Opening Doors to Physical Sample Data Discovery, Integration, and Credit

DO NOT EDIT HERE: NEW VERSION
https://docs.google.com/document/d/1YLO64udvUHRrzHICH2cuf3UHaZNLKi8yQB4zsfNWquw/edit

PLEASE ADD or REMOVE YOUR NAME IN THE APPROPRIATE AUTHOR LIST BASED ON LEVEL YOU WANT TO CONTRIBUTE

Lead Authors (frame paper/organize sections, write section drafts, review and and edit content): Joan Damerow, Andrea Thomer, Natalie Raia, Val Stanley

Contributing Authors (contribute ideas, write portions/use cases, review and edit content): Megan Carter, Neil Byers, John Kunze, Chad Lanctot, Kerstin Lehnert, Dylan O'Ryan, Sarah Ramdeen, Marcella McIntyre-Redden, Stephen Richard, Dave Vieglais, Charles Parker, Elisha M Wood-Charlson, Lesley Wyborn, Rorie Edmunds, Esther Plomp, Erin Robinson, Anne Thessen, Saebyul Choe

Table of Contents

Introduction: the need for physical sample discovery, sharing and citation infrastructure

Physical Samples (for instance, a piece of geologic outcrop, a soil core section, a taxidermied animal taken as a representative of a species, or the residue of some material created through a laboratory experiment (SOSA ontology)) and their associated data are some of the primary building blocks across a wide range of environmental and biological research. Yet despite their importance, finding samples and tracking their use remains persistently challenging. Samples are scattered across numerous repositories, museum collections, and personal collections. Despite efforts to create repositories of sample metadata to improve discoverability (for example SESAR, GBIF?), metadata describing the vast majority of samples is still inaccessible. And though many institutions request that their samples be cited in any paper resulting from their use, this is rarely done in practice, let alone consistently done in a way that would support harvesting these citations to track use over time. Consequently, research that uses samples is rendered less reproducible; samples are made harder to find and reuse; and sample collections are unable to show the impact of their collections and curatorial work.

To try to solve these problems, the Earth Science Information Partners (ESIP) Physical Samples Curation Cluster (a forum for the community supporting physical samples in the Earth, space, and environmental sciences) has been developing best practices for the citation of physical samples in scientific research. This paper was conceived and written as a collaborative activity of this working group. Authors include individual researchers who collect and work with physical samples, curators and collections managers, and cyberinfrastructure providers and developers. We are part of a larger community in ESIP, which is a non-profit organization supported by NASA, NOAA, USGS, and 130+ additional member organizations. The ESIP mission is to support the networking and data dissemination needs of its members and the global Earth science data community by linking the functional sectors of observation, research, application, education and use of Earth science.

The goal of this paper is to share our group's current guidelines for sample citation, as well as to outline broader unmet needs for sample sharing and citation infrastructure. We first present 5 use cases demonstrating key ways in which sample metadata sharing, citation and tracking need to be improved:

1. Document sample provenance, and provide credit for sample collectors and repositories;
2. Enable credit for facilities and laboratories conducting sample analyses;
3. Connect related interdisciplinary sample data and other research outputs;
4. Efficiently cite large number of samples; and
5. Handle legacy collections.

These use cases are drawn from projects undertaken by some of the authors' own work. For each use case, we describe the need for sample citation/tracking for the given use case and provide examples where recommended practices for sample (meta)data publication and citation were applied to the extent possible. We assess remaining problems, and infrastructure needed to make sample data more Open and FAIR.

Second, we present recommended future practices developed by the ESIP Physical Samples Curation Cluster on how to make Samples and associated data more findable, accessible, interoperable and reusable (FAIR) across disciplines. These were developed by our group by discussing our use cases, and by reviewing parallel efforts in other communities (e.g., through efforts such as the new Research Data Alliance (RDA) Complex Citations Working Group, the ESIP Physical Sample Curation Cluster, the Internet of Samples (iSamples Project and the RDA Physical Samples and Collections in the Research Data Ecosystem IG). Our guidance includes recommendations for using Sample identifiers, describing Samples, publishing associated collections of Samples (e.g., a dataset that involves numerous Samples that are part of a research project), and referencing Samples in your journal publications.

In developing recommended future practices, we identified a number of areas where significant infrastructure development is needed, or in which community practices and norms need to change. We conclude by describing these gaps, and the ways researchers, publishers, and infrastructure developers must contribute to making samples open.

We recognize that not all Samples and Sample data can be fully open. Samples that are sensitive or restricted must be protected through appropriate access controls (e.g., permits, access moratoriums). Samples should be as open as possible and as closed as necessary. For Samples related to Indigenous Peoples and lands, authors should consult the CARE Principles for Indigenous Data Governance.

Definitions

Here, the acronym “PID” is shorthand for a “globally unique, persistent, and actionable (resolvable) identifier”. For example, ARKs (Archival Resource Keys), DOIs (Digital Object Identifiers), Handles, and IGSNs (International Generic Sample Numbers) are all kinds of PIDs. In contrast, while UUIDs (Universally Unique Identifiers) are globally unique, and Darwin Core Triplets somewhat unique, they are not PIDs because they are not by themselves actionable.

On the other hand, any identifier that is unique within a given local scope can potentially be “promoted” to a PID by prepending a globally unique string to identify the scope (e.g., the museum, laboratory, or project that assigned the identifier) and then encapsulating (embedding) it in a URL (Uniform Resource Locator). The actionable, encapsulated form of a PID is strongly recommended, but if a shorter form is needed in some situations, a compact form may be acceptable if its prefix (below) remains intact and it comes from a well-known scheme.

The term “compact identifier” means a PID that is displayed without the actionable part (domain name and “https://” or “http://”) in front of it. Instead it consists of a prefix such as (ark:, doi:, handle:) to identify the PID type, followed by a globally unique string within the scope of the prefix. [compact-ids – https://doi.org/10.1101/101279 Uniform Resolution of Compact Identifiers for Biomedical Data, bioRxiv, August 2017] Each PID can be shown in compact format or actionable format (non-compact).

A compact identifier is made actionable (resolvable) by encapsulating it in an implicit or well-known resolver. If you cannot remember the resolver name, encapsulation can be done by prepending “https://n2t.net/” or “https://identifiers.org/” in front of it. Both of these “meta-resolvers” know how to resolve over 600 types of identifiers – ARKs, DOIs, Handles, RRIDs (e.g., n2t.net/rrid:AB_262044). In contexts that support hyperlinks (e.g., HTML, PDF), best practice is to display the compact form to readers and to supply the fully encapsulated URL form beneath it to support unambiguous resolution by browsers and other software agents.

Individual Sample metadata is different from “Sample dataset metadata”. The former tends to be more domain dependent while the latter tends to be more generic, consisting largely of descriptive information such as title, authors, abstract, and methods.

Needs and Current Efforts for Sample Use Tracking

Current Sample publication and citation practices

Describe how people are publishing and citing Samples currently (address if/how they are publishing metadata, datasets, referencing identifiers), and strengths and weaknesses of different options for referencing Samples.
Existing publishing and citing practices for Samples currently constitute a broad web of approaches…. This existing practice has trickle-down effects, including transcription errors for locational data and Sample names from paper to paper. Critical descriptive metadata remains locked in text and figures, with the cumulative knowledge generated on the Sample not easily accessible.

- Ask cluster to provide examples of how Samples are cited
    - Andrea has some examples: natural history and taxonomy - paleontology cites catalog numbers
    - Ask JGI folks to contribute to this
        - [Neil B.]:
            - How Samples are currently mentioned in JGI's context (anecdotal, though I can find some examples):
                - Lab-internal numbers/tags that have no external meaning
                - Plain language descriptions without mention of any standardized metadata (i.e. 'Water Sample 2 from X depth of Y lake')
                - Specific physical Samples from which data being analyzed were derived are not mentioned at all
            - Instead of citing Samples, my impression is that folks are more likely to cite data produced from examples because that is what actually goes into analyses being presented in a given paper. As a result, we need ways to (probably outside the scope of this paper, so can be mentioned maybe in passing):
                - 1) identify data citations
                    - This is messy and hard. Data is cited in myriad ways, most of which do not involve actual identifiers
                - 2) trace citations of data back to physical Samples from which that data is derived.
                    - Each organization likely has means for doing this, but a universal Sample ID system across DOE facilities would likely help bring together this information between different facilities.

- Geological Survey of Alabama and State Oil and Gas Board of Alabama publications tend to reference the oil/gas well name (non-unique) with permit number (unique in ALabama) and/or API number (unique, at least in the US) the first time and then just well name or permit number. Big data tables (think appendices) usually list name and permit number; API is also listed if it's a more oil/gas leaning pub.
    - For example: …"the core from the John Smith 4-13 #1 well (AOGB permit # 34567, API 01-297-88888-00-00)...10 core plugs from the John Smith 4-13 #1 well were analyzed for ...."
      Or "the cuttings from the J.Smith 4-13 well (OGB permit #45678, API 01-203-88888-00-00) showed…fossils were found in the permit #45678 well's cuttings in the Tuscahoma Formation…"
- Non-petroleum related cores and cuttings are more likely to be referenced by owner and well name or project and well name. For example: Town of Flomaton well #2, Jacksonville Fault Project corehole 3b.

Example Use Cases: Sample Citation to Open Sample Data

The following section illustrates specific use cases demonstrating the need for open and FAIR samples to better support science and ensure credit for researchers and institutions. Each use case will provide a.) a brief background on the specific use case needs for samples; b.) how we applied best practices in publishing and citing sample (meta)data to the extent possible; and c.) an assessment of the remaining problems and infrastructure needs.

Use Case 1. Document Sample Provenance, and Credit for Physical Sample Collectors and Repositories

Current practices for sample publication and sample data citation

Citing samples and tracking sample provenance is crucial for giving credit to the repositories and collection managers that curate and manage samples over time. Many of these institutions are regularly asked to show the impact of their collections so as to justify their work and acquire funding. When samples are not cited, they are unable to show the importance and value of these collections, and thus future funding may be in jeopardy. Further, the individual collection managers, repository managers, and other data stewards are less able to document their contributions to science and scholarship (Thessen et al 2019).

Currently, there are several obstacles to tracking sample citations to provide credit for physical sample collectors and repositories. We illustrate this via two case studies: first, work done by author Thomer and collaborators attempting to use text mining to pull natural history sample identifiers from the literature, and then [description of the other use case]…..

Given that instructions for authors are rare, and vague at best, current practices vary significantly. In this paper (DOI: 10.1126/sciadv.add0610), four separate sample repositories are mentioned in the Acknowledgements section. Thirty-three individual samples from marine sediment cores are shown on a map and listed in a table shared as a Supplementary Text pdf. The samples are listed with latitude, longitude, and depth. While this information is indeed useful, the identifiers shown in the table are informally abbreviated, and current archives are not listed for each parent sample, thereby reducing the reproducibility for this study.

Use Case Needs

Applying recommended practices and remaining challenges

Scraping NHM identifiers from the literature: work with the University of Michigan Museum of Zoology

The University of Michigan Museum of Zoology (UMMZ) mammal division contains over 140,000 specimens used in a broad range of scientific studies. Each of these specimens has been assigned a catalog number: a unique identifier within the UMMZ that is associated with both the physical specimen and any associated metadata. To track the use of their collections, mammal division collections staff, led by collection manager Cody Thompson, ask researchers who use the collection to a) include catalog numbers in any subsequent publications, b) acknowledge the use of the collections in any subsequent publications, and c) send the collections staff any papers that result from use of the collections. Thompson and his team maintain a bibliography in Google scholar that lists these papers, as well as papers authored by collection staff (dating back to the early 1900s) and students (https://scholar.google.com/citations?hl=en&user=OmzJW7UAAAAJ).

While the Google Scholar page shows one form of impact -- the papers in this bibliography have received over 91,000 citations, and its h-index is 118 -- it is still just a heuristic of specimen use. Because papers *by* the collection staff are mixed with papers *using* the collection, it does not show the impact of specific specimens over time, and therefore does not precisely show the impact of collections management. In an effort to more precisely show the impact and use of the UMMZ mammal collections, we tried multiple kinds of text mining pipelines to first extract catalog numbers, and then generate metrics to show their use. The results were somewhat underwhelming: of the 1297 papers we analyzed, only 245 included catalog numbers. This was much lower than expected; while we expected the corpus to include papers that didn't include specimen numbers, we didn't expect that it would be almost 90% of the papers. After reviewing the results, we identified the following obstacles to mining specimen citations from legacy literature:

1) Scholars only cite the collection, not the specimens. Often, the UMMZ was thanked in the acknowledgements of the paper, but catalog numbers were not explicitly included in the paper
2) Scholars include catalog numbers in supplementary material, not in the text of the paper. In some cases where we knew specimens were used in analysis, we found that they had been cited in appendices of supplementary material. In future work we could try mining these, but it would be a challenge because they are so heterogeneously formatted.
3) Scholars cite identifiers other than the catalog number. In some cases where genomic analysis had been done, paper authors include GenBank identifiers, or other locally assigned identifiers instead of the catalog number used by the institution.

Smithsonian and Earthchem (mapping to other names that appear in literature; the VG Samples example)

Use Case 2. Credit for Laboratories conducting Analyses (e.g. JGI)

When data becomes publicly available, it takes on a life of its own. A laboratory that publishes data or provides biological samples loses control over provenance as soon as it ends up in the hands of a third party. Approaches that preserve provenance and accumulate metadata in a consistent manner are greatly needed.

Current practices for sample publication and sample data citation

In the microbiology literature, the standard practice for citing samples is to refer to an isolate by a strain identifier or culture collection accession. A strain identifier is not typically unique, and can lead to ambiguity in the literature and public databases. Culture collection accessions, while not participating in a namespace, are more easily identifiable due to their typical use of a three or four letter repository prefix coupled with an integer that uniquely identifies the sample with a specific collection. However, the lack of standards (i.e., accession format, metadata retrieval) among these repositories limits the utility of these accessions in search and indexing. Further complicating the ambiguity of isolates is the potential loss of provenance during transfers among collections, especially for historical samples.

In the genomics literature, the ambiguity of strain identifiers is mitigated by the use of biosample accessions (i.e., a GenBank BioSample). The metadata associated with these identifiers is generally rich, providing information about the source of the isolate (e.g., location, host, organization, personnel), as well as references to associated projects, genome assemblies, and sequence data. A plethora of additional identifier classes exist for biological projects, analyses, and sequence data. These accessions and associated metadata provide near-ideal unique identifiers that lend themselves to efficient retrieval and literature search, although the cardinality of the mapping among these identifiers must be respected in order to prevent ambiguity (i.e., a project may be associated with multiple biological samples). The Genomic Standards Consortium developed a set of recommendations for describing aspects of a genome, (including the biological sample), in a consistent manner. However, consistent

application of the recommendation was never fully realized due to variations in interpretation by authors, editors, and publishers. Future recommendations will need to be clearly and fully articulated to publishers.

The usability of rich BioSample metadata is limited by the use of non-standard references to authors and affiliations. For example, the Joint Genome Institute might be referenced as “JGI,” “DOE Joint Genome Institute,” “DOI-JGI,” or a number of other variations. The use of a curated set of organizations would greatly improve the consistency of indexing and search of affiliations, as well as mapping authors to organizations. Such a resource is currently being developed by ROR and CrossRef.

Author names are historically problematic for search and retrieval, due to the use of initials and lack of care when storing these in databases (e.g., different treatment of diacritics in storage and indexing). Many publishers and compositors rely on private lists of previously-known author-to-affiliation associations, an approach that is likely to be error-prone due to changes in affiliation or multiple affiliations being associated with an author. Fortunately, ORCiD now provides a standard unique identifier for authors, which removes any author ambiguity for publications and data providers that utilize them.

Use Case Needs

- Sample identifiers that lend themselves to efficient indexing (uniqueness, no white space, easily used as part of a URI).
- Consistent methods of search and retrieval of metadata associated with a sample (URL formats, API standards, metadata formats).
- Baseline cross-disciplinary metadata standards (fields, field names, agreement on coded values, namespaces for linking to external resources).
- Consistent association of author (or contributor) and affiliation data, including, potentially, a contributor role type (i.e., person who originally collected a sample) that would indicate the form of credit that should be attributed to an author or institution.
- Propagating provenance that adheres to the above recommendations.
- Clear instructions to authors, editors, and publishers.

Applying recommended practices and remaining challenges

Historical literature and much contemporary literature will remain a challenge for data citation discovery. However, the application of the above recommendations will greatly mitigate the problem for new publications that adopt these practices. At minimum, the consistent and widespread adoption of ORCiD and ROR identifiers in metadata would remove many of the barriers to accurately associating data products and publications with organizations.

Widespread adoption of standards for consistent sample metadata across different institutions may be challenging due to differences in operating procedures and technical or institutional barriers to change, such as mapping lab-internal metadata (i.e., from LIMS) to more standardized metadata fields. However, participating in an ecosystem that enables consistent

tracking of sample and data use may be enough of a motivating factor for many organizations to adopt some of these proactive practices.

Fully solving the current shortcomings in data attribution also requires attention to the historical literature. The JGI has developed a retroactive solution in-part to the data citation problems in the existing literature. The Data Citation Explorer [manuscript in progress] is a web service that implements heuristics for identifying use of genomic data products in published literature, even when those products are not properly cited. Current work on this system involves context analysis to determine not only who is using genomic data, but also how that data is being used to support a publication.

Use Case 3. Connect Interdisciplinary Sample Data and Other Research Outputs

Current practices for sample publication and sample data citation

Use Case Needs

The U.S. Department of Energy's (DOE's) Environmental Systems Science (ESS) program is highly interdisciplinary, and projects have faced a particular challenge in tracking Samples and resulting data that was sent to numerous collaborators and labs, for a variety of analyses, and then compiled, analyzed, and published in numerous files and data systems. Samples are often used to enhance models and predictions of ecological processes and biogeochemical responses to contamination, warming, and other disturbances. A common workflow is to collect a Sample (like soil or water), and then send it out for a combination of microbial metagenomic analyses and a variety of physical/chemical analyses. The process of submitting Samples to different data systems and labs, and then compiling the resulting data is currently inefficient and even prone to error.

Sample PIDs and standard metadata are the foundational elements necessary to coordinate and cross-link across DOE's Biological and Environmental (BER) data systems dealing with interdisciplinary Sample data (Damerow et al. 2021).
For example, the River Corridor and Watershed Biogeochemistry Scientific Focus Area (a project funded by the DOE ESS program) studies hydrologic, biogeochemical, and microbial function within river corridors (Stegen and Goldman 2018). They collected a series of individual surface water (e.g., igsn:10.58052/IEWDR00RT) and sediment Samples (e.g., igsn:10.58052/IEWDR0149) from specific sites (e.g., igsn:10.58052/IEWDR00P4) in the field, extracted DNA and RNA material from these Samples (subSamples/child Samples; e.g., igsn:10.58052/IEWDR00UI), and then sent them to JGI for metagenomic and metatranscriptomic sequencing, and the Environmental Molecular Sciences Laboratory (EMSL) for metabolomics analyses

(Figure X). They obtained raw data from these respective online data systems, and did their own further data processing, analysis, and visualization. They created Sample sets and documented their workflows in the DOE Systems Biology Knowledgebase (KBase; Borton 2022). Analysis and visualizations from the Sample set can then be incorporated into a formally published dataset for long-term preservation and documentation, associated with a DOI in the ESS-DIVE data repository (Toyoda et al. 2020; Goldman et al. 2020). This dataset can then be referenced in the final journal publication associated with the data. All of these entities (i.e. project, site, parent-child Samples, workflow, datasets, journal publication) are important to link and track related Sample data for this interdisciplinary work.

Applying recommended practices and remaining challenges

The ESS-DIVE data repository now has 26 datasets (and counting) that have obtained IGSNs with associated standard metadata, now compiled into a data portal (https://data.ess-dive.lbl.gov/portals/ess-samples). We worked in-depth with three projects that have interdisciplinary data similar to that described above to test how we could track analysis and use of Samples sent to multiple labs and facilities and then published data and associated papers. This started with an evaluation of how well SESAR IGSN metadata and identifiers worked for sample tracking needs for Environmental System Science projects (Damerow et al. 2021). Each project subsequently sent samples for laboratory analyses, conducted their own data processing and analysis, published one or more sample datasets (total of 7 datasets at the time of publication), and published one or more associated papers (Table X). We worked across five laboratories and data systems to ensure that the IGSN was recorded as the original source material sample consistently across all relevant systems, including: 1.) National Microbiome Data Collaborative (NMDC), Joint Genome Institute (JGI), 3.) DOE Systems Biology Knowledgebase (KBase), 4.) Environmental Molecular Sciences Laboratory (EMSL), and 5.) ESS-DIVE data repository.

To document connections between (citations of) related Samples, datasets, workflows, and associated papers we did the following:

- Author Checklist Step 1: For each sample, we documented standard metadata using SESAR IGSN requirements, extended for Environmental System Science samples (Damerow et al. 2021; Damerow et al. 2020; System for Earth Sample Registration ...; System for Earth Sample Registration ...). This included documenting the parent IGSN where relevant, to track connections between related samples.
- Author Checklist Step 2: Each sample was registered for IGSNs through SESAR.
    - Sample relationships were recorded by documenting Parent IGSN.

SESAR landing pages then populate links to all related parent and sibling samples.
    - IGSNs were listed as the source material sample IDs when analyzed at JGI and EMSL.
    - We manually updated sample metadata for individual IGSNs when analyses were completed and published on JGI and ESS-DVE, by providing related URLs for the samples (e.g. doi:10.15485/1603775, doi:10.15485/1729719). These links to sample data are now presented on the sample landing page (e.g. IGSN:10.58052/IEWDR00RF).
- Author Checklist Step 3: Across these projects, we published seven datasets that included Samples with PIDs/IGSNs. Each dataset includes IGSN urls in the dataset metadata (within the methods section), and includes the sample metadata file in the dataset. Each data file in the dataset should contain the IGSNs as the first column. However, some projects opted to use the sample name in the data files and rely on the metadata file for connecting the sample name and IGSN.
    - The ESS-DIVE data repository does not have a specific metadata field currently for samples or related identifiers. But have determined this to be a future need to better support sample tracking. To do this, the data repository can implement related identifiers and provide a user-friendly way to provide a list of samples associated with the dataset (e.g. be automatically extracting the IGSNs from the sample metadata and/or data files).
    - The data repository could enforce IGSNs in sample data files, or provide an automated way to connect sample names to IGSNs as we develop tools for advanced search within data files. Scientists generally want to use their own sample names in their files because IGSNs don't mean anything to them.
    - We also had inconsistency in how people provided the sample metadata file. We did not have automated validation for these files when the datasets were submitted, and some did not include the same metadata file provided to SESAR. We should just ask for the IGSNs, and use the SESAR API to harvest relevant metadata for the samples that would support sample search (e.g. by material type and environmental context). However, we don’t have resources to build this functionality at the present time. 5
- Author Checklist Step 4: The sample datasets in the ESS-DIVE data repository were cited in associated journal publications.
    - We did not cite individual samples in associate papers. This is because it was not practical to do this using current citation practices and

infrastructure.

The Sample PID provides a powerful way to link and exchange relevant scientific information across facilities and data systems. In turn, this supports the FAIR data principles, and improves the ability for future scientists and others to find, compile, interpret, and reuse interdisciplinary biological and environmental data in synthesis research.

Kerstin - people at Lamont who do work earth and bio work with soil scientists

Use Case 4. Cite Large Number of Samples

Current practices for sample publication and sample data citation

The Interdisciplinary Earth Data Alliance (IEDA2) is a collaborative data infrastructure funded by the NSF via a cooperative agreement and consists of three complementary data systems – EarthChem, LEPR/traceDs, and the System for Earth and Extraterrestrial Sample Registration (SESAR). This use case focuses on interactions between EarthChem, which offers multiple data services geared towards geochemical research, and SESAR, which provides IGSN minting and sample discovery services for a broad range of earth and space science samples. EarthChem provides data publishing services through the EarthChem Library (ECL), and data discovery and analysis through PetDB and the EarthChem Portal. The EarthChem Portal operates as a collective system where users can search for geochemical datasets from PetDB, SedDB, NAVDAT, MetPetDB GEOROC, USGS National Geochemical Database, and GANSEKI.

IGSNs can be used by researchers to permanently store metadata about their samples and link their samples to published datasets and publications.

IGSNs minted through SESAR can be added to a dataset in ECL via two methods:

(1) They can be added by the user as related information along with volcano and cruise DOI lookup in the submission
(2) They can be added as part of data submission using the ECL template. Before a dataset is published in ECL, curators perform an IGSN file check, where the system ingests any IGSNs listed in the data file.

Datasets for metamorphic and igneous rock samples are ingested into the PetDB synthesis through datasets published in ECL. The PetDB data curator manually adds the sample metadata including IGSNs, however these are not officially ingested into the system and IGSNs do not redirect to the sample landing page. This portal has IGSN search capability for geochemical datasets. However the search is similar to PetDB where the IGSN is not linked.

- As of November 2023, of EarthChem Library's 1,336 published datasets, 354 included links to IGSN's.
- To do: include a few examples of linked datasets
    - Ex. 1 Dataset: https://doi.org/10.26022/IEDA/112300.
    - Ex. 1 Sample (parent): https://app.geosamples.org/sample/igsn/10.58052/IENHR0002
    - Ex. 1 Sample (child): https://app.geosamples.org/sample/igsn/10.58052/IENHR0064

Use Case Needs

Applying recommended practices and remaining challenges

There are several challenges in linking IGSNs to EarthChem. Many of these are infrastructure challenges, where better development and connection is needed between the systems and the confusion in IGSN name use (with or without prefix). First challenge is in linking datasets and publications on IGSN landing pages in SESAR. Currently, SESAR requires authors to either provide this information at sample registration or to update this information as datasets and publications are produced. Including the IGSN in ECL does not allow for backwards integration at this time.

Second, in ECL, only IGSNs that are minted through SESAR are recorded. IGSNs from other allocating agents are not verified under the ECL system. Part of this issue is a result of the addition of the prefix during the IGSN e.V. and DataCite partnership. Since IGSNs are now considered DOIs, the ECL developer needs to manually include the verified prefix as part of the code for IGSN validation.

Third, PetDB and ECP have limited linking to IGSNs. Users can search for datasets based on IGSNs, however the IGSN themselves are not linked to the sample landing page in SESAR. A separate search in SESAR is needed for sample metadata. Additionally, PetDB and ECP at this time have limited search capabilities. Grandfathered IGSN (without the prefix) search is successful, however the same IGSN with the prefix does not exist, as the system pulls static information from previously published data and metadata.

- Example: IGSN:ODP02OBQF vs. IGSN:10.60471/ODP02OBQF (https://app.geosamples.org/sample/igsn/10.60471/ODP02OBQF&header=1)

- **Recommendations and Best Practices**
    - Infrastructure Updates
        - Automatic reference and updates to related URL field for sample profiles that are cited in publication
        - Linking between grandfathered IGSNs with unregistered IGSNs (with prefix) is needed within the systems. There are developments in progress for ECP and PetDB for a new interface and possible linking functionality to IGSNs.
        - There should be an option to download IGSN list and weblinks on submission landing pages in ECL as a separate document from the dataset itself.

        - Clearer communication of the prefix use is needed to users. Many authors are willing to use and cite IGSNs. To make this information more accessible, SESAR and other allocating agents should have a “how to cite this sample and IGSN” section that gives direct reference authors can use in their manuscript.

- Dave Vieglais - iSamples work on RDA complex citation (reliquary)

Use Case 5. Dealing with legacy collections

- Rorie on legacy identifiers in literature
- Possibly something from Val?

Samples that were collected and stored before modern curation management systems and methods are considered legacy collections. Often these samples have limited metadata and ambiguous history that have legal and social implications.

Recommended Future Practices for Sample Citation

Related Identifiers and Connection Metadata

[ewc] - Samples with PIDs can be linked to other identifiers using defined relationship types. This includes other samples with PIDs (parent-subsample as “IsPartOf”, or parent-child as “IsDerivedFrom”, and data sets derived from the sample set (data set DOI “Cites”). Connecting sample PIDs to all downstream sample/research products by “PID+relationship type” enables DataCite to automatically create and track directional linkages.

Compact Citation for Large Numbers of Samples

[ewc] - The linking of PIDs is relatively straightforward when researchers use a single set of samples with the same collection metadata (sample owner, collection source, etc). Depending on the science question, researchers may need to mix and match sample collections into new sample sets for analysis. These new sample sets become a novel, downstream research product for each of the original sample collections, in part because they have a new research purpose that is derived from or benefited by the original collection. The recommendation is for researchers to work at the sample collection facility or data platform where they will release their results to develop a new collection PID that “Cites” the original sample PIDs. This new collection PID should have an HTML landing page that lists which samples from each collection were used.

Community Needs

Scientists - Adoption of Standard Sample IDs and Metadata

Scientists should cultivate a shared culture that supports procurement and usage of Sample IDs as standard scientific best practice. This can be cultivated through the mentorship and training of early career researchers, regular adoption of Sample IDs as a data management tool in data management plans, and guidance and support for retiring researchers and their collections, for instance. Sample repositories and managed collections can play supporting roles by providing opportunities for training and Q&A via webinars, workshops, conference booths, office hours, and funding agencies should support these important forms of outreach and training.

Discipline-specific metadata standards

Publishing - Journal and Data Publishers

In the same way that journal publishers provide data and software citation guidance, publishers should provide explicit author instructions on where and how to cite samples in publications. This includes information about how to encode sample IDs so that they become linked in the publication process. This guidance should outline procedures for all components of a paper (i.e., how to cite sample ID's in line in text, in tables, and how they should appear in Data Availability statements or sections).

Development, training, and uptake of editorial checks for ensuring consistent linking of Sample IDs is a significant undertaking. Editor and reviewer guidelines can play a supporting role in encouraging authors to use Sample IDs where appropriate.

Infrastructure Needs

Sample IDs and Metadata Management (iSamples)

Create Complex Data Citations

Track use of Samples

How to Plan and Publish Research that includes Physical Samples

These guidelines offer guidance for describing and referencing Samples used in scientific publications, with links to additional information and examples to assist authors in meeting these requirements and recommendations.

Step 1. Use Identifiers for Samples

Sample identifiers are typically used to track a Sample as it moves through its lifecycle and physically from location to location. In particular, they enable a source Sample to be easily recognized when it is subSampled or undergoes multiple analyses by one or more research groups. However, they are also essential for linking a Sample to other, associated research outputs, as well as ensuring that researchers are appropriately credited for their contribution to collecting and managing the Sample.

Particularly in cases where Samples will be stored long-term, analyzed in multiple labs, and/or used in multiple publications, we recommend assigning Samples globally unique, persistent, and actionable identifiers (PIDs) that resolve to a landing page containing rich metadata describing the Sample. This metadata can be readily accessed and enhanced over time through automated or manual processes as new information is created. Acceptable examples include: Handles, Archival Resource Keys (ARKs), Digital Object Identifiers (DOIs), and International Generic Sample Numbers (IGSN IDs). PIDs can encapsulate or supplement human-readable local Sample identifiers, and enable tracking of the provenance and use of material Samples, as well as capture the hierarchical (parent–child) relationships among them. Importantly, PID metadata can include information leading to associated data, papers, researchers, funders, and more (which ideally should also all have their own PIDs).

Samples are often assigned locally unique identifiers encoded with meaningful information to enable Sample collection and process management. While useful, when that information is subject to change (e.g., the project, funder, or department name), it threatens the stability of the original identifiers because researchers will tend to update identifier strings to match, and to abandon the old strings. Characteristics based on analysis or computation (e.g., composition, type), while unlikely to change, are actually unstable enough over time (e.g., re-analysis with updated algorithms and technology) to be unreliable as identifier strings.

Some of these issues are addressed by using strings such as Darwin Core Triplets, UUIDs, and ULIDs (Universally Unique Lexicographically Sortable Identifiers) that are considered to be effectively globally unique. However, unless promoted to PIDs, such identifiers do not allow easily monitoring Sample usage across data sources and publications or access to associated metadata.

When you are unable to assign PIDs to your Samples, best practice is to assign distinct, opaque identifiers (e.g., UUIDs, ULIDs, Darwin Core Triplets, RRIDs) and standard metadata. In such cases, associated datasets should consistently identify Samples using the unique Sample identifiers; do not use partial or approximate variations of the identifier in different files and publications.

Note that this guidance concerns the assignment of new Sample identifiers. Ideally, you should not assign new identifiers to Samples that already have PIDs. Existing identifiers (whether PIDs or not) should always be preserved. [tdwg-recs – Community Next Steps for Making Globally Unique Identifiers Work for Biocollections Data, 2015, DOI: 10.3897/zookeys.494.9352]. In cases with usable pre-existing PIDs, please consult the recommendations in the 'Reference Samples Used' section.

How to Assign and Use Sample PIDs

As an individual researcher, you may obtain PIDs from an organization that registers the PIDs and provides long-term support for maintaining Sample metadata and landing pages. This may include:

- **Internally within your institution**. Depending on its organizational structure, PIDs are typically registered by a centralized information service or data repository within an institution (e.g., the university library). This is very often the case when an institution is registering DOIs or IGSN IDs via DataCite. PID registration is sometimes also possible through your Sample repository or even through an individual laboratory, which may then have tools or processes in place to assign PIDs during standard analysis or archival workflows.
- **Via an external organization**. Several DataCite Members offer free DOI and/or IGSN IDs registration services to individual researchers (e.g., the System for Earth Sample Registration [SESAR]). SESAR currently provides the most open and user-friendly system for individual researchers to register Sample PIDs, including tools to register, resolve, and manage PIDs and metadata (see next section for guidance about Sample metadata). IGSN IDs are interdisciplinary and suitable for all Sample types. Likewise, services also exist that offer free ARK or Handle registration (e.g., EZID, operated by the California Digital Library).

It is important to use these Sample PIDs in your associated datasets and papers. PIDs for Samples should be included in the text or tables of journal articles, and/or in any associated dataset(s) using the compact format with actionable hyperlinks underneath, or by providing the full URL (e.g., IGSN:10.58052/IEGRW002B; see Reference Samples section for more detail).

Note that a PID should be registered only by the Sample's owner. If you are not the owner (e.g., you are working with a Sample on loan from a repository), you should request the owner to assign a PID to the Sample. This includes any child subSamples split off from a parent.

Step 2. Describe Samples

Sample metadata includes basic information about the Sample--what it is, when and where it was collected, who collected it, who currently owns it, and the current physical location if the Sample is accessible. Detailed metadata can help other researchers find, access, and/or integrate Sample (meta)data for synthesis or new research.

Some disciplines have specific metadata templates that include both general metadata listed above, along with additional discipline-specific requirements and recommendations for describing physical Samples.

We recommend that you create a metadata file/table that describes basic characteristics of the Samples used in your paper, and how they were collected and archived (if applicable). Use an existing community standard and associated template, according to your Sample type:

- Geoscience Samples - SESAR IGSN metadata
    - Batch Sample registration tutorial
- Ecosystem sciences Samples - IGSN for Environmental System Science metadata, which is compatible with SESAR IGSN with some additional/revised terms for interdisciplinary biological and environmental Samples
- Biodiversity collections or species occurrence records - Darwin Core standard
- Genomics Samples - Minimum Information about any Sequence (MIxS) standard

Example metadata that should be included for each Sample to enable Sample search, integration, and reuse includes:

- **Sample Identifier**: a unique compact identifier, ideally a resolvable PID, that can be used to identify the Sample and link it to other metadata. The identifier type should be apparent in the compact identifier's prefix, for example, "igsn" in "igsn:10.58151/NHB000QPS", with a hyperlink.
- **Sample Name**: a project-specific unique Sample name, which often includes meaningful codes that can fit on a Sample label (e.g. combination of location, Sample type, and date; StrawCr-Soil-20230530.
- **Sample Type**: The basic form of the object that is registered (e.g., core, individual Sample, organism).
- **Material**: The material or materials that compose the Sample (e.g., soil, sediment, rock, water).
- **Collection Date**: Date that the Sample was collected, in ISO 8601 format (YYYY-MM-DD HH:MM:SS, e.g. 2023-09-22 15:00:00). Note that this may include time as well, or a date range.

- **Collector or Chief Scientist**: Person(s) who collected or own(s) the Sample. Chief Scientist can represent the Principal Investigator of a project. Ideally, the collector(s) or chief scientist(s) would be associated with an ORCID.
- **Local Environmental Context, or other Context Category (If applicable):** Local physiographic feature or environment type from which a Sample was collected. For example lake water, seamount, Cathedral Peak granite, soil biome, submarine hydrothermal vent, atmospheric particulates, intertidal zone, abyssal plain, stream bed load, contact metamorphic aureole.
- **Geographic Coordinates (if applicable):** Sample latitude and longitude in WGS84 decimal degrees.
- **Collection Locality (if applicable):** A description of the location.
- **Minimum Depth in Meters (if applicable):** Depth, or minimum depth (if taken over a range) at which a Sample was collected, below ground or under water.
- **Maximum Depth in Meters (if applicable):** Maximum depth at which a Sample was collected, below ground or under water.
- **Collection Method:** Description of the collection method for the Sample. Include any important terms and details for potential users to understand how your Sample was collected.
- **Related Identifier (If applicable):** Persistent Identifiers assigned to related resources, which may include a specific type of identifier (e.g., Parent IGSN, Site ID, Associated Paper DOI), or a general related identifier and a relation type (see DataCite's Related Identifier description).
- **Current Archive (if applicable)**: Sample repository or collection where the Sample is deposited. E.g. *La Brea Tar Pits Museum, Los Angeles, CA*
- **Permits (if applicable):** If your Samples have been collected from Indigenous lands or other sensitive sites (National Parks, monuments etc), please specify the permit and/or acknowledge that these Samples were legally collected.
- **Sample Archive Manager/Curator** (if applicable): Person(s) who manage the physical Sample repository/collection.

Step 3. Archive Physical Samples and Data

3a. Archive Physical Samples and Metadata

Check with your institution or funding agency to determine any available resources for archiving physical Samples.

If you have access to a Sample repository: work with collection managers to archive your physical Samples according to the requirements of that repository (provided the Samples were not destroyed in analysis).

If you do <u>not</u> have access to a Sample repository AND if your Samples were not destroyed in analysis

Treat your Samples as a personal collection. It's beyond the scope of this document to provide detailed guidance for the curation of a personal collection. Samples should be preserved for a minimum of 5 years, or as long as they are viable, and up to as long as funding permits).

If you obtained Sample PIDs, Sample metadata may already be archived and associated with your individual Sample PIDs (e.g., SESAR), which you can manage and update over time as needed. In this case, you may not need to take further steps to archive Sample metadata, other than to summarize this information within your paper's methods section and to reference your Samples as discussed in Step 4 below.

If you have not obtained Sample PIDs associated with standard metadata, it is important to archive standard metadata as part of a Sample dataset (recommended if you use more than ~10 Samples, see step 3b), or in your associated paper (may be preferred if you use fewer than 10 Samples, see step 4).

In the case of natural history collections, Sample metadata may also be stored in the long-term collection's database (e.g., UC Berkeley Museum of Vertebrate Zoology database) that in some cases are published online. For biodiversity records, these collections databases are then sent to, integrated and published in the Global Biodiversity Information Facility (www.gbif.org).

3b. Archive Sample Data

Ideally, submit and publish Sample data in a domain or institution-specific data archive, such as EarthChem, Environmental Data Initiative (EDI), ESS-DIVE, and/or National Center for Biotechnology Information (NCBI). If there is no relevant domain or institution specific archive, you can use generalist repositories, such as Dryad, Zenodo, or Figshare.

Specific recommendations for your Sample dataset:

1. Include a standard Sample metadata file that describes each Sample. Include the original (earliest) identifier if you are using a legacy Sample.
2. Include a list of Sample PIDs in the appropriate standard format (e.g., IGSN:10.58052/IEGRW002B) within your dataset metadata (information to describe your dataset, such as title, authors, abstract, methods). The specific metadata field or fields that you use to provide Sample PIDs may vary depending on the data archive, but could be a field specific to Samples, or a free-text field such as methods. This will help ensure that your Sample PIDs are discoverable in dataset searches.
3. All Sample-related files in your dataset should consistently use the appropriate Sample Identifier (e.g., have a column/field for Sample ID, or IGSN).

See AGU Guidelines for data and software sharing for more general information about archiving data.

Step 4. Reference Samples Used in Your Paper

The goal of referencing Samples consistently in your papers and associated datasets is to help efficiently manage/track and connect related Samples and data across data systems, ensure that potential data users can access the physical Samples used, and assign appropriate credit when data is used; it is important to credit original Sample and data collectors, as well as curators of Samples archived in physical repositories.

We recommend that you include Sample identifiers in the text and/or within one or more table(s) of your papers (e.g., in the methods section, or a Sample availability statement). To do this, provide Sample PIDs in their standard compact format with hyperlink (i.e., IGSN:10.58052/IEGRW002B). If using other identifiers, list them consistently in your paper, exactly as they appear in any associated datasets or files. Do not use approximate variations for the same identifier. If you have fewer than 10 Samples, you may cite them using the standard identifier format listed above within your Sample or data availability statement.

Please note that the identifiers should be associated with standard metadata, as described in Steps 3a and 3b. The minimal metadata listed in Step 3a provides essential information necessary to find, access, integrate, and reuse Sample data. So if you have not obtained Sample PIDs with associated standard metadata, it is essential to provide standard metadata within a table of your paper as demonstrated in examples below (e.g., methods, data availability statement), and/or as a standard Sample metadata file (e.g., using a domain-specific Sample metadata template, section 3b) in an associated dataset.

If you have published an associated dataset, cite this dataset in the reference section of your paper, and/or the data availability statement.

As mentioned previously, specific recommendations may evolve as further infrastructure to share Samples and track citations is developed.

Examples

Samples from Museum of Great Samples. Samples available upon request. Contact collections@museumsofgreatSamples.edu to request Samples. Sample metadata provided below.

Sample Identifier	Sample name	Sample type	Material	Collection date	Collector	Local context	locality	Related identifiers	Current archive
IGSN:slk222-046537	Really cool rock 1	Rock	Dark purple chunk of leaverite	2010-10-29	https://orcid.org/0000-0099-9999-530	mineralized vein	SMR station 2020-018	AZGS field Sample number:	Museum of Great Samples https://ror

					2			smr2010-398	.org/someorg7523824
IGSN:slk222-046538	C. hesternus prox left ulna	Fossil	Mineralized bone	2010-10-29	https://orcid.org/0000-0099-9999-5302	Asphalt seep	Rancho La Brea, Project 23, Box 12	RLP catalog number 101933	Museum of Great Samples https://ror.org/someorg7523824
IGSN:slk222-046539	Really cool thin section	Rock	apatite	2010-10-29	https://orcid.org/0000-0099-9999-5302	mineralized vein	SMR station 2020-018	AZGS field Sample number: smr2010-400	n/a, destroyed in analysis

---

Samples from J. Damerow's personal collection. Samples destroyed in analysis. Sample metadata is available at www.repository.edu/filename.html

---

Smithsonian National Museum of Natural History, Department of Mineral Sciences. Samples available upon request. Contact Leslie Hale (halel@si.edu) to request Samples. Sample metadata is available at www.repositoryurl.com

[Author Checklist for Sample-Related Papers]

Recommendations to AGU

- Add a “Sample availability statement” at the end of papers that use physical Samples. This should be a dedicated section where authors can provide a short narrative summarizing where and how others may access their Samples, and that links to a supplemental Sample metadata file including the Cross-Discipline Metadata we outline in section X.
- Embed Sample metadata in files for crawling/harvesting. This is essential to tracking Sample use across papers, and therefore making it possible for repositories and researchers to get credit for their curatorial work, and also better enabling Sample reuse and study reproducibility. While there aren’t currently infrastructures to harvest Sample citations, infrastructures aren’t possible without a consistent method of exposing Sample metadata. Publishers must break this chicken-or-egg conundrum by surfacing Sample metadata and thereby enabling the development of Sample citation tracking infrastructure.

Methods

References

Borton, Kayla. 2022. “KBase Narrative - GROWdb US River Systems - Samples.” Systems Biology Knowledgebase (KBase). https://doi.org/10.25982/109073.30/1895615.

Damerow, Joan E., Charuleka Varadharajan, Kristin Boye, Eoin L. Brodie, Madison Burrus, K. Dana Chadwick, Robert Crystal-Ornelas, et al. 2021. “Sample Identifiers and Metadata to Support Data Management and Reuse in Multidisciplinary Ecosystem Sciences.” *Data Science Journal* 20 (1): 11.

Goldman, Amy E., Shai Arnon, Edo Bar-Zeev, Rosalie K. Chu, Robert E. Danczak, Rebecca A. Daly, Dillman Delgado, et al. 2020. “WHONDRS Summer 2019 Sampling Campaign: Global River Corridor Sediment FTICR-MS, Dissolved Organic Carbon, Aerobic Respiration, Elemental Composition, Grain Size, Total Nitrogen and Organic Carbon Content, Bacterial Abundance, and Stable Isotopes (v8).” Environmental System Science Data Infrastructure for a Virtual Ecosystem; River Corridor and Watershed Biogeochemistry SFA. https://doi.org/10.15485/1729719.

Stegen, James C., and Amy E. Goldman. 2018. “WHONDRS: A Community Resource for Studying Dynamic River Corridors.” *mSystems* 3 (5): e00151–18.

Toyoda, Jason G., Amy E. Goldman, Rosalie K. Chu, Robert E. Danczak, and Rebecca A. Daly. 2020. “WHONDRS Summer 2019 Sampling Campaign: Global River Corridor Surface Water FTICR-MS, NPOC, and Stable Isotopes.” https://data.ess-dive.lbl.gov/view/doi:10.15485/1603775.