger connected to 9 sapflow and 5 soil moisture sensors (BR-Ma2 E-field log\_20160501.xls) has information on locations where the sensors were deployed, details on the sensors and their configuration, and raw data. No data processing or QA/QC was done on the raw data packages. Processing steps will be done by the user before sharing the data.

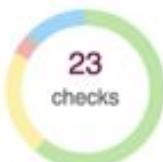
**Keywords \***  
Keywords that should be associated with this data package to enable thematic searches.  
*Search for a keyword from the list or write in your own. Tab or click enter to add to the list below.*

# Package-Level Metadata and Data Review

## Metadata Quality Report

### Manual review process

After running your metadata against our standard set of metadata, data, and congruency checks, we have found the following potential issues. Please assist us in improving the discoverability and reusability of your research data by addressing the issues below.



Quality suite: DataONE Metadata Completeness Suite v1.0 [+](#)

Identification: 88% complete

Discovery: 100% complete

Interpretation: 100% complete

▶ Passed 14 checks out of 20 (informational checks not included).

▶ Warning for 5 checks. Please review these warnings.

▼ Failed 1 check. Please correct these issues.

☒ More than one license was found which was an unexpected state.



identification REQUIRED FAILURE

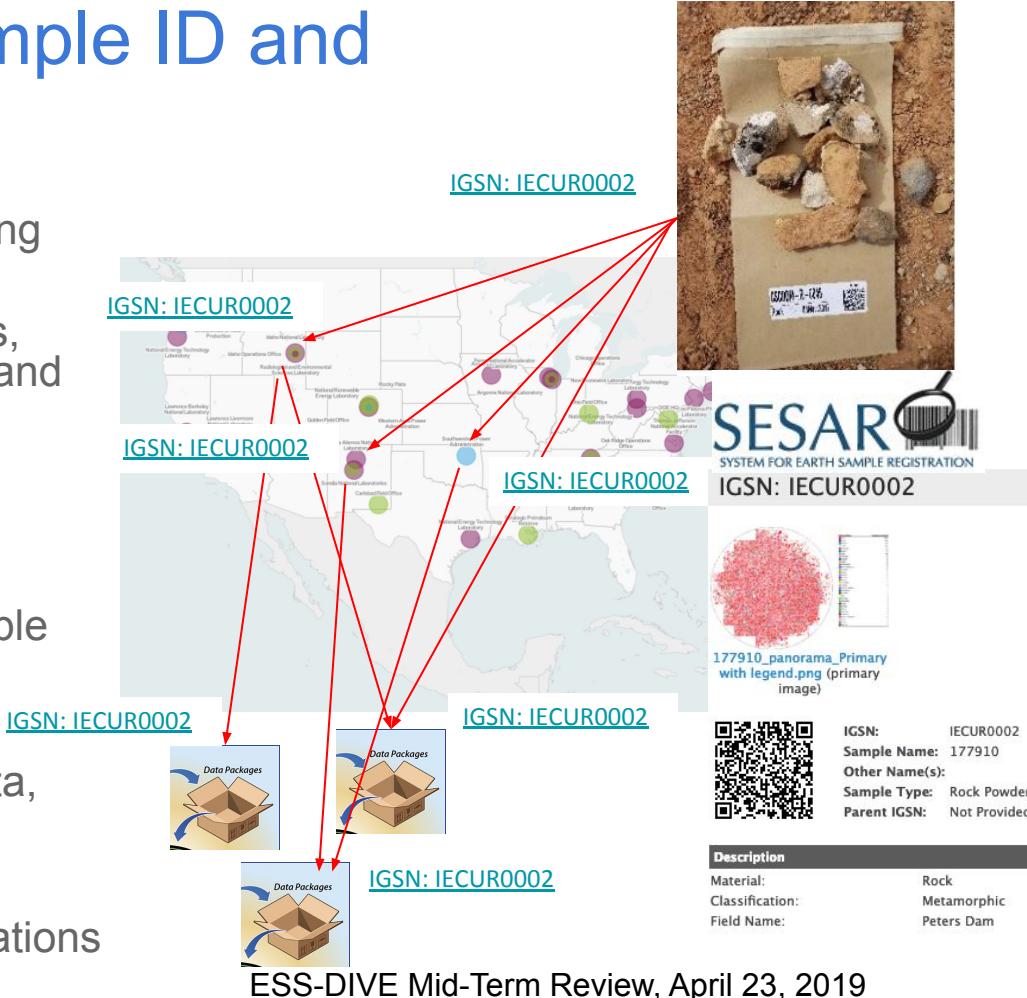
# Community Need: Sample ID and Tracking

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



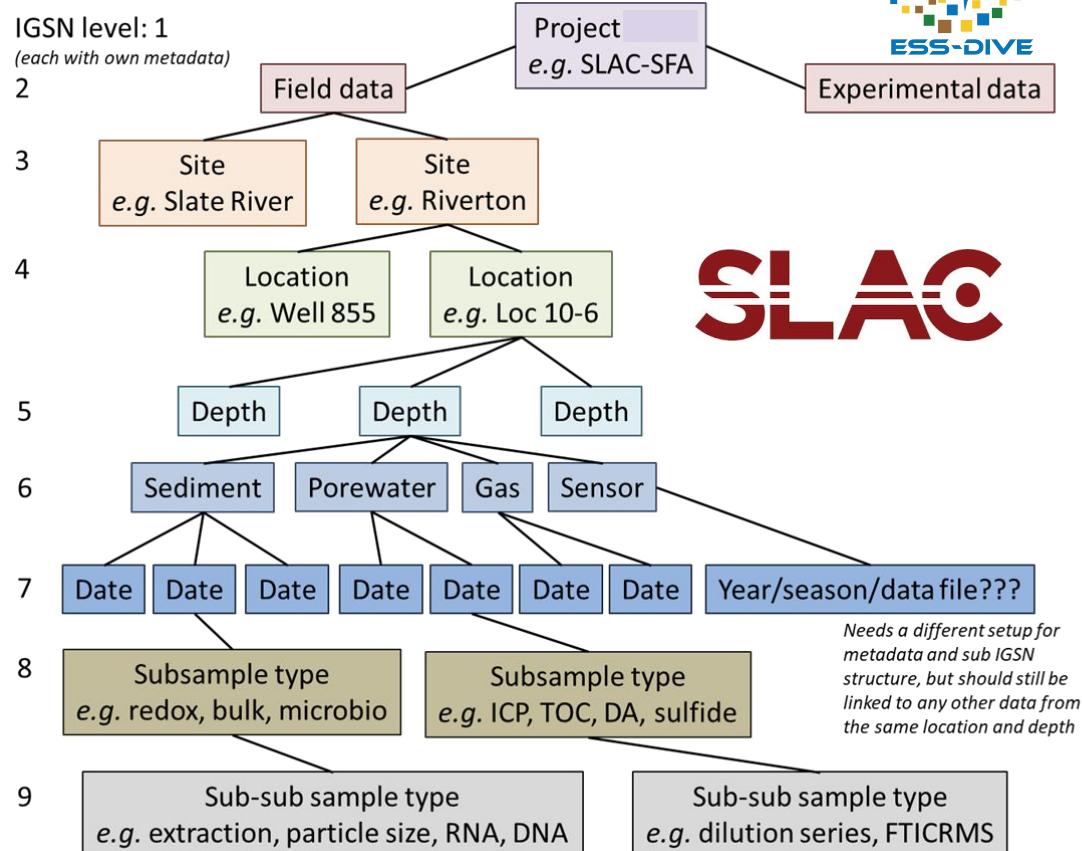
# Pilot IGSN and Sample Tracking



SLAC-SFA, SBR SFAs,  
WHONDRS

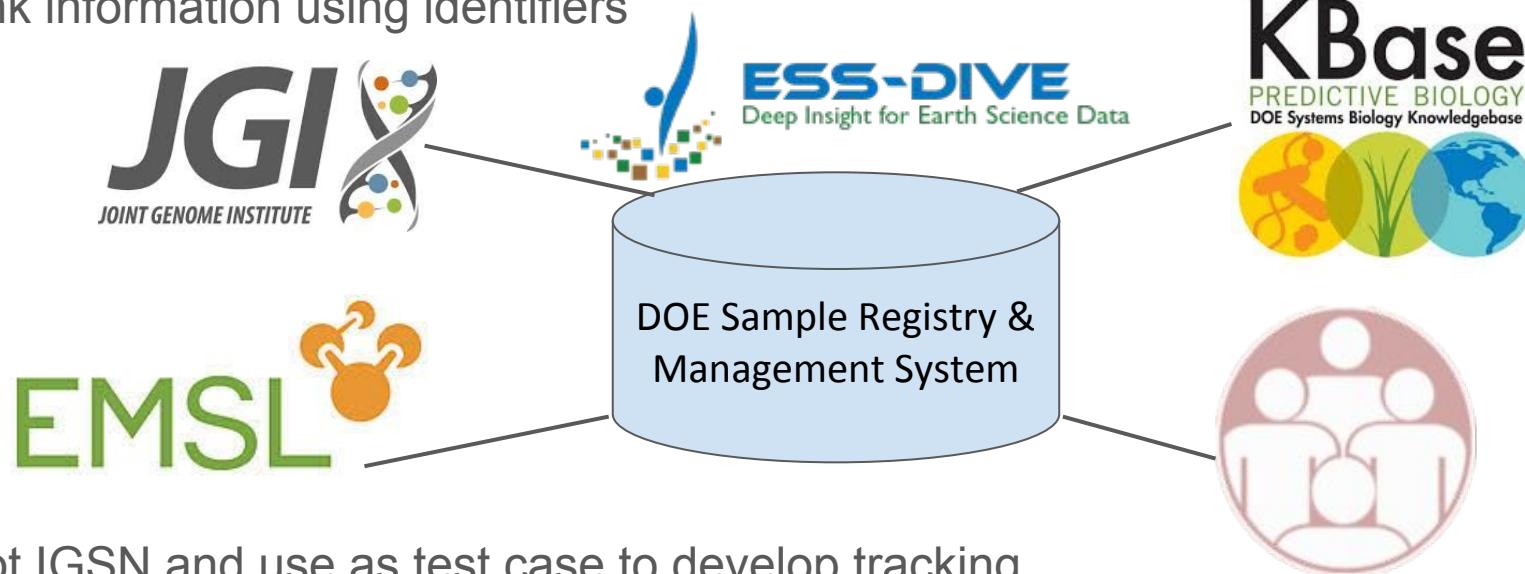
**Register IGSNs:** Decide what gets IGSN, sample relationships, and metadata needed

**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**

# User Services

Provide key front-end capabilities for ESS users.

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

Community  
Driven

Four major interfaces and support services

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

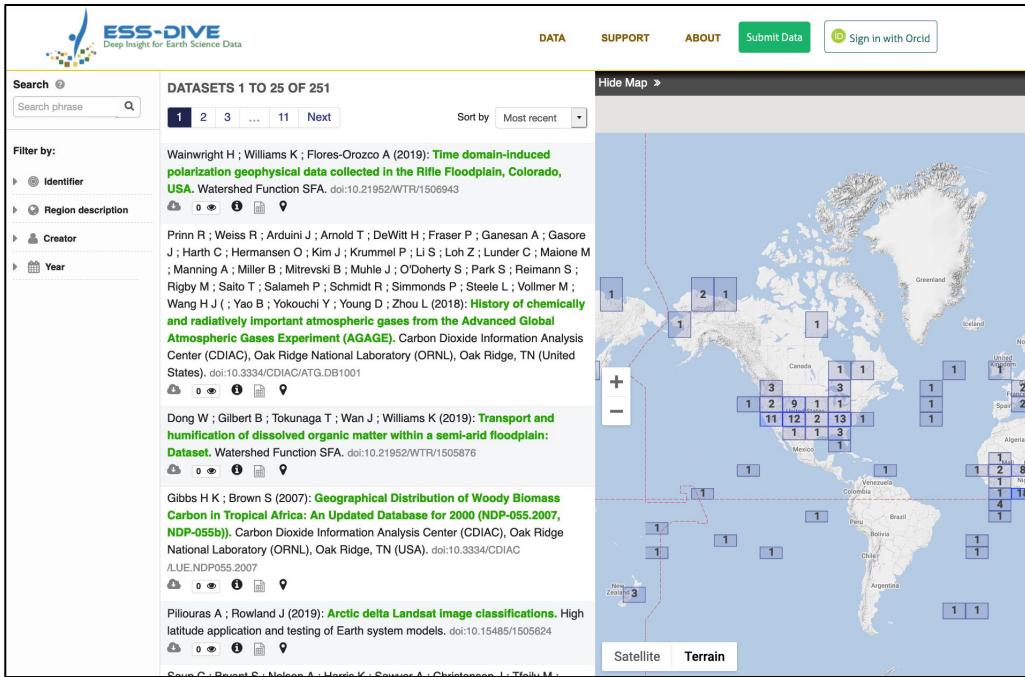
# ESS-DIVE Data Portal Launched on April 1, 2018



## 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

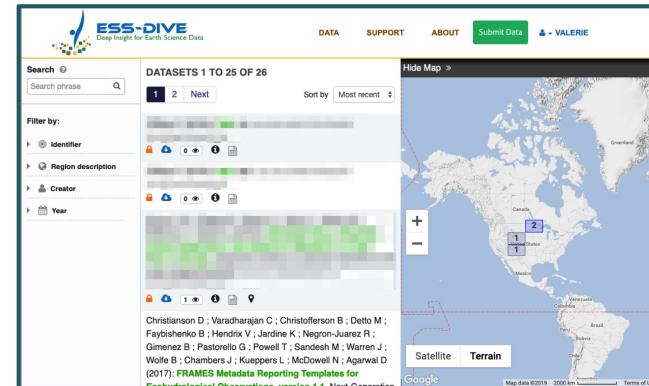
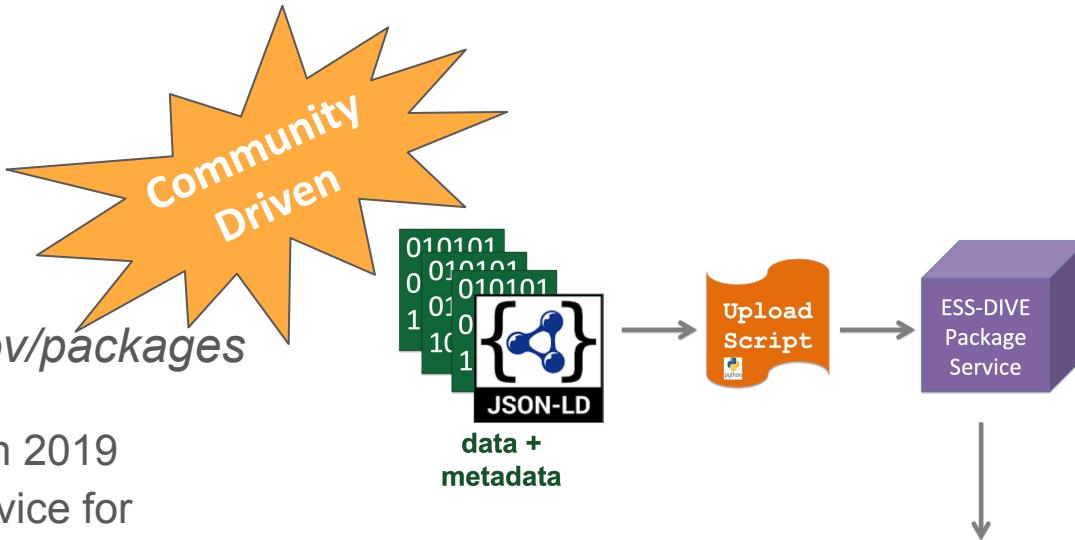


# User Services

## 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.

**ESS-DIVE Package Service 1.0.0 OAS3**

[/schema.json](#)

The ESS-DIVE Package Service store data packages and then restore data packages and then re

This is an API service release for LDs. You will be able to use the A

- [Submit your data package](#)  
You can upload data files.
- [Submit your JSON-LD schema](#)  
The ESS-DIVE JSON-LD Schema
- [Get a list of data packages](#)
- [Search for a single data](#)

This documentation details the E follows:

**Schemas**

**Test your upload scripts at:**  
<https://api-sandbox.ess-dive.lbl.gov>

Terms of service  
ESS-DIVE - Website  
Send email to ESS-DIVE

```

#type          string
            const: Dataset
            default: Dataset
            > {...}
            string
            minLength: 1
            (Title) A brief but meaningful title for this data package.

name*          string
            > {...}
            string
            minLength: 1
            (Name) A unique name for this data package.

description*  string
            > {...}
            string
            minLength: 1
            (Description) A detailed description of this data package.

creator*       string
            > {...}
            string
            minLength: 1
            (Creator) The creator or source of this data package.

datePublished* string
            pattern: `^((012)\d(3)(-(01-9)|1[0-2])-(01-9)|(12)\d(3|01))?\$`  

            (Publication Date) Specify a custom date or year when this data package can be made  

publicly available. If this is not specified, it will default to the current date. The  

value should either be a four digit year (YYYY) or a full date in the ISO format (YYYY-MM-DD).

```

# 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

A screenshot showing two parts of the ESS-DIVE GitHub repository. On the left, the main repository page for "essdive-package-service-examples" is shown, featuring a repository icon, a count of 1 repository, 0 people, and 0 projects. It includes a search bar and filters for "Type: All" and "Language". Below the header, there's a section for "essdive-package-service-examples" with a link to "Coding examples for testing on ESS-DIVE Sandbox package service". A note indicates it was last updated 7 days ago. At the bottom of the page, there are links to "© 2019 GitHub, Inc.", "Terms", "Privacy", "Security", and "Status". On the right, a separate page titled "ESS-DIVE Package Service Tutorial" is displayed, with the subtitle "Getting started with ESS-DIVE Package Service 1.0.0 March 2019". This page contains detailed information about the service, including its purpose, how it uses JSON-LD, and its benefits. It also provides links to "coding examples" and a "Get Access" section. The overall layout is clean and professional, typical of a scientific software documentation site.

# User Interfaces

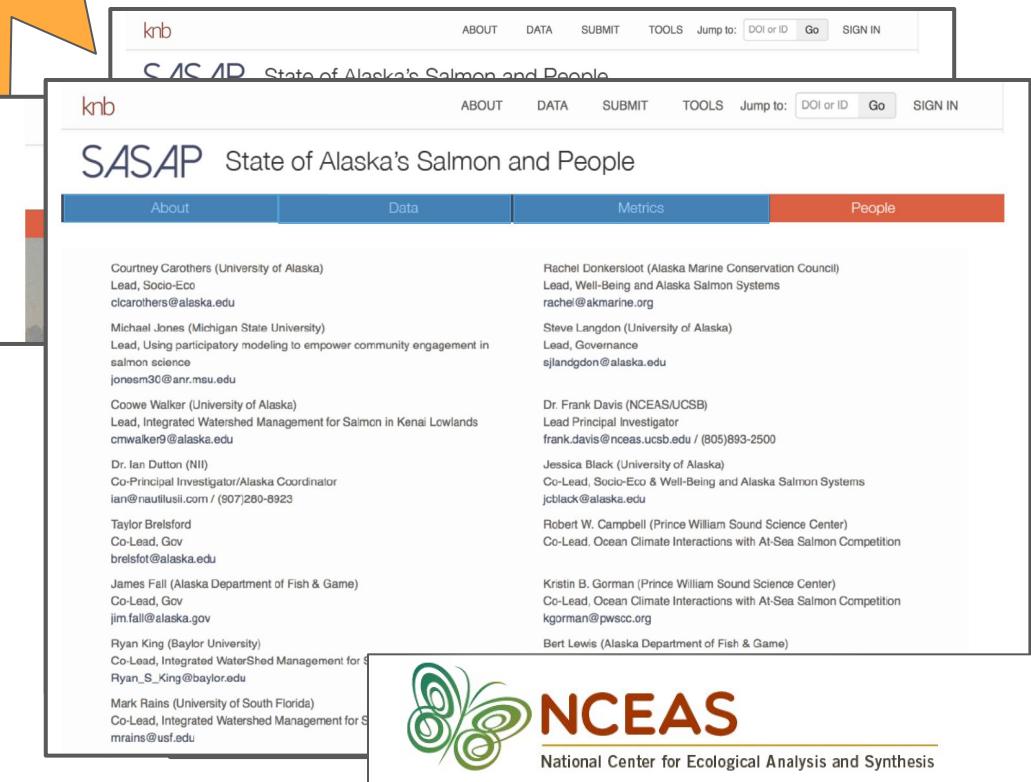
## Web Portal

<http://data.ess-dive.lbl.gov>

## Project Spaces

- A **beta** version has been released 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.





The screenshot displays the ESS-DIVE web portal interface. At the top, there are two project spaces: "CACAD - State of Alaska's Salmon and People" and "SASAP - State of Alaska's Salmon and People". The "SASAP" space is currently active. Below the header, there are tabs for "About", "Data", "Metrics", and "People". The "People" tab is selected, showing a list of project members:

Name	Role	Organization	Contact Information
Courtney Carothers	Lead, Socio-Eco	University of Alaska	cicarother@alaska.edu
Michael Jones	Lead, Using participatory modeling to empower community engagement in salmon science	Michigan State University	joneen30@anr.msu.edu
Coowe Walker	Lead, Integrated Watershed Management for Salmon in Kenai Lowlands	University of Alaska	cmwalker9@alaska.edu
Dr. Ian Dutton	Co-Principal Investigator/Alaska Coordinator	NII	ian@nautilus.us.com / (907)280-8923
Taylor Brelsford	Co-Lead, Gov		brelstof@alaska.edu
James Fall	Co-Lead, Gov	Alaska Department of Fish & Game	jim.fall@alaska.gov
Ryan King	Co-Lead, Integrated WaterShed Management for Salmon	Baylor University	Ryan_S_King@baylor.edu
Mark Rains	Co-Lead, Integrated Watershed Management for Salmon	University of South Florida	mrains@usf.edu
Rachel Donkersloot	Lead, Well-Being and Alaska Salmon Systems	Alaska Marine Conservation Council	rachel@akmarine.org
Steve Langdon	Lead, Governance	University of Alaska	sjlangdon@alaska.edu
Dr. Frank Davis	Lead Principal Investigator	NCEAS/UCSB	frank.davis@nceas.ucsb.edu / (805)893-2500
Jessica Black	Co-Lead, Socio-Eco & Well-Being and Alaska Salmon Systems	University of Alaska	jcbblack@alaska.edu
Robert W. Campbell	Co-Lead, Ocean Climate Interactions with At-Sea Salmon Competition	Prince William Sound Science Center	
Kristin B. Gorman	Co-Lead, Ocean Climate Interactions with At-Sea Salmon Competition	Prince William Sound Science Center	kgorman@pwssc.org
Bert Lewis	Co-Lead, Ocean Climate Interactions with At-Sea Salmon Competition	Alaska Department of Fish & Game	

**NCEAS**  
National Center for Ecological Analysis and Synthesis

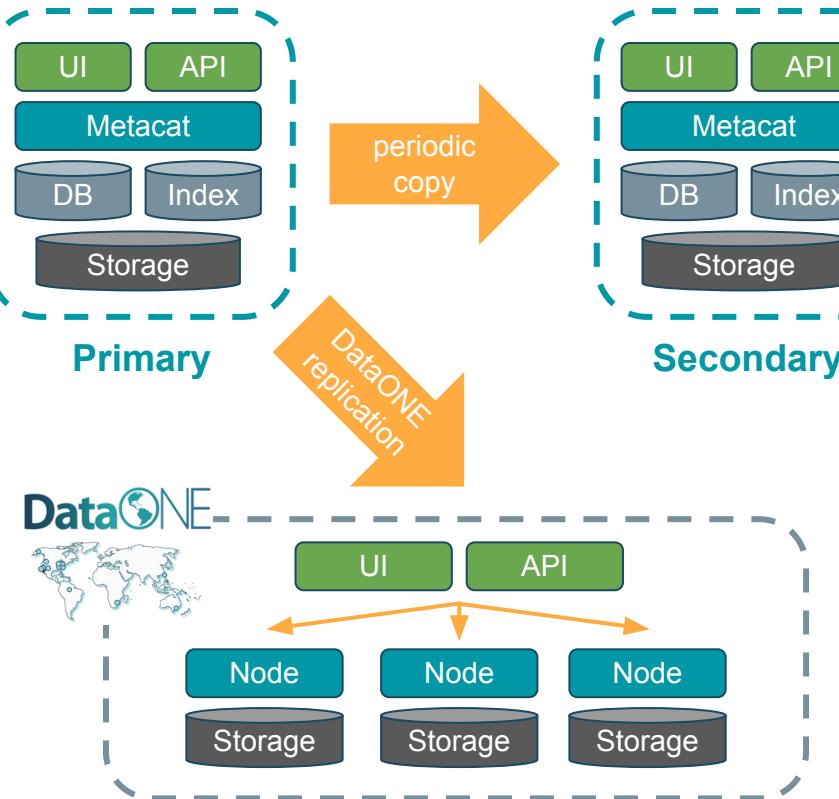


# Preservation Services

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