ESS-DIVE Community Priorities

March 25, 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

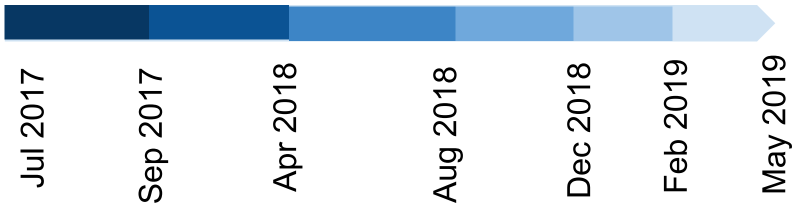
2018 Apr – ESS-DIVE live

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

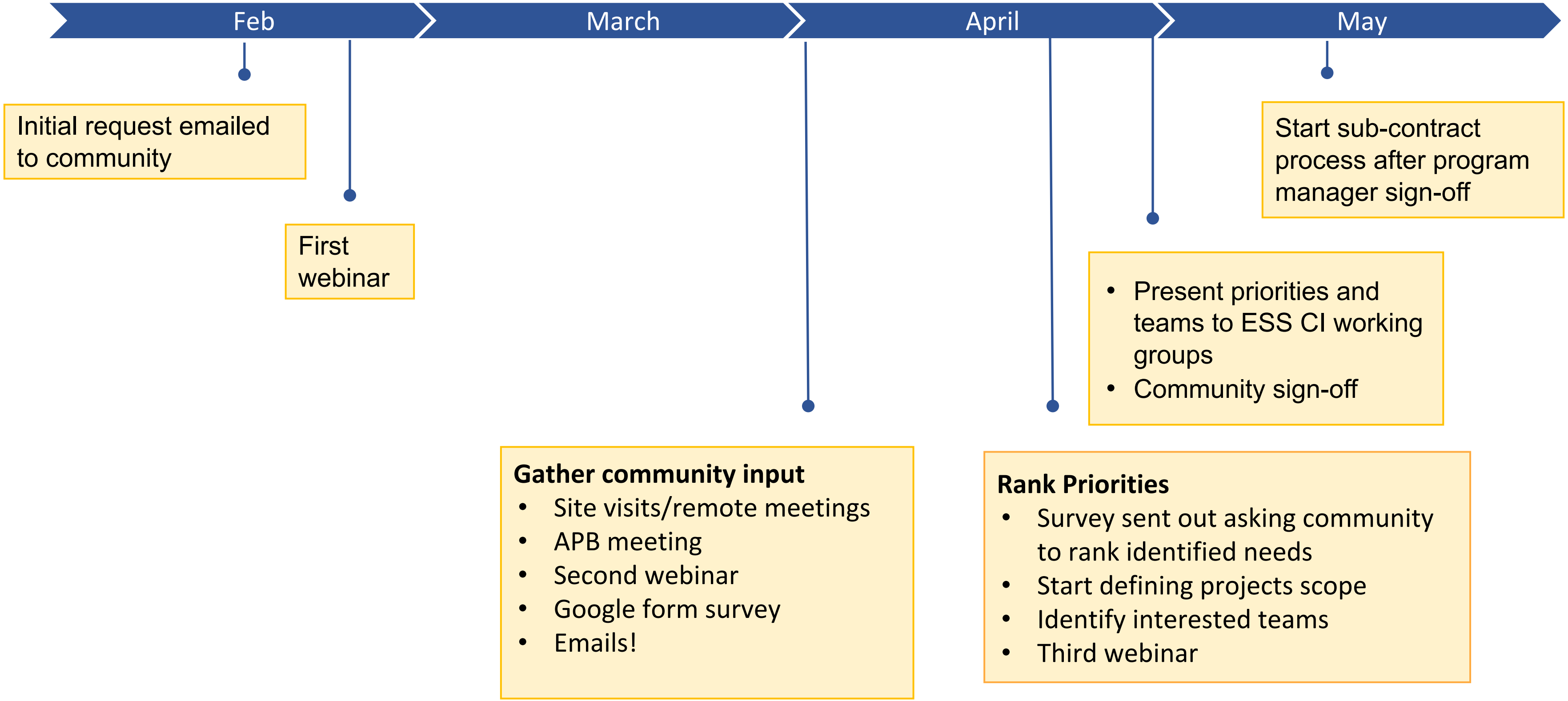
2019 Jan – Visit to PNNL

2019 Mar - Visits to ORNL, LLNL

2019 May – ESS CI and PI Meeting

+ Many conferences, workshops etc.

Process for allocating community funds



Feb

March

April

May

Initial request emailed to community

First webinar

Gather community input

- Site visits/remote meetings
- APB meeting
- Second webinar
- Google form survey
- Emails!

Rank Priorities

- Survey sent out asking community to rank identified needs
- Start defining projects scope
- Identify interested teams
- Third webinar

- Present priorities and teams to ESS CI working groups
- Community sign-off

Start sub-contract process after program manager sign-off

Summary: 2019 ESS-DIVE Roadmap

ESS-DIVE

Jan - March | April - June | July - Sept | Oct - Dec

Jan - March

- API for data package submission

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

April - June

- 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

July - Sept / Oct - Dec

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

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/)

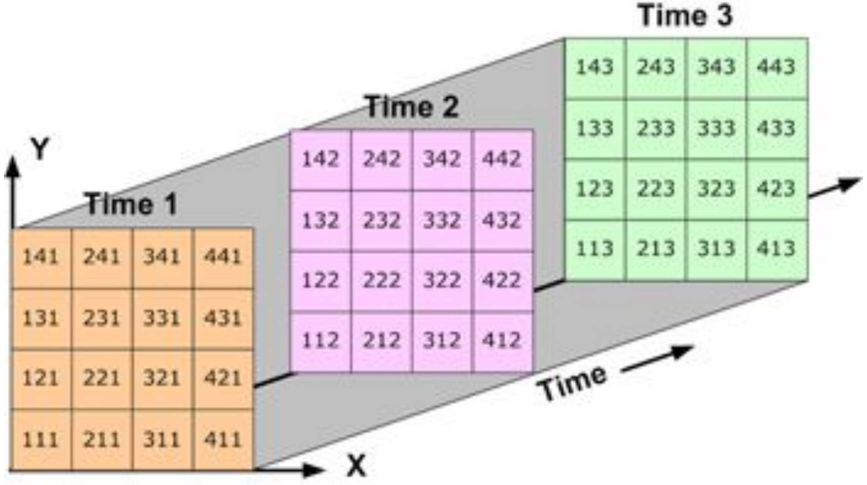
- The ESS-DIVE metadata schema is a restricted subset of https://schema.org/Dataset specification
- This covers all of the fields that ESS-DIVE collects from users (see ESS-DIVE JSON-LD Schema Proposal)
- JSON-LD is recommended by DataCite for package submission.
- JSON-LD has broad tool support and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



Time 1

141	241	341	441
131	231	331	431
121	221	321	421
111	211	311	411

Time 2

142	242	342	442
132	232	332	432
122	222	322	422
112	212	312	412

Time 3

143	243	343	443
133	233	333	433
123	223	323	423
113	213	313	413

Standardize Sample Identification and Tracking

Sample Tracking and IGSN

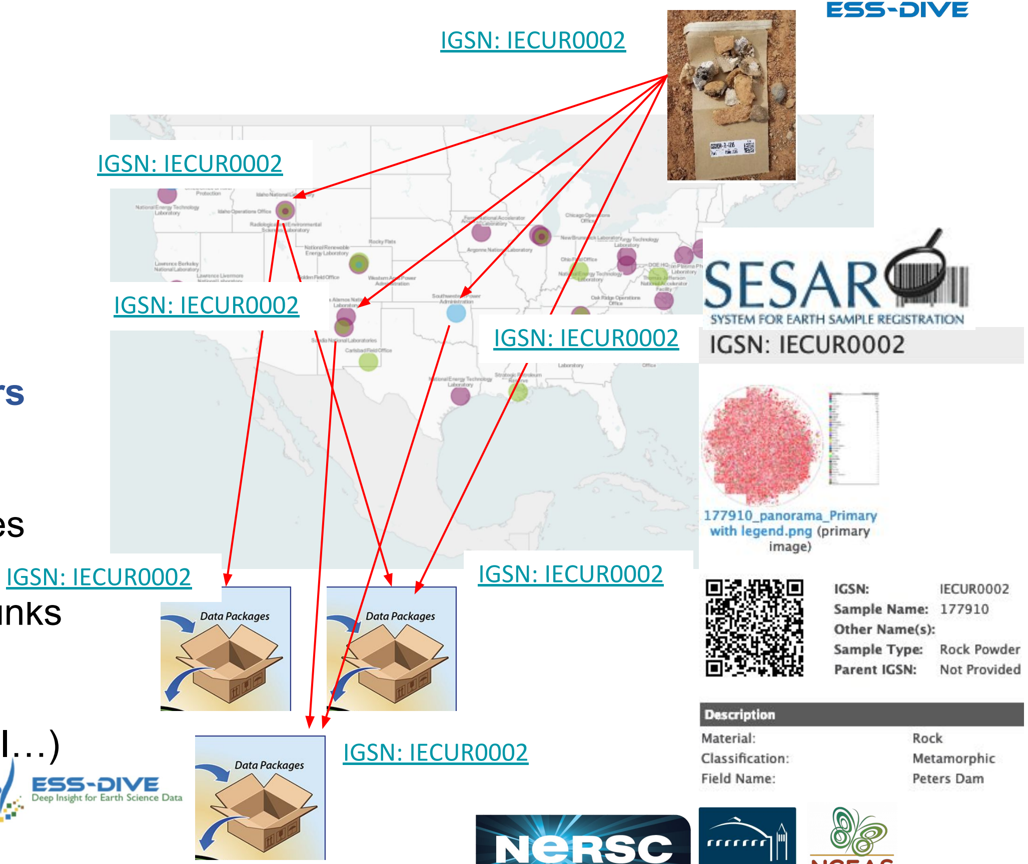
ESS-DIVE

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, lınks to online metadata profiles
- Facilitate advanced searches
- Link to other important identifiers (IGSN, DOI…)



IGSN: IECUR0002

177910_panorama_Primary with legend.png (primary image)

IGSN:	IECUR0002
Sample Name:	177910
Other Name(s):	
Sample Type:	Rock Powder
Parent IGSN:	Not Provided

Description

Material:	Rock
Classification:	Metamorphic
Field Name:	Peters Dam

ESS-DIVE Deep Insight for Earth Science Data

U.S. DEPARTMENT OF ENERGY | Office of Science

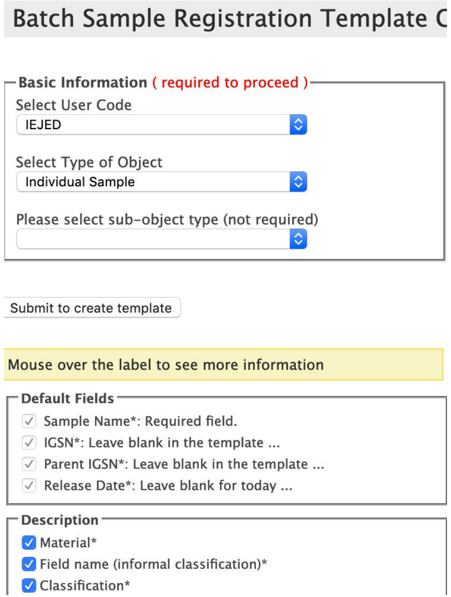
NERSC · BERKELEY LAB · NCEAS

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



Batch Sample Registration Template C

Basic Information (required to proceed)

Select User Code: IEJED

Select Type of Object: Individual Sample

Please select sub-object type (not required):

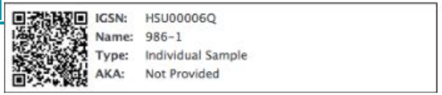
Submit to create template

Mouse over the label to see more information

Default Fields
- ✓ Sample Name*: Required field.
- ✓ IGSN*: Leave blank in the template ...
- ✓ Parent IGSN*: Leave blank in the template ...
- ✓ Release Date*: Leave blank for today ...

Description
- ☑ Material*
- ☑ Field name (informal classification)*
- ☑ Classification*

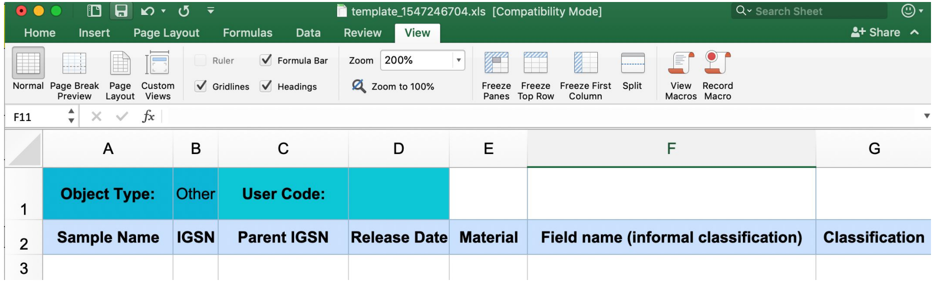
4. Print IGSN labels, using SESAR template
http://www.geosamples.org/help/l



IGSN: HSU00006Q
Name: 986-1
Type: Individual Sample
AKA: Not Provided

5. Collect samples and metadata

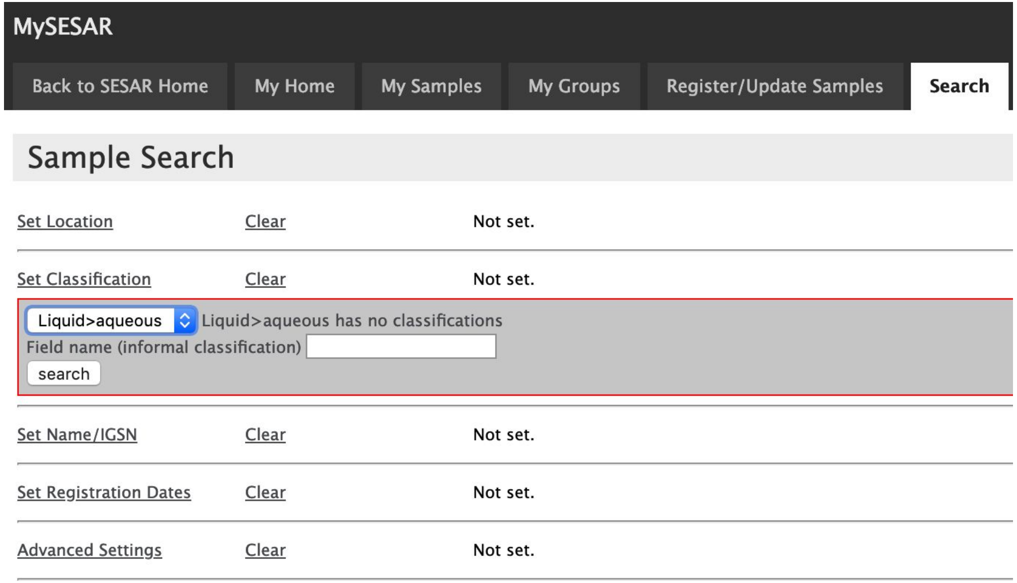
6. Batch upload completed metadata in the customized SESAR spreadsheet template



	A	B	C	D	E	F	G
1	Object Type:	Other	User Code:				
2	Sample Name	IGSN	Parent IGSN	Release Date	Material	Field name (informal classification)	Classification
3							

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



MySESAR

Back to SESAR Home | My Home | My Samples | My Groups | Register/Update Samples | Search

Sample Search

Set Location	Clear	Not set.
Set Classification	Clear	Not set.

Liquid>aqueous Liquid>aqueous has no classifications

Field name (informal classification)

search

Set Name/IGSN	Clear	Not set.
Set Registration Dates	Clear	Not set.
Advanced Settings	Clear	Not set.

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

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

Data Access: Hierarchical Data Package Support

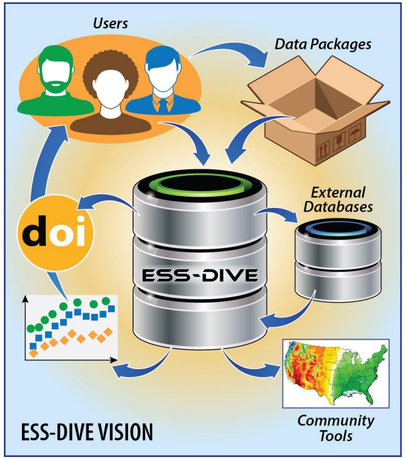
- **Underlying data layout and metadata scheme:** Scheme should allow data packages with explicit hierarchical ("folder") layout
- **Ingest mechanism and API support:** Right now users are just bundling into a single data file e.g. tar or zip. Need to be able to preserve hierarchy in metacat
- **UI presentation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

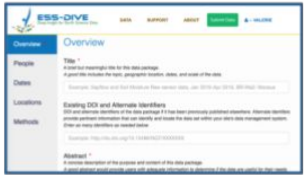
Standards enable extraction of data in packages



User

Assembles data packages:
Data files with metadata

ESS-DIVE User Services



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

Existing DOI and Alternate Identifiers

Abstract *
A concise description of the purpose and content of this data package

- 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