Sample Tracking and Analysis Workflows

Led By: Joan Damerow
Community Engagement Lead Scientist for ESS-DIVE
Lawrence Berkeley National Lab
Earth & Environmental Sciences Area

Session Overview

- **Sample tracking challenge**
- IGSNs for sample tracking and linking
- **Discussion:** Sample tracking use cases
- **Activity:** Sample relationships / journey map (if time)

Takeaways from this session

- Identifiers are essential to:
  - Track metadata and data over time
  - Link related multidisciplinary data
- Learn about **ESS sample tracking use cases**

A common approach for sample identifiers and metadata will enable more effective sample planning, tracking, discovery, and reuse.

Reporting format tutorial

Sample ID and Metadata (Joan Damerow)

ESS-DIVE Sample ID and Metadata Reporting Format (IGSN-ESS) v1.0.0

ESS-DIVE recommends registering samples for Global Sample Numbers (IGSNs) through the System for Earth Sample Registration (SESAR). IGSNs are associated with standardized metadata to characterize a variety of different samples and their collection details. These sample identifiers facilitate sample discovery, tracking, and reuse; they are especially useful when sample data is shared with collaborators, sent to different labs or user facilities for analyses, or distributed in different data files, datasets, and/or publications.

https://ess-dive.gitbook.io/sample-id-and-metadata/

DATA SCIENCE JOURNAL

Reading: Sample Identifiers and Metadata to Support Data Management and Reuse in Multidisciplinary Ec...

Share:

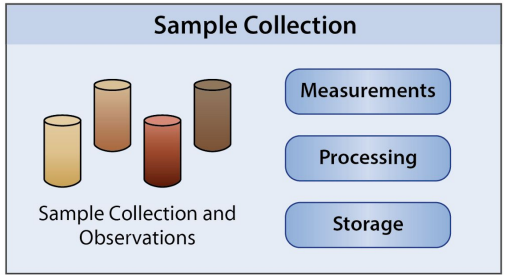
Research Papers

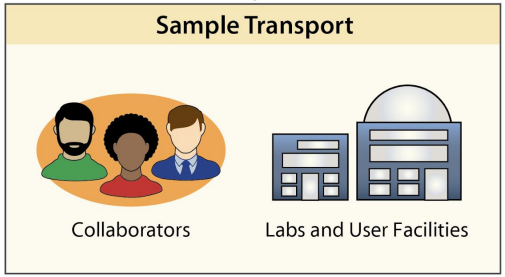
Sample Identifiers and Metadata to Support Data Management and Reuse in Multidisciplinary Ecosystem Sciences

http://doi.org/10.5334/dsj-2021-011

Sample Tracking Challenge

Address Community Challenge - Sample Tracking





Challenge

Lack of a practical, standardized sample tracking system

DISCUSSION

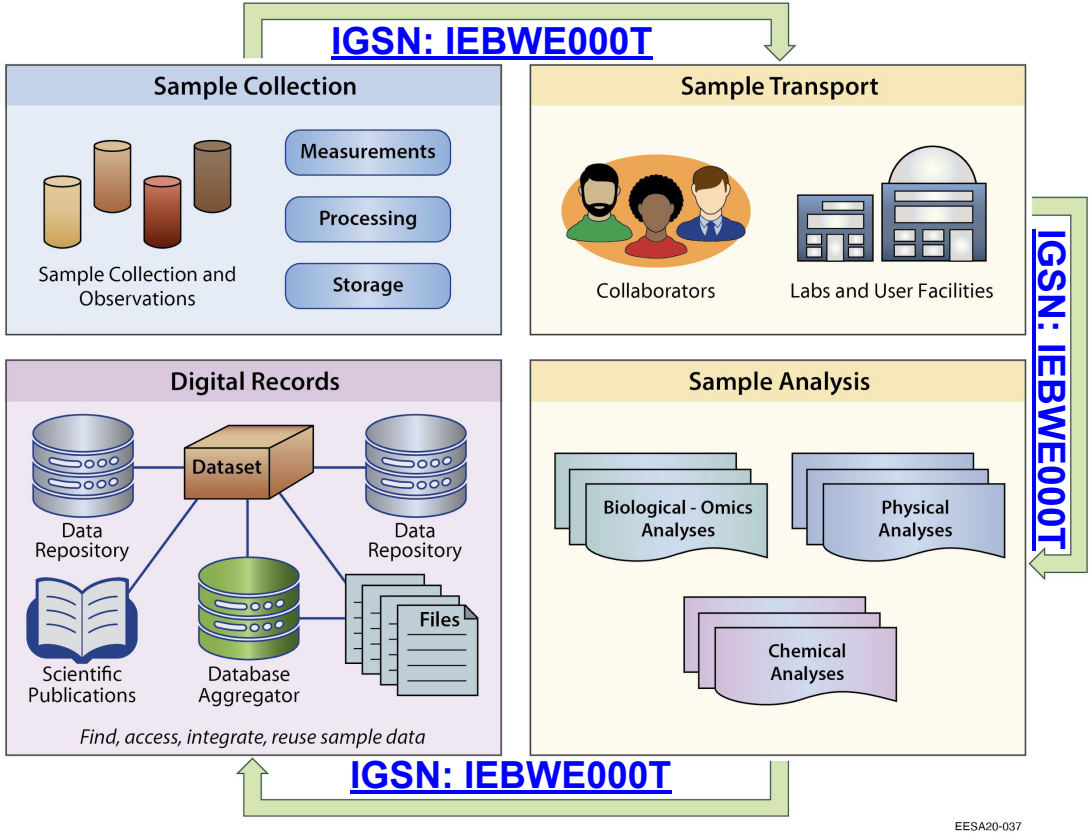
What is your sample tracking approach?

What is your process for tracking samples sent to different labs, collaborators, etc. for analyses and publication? What do your sample names look like, and how do you use them?

What (if any) challenges have you encountered with sample tracking?

Example: **Biological and Environmental Research (BER):** process for submitting samples and related data to EMSL, JGI, KBase, NMDC, and/or ESS-DIVE for analyses and publication?

Address Community Challenge - Sample Tracking



Challenge

Lack of a practical, standardized sample tracking system

Solution

International Geo/General Sample Numbers **(IGSNs)**

Terminology check: identifiers and metadata

Unique Identifier

Provides a meaningful, project-specific unique ID to organize your data

Sample Name: RockCr001_2021-05-25

Metadata

Descriptive information about data

Sample Type: Water
Feature: Stream
Location: Rock Creek, Crested Butte, CO

Persistent Identifiers

Globally unique IDs with permanent link/landing page, associated metadata

ORCiD: People
DOI: Data, publications

IGSN: Samples IEWFS000U

Persistent IDs: Landing Pages

Soil Sample Landing Page

https://app.geosamples.org/sample/igsn/IEWFS0001

IGSN: IEWFS0001

IGSN:	IEWFS0001
Sample Name:	115
Other Name(s):	
Sample Type:	Core Section
Parent IGSN:	Not Provided

Description

Material:	Soil
Classification:	Not Provided
Field Name:	Not Provided
Description:	Soil cores that were collected seasonally during autumn, winter, snowmelt, and spring at a high altitude field site which is predominately montane meadow
Age (min):	Not Provided
Age (max):	Not Provided
Collection Method:	Manual>Hammer
Collection Method Description:	Soil cores were collected using soil bulk density corer attached to a slide hammer
Size:	Not Provided
Geological Age:	Not Provided
Geological Unit:	Not Provided
Comment:	Not Provided
Purpose:	Not Provided

Geolocation

Latitude (WGS84):	38.917216053
Longitude (WGS84):	-106.955994698

Relevant Links:

- http://identifiers.org/gold:Gp0321263: Soil microbial communities from the East River watershed near Crested Butte, Colorado, United States - Metgenomes (Genomes Online Database, GOLD)
- http://identifiers.org/gold:Gp0396393: Soil microbial communities from the East River watershed near Crested Butte, Colorado, United States - Metatranscriptomes (Genomes Online Database, GOLD)
- https://doi.org/10.15485/1577267: Dataset: Soil Nitrogen, Water Content, Microbial Biomass, and Archaeal, Bacterial and Fungal Communities from the East River Watershed, Colorado collected in 2016-2017.
- https://doi.org/10.21952/WTR/1573029: Dataset for sample collection metadata

Sorensen P ; Brodie E ; Beller H ; Wang S ; Bill M ; Bouskill N (2019): Soil Nitrogen, Water Content, Microbial Biomass, and Archaeal, Bacterial and Fungal Communities from the East River Watershed, Colorado collected in 2016-2017. Watershed Function SFA. doi:10.15485/1577267

Citations 0 | Downloads 0 | Views 0 | Copy Citation | Assessment report

Files in this dataset Package: ess-dive-4ca8bf6d5ba818f-20210430T014715527688

Name		File type	Size	Login to Download
Metadata: Soil_Nitrogen_Water_Content_Microbial_Biomass_and_Archaeal_Bacterial_and_Fungal_Communities_from_the_East_River_Watershed_Colorado_collected_in_2016_2017.xml		EML v2.1.1	10 KB	Download
2017_East_River_Pumphouse_Extractable_Soil_N_Pools__1_.csv	More info	Microsoft Excel	48 KB	Download
2017_East_River_Pumphouse_Archaea_and_Bacteria_Life_Strategies__1_.csv	More info	Microsoft Excel	5 MB	Download
2017_East_River_Pumphouse_Fungal_Life_Strategies__2_.csv	More info	Microsoft Excel	448 KB	Download

Related References

Sample Metadata:Sorensen P ; Brodie E ; Beller H ; Wang S ; Bill M
Metadata for Soil Cores from the East River Watershed, Colorado co
doi:10.21952/WTR/1573029

ESS-DIVE Dataset: Soil Measurements

https://doi.org/10.15485/1577267

Using Sample PIDs/IGSNs for Sample Tracking and Linking

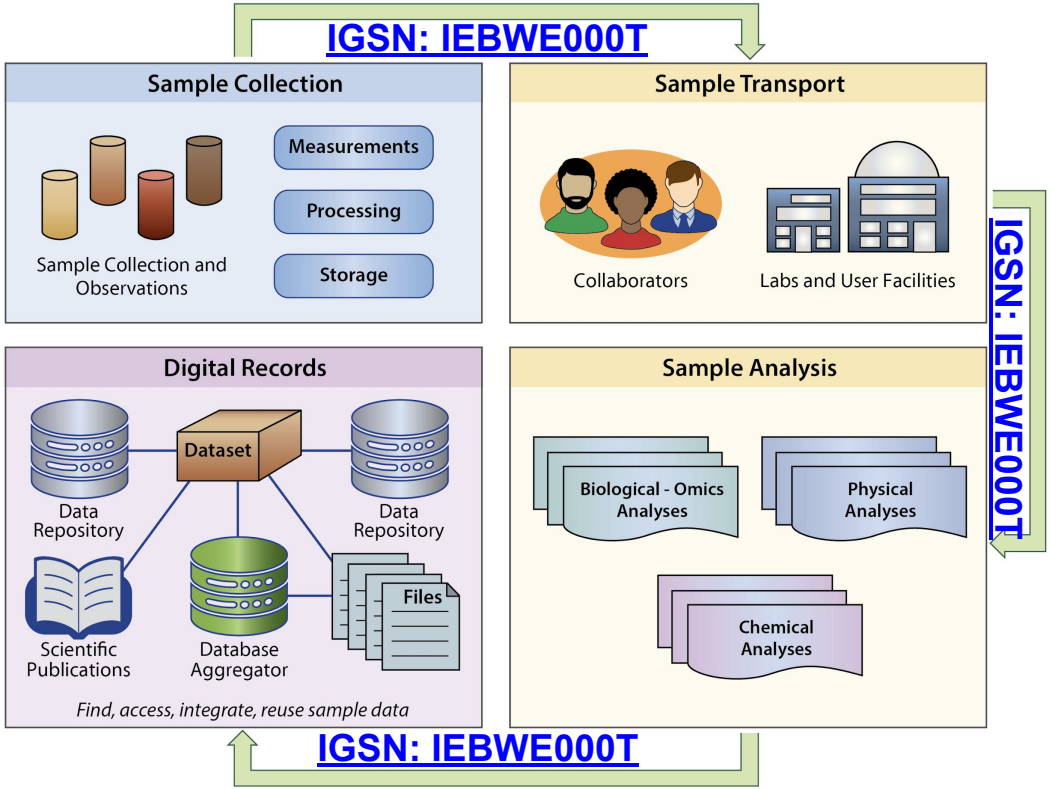
When do you need persistent IDs for samples?

1.) Multiple datasets, journal publications

2.) Collaborators work on same samples

3.) Multiple labs for analyses

4.) Sample-related data in different repositories



EESA20-037

Benefits of using IGSNs Across Facilities and Data Systems

Persistent Identifier Benefits

1. **Link** and **expand** access pathways
2. **Avoid duplication** of information across platforms
3. Interpretation and **reuse**

Linking related interdisciplinary data

IGSN: IEWFS0001

SESAR

IGSN:	IEWFS0001
Sample Name:	115
Other Name(s):	
Sample Type:	Core Section
Parent IGSN:	Not Provided

Description

Material:	Soil
Classification:	Not Provided
Field Name:	Not Provided
Description:	Soil cores that were collected seasonally during autumn, winter, snowmelt, and spring at a high altitude field site which is predominately montane meadow
Age (min):	Not Provided
Age (max):	Not Provided
Collection Method:	Manual>Hammer
Collection Method Description:	Soil cores were collected using soil bulk density corer attached to a slide hammer
Size:	Not Provided
Geological Age:	Not Provided
Geological Unit:	Not Provided
Comment:	Not Provided
Purpose:	Not Provided

Geolocation

Latitude (WGS84):	38.917216053
Longitude (WGS84):	-106.955994698

Relevant Links:

- http://identifiers.org/gold:Gp0321263: Soil microbial communities from the East River watershed near Crested Butte, Colorado, United States - Metgenomes (Genomes Online Database, GOLD)
- http://identifiers.org/gold:Gp0396393: Soil microbial communities from the East River watershed near Crested Butte, Colorado, United States - Metatranscriptomes (Genomes Online Database, GOLD)
- https://doi.org/10.15485/1577 Microbial Biomass, and Archaeal the East River Watershed, Colora
- https://doi.org/10.21952/WTR/ metadata

Soil Sample Landing Page

Sorensen P ; Brodie E ; Beller H ; Wang S ; Bill M ; Bouskill N (2019): Soil Nitrog Content, Microbial Biomass, and Archaeal, Bacterial and Fungal Communities fr Watershed, Colorado collected in 2016-2017. Watershed Function SFA. doi:10.15

ESS-DIVE Deep Insight for Earth Science Data

Citations 0 | Downloads 0 | Views 0 | Copy Citation | Assessment report

Files in this dataset Package: ess-dive-4ca8bf6d5ba818f-20210430T014715527688

Name		File type	Size	Login to Download
Metadata: Soil_Nitrogen_Water_Content_Microbial_Biomass_and_Archaeal_Bacterial_and_Fungal_Communities_from_the_East_River_Watershed_Colorado_collected_in_2016_2017.xml		EML v2.1.1	10 KB	Download
2017_East_River_Pumphouse_Extractable_Soil_N_Pools__1_.csv	More info	Microsoft Excel	48 KB	Download
	More info	Microsoft Excel	5 MB	Download
	More info	Microsoft Excel	448 KB	Download

re items in this data set

ESS-DIVE Dataset: Soil Measurements

The synchronization of microbial and plant phenology in a mountainous watershed and its importance for nutrient retention under changing hydrologic regimes.

Description The goal of the study is to observe the activation of microbial metabolic potential beneath the snowpack during winter and during the snowmelt period, as well as advanced characterization of the chemistry of carbon and nutrient transformations and assimilation by microorganisms and vegetation in response to earlier snowmelt timing.

Metabolomics: 52 | Metatranscriptome: 45 | Metagenome: 48 | Organic Matter: 1007

nmdc National Microbiome Data Collaborative

Eoin Brodie
Principal investigator
https://eesa.lbl.gov/profiles/eoin-brodie/
https://orcid.org/0000-0002-8453-8435
https://watershed.lbl.gov/

National Microbiome Data Collaborative: Study Page

Project Information | Sequencing Information | Sample Information

GOLD Genomes Online Database

JGI JOINT GENOME INSTITUTE

PROJECT INFORMATION	
GOLD Project ID	Gp0321263
Project Name	Soil microbial communities from the East River watershed near United States - ER_DNA_115
Other Names	
Legacy GOLD ID	
NCBI BioProject Name	Soil microbial communities from the East River watershed near Crested Butte, Colorado, United States - ER_DNA_115
NCBI BioProject ID	5185
NCBI BioProject Accession	PRJN
NCBI Locus Tag	EVQ
NCBI BioSample Accession	SAM
PI	Eoin

JGI Sequencing Projects: Microbial Communities

Sample Tracking Use Cases

DISCUSSION

Where is your sample data stored or published?

Where is your resulting sample data currently stored or published and archived? (clarify the project(s) these questions refer to)

a.) Examples of places the data may be: Paper only; Personal files; In one or more databases; One or more published datasets (if so, how many)? Is the data published in different archives/public databases?

b.) Is the data clearly linked in some way, or is it currently disconnected? If it is linked, how? (e.g. in the paper, links on the dataset landing page/metadata)

DISCUSSION

What tools do you think would be useful for your sample management and tracking?

DISCUSSION

What is needed for sample data interpretation and reuse?

What would other future researchers need to do to be able to compile and link the related data from your project?

WFSFA Use Case: Pilot Using Barcode Labels in Sample Workflows

SESAR
SYSTEM FOR EARTH SAMPLE REGISTRATION

Create Standard Sample IDs

1. Plan and assign meaningful unique sample names for each sample
2. Each sample name is registered for an IGSN
3. Sample name and IGSN always associated together

IEWFS0040;EBC_2
021-03-25_AM

Generate Barcode Labels

1. Choose recommended durable labels to fit sample containers
2. Generate csv with list of sample names
3. Format barcode label sheet using R scripts / baRcodeR

Scan Samples During Workflows

1. Scan labels into inventory COC, other spreadsheet, or instrument software during workflow to record sample names.

→ Reduce processing time and eliminate manual entry error

Next Steps

1. Scripts to automate ID and labels creation, update IGSN metadata
2. Integrate Sample Names and IGSNs into all data files, reports, and database
3. Use IGSNs and metadata to link related data when published online

Organize your sample campaign:

Create a sample journey map, with identifiers

Link related samples using identifiers

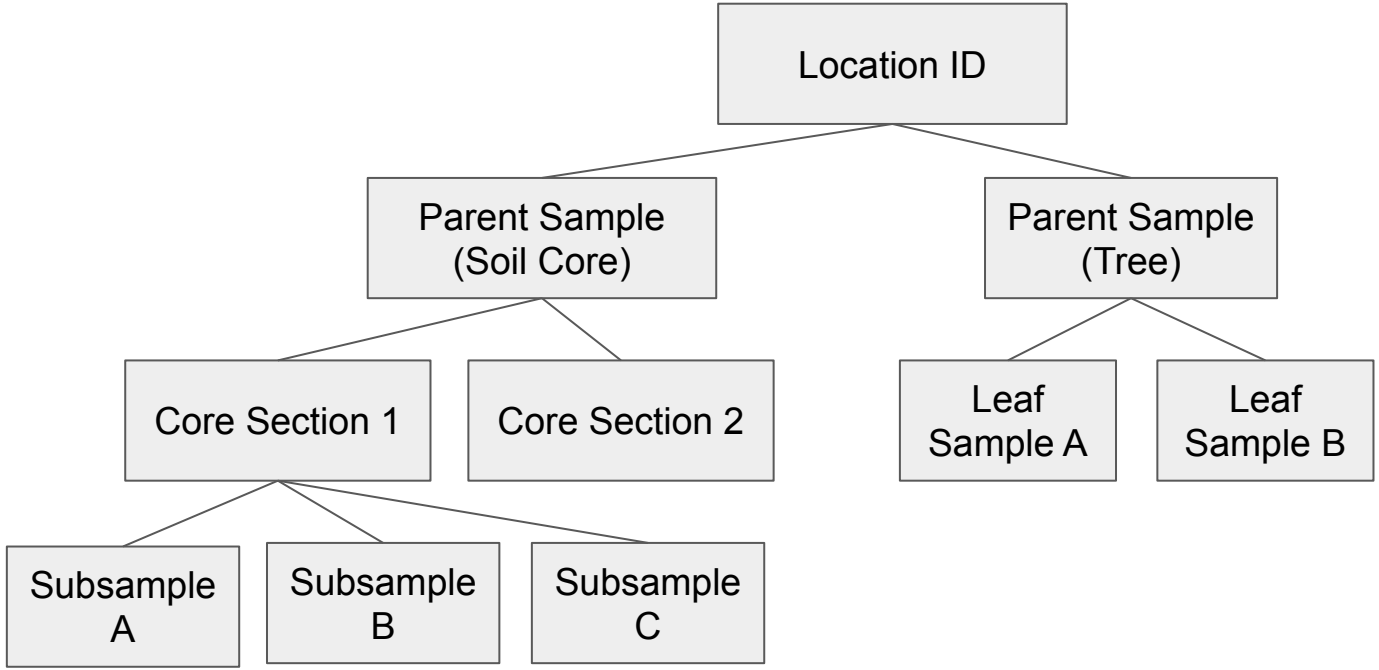
Collection ID

Sampling Event ID

Location ID

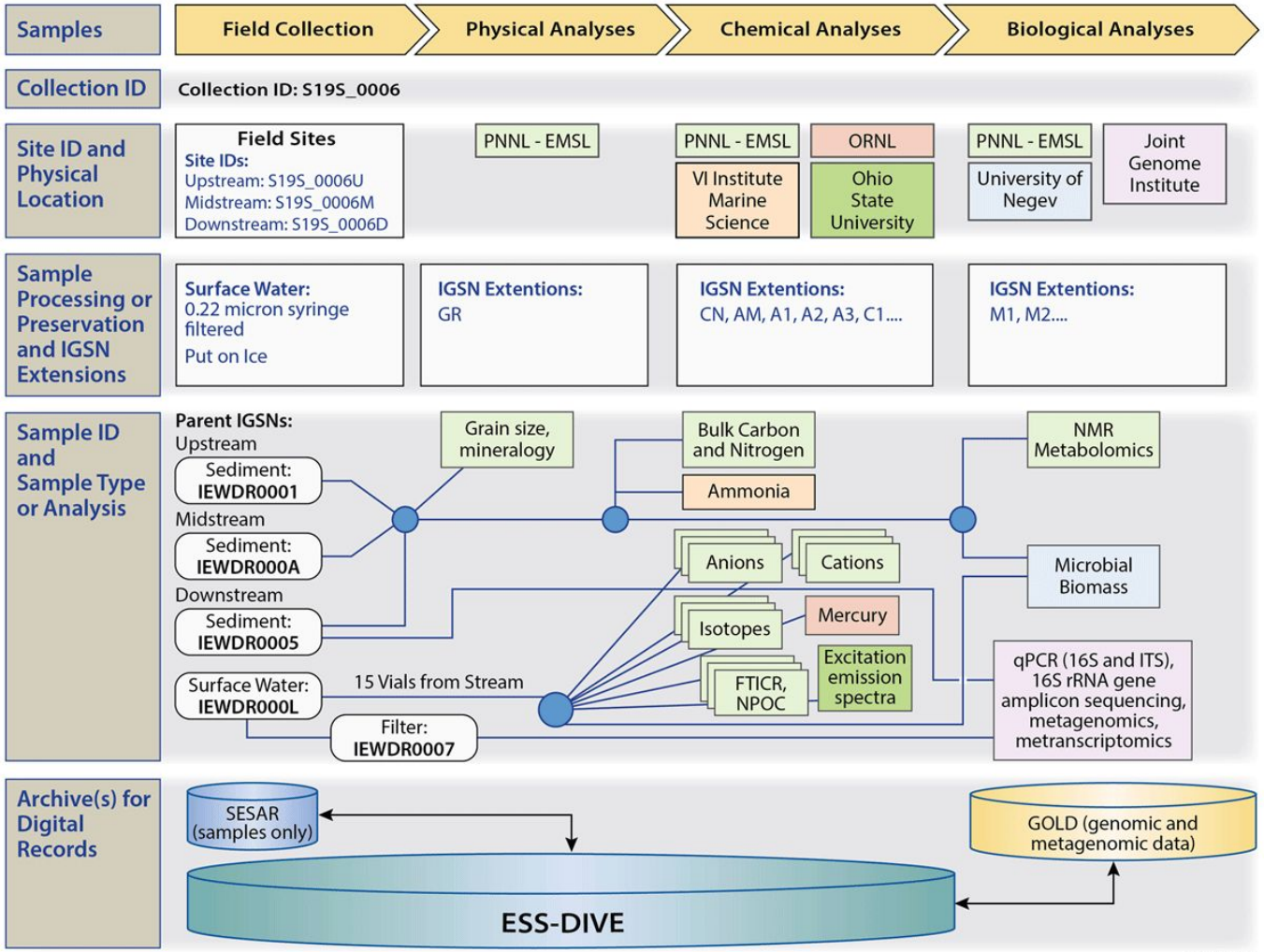
Parent Sample

- Each sample: record parent



Example sample journey map

Use sample journey map as a tool to decide how to organize and assign sample IDs



Activity: Create a sample journey map

What to include:

- Related entities: locations, sampling events
- Related samples:
  - parent samples, subsamples, replicates
  - other sample types (e.g. plant, water)
- Analyses: type, location/lab
- Assign project-specific sample names

Link to sample journey template: http://bit.ly/SampleJourneyMap

Conclusions

IGSNs enable tracking samples and exchanging related information:

- Over time
- Across data systems

Working across DOE BER data systems now:

- Link to related metadata and data
- Determine most useful approach
- **Make complex sample tracking easier**

Sample Relationships - Use Cases Activity

Link related samples using identifiers

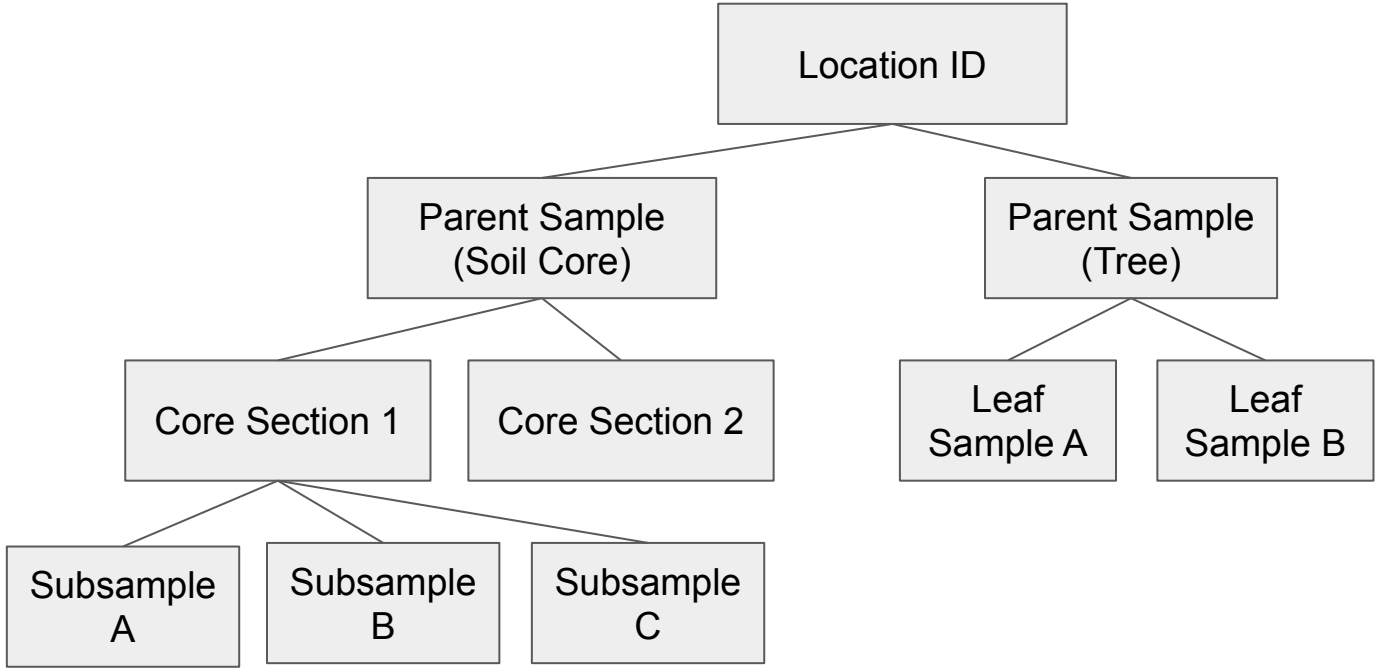
Collection ID

Sampling Event ID

Location ID

Parent Sample

- Each sample: record parent



Link related samples using identifiers

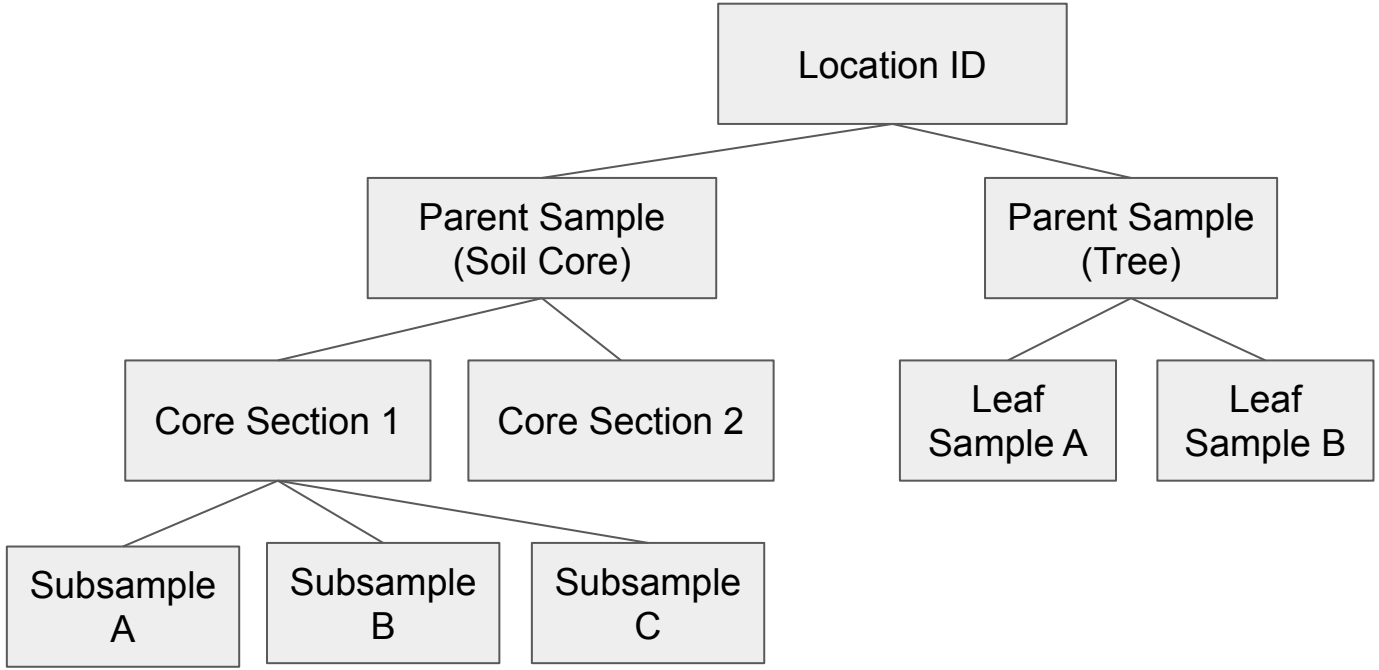
Collection ID

Sampling Event ID

Location ID

Parent Sample

- Each sample: record parent



Link related samples using identifiers

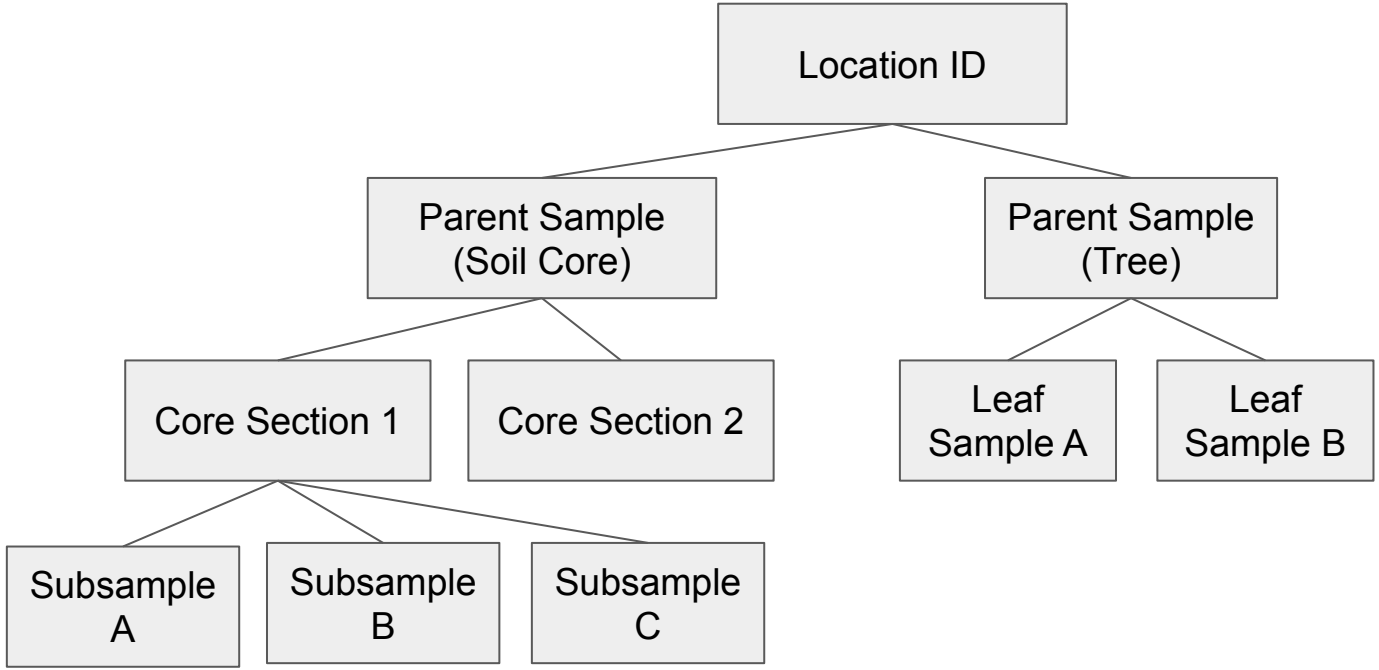
Collection ID

Sampling Event ID

Location ID

Parent Sample

- Each sample: record parent



Link related samples using identifiers

Collection ID

Sampling Event ID

Location ID

Parent Sample

- Each sample: record parent