



ESS-DIVE Data Repository: Updates

April 29, 2019

ESS Cyberinfrastructure Working Group Meeting



ESS-DIVE

Deep Insight for Earth Science Data

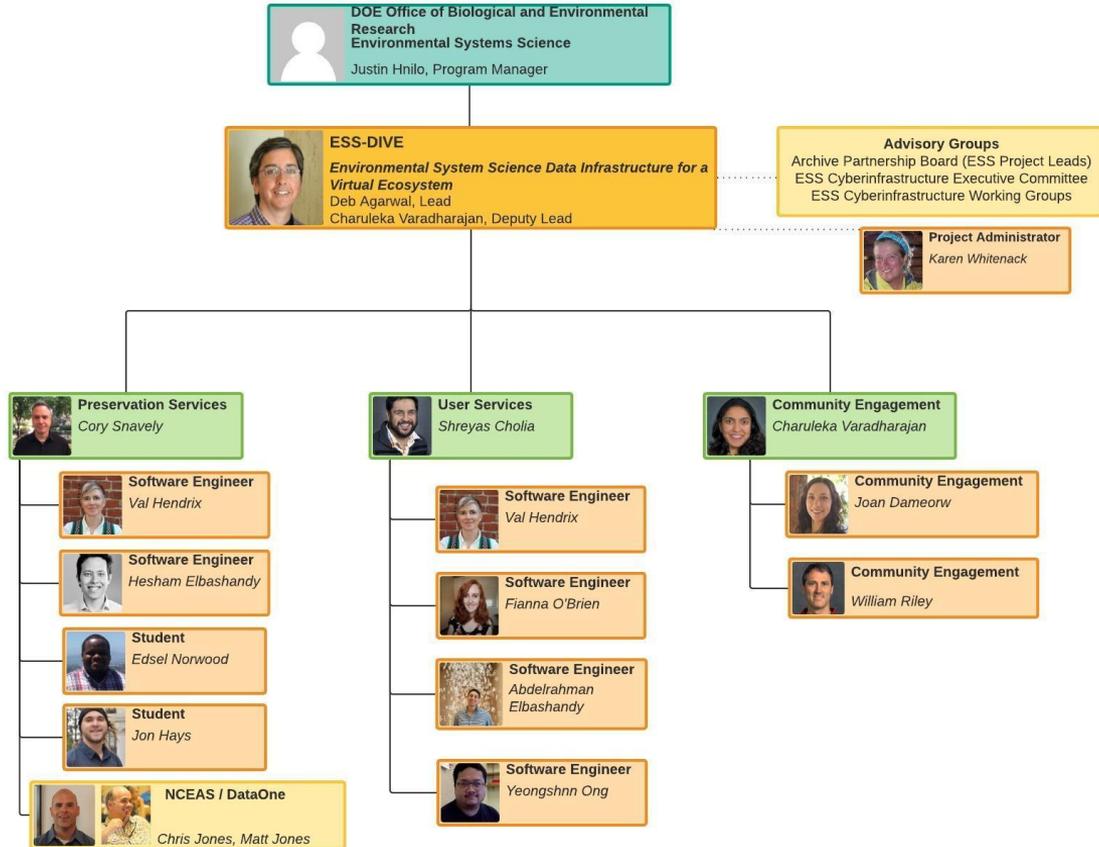


U.S. DEPARTMENT OF
ENERGY

Office of
Science



ESS-DIVE team has expanded



Environmental Scientists

Data Scientists

Software Engineers

Digital Librarians

Three Pronged Approach to Repository



User Services



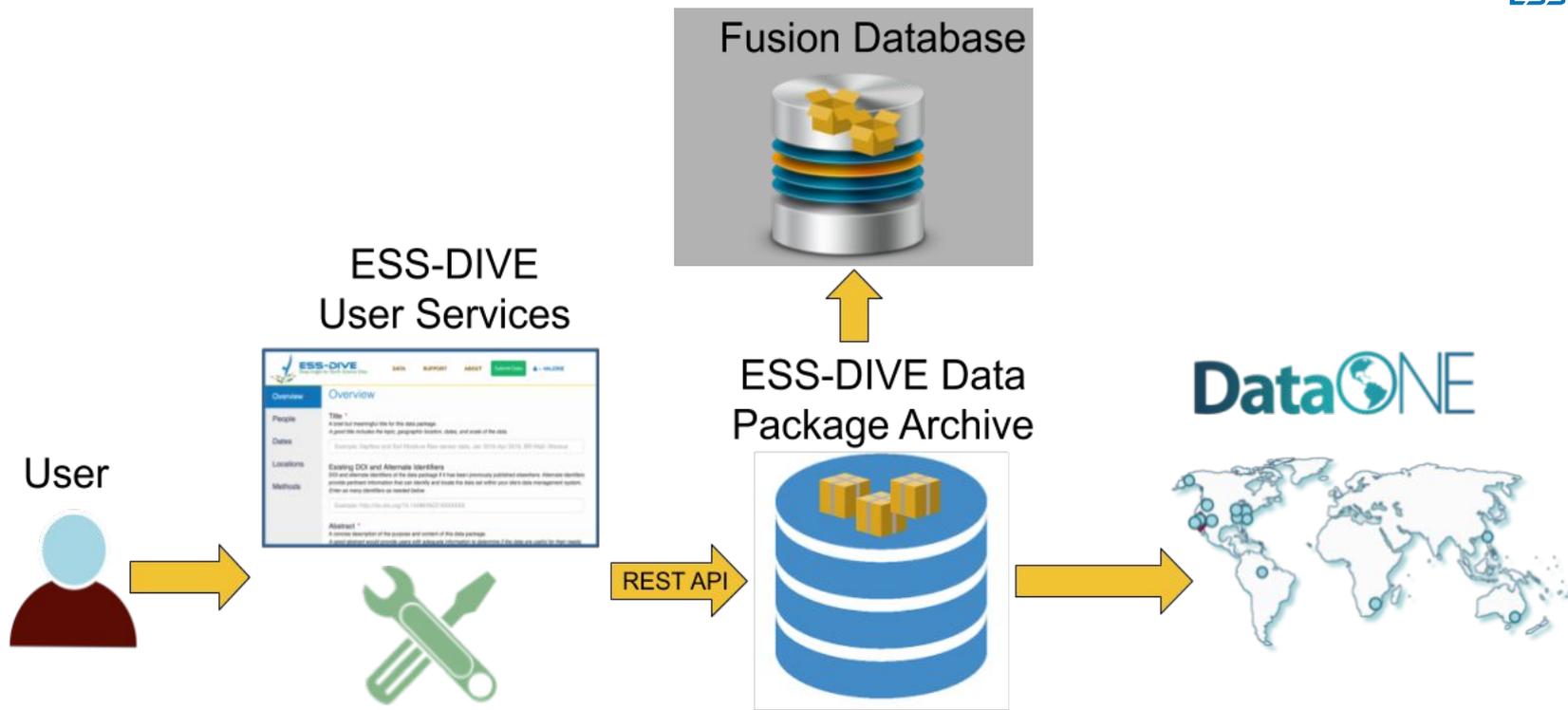
Data Repository



Community Engagement



ESS-DIVE Archiving Features



ESS-DIVE Implementation Timeline



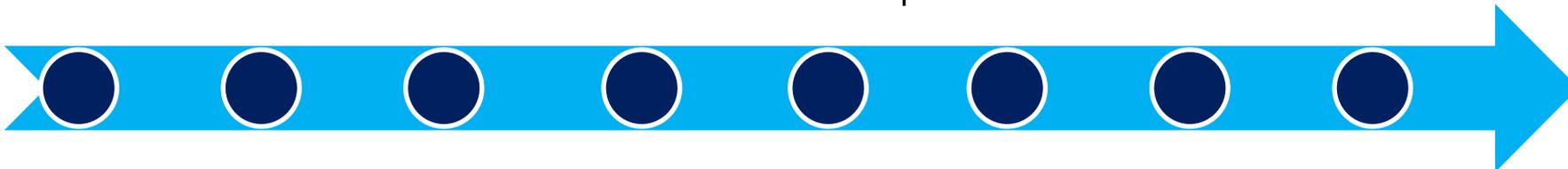
July 2017
Project Start

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

May-Dec 2017

Jan - Jun 2018

Jul - Dec 2018

Jan - Jun 2019

ESS CI/PI Meeting

ORNL/
OSTI Visit

SLAC SFA/
Stanford Visit

Package metadata

ESS CI/PI Meeting

PNNL/EMSL
/ARM Visit

Pre-launch
APB meeting

2nd APB
meeting

Package API
Demo

PNNL/
EMSL Visit

API Preview

Community
Funds

LLNL/
ESGF Visit

ORNL/
OSTI Visit

SLAC SFA/
Stanford Visit

4th APB
meeting

ESS CI/PI Meeting

Key: ESS Activities

ESS CI Working Groups

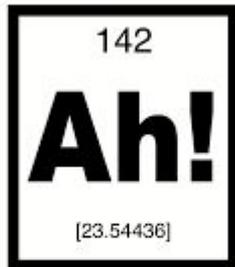
Partner Site Visits

Archive Partnership Board

Monthly Webinars

Visits to conferences, meetings, workshops not shown

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



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/XXXXXXX>

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 us

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 sensor was installed. No data processing or QA/QC was done on the raw data packages. Proc

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.

Automated metadata checks in the web form and API



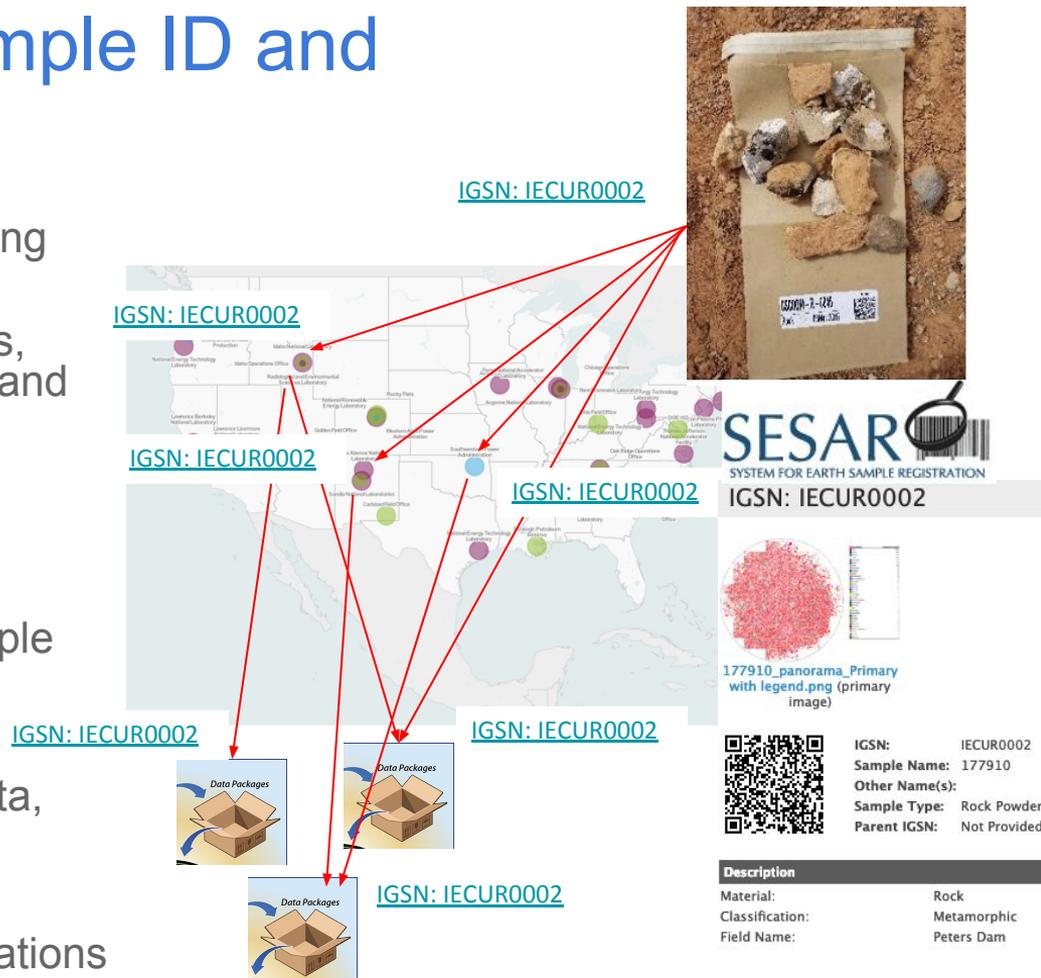
Community Need: Sample ID and Tracking

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



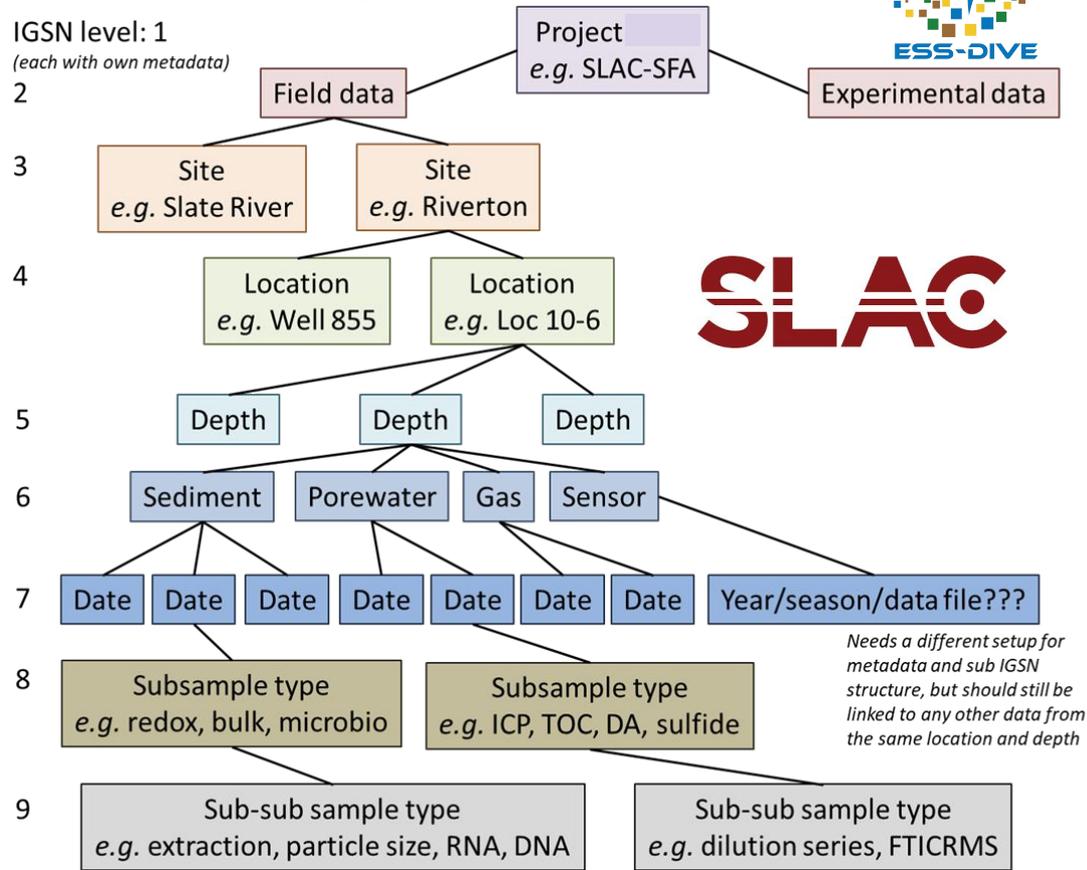
Pilot IGSN and Sample Tracking



SLAC-SFA, SBR SFAs,
WHONDRS

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

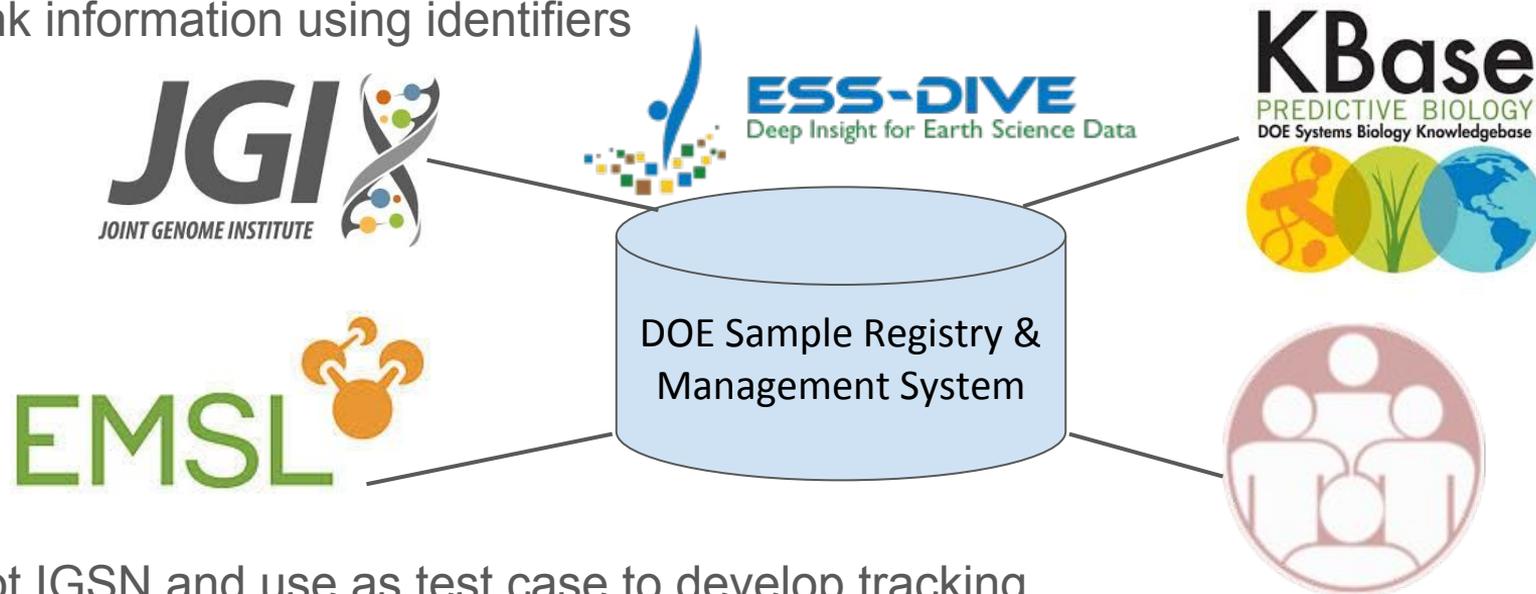
Develop workflows: Fit IGSN and metadata collection into planned field and lab workflows



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

A screenshot of the ESS-DIVE Data Portal website. The header includes the ESS-DIVE logo, navigation links for DATA, SUPPORT, ABOUT, and a Submit Data button, along with a Sign in with Orcid button. The main content area is divided into a search and filter sidebar on the left, a central list of datasets, and a map on the right. The dataset list shows entries with titles, authors, and DOIs, such as 'Time domain-induced polarization geophysical data collected in the Rifle Floodplain, Colorado, USA' and 'Transport and humification of dissolved organic matter within a semi-arid floodplain: Dataset'. The map on the right shows a world map with numbered markers (1, 2, 3) indicating the geographic locations of the datasets.

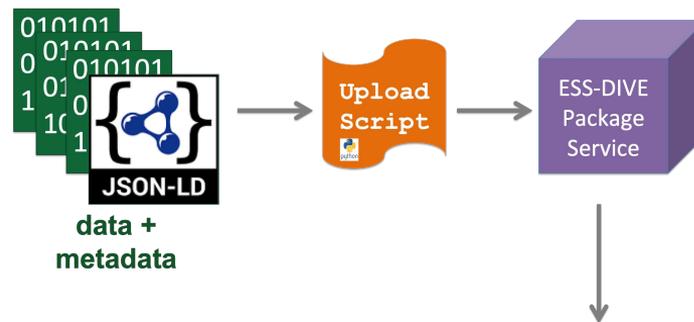
User Services

Community
Driven

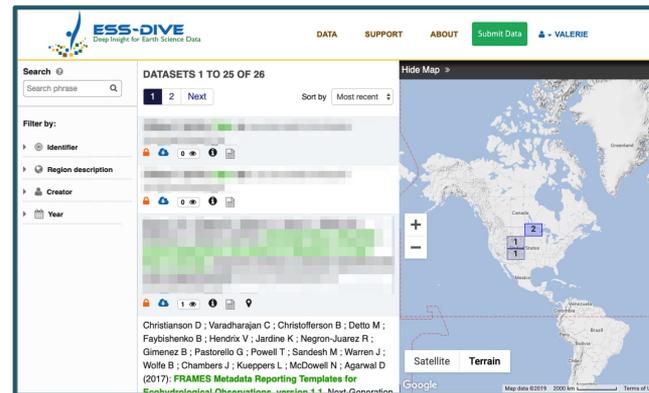


Package Service API

<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.



ESS-DIVE Package Service 1.0.0 GA33

[/schema.json](#)

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

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

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

This documentation details the E... follows:

GET /packages/{identifier} Get a data package by its identifier

GET /packages List data packages

POST /packages Submit data package metadata

Schemas

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

```

@type string
const: Dataset
default: Dataset
alternateName > {...}
name* string
minLength: 1
(Title) A brief but meaningful title for this data package.
description* > {...}
creator* > {...}
datePublished* string
pattern: ^([012]\d{3})(--([01-9]|[10-2])--([01-9]|[12]\d|3[01]))?S
(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 by a four digit year (YYYY) or a full date in the ISO format (YYYY-

```



User Services

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

Why GitHub? ▾ Enterprise Explore ▾ Marketplace Pricing ▾

Sign in Sign up

ESS-DIVE

Environmental Systems Science Data Infrastructure for a Virtual Ecosystem
Berkeley, CA USA <http://ess-dive.lbl.gov> ess-dive-support@lbl.gov

Repositories 1
People 0
Projects 0

Type: All ▾
Language

[essdive-package-service-examples](#)

Coding examples for testing on ESS-DIVE Sandbox p...

- Java Updated 7 days ago

ESS-DIVE Package Service Tutorial
Getting started with ESS-DIVE Package Service 1.0.0
March 2019

ESS-DIVE Package Service Tutorial

Getting started with ESS-DIVE Package Service 1.0.0

The ESS-DIVE Package Service is a service that enables projects to programmatically store data packages with ESS-DIVE. This is an alternative to using the [ESS-DIVE web portal](#) form for data uploads. This service encodes metadata using the [JSON-LD](#) specification. JSON-LD is a schema to encode linked Data using JSON, and in the future will be [used by Google](#) to index metadata for searches. The use of the standardized JSON-LD schema will dramatically increase the visibility of data packages, and also enable projects to create one-time code that can be reused for periodic uploads of data packages to ESS-DIVE.

The ESS-DIVE Package service, allows you to test submissions of [JSON-LD](#) data package *metadata* to ESS-DIVE's sandbox instance, to check whether metadata curated by projects are mapped correctly onto ESS-DIVE's data package metadata schema. Data package metadata refers to the top level metadata that enables a data package to be "discoverable" in search results. Examples of top-level metadata include the title, abstract, authors, variables and keywords. Other file-level metadata, such as those that describe the data file structure or variables are not included in this service.

You can get access to functioning coding examples on ESS-DIVE package service repository.

Provide feedback on this service to ess-dive-support@lbl.gov.

Get Access	2
Preparation	2
Coding Examples	3
R	3

© 2019 GitHub, Inc. [Terms](#) [Privacy](#) [Security](#) [Status](#)

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.

A screenshot of the SASAP web portal interface. The page title is "SASAP State of Alaska's Salmon and People". The navigation bar includes "About", "Data", "Metrics", and "People". The main content area lists project leads and coordinators, including Courtney Carothers, Michael Jones, Coowe Walker, Dr. Ian Dutton, Taylor Brelsford, James Fall, Ryan King, Mark Rains, Rachel Donkersloot, Steve Langdon, Dr. Frank Davis, Jessica Black, Robert W. Campbell, Kristin B. Gorman, and Bert Lewis. The NCEAS logo and name are visible at the bottom right of the screenshot.

knrb ABOUT DATA SUBMIT TOOLS Jump to: DOI or ID Go SIGN IN

SASAP State of Alaska's Salmon and People

knrb ABOUT DATA SUBMIT TOOLS Jump to: DOI or ID Go SIGN IN

SASAP State of Alaska's Salmon and People

About Data Metrics People

Courtney Carothers (University of Alaska)
Lead, Socio-Eco
ccarothers@alaska.edu

Michael Jones (Michigan State University)
Lead, Using participatory modeling to empower community engagement in salmon science
jonesm30@anr.msu.edu

Coowe Walker (University of Alaska)
Lead, Integrated Watershed Management for Salmon in Kenai Lowlands
cmwalker9@alaska.edu

Dr. Ian Dutton (NII)
Co-Principal Investigator/Alaska Coordinator
ian@nautilusii.com / (907)280-8923

Taylor Brelsford
Co-Lead, Gov
brelsfot@alaska.edu

James Fall (Alaska Department of Fish & Game)
Co-Lead, Gov
jim.fall@alaska.gov

Ryan King (Baylor University)
Co-Lead, Integrated WaterShed Management for
Ryan_S_King@baylor.edu

Mark Rains (University of South Florida)
Co-Lead, Integrated Watershed Management for S
mrains@usf.edu

Rachel Donkersloot (Alaska Marine Conservation Council)
Lead, Well-Being and Alaska Salmon Systems
rachel@akmarine.org

Steve Langdon (University of Alaska)
Lead, Governance
smlandgdon@alaska.edu

Dr. Frank Davis (NCEAS/UCSB)
Lead Principal Investigator
frank.davis@nceas.ucsb.edu / (805)893-2500

Jessica Black (University of Alaska)
Co-Lead, Socio-Eco & Well-Being and Alaska Salmon Systems
jblack@alaska.edu

Robert W. Campbell (Prince William Sound Science Center)
Co-Lead, Ocean Climate Interactions with At-Sea Salmon Competition

Kristin B. Gorman (Prince William Sound Science Center)
Co-Lead, Ocean Climate Interactions with At-Sea Salmon Competition
kgorman@pwsc.org

Bert Lewis (Alaska Department of Fish & Game)

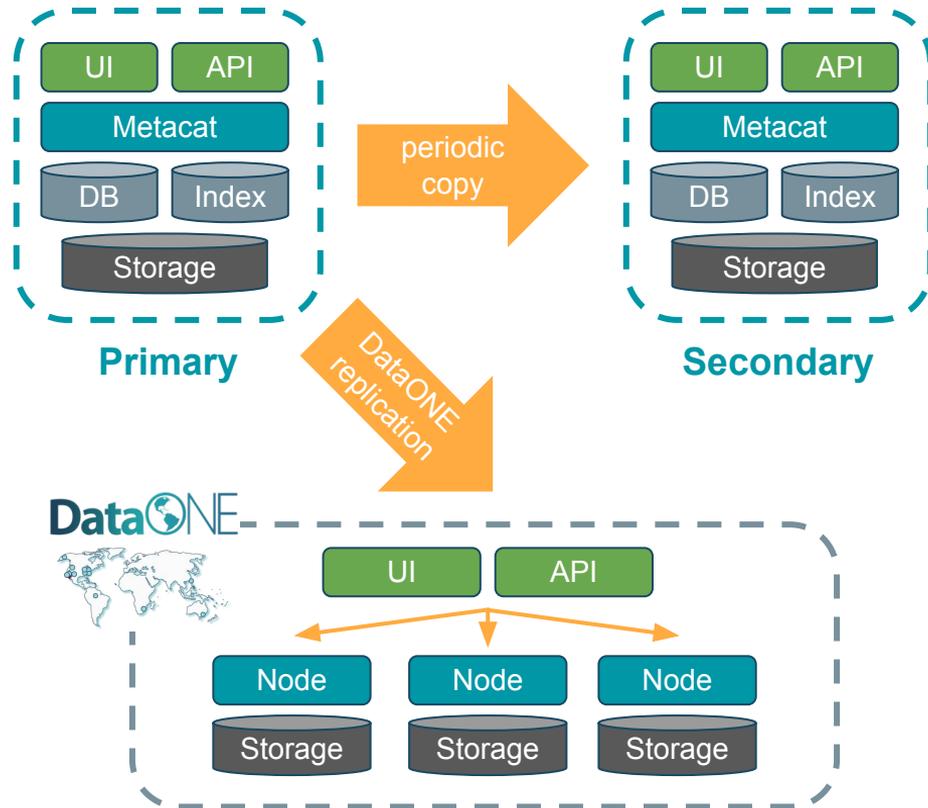
NCEAS
National Center for Ecological Analysis and Synthesis

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