ESS-DIVE Community Priorities

April 22, 2019

ESS-DIVE
Deep Insight for Earth Science Data

U.S. DEPARTMENT OF ENERGY | Office of Science

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB
Lawrence Berkeley National Laboratory

NERSC

NCEAS

Community Partnership to Build Capabilities

- Upto **$1 M of community funds** are available for projects to partner with ESS-DIVE to **build features or implement standards**
- Funds allocated to **community priorities** for ESS-DIVE
- Projects/Labs encouraged to form **collaborative teams** to facilitate community input
- **Deliverables** will be associated for each award

Project Timeline

Implementation

2017 Jul – Project start

2017 Sep – Old archive transferred

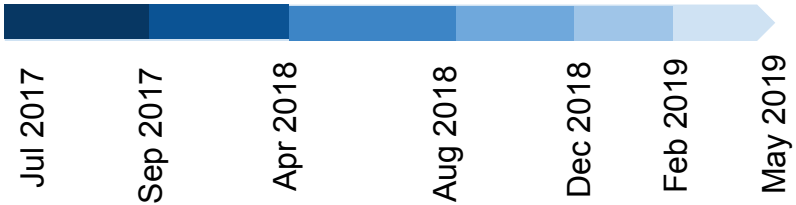
2018 Apr – ESS-DIVE live

2018 Aug – Join DataONE

2018 Dec – Prototype API

2019 Feb – ESS-DIVE/NCEAS Meeting

2019 May – Data upload API released



Community engagement

2017 May – ESS CI and PI Meeting

2017 Jul – Visit to ORNL and OSTI

2017 Dec – Visit to SLAC/Stanford

2018 Mar – Archive Partnership Board Meeting

2018 May – ESS CI and PI Meeting

2018 Jul – Visit to PNNL

2018 Jul – Archive Partnership Board Meeting

2018 Nov – Archive Partnership Board Meeting

2019 Dec - Monthly community webinar kickoff

2019 Jan – Visit to PNNL

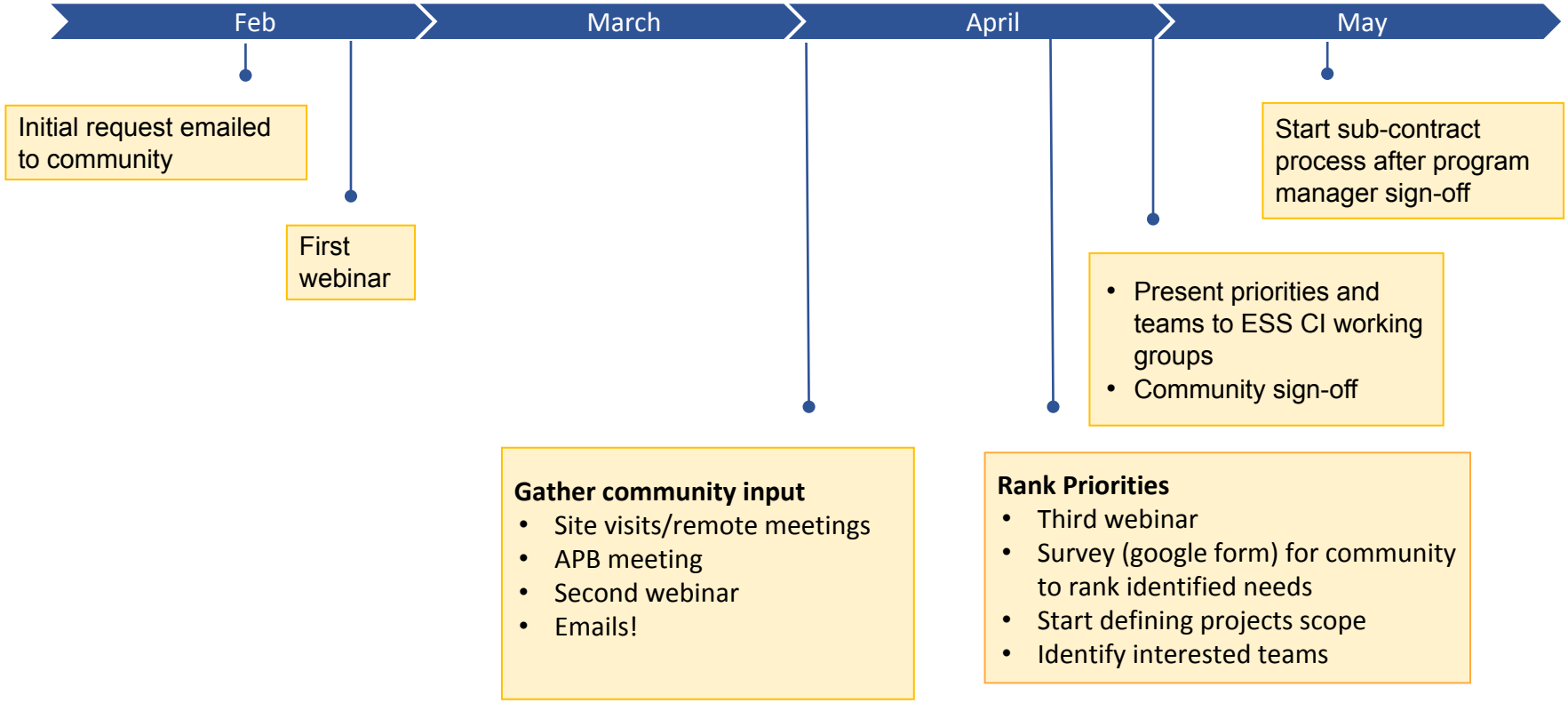
2019 Mar/Apr - Visits to ORNL, LLNL, SLAC, community webinars

2019 May – ESS CI and PI Meeting

\+ Many conferences, workshops etc.

U.S. DEPARTMENT OF ENERGY | Office of Science | BERKELEY LAB | NERSC | NCEAS

Process for allocating community funds



Summary: 2019 ESS-DIVE Roadmap

- API for data package submission

- Sample ID and metadata research
- Webinars, meetings and surveys to identify community priorities

- API for data package access and download

- Finalize community priorities
- ESS PI/CI Meeting training and outreach
- Work with interested projects on identifying sample tracking needs
- IGSN sample registration testing

On current roadmap

- Project spaces: ESS PI custom data package admin support
- Data usage reporting
- Large file data upload
- Automated data quality reports
- Implementation of file-level metadata standards/fusion database
- **Other Community-identified priorities**
- Support for globally unique sample IDs
- Links to external archives
- Connection with EMSL/KBase

- File-level metadata for select data types
- Sample ID and metadata recommendations
- Ongoing monthly webinars, tutorials and site visits
- Data Management Training

Key:

Infrastructure Development

Standards & Engagement*

ESS-DIVE Roadmap Planning: Items to Consider

PROJECT SPACES

- Admin Support
- Metrics and data usage notifications

DATA INGEST/EXPORT IMPROVEMENTS

- Utilizing the REST API to upload data
- Other Bulk Data Transfer (Globus etc.)
- DOI harvest/Link to data on other archives

STANDARDS DEVELOPMENT

- Sample IDs and Tracking, Sample Metadata
- File-level Metadata
- netCDF file representations

CONNECTION WITH DOE FACILITIES

- EMSL, KBase, ARM, JGI etc.

DOE MODEL DATA WORKSHOP

HIERARCHICAL DATA SUPPORT

- Ingest and API support, hierarchical representation, metadata schema

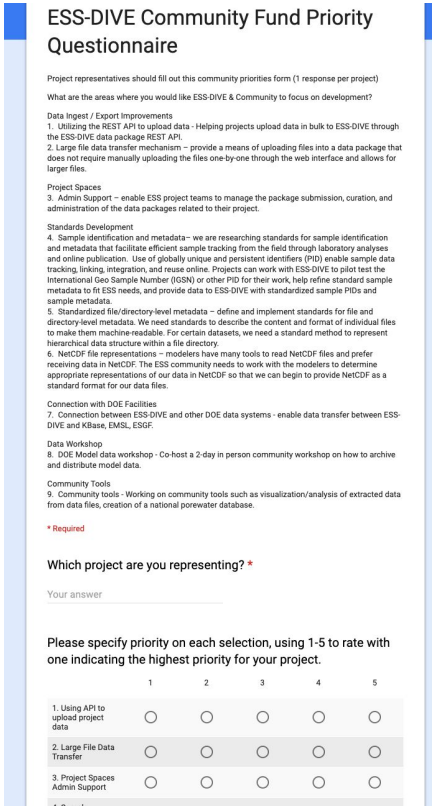
FUSION DATABASE

- Faceted search for properties and generalized search across datasets
- Support for data visualization
- Depends on community development and adoption of data standards

Survey: Google form to rank priorities

https://docs.google.com/forms/d/e/1FAIpQLSfwminMNXcqdNt0A9aTJL7WCzX8i55y1si2bOUfJJ-iIBNvNQ/viewform

- Purpose
  - Rank community priorities
  - Find who is willing to work with ESS-DIVE on building/implementing priorities
- Ideally 1 per project
  - Enter N/A if not applicable
- Fill by Saturday, April 27th
- Present responses at the ESS Cyberinfrastructure Working Group Meeting on April 29th



ESS-DIVE Community Fund Priority Questionnaire

Project representatives should fill out this community priorities form (1 response per project)

What are the areas where you would like ESS-DIVE & Community to focus on development?

Data Ingest / Export Improvements
1. Utilizing the REST API to upload data - Helping projects upload data in bulk to ESS-DIVE through the ESS-DIVE data package REST API.
2. Large file data transfer mechanism – provide a means of uploading files into a data package that does not require manually uploading the files one-by-one through the web interface and allows for larger files.

Project Spaces
3. Admin Support – enable ESS project teams to manage the package submission, curation, and administration of the data packages related to their project.

Standards Development
4. Sample identification and metadata– we are researching standards for sample identification and metadata that facilitate efficient sample tracking from the field through laboratory analyses and online publication. Use of globally unique and persistent identifiers (PID) enable sample data tracking, linking, integration, and reuse online. Projects can work with ESS-DIVE to pilot test the International Geo Sample Number (IGSN) or other PID for their work, help refine standard sample metadata to fit ESS needs, and provide data to ESS-DIVE with standardized sample PIDs and sample metadata.
5. Standardized file/directory-level metadata – define and implement standards for file and directory-level metadata. We need standards to describe the content and format of individual files to make them machine-readable. For certain datasets, we need a standard method to represent hierarchical data structure within a file directory.
6. NetCDF file representations – modelers have many tools to read NetCDF files and prefer receiving data in NetCDF. The ESS community needs to work with the modelers to determine appropriate representations of our data in NetCDF so that we can begin to provide NetCDF as a standard format for our data files.

Connection with DOE Facilities
7. Connection between ESS-DIVE and other DOE data systems - enable data transfer between ESS-DIVE and KBase, EMSL, ESGF.

Data Workshop
8. DOE Model data workshop - Co-host a 2-day in person community workshop on how to archive and distribute model data.

Community Tools
9. Community tools - Working on community tools such as visualization/analysis of extracted data from data files, creation of a national porewater database.

* Required

Which project are you representing? *

Your answer

Please specify priority on each selection, using 1-5 to rate with one indicating the highest priority for your project.

	1	2	3	4	5
1. Using API to upload project data	○	○	○	○	○
2. Large File Data Transfer	○	○	○	○	○
3. Project Spaces Admin Support	○	○	○	○	○

Topic of choice

Project Spaces: Administrative Management

Project Spaces: Initially project management interface for use by ESS PIs and designates.

- Allow PIs to manage the list of people authorized to upload data
- Allow designates to:
  - Upload data on behalf of project members
  - Manage data packages for their project
  - Manage the data package publication process for their project.
- Contains metrics and notifications on data usage

Data Ingest/Export Improvements

- **Using the REST API:** Enabling projects to utilize the REST API to do a one-time bulk upload of their data to ESS-DIVE
- **Alternate Data Transfer Mechanism:** Scalable user-facing ingest using large data transfer tool (e.g.Globus).
- **Data Citation Harvesting:** Import data package by harvesting metadata for a given DOI
- **Link to other archives:** Enabling connections to data that exists on other recognized repositories without transferring data over

ESS-DIVE Package Service API: Data Ingest

*The **ESS-DIVE Package Service** is a more general interface than the ESS-DIVE repository. Via this service, organizations can **write code to store data packages** and then **reuse** the code to upload other data packages in the same or different repositories.*

JSON for Linked Data (JSON-LD)

JSON-LD (JavaScript Object Notation for Linked Data), is a method of encoding Linked Data using JSON (see https://json-ld.org/)

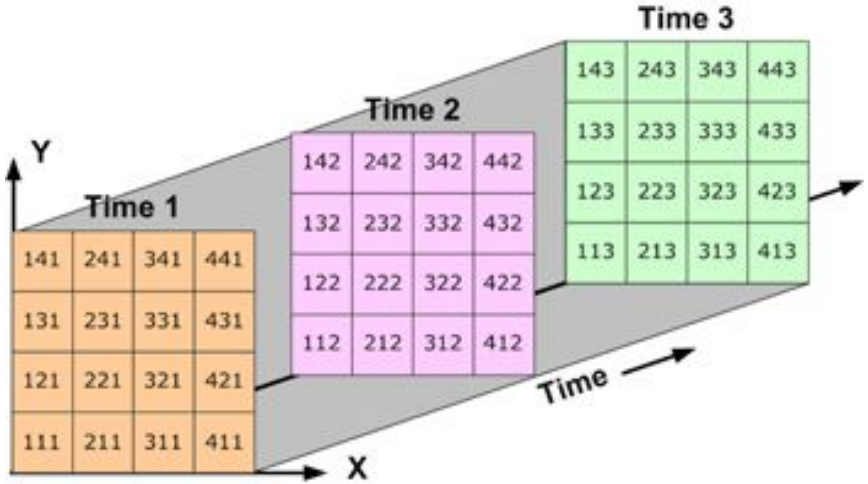
- The ESS-DIVE metadata schema is a restricted subset of https://schema.org/Dataset specification
- This covers all of the fields that ESS-DIVE collects from users (see ESS-DIVE JSON-LD Schema Proposal)
- JSON-LD is recommended by DataCite for package submission.
- JSON-LD has broad tool support and can be embedded in landing pages for harvesting by DataCite and indexing by Google.

File-Level Metadata Standards

- File-level metadata standards that fit **diverse ESS data** and community needs.
- Evaluate the various formats in use by ESS projects and to work with the ESS community **to identify, adopt, and define standards** for the file-level metadata.
- **Variables** move down to file level with more specific information, making file level metadata more usable.
- Support for **automatic metadata extraction** directly from files

netCDF Standards

- Accepted self-describing format for scientific data
- Leverage existing tools – e.g. iLAMB, ORNL DAAC for automatically parsing netCDF files
- Positions ESS-DIVE to handle modeling data in the next phase



Standardize Sample Identification and Tracking

ESS-DIVE
U.S. DEPARTMENT OF ENERGY | Office of Science
NERSC
BERKELEY LAB
NCEAS

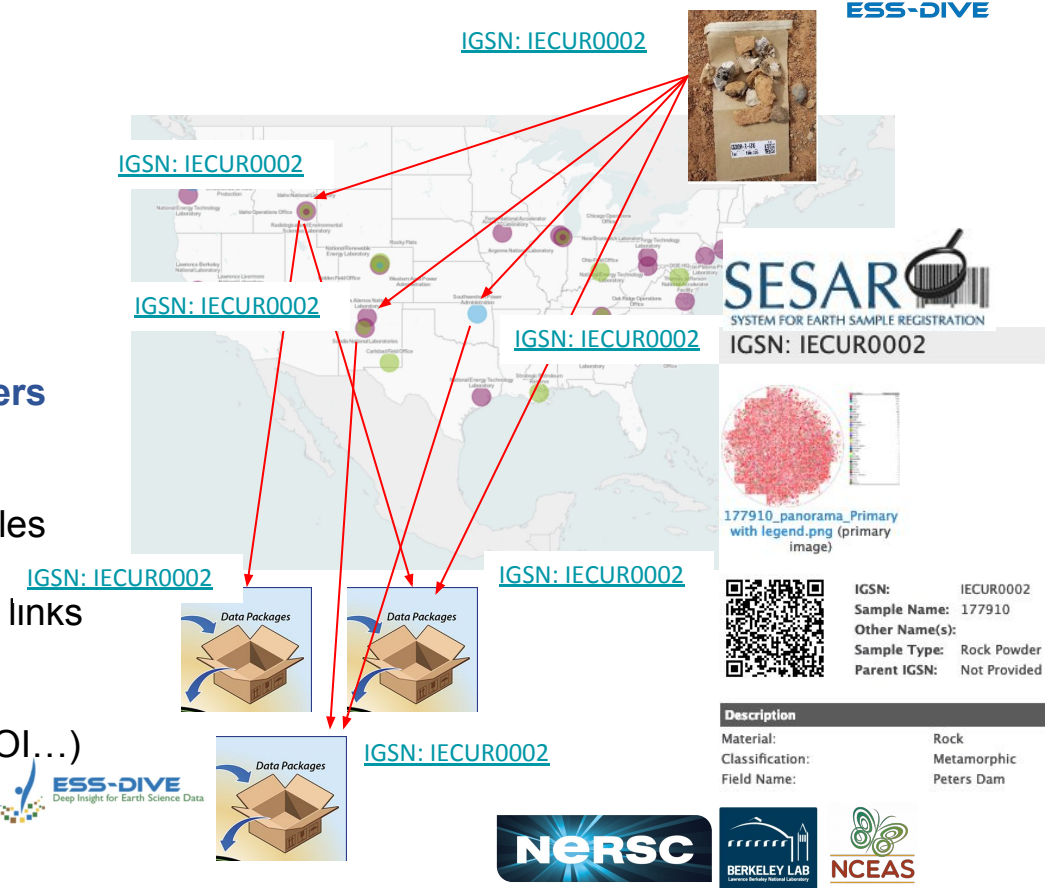
Sample Tracking and IGSN

Challenge: Tracking samples from field to dataset publication

- Need an efficient, practical, standardized sample tracking system for field, lab, and online
- Integrating data effectively online requires globally unique, persistent identifiers

Solution: International Geo Sample Numbers (IGSNs) for ESS samples

- Physical samples, sample feature (site, borehole), aggregate of samples, subsamples
- Example IGSN: IECUR0002
- Standardized sample metadata: templates, links to online metadata profiles
- Facilitate advanced searches
- Link to other important identifiers (IGSN, DOI…)



ESS-DIVE
Deep Insight for Earth Science Data

U.S. DEPARTMENT OF ENERGY | Office of Science

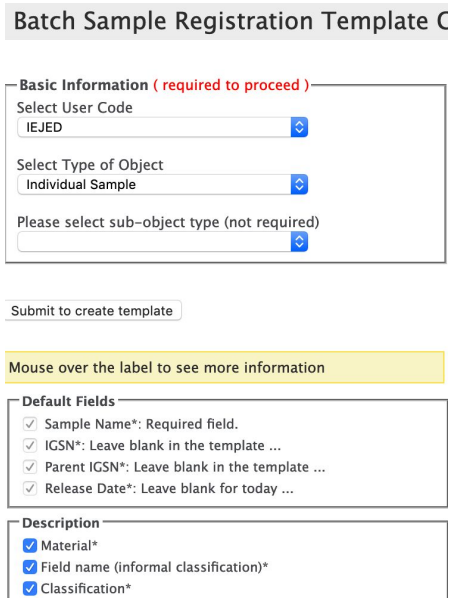
NERSC | BERKELEY LAB | NCEAS

Example Workflow:

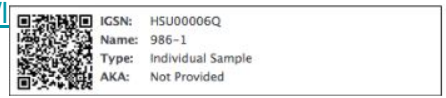
1. Login and select a user code
http://www.geosamples.org/getigsn

2. Template creator for customized excel template with appropriate metadata
https://app.geosamples.org/create_template.php

3. Batch upload basic metadata to register samples, get IGSNs

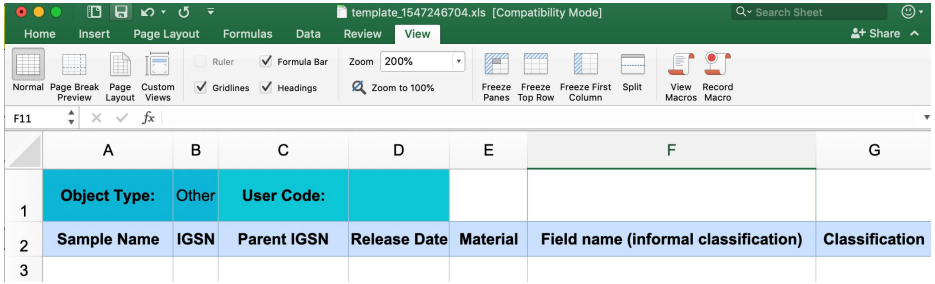


4. Print IGSN labels, using SESAR template
http://www.geosamples.org/help/l



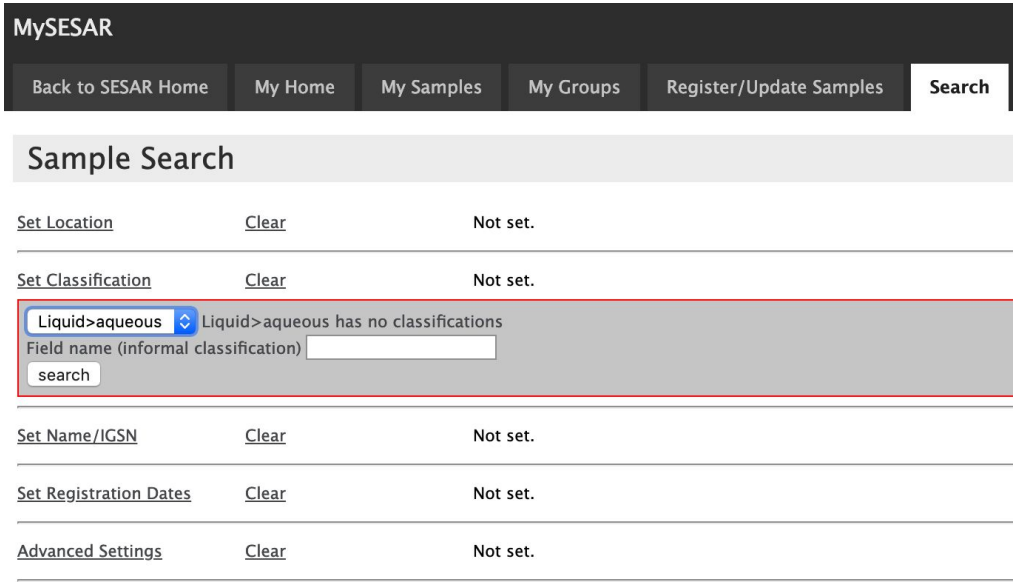
5. Collect samples and metadata

6. Batch upload completed metadata in the customized SESAR spreadsheet template



7. Manage and publish sample data
- IGSN used with all records involving sample data, processing, results
- Updates as needed in SESAR catalogue
- Submit datasets with IGSNs to ESS-DIVE

IGSN Sample Data Search and Linking



SESAR catalogs metadata profiles, and provides access via the Global Sample Search

Increase data discovery - links to current archive for data

- Does similar data exist?
- Find datasets for integration
- Find collaborators
- Grant proposals

IGSN is a “Related identifier” in DataCite metadata

Link samples to other identifiers: IGSN, publications (DOI), datasets (DOI), researchers (ORCID), sensors, funding (FundRef#)

Link to YouTube video presentation on IGSN

U.S. DEPARTMENT OF ENERGY | Office of Science
NERSC | BERKELEY LAB | NCEAS

Summary of Benefits

- Make process of naming and tracking samples easier
- Avoid ambiguity, track history of samples, online metadata catalogue
- Facilitate advanced data searches: integrate samples with certain attributes across datasets
- Cite and track data usage at the sample level
- Link samples to other important identifiers

ESS-DIVE

Work with project teams to implement IGSNs, workflow guides for optimized sample registration and tracking, feedback on the process

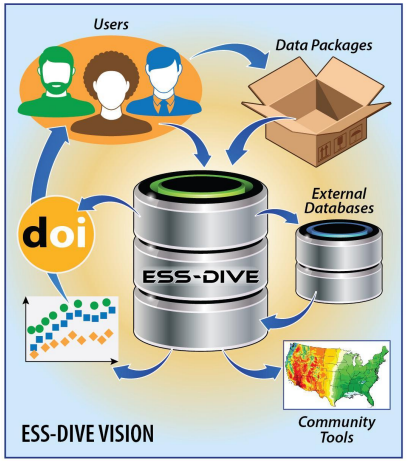
White Paper: Globally unique sample identifiers to support data management, reuse, and attribution

U.S. DEPARTMENT OF ENERGY | Office of Science
NERSC | BERKELEY LAB | NCEAS

Data Access: Hierarchical Data Package Support

- **Underlying data layout and metadata scheme:** Scheme should allow data packages with explicit hierarchical ("folder") layout
- **Ingest mechanism and API support:** Right now users are just bundling into a single data file e.g. tar or zip. Need to be able to preserve hierarchy in metacat
- **UI presentation and editing:** How do hierarchical packages get represented in MetacatUI, for both display and editing?

Fusion Database



Fusion Database

- Queries within data
- Data integration
- Support for visualizations

Standards enable extraction of data in packages

User

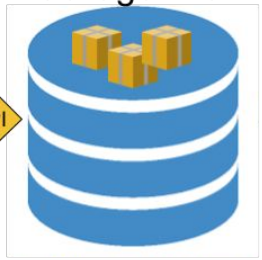
Assembles data packages:
Data files with metadata

ESS-DIVE User Services

- Data Package upload via:
  - User Interface
  - Programmatic tools that use a REST API
- Project spaces for customized data sharing

REST API

ESS-DIVE Data Package Archive



- Primary data storage at NERSC, LBNL
- Data and metadata regularly backed up

DataONE

- Replication, auditing and redundancy achieved as a DataONE member node
- Federation with DataONE network enables cross-catalog search

Fusion Database

Fusion Database for deeper data indexing and cross dataset comparisons

- Develop fusion DB capabilities through a NoSQL schema-free DB layer
- Support for faceted search for properties within the dataset
- Support for search across datasets
- Integration of external datasets and APIs