ESS-DIVE January 2020 Webinar:

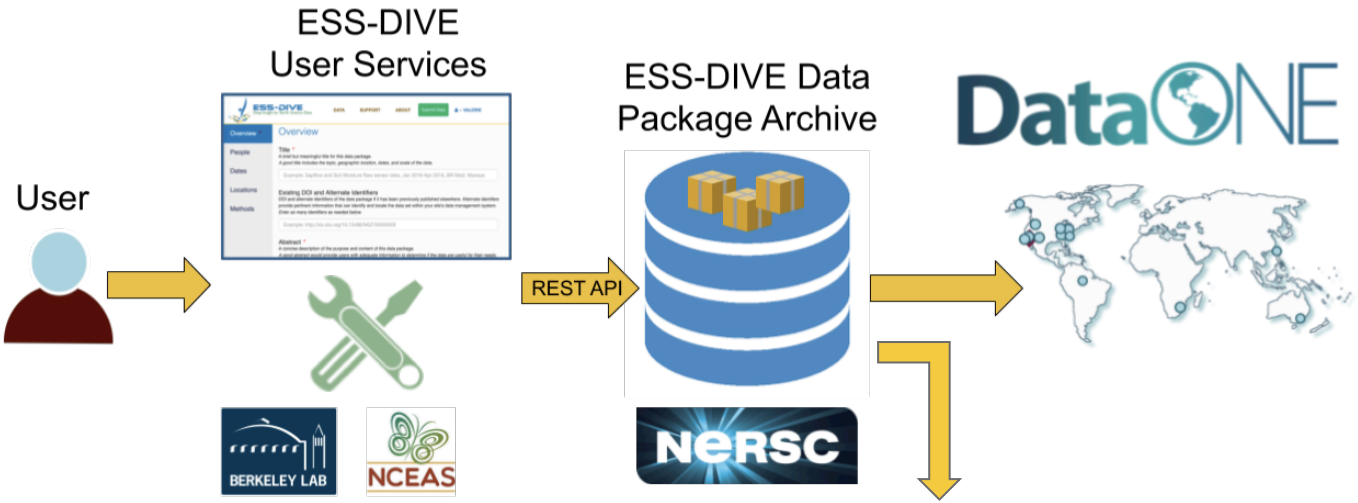
Community use of persistent sample identifiers and standard metadata to support sample management, tracking, and search

Joan Damerow, *Deb Agarwal, Kristin Boye, Eoin Brodie, Madison Burrus, Shreyas Cholia, Hesham Elbashandy, Ricardo Eloy Alves, Kim Ely, Amy E Goldman, Val Hendrix, Zarine Kakalia, Ken M Kemner, Annie B Kersting, Katharine Maher, Nancy Shiao-Lynn Merino, Fianna O'Brien, Zach Perzan, Emily Robles, Cory Snavely, Patrick Sorensen, James Stegen, Pamela Weisenhorn, Karen Whitenack, Mavrik Zavarin and Charuleka Varadharajan*

ESS-DIVE

Deep Insight for Earth Science Data

ESS-DIVE: Toward a Fusion Database



Future: Fusion Database
Advanced data search and integration across datasets

Interoperable, standardized metadata needed for Fusion DB`

Address Community Need - Samples

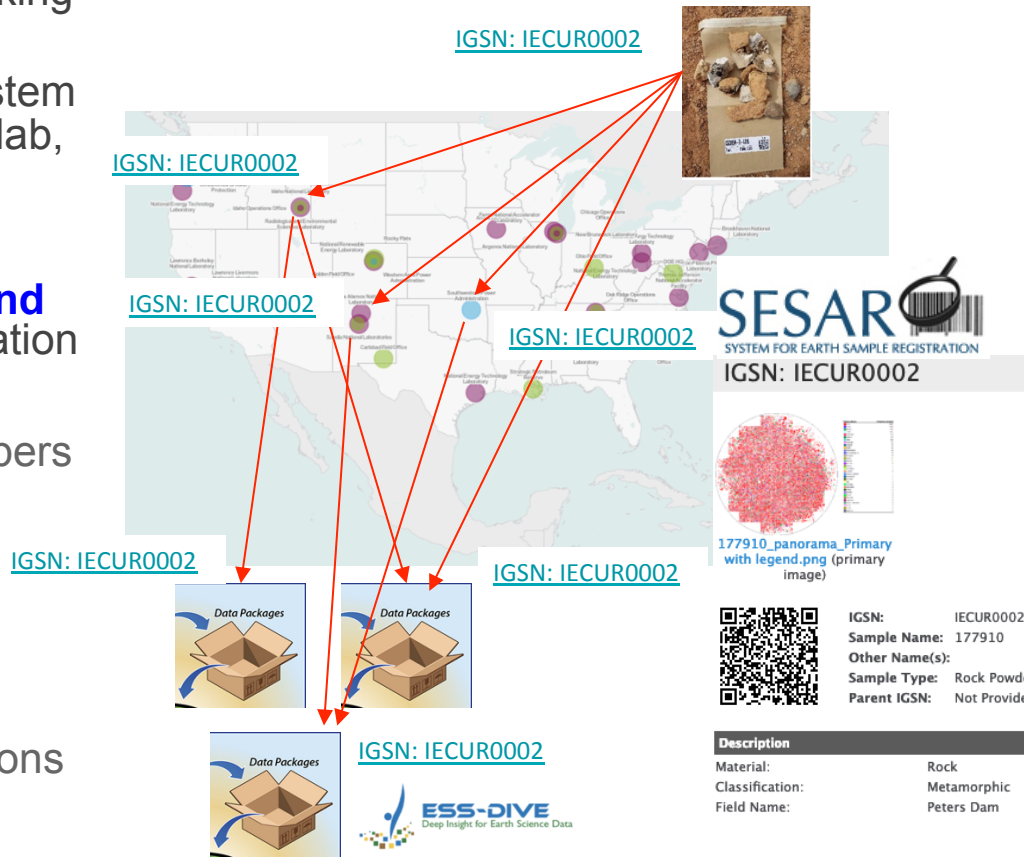
Challenge: Sample identification and tracking

Need practical, standardized sample identification, description, and tracking system to **integrate and publish data** from field, lab, DOE user facilities

Support Data Linking/Interoperability and Reuse: Advanced data search and integration

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

- SESAR provides standardized core sample metadata, templates
- Linking to other samples, online metadata profiles, datasets, publications



Recent and Upcoming Sample-Related Activities

May 2019 - Start IGSN pilot test, gathered first round feedback through Nov 2019

Oct 2019 - National Microbiome Data Collaborative (NMDC) Ontology workshop

Dec 2019 - AGU: IGSN Community Meeting

Jan 2020 - Earth Science Information Partners (ESIP):
Defining the Bull's Eye of Sample Metadata

Jan 2020 - USGS: National Digital Catalog Sample Collections Metadata workshop

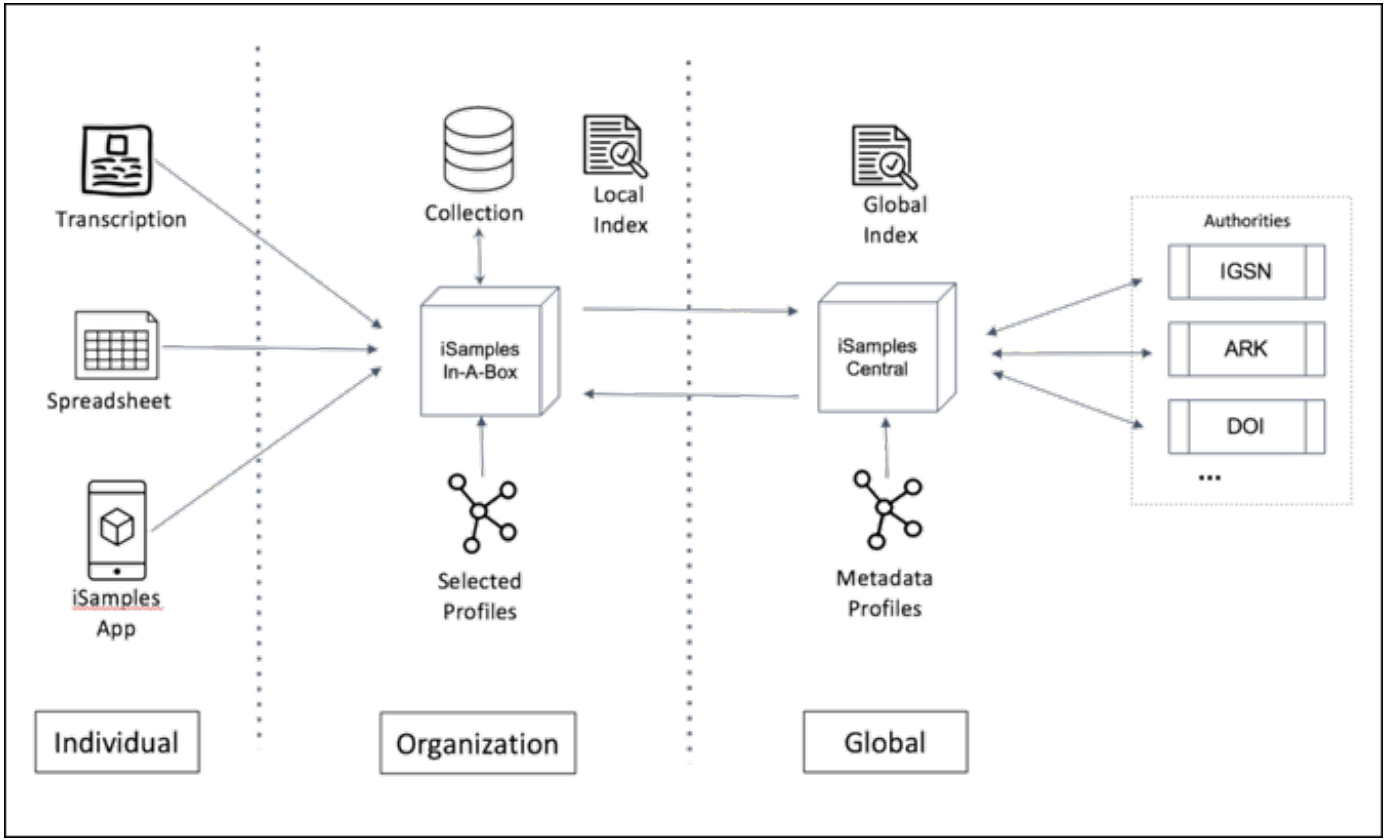
March 2020 - Research Data Alliance (RDA) 15th Plenary Meeting: Identifying physical samples and linking them to the global research data ecosystem

Internet of Samples: Toward an Interdisciplinary Global Infrastructure for Material Samples

iSamples

Neil Davies
John Deck
Sarah Kansa
Kerstin Lehnert
John Kunze
Sarah Ramdeen
Dave Vieglais
Ramona Walls
Raymond Yee

Proposal: Oct 2019
NMDC and ESS-DIVE submitted letters of collaboration

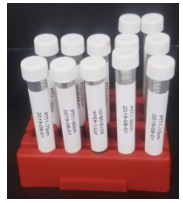


Sample PID and Metadata Pilot Test

Diverse, Interdisciplinary Projects →

Biogeochemical responses to contamination, warming, disturbance

SLAC SBR SFA - Kristin Boye, Zach Perzan
LLNL SBR SFA - Mavrik Zavarin, Nancy Merino
PNNL SBR SFA - James Stegen, Amy Goldman
ANL SBR SFA - Pamela Weisenhorn
LBNL Watershed Function SFA and TES SFA FICUS - Eoin Brodie, Patrick Sorenson, Ricardo Eloy Alves
BNL TES SFA NGEE Tropics - Kim Ely



Argonne NATIONAL LABORATORY
SLAC
BERKELEY LAB
Pacific Northwest NATIONAL LABORATORY

Pilot Test Methods

Object Type:	Individual Sample	User Code:						
Sample Name	IGSN	Parent IGSN	Material	Sample description	Collection method	Latitude	Longitude	Location des

SESAR
SYSTEM FOR EARTH SAMPLE REGISTRATION

genomic STANDARDS consortium

ESS-DIVE

Field test of IGSN PID and metadata

IGSN

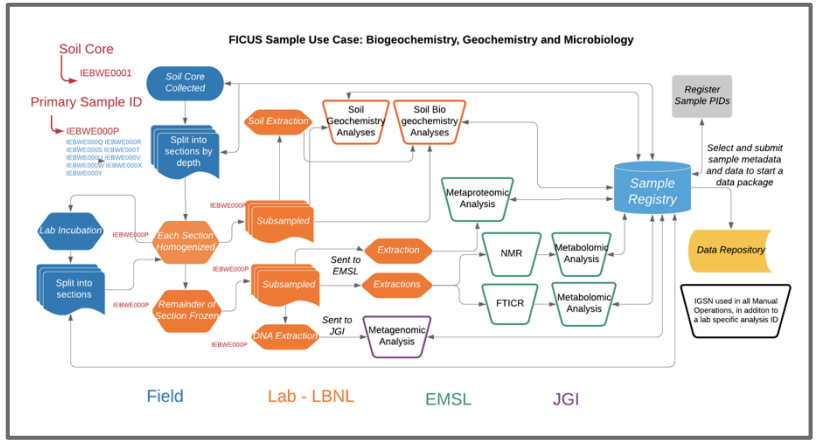
Sample Metadata Research and Crosswalk

Biodiversity Information Standards TDWG

Publish data packages in ESS-DIVE

1 2 3 4 5 6

Map sample relationships and history



PID for sample tracking and publication

IGSN

University of California CDL California Digital Library

Tools to support sample management and reuse.

Basic Sample Metadata for Pilot Test

Required Fields

- PID, and Parent PID where relevant
- Sample Name (unique)
- Chief Scientist, Collector
- Sample type: Object type (controlled), Material (controlled)
- Collection date (ISO)
- Latitude, Longitude (WGS 84), if relevant
- Project (controlled)
- Collection method
- Bio: Scientific Name
- Processing details (required for certain analyses

Recommended Fields

- NEW: Site ID, Collection Event ID
- Coordinate uncertainty or Instrument used
- Location description
- Sample description
- Sample type: Classification (controlled)
- Primary Physiographic feature (ENVO)
- Preservation method
- Purpose
- Collection time
- Depth (min, max), Depth scale (required if depth entered)
- Elevation
- Physiographic Feature (local environment)
- NEW: Biome (broader ecosystem)

Crosswalk: https://docs.google.com/spreadsheets/d/1pV1zmqRFbAhA9kxcbOnz2RDWO8siXjnlfL2VhYJHDiM/edit#gid=0

Sample Metadata Research - Existing standards and templates

General: Dublin Core, DataCite, ISO Observations and Materials

Geoscience: IGSN Descriptive Metadata Schema, SESAR IGSN template, USGS, EPA

Biology: Darwin Core, Minimum Information about any sequence (MIxS) and Environmental Packages

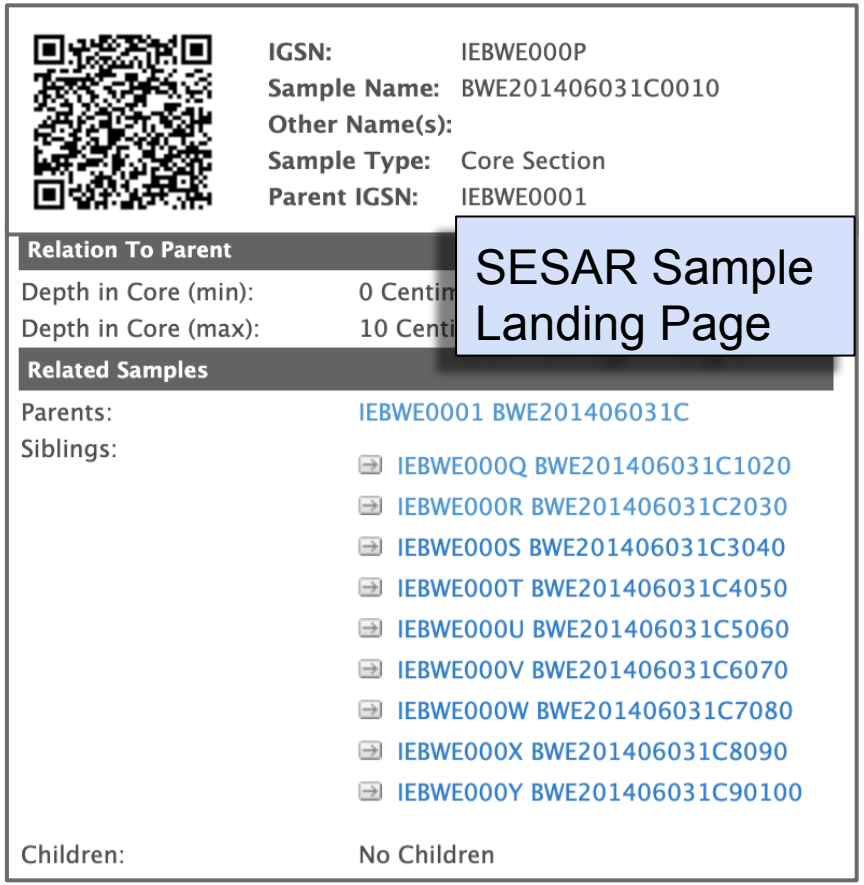
Other Relevant Ontologies or Shared Vocabularies: e.g. ENVO and GAZ

Crosswalk: https://docs.google.com/spreadsheets/d/1pV1zmqRFbAhA9kxcbOnz2RDWO8siXjnlfL2VhYJHDiM/edit#gid=0

2,476 Registered IGSNs

User Feedback: Sample Registration

- More automation for large campaigns
  - Field metadata collection apps
  - Scripts to standardize more quickly
  - Convert to xml or json and use API
  - Inherit relevant metadata from parent
- Describe collection - Landing pages for collections



GFZ

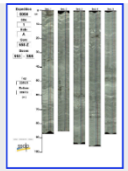
Helmholtz Centre Potsdam

Helmholtz Centre Potsdam
GFZ German Research Centre for Geosciences

ESS-DIVE

General Identifiers

Program:	ICDP
Expedition:	ICDP 5054
Type:	Core
Name:	5054_1_A_658_Z
IGSN:	ICDP5054ECYD101
Parent IGSN:	ICDP5054EHW1001
Release Date:	2017-3-1



Sampling Location

Latitude:	63.40163
Longitude:	13.202917
Coordinate System:	WGS84
Elevation:	-1735.452
Final Depth:	-1741.562
Location Type:	N/A
Location Name:	Åre, Jämtlands län, Sweden
Location Description:	COSC-1 is located in the vicinity of the abandoned Fröå mine
Country:	Sweden
Province:	Jämtlands län
County:	N/A
City:	Åre

Geology

Material:	Rock
Rock Classification:	N/A
From Corrected Depth:	2257.962
To Corrected Depth:	2264.072
Depth Reference:	meter below ground level
Geological Age:	mid-paleozoic
Geological Unit:	N/A

Methods

MSCL	yes
XRF	yes
Lithological Description	yes
Core Overview	yes
Core Section Scan	yes
Core Catcher Scan	no

Drilling

Drilling Method:	Coring>RockCorer wireline diamond coring, HQ and NQ bit size
Operator:	Lund University, Engineering Geology Larsson Drilling Consulting AB
Funding Agency:	Swedish Research Council (Vetenskapsrådet)
Total Length:	2400.1m

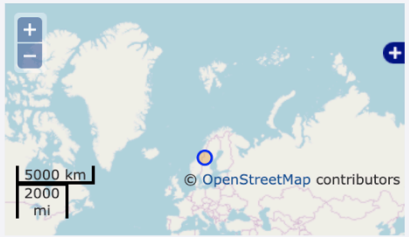
Sample Family

- 5054_1_A
  - 5054_1_A_1_Z
  - 5054_1_A_2_Z
  - 5054_1_A_3_Z
  - 5054_1_A_4_Z
  - 5054_1_A_5_Z
  - 5054_1_A_6_Z

=Hole, =Core, =Core-Section, =Core-Sample

The Sample Family shows a sub-sampling graph. Select entries to navigate samples. Core-Samples are issued to scientists on request. The naming convention for a Core-Sample is: *Expedition_Site_Hole_Core_Section,from-to(cm)*. Hole, Core, and Core-Section are following the same schema respectively.

Location Map



Drilling Start/End: 2013-09-05 / 2014-08-26 *
Latitude: 63.40163 * Longitude: 13.20292 *
Åre, Jämtlands län, Sweden

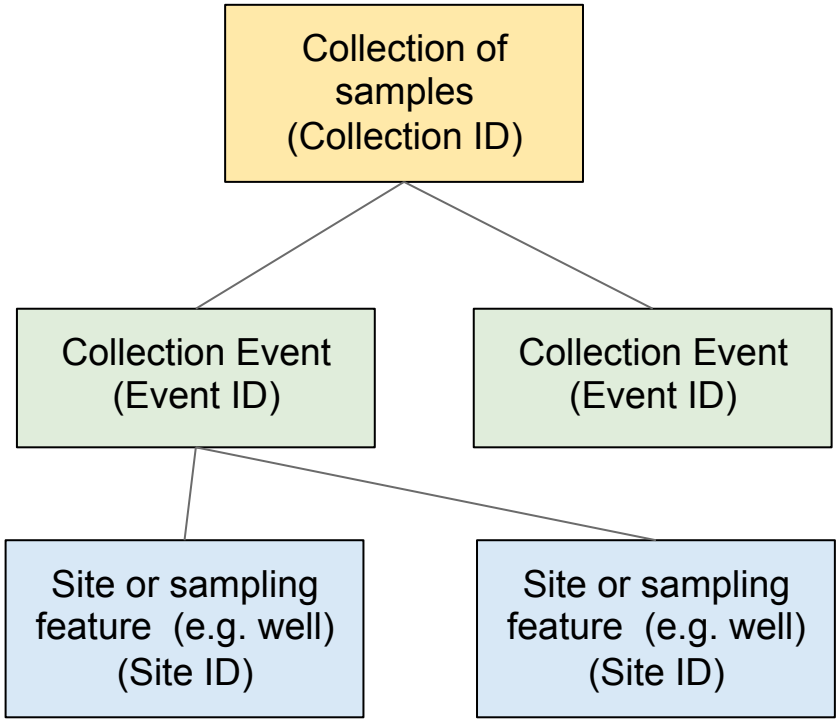
Publications & Datasets

Lorenz, H., Rosberg, J.-E., Juhlin, C., Bjelm, L., Almqvist, B. S. G., Berthet, T., … Tsang, C.-F. (2015). COSC-1 – drilling of a subduction-related allochthon in the Palaeozoic Caledonide orogen of Scandinavia. Scientific Drilling, 19, 1–11. doi:10.5194/sd-19-1-2015

Lorenz, Henning; Rosberg, Jan-Erik; Juhlin, Christopher; Bjelm, Leif; Almqvist, Bjarne; Berthet, Théo; Conze, Ronald; Gee, David G.; Klonowska, Iwona; Pascal, Christophe; Pedersen, Karsten; Roberts, Nick;

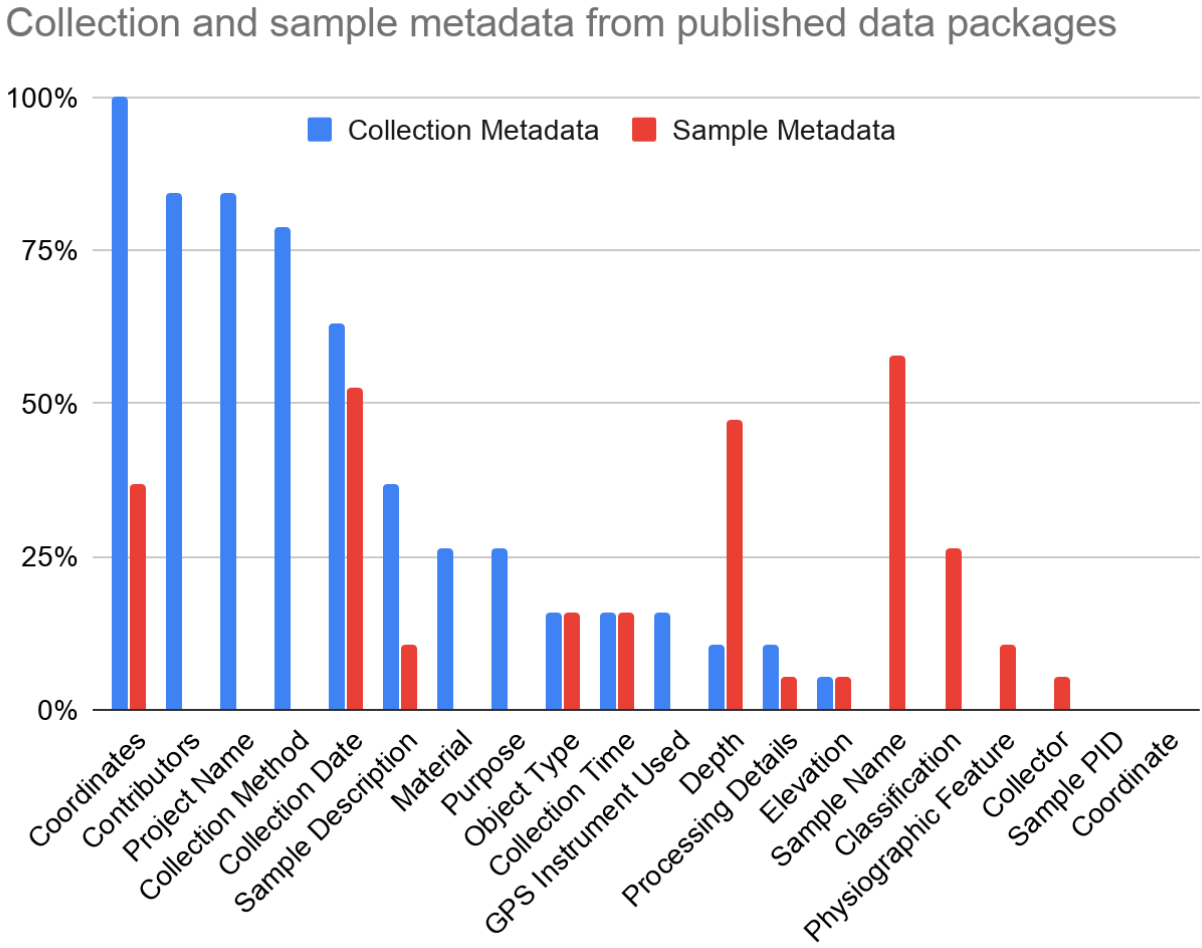
Proposed Option: Collections, Events, Sites

- Allow separate entry and link with unique identifiers: flexible
- Identifiers - from Darwin Core & IGSN
  - Collection ID
  - Collection Event ID
  - Site ID
  - Related ID
  - Parent IGSN (subsamples)
- Scripts to transfer relevant metadata for IGSN sample template
- Test for large campaigns



Metadata in Existing, Published Data Packages

- Searches for soil, water, leaf samples
- Pangaea, ORNL DAAC, ESS-DIVE, DataONE
- 19 data packages
- Consider separate requirements for collection/samples



User Feedback: IGSNs for Subsamples

Metadata
Parent ID
Child ID

☑	Soil Core	Collection of water samples at one place and time	→ Collection of highly related samples	**IEBWE0001**
☑	Core Section	Varying depths	→ Primary sample	**IEBWE000P** **IEBWE000Q**
?	Subsamples sent for analyses	Each individual container sent for analyses	→ Subsamples/ Containers	**IEBWE000R** **IEBWE000S** **IEBWE000T**

Proposed Option: ID Extensions for Subsamples

Add extensions for subsamples and record parent-child relationships in metadata

☑	Soil Core	Collection of water samples at one place and time	→ Collection of highly related samples	**IEBWE0001**
☑	Core Section	Varying depths	→ Primary sample	**IEBWE00011M** **IEBWE00012M**
?	Subsamples sent for analyses	Each individual container sent for analyses	→ Subsamples/ Containers	**IEBWE00012MA** **IEBWE00012MB** **IEBWE00012MC**

ARK Qualifiers for Subsamples

ARKs: containment qualifiers

☑	Soil Core	Collection of water samples at one place and time	→ Collection of highly related samples	ark:/12025/654xz321
☑	Core Section	Varying depths	→ Primary sample	ark:/12025/654xz321/1m ark:/12025/654xz321/2m
?	Subsamples sent for analyses	Each individual container sent for analyses	→ Subsamples/ Containers	ark:/12025/654xz321/1m/v1 ark:/12025/654xz321/1m/v2 ark:/12025/654xz321/1m/v3

User Feedback: IGSN Collection Metadata

Some metadata not previously recorded supports integration and reuse

- Sample relationships - parent IGSN,
- Sample Type

Additional classification needed: plant (biology), water

SESAR IGSN Liquid>aqueous informal classifications

User Feedback: IGSN Metadata

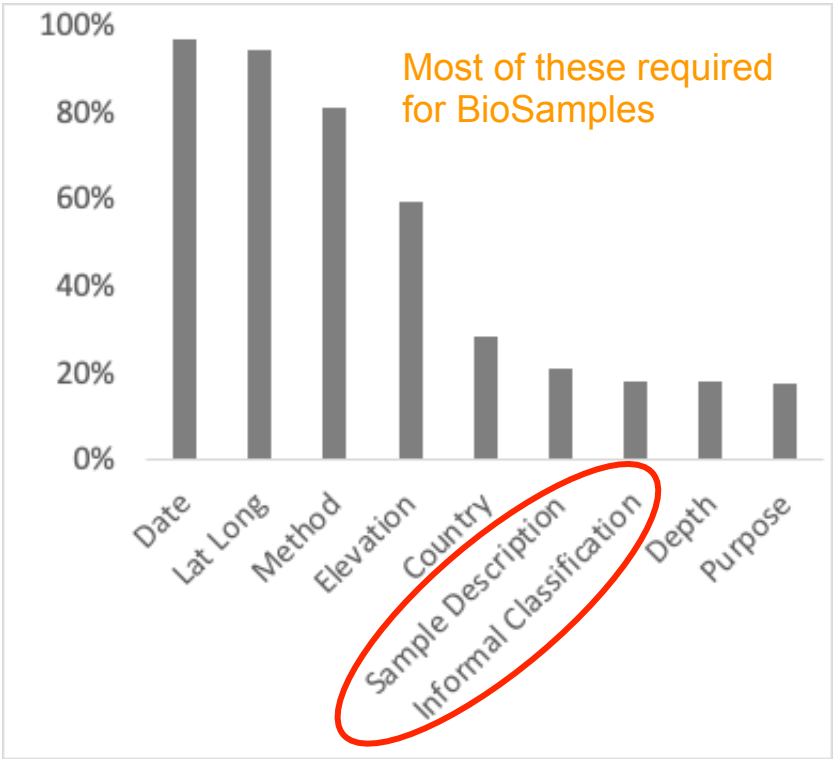
Some metadata not previously recorded supports integration and reuse

- Sample relationships - parent IGSN,
- Sample Type

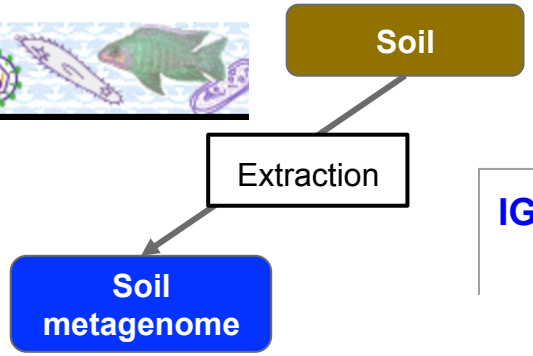
Additional classification needed: plant (biology), water

Key Need: Linking Geo and BioSamples

General Metadata provided for liquid>aqueous samples



IGSN and BioSample Interoperability

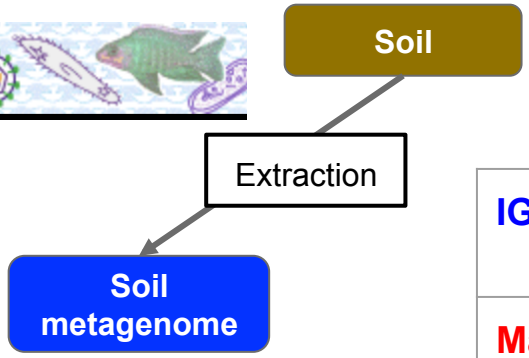


IGSN: IEBWE000P ⟷ **BioSample: SAMN04007148**

IGSN	**Source_material_id = IGSN Link**

- Need links between IDs

IGSN and BioSample Interoperability



IGSN: IEBWE000P ⟷ **BioSample: SAMN04007148**

- Need links between IDs
- IGSN metadata → BioSample MIxS/MIMS, Env Packages
- Use of standard ontologies
- Classifications for water samples
- Add biome, sample processing

IGSN	Source_material_id = IGSN Link
Material	env_medium* = ENVO; isolation_source
populate from material	organism* (e.g. soil metagenome)
Physiographic feature	Env_local_scale* = ENVO
	Env_broad_scale* = ENVO
Country; Locality	Geo_loc_name* = GAZ
	samp_mat_process

ESS-DIVE Plans for Sample PID registration and management

Near term - Continue to use SESAR to register sample PIDs, with options:

- Register samples and update metadata through SESAR portal or API
- Customize SESAR IGSN template with some additional fields and picklists
- Offline generation of IDs that you can register later, ID extensions
- Document recommended practices and metadata template

Address Efficiency

- Sample collection app to obtain standardized info in the field
- Separate collection, site, event metadata, scripts transfer to samples

Continue discussions with OSTI about administering DOE sample PIDs

ESS-DIVE Plans for Fusion Database and Sample Tracking

Prototyping for Fusion database

Need pilot testers to publish sample related datasets in ESS-DIVE

Advanced search and integration - sample metadata parsing/indexing

Sample tracking - connect relevant collection metadata and data using PIDs

- Work with DOE user facilities

Plans to Address Geo and BioSample Interoperability

Scripts to generate MIxS template for registered IGSNs

- Link IGSN and BioSample IDs: IGSN URL in BioSample metadata, vice versa

Paper on pilot test and geo/bio sample interoperability, recommended practices

Collaborate with broader community: NMDC, GSC, IGSN, DOE user facilities, USGS, sample groups at ESIP and RDA

Sample Metadata Crosswalk

https://docs.google.com/spreadsheets/d/1pV1zmqRFbAhA9kxcbOnz2RDWO8siXjnlfL2VhYJHDiM/edit#gid=0

Meetings/form to get detailed feedback on sample metadata requirements

Elements that should be required

Elements that are more appropriate or feasible at a collection-level