Standardize Sample Identification and Metadata

Joan E. Damerow, Charu Varadharajan

May 29, 2019

Digitizing and integrating species occurrence records from museum collections, citizen science, literature, field notes over the past

200 years

1.) Integrate similar data

2.) Assess data quality

3.) Link data - location and date

What is your motivation for standardizing sample identifiers and metadata?

What do you want to be able to do? Describe your potential use cases.

How do you want to search for sample data? What fields do you want to search?

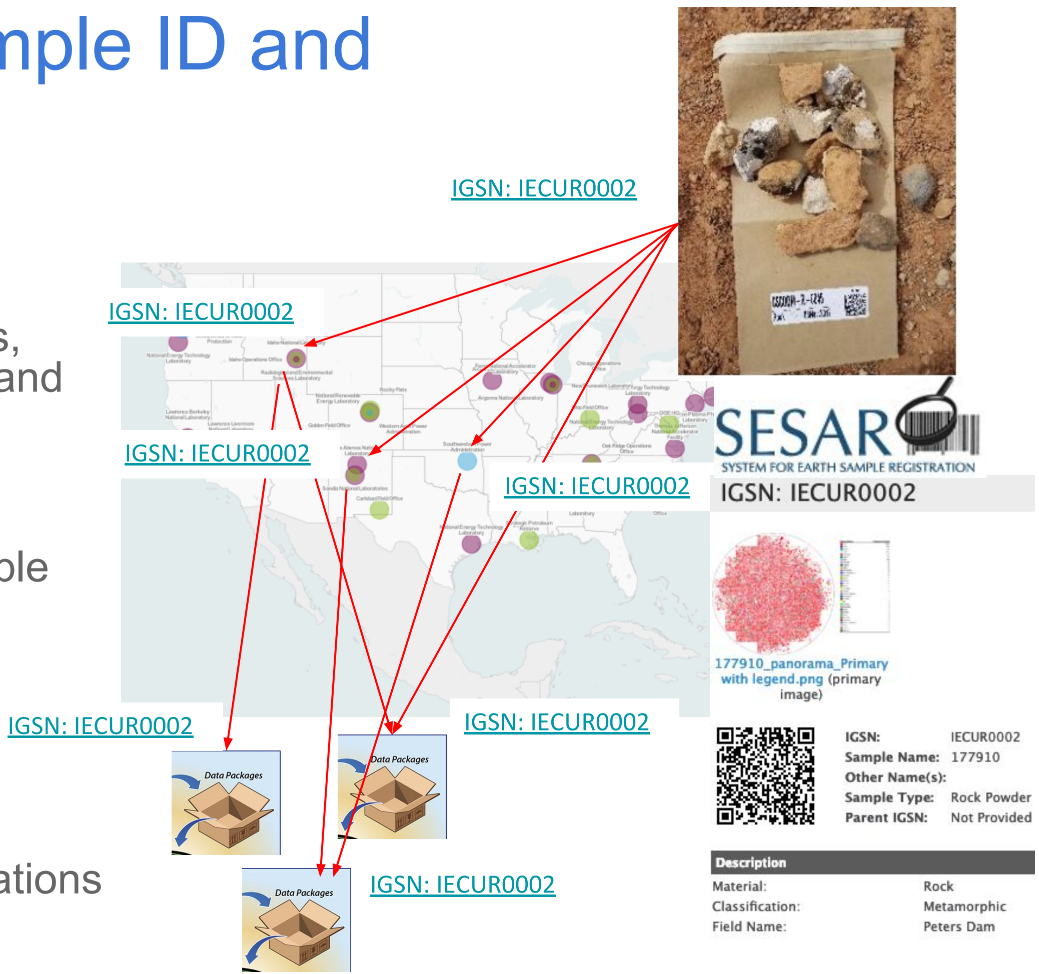
Community Need: Sample ID and Tracking

Challenge: 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 metada templates
- Linking to other samples, online metadata profiles, datasets, publications



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

IGSN Adoption

~ 7 million samples registered

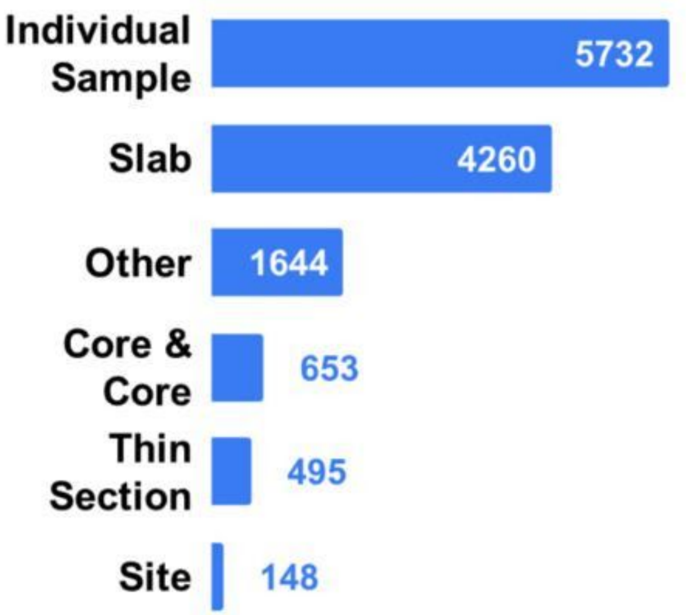
24 allocating agents globally

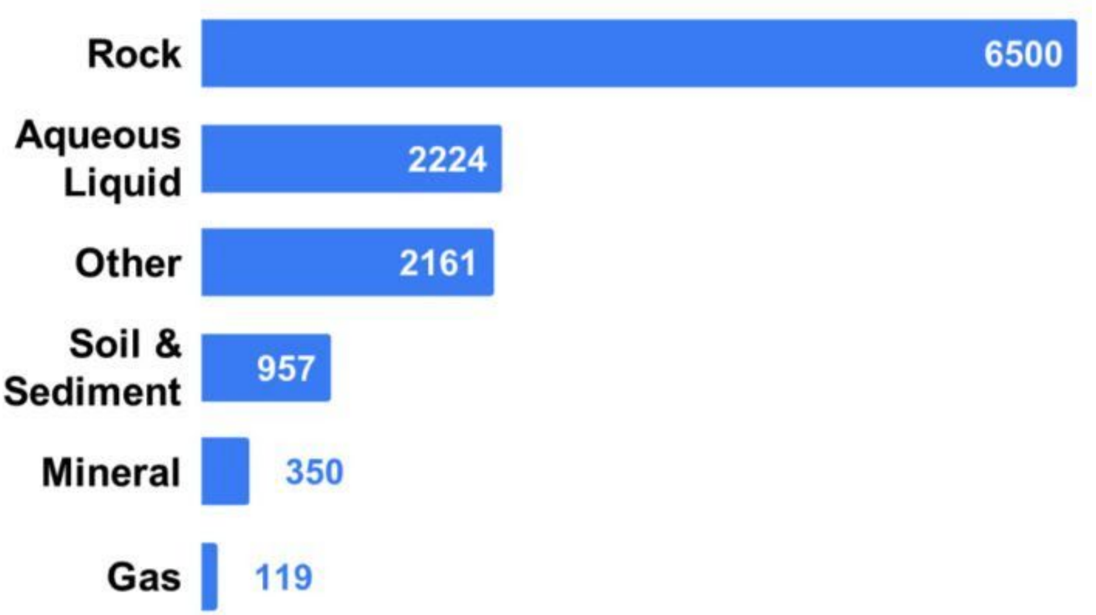
- USGS, geological surveys of UK, Australia, Korea
- Large data service providers such as ARDC, SAEON
- Research organizations and national labs (GFZ Potsdam, CSIRO
- Universities (Columbia, Curtin, Oregon State)

Sampling campaigns: International Ocean Drilling Program (3.5 million samples), US Critical Zone Observatories

Smithsonian Institution

Sample and material types registered in SESAR since Feb 2019





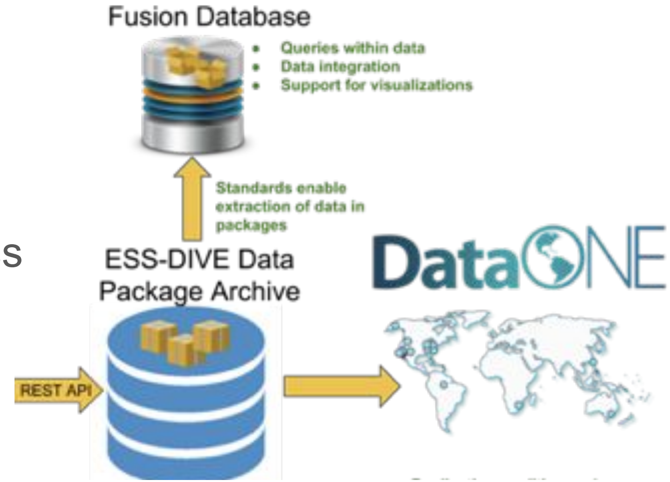
Standardization needed for advanced searches

Indexing by sample data in ESS-DIVE and potentially DataONE: search for datasets that contain sample IDs, locations, certain sample types, dates

Resource maps that link samples to other samples, subsamples, datasets, and publications

Fusion database: Advanced faceted search within and across datasets

File-level metadata: Machine-readable content and format descriptions



Fusion Database
Queries within data
Data integration
Support for visualizations
Standards enable extraction of data in packages
ESS-DIVE Data Package Archive
DataONE
REST API

Standardizing Sample Data - Summary

- Make process of naming and tracking samples easier
- Avoid ambiguity, track history of samples
- Link samples to other important identifiers…
- Increase discoverability of samples: Link to ESS-DIVE Dataset DOI
- Facilitate advanced data searches: indexing by IGSN or sample type in ESS-DIVE and DataONE, integrate samples with certain attributes across datasets
- Cite and track data usage at the sample level

IGSN and Sample Metadata Standards Pilot Test

Develop and test an interoperable system for sample identification, tracking, and metadata standardization

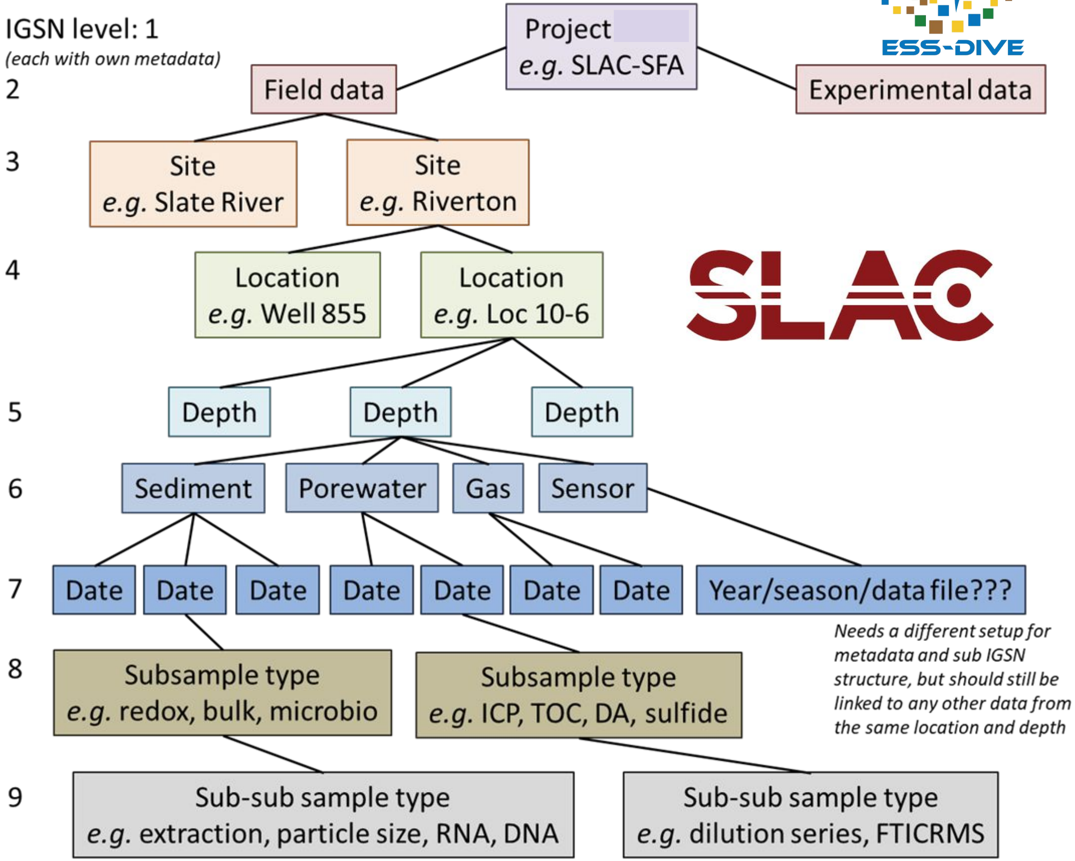
Pilot IGSN and Sample Tracking

SFAs: SLAC, PNNL, LLNL, ANL, LBNL

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

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

Goals: **1.) IGSN or other PID?**
2.) Decide on core metadata standards
3.) Develop standards for specific sample types



Needs a different setup for metadata and sub IGSN structure, but should still be linked to any other data from the same location and depth

Pilot IGSN - Project Sample Types

SLAC: sites (location, wells), sediment, ground powder sample, sand/silt/clay, particulate organic matter, mineral, colloidal fraction, ground water, pore water, plant biomass, plant roots, plant seeds, microbial, synthetic mineral, gas

PNNL: surface water, pore water, sediment, filters from pore water,

LLNL: surface water, sediment

ANL: soil core, soil surficial, surface water, plants, plant-associated soil, floc (biofilm and inorganic), gas, ground water

LBNL: soil and derivatives

BNL?: tree occurrences?, leaves

Evaluate the IGSN Pilot Test

ESS-DIVE

Question: IGSN or another identifier? Allocating agent?

How much extra effort does it require to register for IGSNs through SESAR?

Did IGSN improve ability to track samples or have any other immediate benefit for projects?

Are there future benefits of using IGSN or some other PID for samples?

Are we able to track samples effectively using parent-child relationships or is it better to use related identifiers and relation types?

Is the resulting sample data more FAIR? Findable, accessible, interoperable, reusable?

Discussion: Interoperable, Core Sample Metadata Elements

Identifier: PID (IGSN), Sample name (free)

Sample & material type: controlled

Sample description: free

Location: name, coordinates (WGS 84 decimal degrees); elevation

Collector: name, ORCID

Project Name: controlled

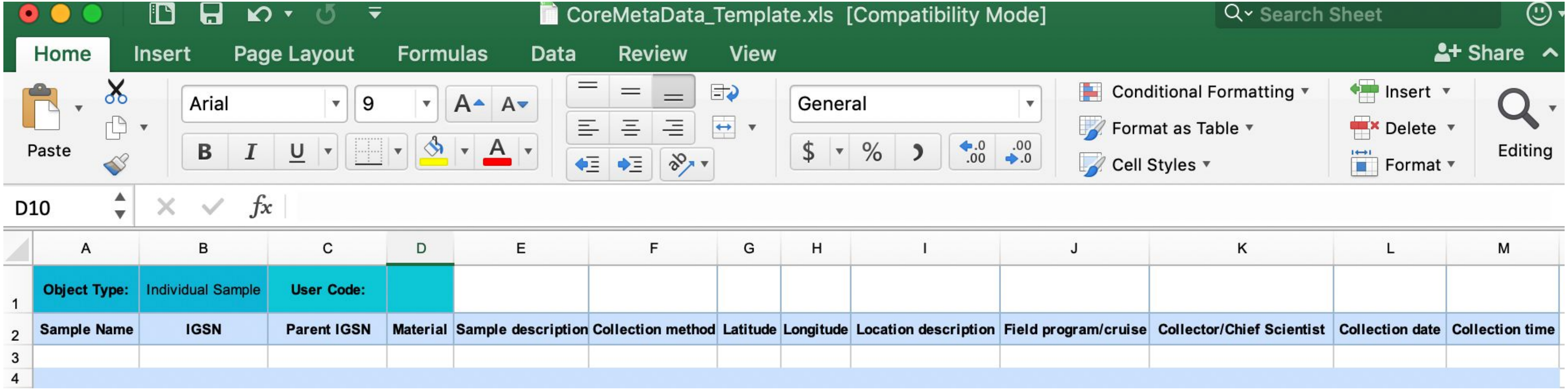
Collection date/time: (ISO 8601 format, YYYY-MM-DD, HH:MM:SSZ)

Collection method: controlled, free description

Sample Relationships: Parent Identifer, Related Identifier, Relation Type (controlled)

Supplemental Metadata & Data: link to records

Example Template - Core Sample Metadata



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Object Type:	Individual Sample	User Code:										
2	Sample Name	IGSN	Parent IGSN	Material	Sample description	Collection method	Latitude	Longitude	Location description	Field program/cruise	Collector/Chief Scientist	Collection date	Collection time
3													
4													

Discussion: Interoperable across standards

Link to DRAFT core sample metadata cross-walk

Dublin Core - general vocabulary terms to describe digital resources

Ecological Metadata Language (EML)

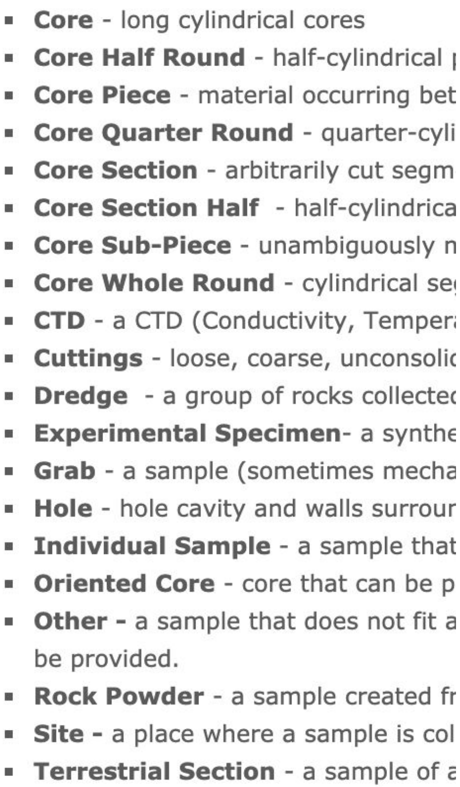
Minimum Information about any (x) Sequence (MIxS), Environmental Packages

ISO Observations and Measurements, Specimen model

Darwin Core - extension of Dublin Core for biodiversity informatics, share information on biological diversity

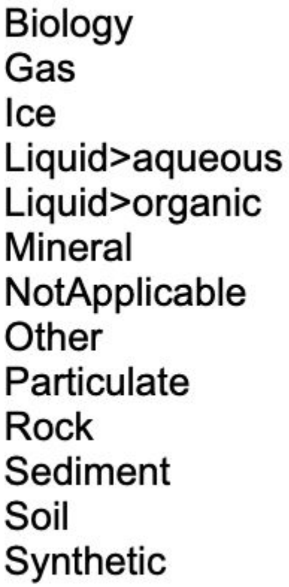
Characterize sample types

SESAR Object Type:



- **Core** - long cylindrical cores
- **Core Half Round** - half-cylindrical |
- **Core Piece** - material occurring bet
- **Core Quarter Round** - quarter-cyli
- **Core Section** - arbitrarily cut segm
- **Core Section Half** - half-cylindrica
- **Core Sub-Piece** - unambiguously n
- **Core Whole Round** - cylindrical se
- **CTD** - a CTD (Conductivity, Temper
- **Cuttings** - loose, coarse, unconsolic
- **Dredge** - a group of rocks collectec
- **Experimental Specimen**- a synthe
- **Grab** - a sample (sometimes mecha
- **Hole** - hole cavity and walls surrour
- **Individual Sample** - a sample that
- **Oriented Core** - core that can be p
- **Other -** a sample that does not fit a be provided.
- **Rock Powder** - a sample created fr
- **Site -** a place where a sample is col
- **Terrestrial Section** - a sample of a

Material:



Biology
Gas
Ice
Liquid>aqueous
Liquid>organic
Mineral
NotApplicable
Other
Particulate
Rock
Sediment
Soil
Synthetic

What other terms do we need?

Type of site/sample feature:
E.g. well, piezometer

Further characterize water samples: surface water, pore water, groundwater

Discussion: Metadata elements and terms for specific sample types?

ESS-DIVE

Depth

Physiographic feature (e.g. stream, aquifer, floodplain)

Additional terms for water sample types: e.g. surface water, pore water, groundwater

Size and unit: size of the registered object

Temperature, pH

Vegetation type

Characterististics of physiographic feature (hydrogeomorphology, dominant sediment type, etc.)

Morphological characteristics of the sample

Approach: Define required metadata versus supplemental metadata?

Next Stage: Attribute fields for sample data processing, QAQC, Analysis

Any ideas on approach for this?

Set up working groups? Sample types, sample processing, QAQC, analysis

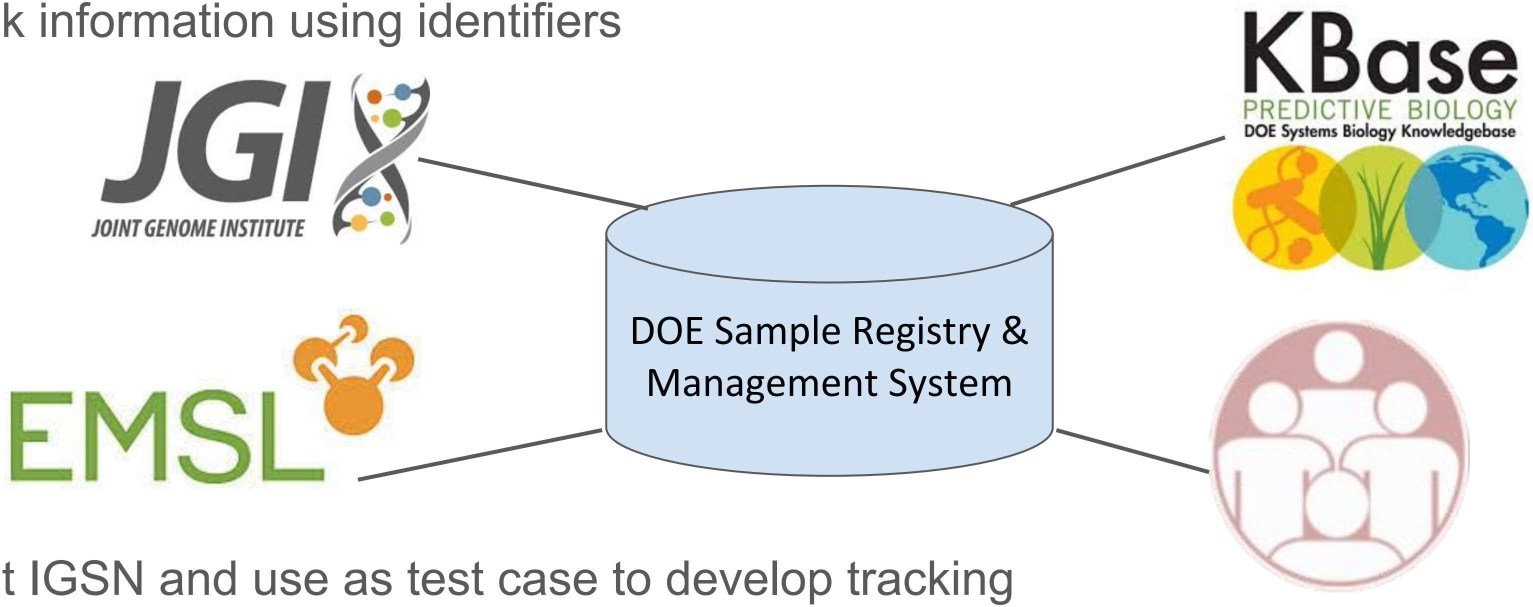
Community funds sample standards?

ESS-DIVE

Deep Insight for Earth Science Data

Sample Identification - User Facilities

Need central DOE system to register samples, obtain PIDs, add metadata, and link information using identifiers



DOE Sample Registry & Management System

Pilot IGSN and use as test case to develop tracking system across facilities