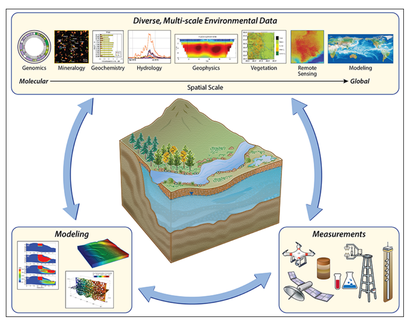
Environmental System Science Data Infrastructure for a Virtual Ecosystem (ESS-DIVE)

Valerie Hendrix (LBNL/CRD), Deb Agarwal (LBNL/CRD), Charuleka Varadharajan (LBNL/EESA), Shreyas Cholia (LBNL/CRD/NERSC), Cory Snavely (NERSC), Fianna O'Brien (LBNL/CRD), Joan Damerow (LBNL/EESA), Hesham Elbashandy (LBNL/CRD), Chris Jones (NCEAS), Matt Jones (NCEAS), Sara Studwell (OSTI), Crystal Sherline (OSTI) and Karen Whitenack (LBNL/CRD)

ESS-DIVE Vision

*ESS-DIVE addresses an aspect of tomorrow's **superfacility** by enabling data archiving, reuse, search and lifecycle management.*

Long-term Vision



Data Diversity

ESS-DIVE's long-term vision is to be a **data archive that enables science** through community engagement.

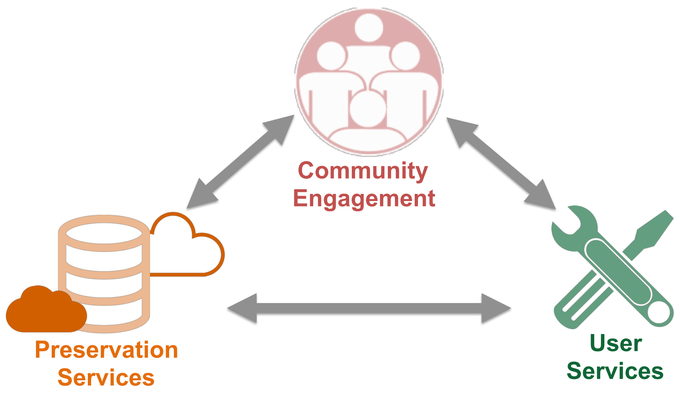
- Enable projects to archive with ESS-DIVE starting from data collection.
- Implementation of FAIR principles, data standards and DOIs issued for all data packages.
- Users able to search and retrieve data.
- Extensive search and visualization capabilities.

Approach

Our team is an **interdisciplinary collaboration**:

- Data scientists and software engineers who design and implement the system.
- Digital librarians who develop practices and provide guidance for long-term data preservation, publication and reuse.
- Environmental scientists to engage the community and provide use cases for archive design.

ESS-DIVE Focus Areas



Community Engagement partners with the scientific community to build a standards-based data archive.

User Services provides key front-end capabilities for users such as data packaging services, search, and integrated access to data.

Preservation Services manages long-term data package preservation and availability.

About ESS-DIVE

Information: https://ess-dive.lbl.gov

Data Archive: https://data.ess-dive.lbl.gov

	DOE Projects	Storage Capacity	Public Datasets	Temporal Coverage
Year 1	4*	20 TB*	240	1841 to 2018

* growth expected in the next several years based on community need

User Services and Support

Researchers directly interact with the data preservation infrastructure through the User Services interfaces.

Data Management	Publication workflow	REST Services	DataONE Federation
• Search • Data Upload/ Download • Metadata Editing • Citations and Metrics	• DOI Generation • Metadata and Data Quality Check	• JSON metadata specification • Data package upload	• Single Sign On via ORCID • Data Replication • Data Auditing

Data Management provided by web portal is built on Metacat and Metacat UI provided and maintained by DataONE and NCEAS

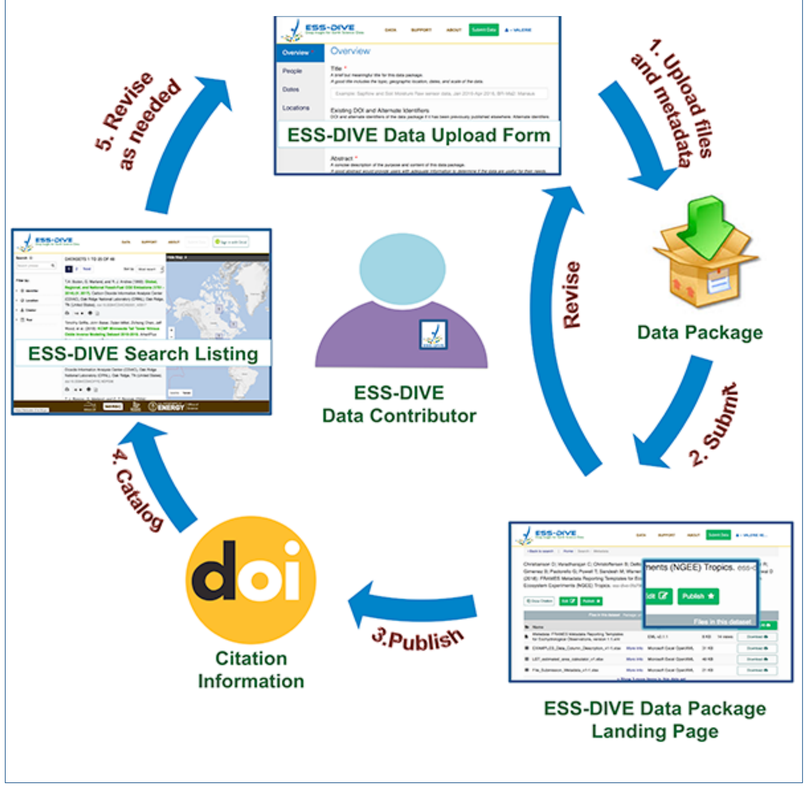
Publication Workflow enables scientists to create data packages privately and then published. The data package metadata and data are validated and curated before publication.

REST Services for uploading data packages which makes use of the latest metadata standards (JSON-LD).

DataONE Federation provides Single Sign On via Orcid, federated search across DataONE, data replication and auditing.

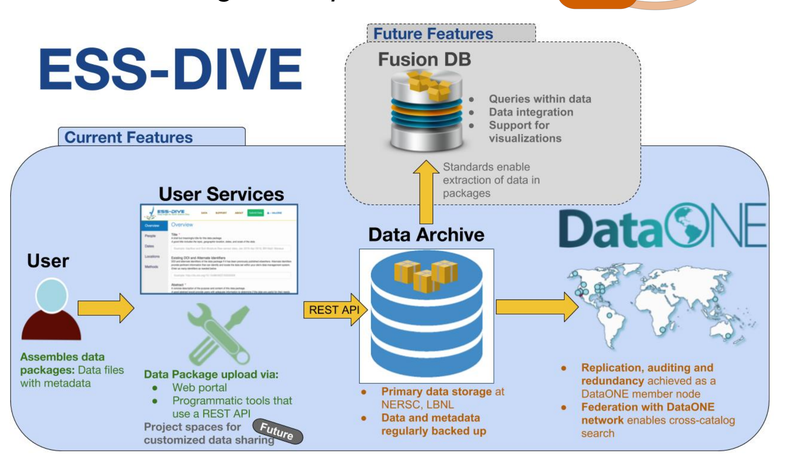
Publication Workflow

ESS-DIVE's publication life cycle allows users to iteratively curate, package, and update their data and metadata.



Data Preservation

Preservation layer consists of software systems, storage, and data management processes.



Microservice Architecture at NERSC

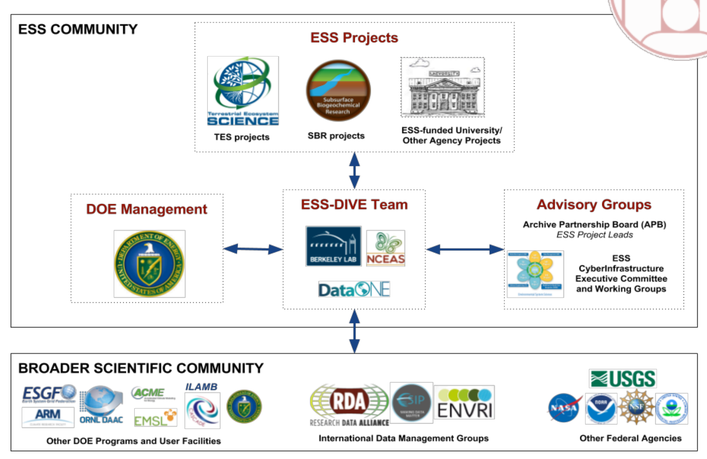
- Built on NERSC's *Spin* platform; a *Docker infrastructure* for web and network services.
- *Integrated with NERSC* storage systems, HPC, and network.
- Multiple instances *easy to "spin up"* using isolated containers.

Fusion Database will extend metadata search capabilities

- **Automatic Parsers** parse data files, and store the results.
- **Integrated Search** index/join scientific data with the metadata using a scalable and flexible database.
- **Custom Data Download** sub-set one or several archived data packages that meet search criteria specified by the user.
- **Cross-archive Search** enable search across data and affiliated archives.

Building what the Community Wants

Input from the scientific community drives our feature pipeline.



July 2017 Project start
Sept 2017 Host the legacy ESS archive (CDIAC)
March 2018 1st Archive Partnership Board Meeting
April/May 2018 Accepting new data
July 2018 2nd Archive Partnership Board Meeting
Aug 2018 DataONE Member
Dec 2018 BETA version of data package service API

Community visits to ORNL, OSTI, SLAC, Stanford, PNNL and ESS Cyber-infrastructure Working Group

***Funded by:* DOE BER Data Management program within the Climate and Environmental Science Division**

Partners:

National Center for Ecological Analysis and Synthesis *(NCEAS) - MetacatUI and DataONE*

DOE Office of Scientific and Technical Information *(OSTI) – Transition DOIs from prior archive*

DataCite *– Consultations and transfer of DOIs*

National Energy Research Scientific Computing facility *(NERSC) – Hosting archive*

CDIAC *– Carbon Data Information Analysis Center*

For more information: ess-dive@lbl.gov

Varadharajan, C. (2019, January 08). Launching an Accessible Archive of Environmental Data. Retrieved from https://eos.org/project-updates/launching-an-accessible-archive-of-environmental-data