Publishing Open Research Using Physical Samples: Guidance for Authors

Suggested Citation

Damerow, Joan; Raia, Natalie; Stanley, Val; Byers, Neil; Choe, Saebyul; Edmunds, Rorie; Kunze, John; Lehnert, Kerstin; McIntyre-Redden, Marcella; Mungall, Chris; O'Ryan, Dylan; Parker, Charles; Plomp, Esther; Ramdeen, Sarah; Richard, Stephen; Vieglais, Dave; Wood-Charlson, Elisha; Thomer, Andrea. (2024). *Publishing Open Research Using Physical Samples: Guidance for Authors v4*. Earth Science Information Partners. Online resource. https://doi.org/10.5281/zenodo.18854424

Table of Contents

Publishing Open Research Using Physical Samples: Guidance for Authors 1
- Suggested Citation 1
- Table of Contents 1
- Introduction 1
- Step 1. Describe Samples with Rich Metadata 2
- Step 2. Assign and/or Use Identifiers for Samples 2
  - 2a. Samples from long-term physical collections 3
  - 2b. Samples used in multiple analyses or publications 3
  - 2c. Subsamples sent to laboratories for analyses 3
  - 2d. Samples used only once 4
- Step 3. Publish and Cite Samples in Datasets 4
- Step 4. Cite Sample Identifiers in Paper 5
- Appendices 5
  - Appendix A: Why you should describe samples with rich metadata (Step 1) 5
  - Appendix B: Why you should assign and/or use identifiers (Step 2) 6
  - Appendix C: Why you should publish and cite samples within datasets (Step 3) 7
  - Appendix D: Why you should reference samples with consistent formatting (Step 4) 8

Introduction

Physical samples and their associated data are foundational elements of Earth, extraterrestrial, and environmental science research. By "samples," we mean physical subsets or extracts from some feature of scientific interest -- for example, a chip from a 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. Samples are used in a huge range of scientific domains and studies and are often collected and curated at great financial costs; thus, facilitating their access and use is paramount. Their metadata need to be accessible, understandable, and as open as possible for reuse to support transparency and reproducibility. Additionally, sample repositories need to be able to track the demonstrated use of samples to better manage and show the impact of their collections.

One important step toward facilitating physical sample access and use is the consistent citation and referencing of samples in academic papers. When samples are used as the basis of a scientific study, clear metadata describing the samples should be included in any publications and datasets resulting from that study. By describing and citing samples used in scientific analyses and papers, we make science more reliable and cost effective. We also make it easier for samples to be found and reused, and for sample collectors and curators to get credit for their work.

In this document, we present guidelines for authors publishing open research using physical samples. These guidelines were developed by the ESIP Physical Sample Curation Cluster to help authors of scientific papers make their sample-based studies Open, and Findable, Accessible, Interoperable, and Reusable (FAIR) to advance sample-based science in the future.

Step 1. Describe Samples with Rich Metadata

Describe key characteristics and collection details of the samples used for the paper, often by providing a sample metadata file or table. This can be a csv file with sample Identifiers as rows, and metadata fields as columns, including information on sample type, how and where it was collected, by whom, and where archived (if applicable). Use a domain-specific standard or community reporting format relevant for your sample type, such as:

- Earth science samples: System for Earth and Extraterrestrial Sample Registration (SESAR) International General Sample Number (IGSN) metadata (Quick guide)
- 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, Metadata guide for Minimum information about a metagenome sequence (MIMS).
- Interdisciplinary samples that may span multiple standards: see further reading below for recommendations.

Further reading: Why you should describe Samples with rich metadata, and additional examples

Step 2. Assign and/or Use Identifiers for Samples

Use sample Identifiers, ideally Persistent Identifiers (PIDs), to track samples and associated data (See Appendix 2). You can assign Sample PIDs, or use existing PIDs — some institutions or data systems may assign them for you. Identifiers and specific steps for their assignment may vary depending on your use case, as outlined below.

2a. Samples from long-term physical collections

If contributing to or using samples from a long-term physical collection, such as those managed by museums or otherwise professionally curated, **consult with the sample manager to obtain new or existing identifiers (ideally PIDs) for the samples.**

> *Example 2a-1:* Dredge rock archived in the collection at Oregon State University Marine and Geology Repository (igsn:10.58052/OSU-KM1609-D1-5) and subSample from that rock sent out to a researcher for destructive analysis (igsn:10.58052/OSU-KM1609-D1-5.AK1)

2b. Samples used in multiple analyses or publications

For samples that may not be preserved long-term, but will be used in multiple analyses or publications, **register samples for PIDs from a recognized PID allocating agent,** which could be through your institution or other examples provided below. For work involving subsamples (“child” sample) sent to multiple labs, see Step 2c.

> *Example 2b-1:* PIDs may be registered through a centralized information service (e.g. University library), data repository, or laboratory within your institution.
>
> *Example 2b-2*: For Earth science and other samples, the System for Earth and Extraterrestrial Sample Registration (SESAR) provides IGSN IDs. IGSN IDs are interdisciplinary and include standard metadata suitable for any sample type. SESAR provides instructions on how to register samples with standard metadata.
>
> *Example 2b-3*: University of California EZID, operated by California Digital Library, provides ARKs.

2c. Subsamples sent to laboratories for analyses

For samples that you are sending to a laboratory for analysis, **provide your source material sample PID for each sample** (which will represent the parent sample collected from the field or lab). You may also provide an identifier for the subsample (can be a project-unique sample name or a child PID), and/or determine if the laboratory assigns PIDs or other sample identifiers as part of their workflow when samples are submitted to their system. **When a dataset or paper includes results of these laboratory analyses, cite the laboratory sample IDs/PIDs in your dataset (Step 3) and/or paper**

(Step 4).

Example 2c-1: Laboratory with online data system, Joint Genome Institute (JGI)

Example 2c-2: Analytical identifier for genomic samples, BioSample Accession Numbers.

2d. Samples used only once

For samples that will not be stored or used more than once (e.g. in multiple publications), a PID is recommended but not required. At minimum, **use a identifiers that are unique within your project and/or research team (e.g., a UUID), and use them consistently across all of your data files and in your publication.** Publish sample metadata along with data, as described in Step 3.

Further reading: Why you should assign and/or use identifiers, and additional examples

Step 3. Publish and Cite Samples in Datasets

Publish a dataset that includes your sample identifiers (ideally PIDs) and associated data; see existing guidance on how and where to publish datasets. If your samples have PIDs, include them in your dataset(s) metadata, and include a sample PID column (with a header such as “IGSN” or “sample_PID”) within all data files containing sample data. We recommend including your sample metadata file describing samples (from Step 1) as part of your dataset. However, if your samples have PIDs that are already registered with standard metadata and are readily accessible, including an additional copy with your dataset may not be necessary.

Example 3-1: A dataset with geochemistry data for water and sediment samples, which includes both a sample metadata file and data files that include a column with PIDs for every sample examined in a paper, https://doi.org/10.15485/1923689.

Example 3-2: A dataset with stable isotope data for rock samples, which includes sample PIDs in the dataset metadata and in a designated column in the data files, https://doi.org/10.26022/IEDA/112300.

Example 3-3: You may also publish/cite a BioProject when using genomics data generated from samples, which is an umbrella entity with links to multiple samples and associated data (https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA174172); or cite one or more BioSample(s), which is more precise and includes links to associated data.

Then **cite the dataset in the reference section of your paper and/or include it in your data availability statement.**

Example 3-4: A paper that references two datasets with environmental and molecular data for sediment and water samples in the data availability statement and reference section, https://doi.org/10.3390/metabo10120518.

> *Example 3-5:* A paper that cites metabolomics and metagenomics data for samples in *Nature Scientific Data's* Data Record and References sections, https://doi.org/10.1038/s41597-024-03069-7.

Further reading: Why you should publish and cite samples in datasets, and additional examples

Step 4. Cite Sample Identifiers in Paper

If referring to samples within the text and/or table(s) of your paper, use sample identifiers in a **consistent standard format** where relevant to address methods or findings.

> *Example*: Provide sample PIDs in their standard compact format with hyperlink (igsn:10.58052/IEGRW002B), or the full url (https://doi.org/10.58052/IEGRW002B).

> *Example*: Provide sample PIDs in table alongside laboratory identifiers and summary of geochemical results (Table 2, Appendix C, Appendix D1, and Appendix D2 in https://doi.org/10.1016/j.gca.2013.08.001).

Further reading: Why you should reference sample identifiers in your paper, and additional examples

Appendices

Appendix A: Why you should describe samples with rich metadata (Step 1)

Metadata is "data about data" – it is the information needed to find, integrate, and reuse samples and associated data. Recording this information at the beginning of your projects, even if it feels trivial or obvious to you (e.g., the material type is "soil"), will increase the accessibility of your samples when reusing or sharing samples in the future. Your metadata may be important to a future researcher searching for and integrating data across studies.

> *Note 1:* What some disciplines consider to be data, such as biogeochemical measurements, might be considered metadata in a new context; for example, when submitting samples for 'omics analysis using the MIxS standard.

Why you should use an existing metadata standard: There are differences in what metadata is required and useful in different fields and repositories. We recommend that you use the standard required by or most relevant for your scientific field. These metadata standards usually provide quick reference guides and templates in spreadsheet or table format (Excel, CSV, TSV). These resources can help you quickly format your sample metadata in a way that people can easily understand and use it. Providing standard metadata will also make your sample metadata more machine-readable, enabling more specific data search and integration tools

> *Note 2:* The SESAR IGSN template, in particular, will also enable you to easily obtain PIDs (see Step 2), and the metadata provided will always be associated with the sample

PID in the future. Using this template early in your work will make Step 2 easy, and will ensure you have captured enough metadata to make your sample re-usable.

> *Example 2-1*: A dataset including both a sample metadata file using the ESS-DIVE SESAR IGSN template (https://doi.org/10.15485/1923689).

Note 3: When working with samples from long-term physical sample repositories/ collections, contact the managers of these repositories/collections. If you are using existing samples, they may be able to provide you with useful sample collection metadata. If you are creating new metadata, they may want to include your sample metadata in the collection's database.

> *Example 4-1*: Collection databases with sample metadata are sometimes published online, such as the UC Berkeley Museum of Vertebrate Zoology database. For biodiversity records, these collections database records are sent to, integrated with, and published in the Global Biodiversity Information Facility.

Appendix B: Why you should assign and/or use identifiers (Step 2)

Persistent Identifiers (PIDs) are globally unique, associated with standard metadata that are accessible online by both humans and computers, and have a landing page where you can link and exchange relevant information about the samples. For example, DOIs are PIDs that you may be familiar with for journal articles and dataset publications, and can also be applied to samples. Sample PIDs are particularly useful to track information about long-term samples, or samples used in multiple analyses (including subsamples sent to multiple labs) or publications.

Other Identifiers, such as **Universally Unique IDs (UUIDs)** and Darwin Core Triplets, are considered effectively globally unique. However, they are usually not associated with standard metadata, nor are they readily web-accessible unless they are modified to be URLs and maintained over time by an institution committed to long-term preservation. Therefore, they cannot be used to link and exchange information about samples.

Sample Names are project-specific, often meaningful, identifiers that scientists use in their collection and analysis workflows. These identifiers are highly useful for project workflows, and are associated with PIDs as metadata, but cannot be used for data integration and reuse. This is because they are not globally unique and are often inconsistently used (for example use of shorthand names, or other variations on the same sample name used by different people).

2a. Why you should consult with the sample manager to obtain new or existing PIDs for the samples in long-term collections: Using PIDs from long-term physical collections allows museums and sample managers to keep track of sample re-use and publication outputs, and ensures that scientists can locate all metadata, data, and publications associated with these samples.

> *Note 1:* PIDs should be included in the sample metadata file provided by the sample

manager in Step 1. If no PIDs are present, consult the sample manager to determine if a sample PID exists, and if not, request that sample PIDs be provided in the future.

Note 2: If you are taking a subsample ("child" sample) of a sample with an existing PID ("parent" or "source" sample), work with the sample manager to obtain new related PIDs as needed. It is important to provide the parent sample PID in the child subsample metadata (in Step 1).

2b. Why you should use sample PIDs for projects with multiple analyses: Using PIDs will make it easier to integrate data from multiple analyses and publications, and will make your work more discoverable and reproducible.

2c Why you should provide the laboratory your source material sample PID: This is particularly important if you have sent subsamples from the same source material sample to multiple labs with online data systems. The source material and subsequent analyses can then be linked through parent–child relationships to enable tracking the provenance of analyses and use of the original source sample.

Note 3: For genomics and other 'omics samples, using the MIxS metadata terms, you can provide the source sample PID in the "source material ID" field.

2d Why you should use a reasonably unique identifier for samples used only once: Using consistent identifiers will make your research more reproducible and will make data and sample management easier.

Note 4: You can still assign PIDs to these types of samples if you wish. The SESAR database, for example, allows you to register and assign PIDs for samples that are destroyed in analysis or not preserved in the long term.

Appendix C: Why you should publish and cite samples within datasets (Step 3)

Many journals and funding agencies require that metadata and data supporting research findings be published in a recognized data repository for long-term preservation and access. For sample-related research to be FAIR, it is important to ensure that sample metadata (described in Step 1) is readily accessible as part of your published dataset(s). If you have obtained sample PIDs, then you can simply include the list of relevant sample PIDs using the appropriate standard compact format (igsn:10.58052/IEGRW002B) within your dataset metadata.

By consistently using sample PIDs in your datasets, you will make it possible for others to discover and reuse your physical samples, integrate and reuse any data derived from your samples, and track use and provenance of your samples over time.

Note 1: The specific metadata field that you provide your list of sample PIDs may vary depending on the data repository.

Example 1-1: EarthChem dataset with sample PIDs included in a metadata field specific to samples (https://doi.org/10.26022/IEDA/112300).

Example 1-2: Sample PIDs included in a free-text field such as methods (https://doi.org/10.15485/1895159).

Note 2: If you obtain sample PIDs (see Step 2), sample metadata will already be archived and associated with your individual sample PIDs. You may still wish to include a file listing all the PIDs used in your study.

Example 2-1: An example SESAR landing page for a water sample that has been assigned a PID (igsn:10.58052/IEGRW002B).

Example 2-2: A paper that includes a supplementary file listing IGSN IDs (https://doi.org/10.1029/2022GL101903).

Note 3: If you did not obtain PIDs, you should still create a table or file containing the metadata from Step 1 and include that as part of your final data deposit.

Example 3-1: An example dataset that includes a sample metadata file, see file titled “enzymes_SampleMetadata.csv” (https://doi.org/10.15485/1830417).

Note 4: There are a wide range of domain or institution-specific data archives you can use. When selecting a data repository for deposit, refer to guidance available from your institution, funder or research community. If there is no relevant archive, you can use generalist repositories. The Registry of Research Data Repositories provides a good resource to search existing data repositories.

Example 4-1: Domain-specific repositories include EarthChem, Environmental Data Initiative (EDI), ESS-DIVE, and/or National Center for Biotechnology Information (NCBI).

Example 4-2: Generalist repositories include Dryad, Zenodo, or Figshare. See this Generalist Repository Comparison Chart (https://doi.org/10.5281/zenodo.7946938).

Appendix D: Why you should reference samples with consistent formatting (Step 4)

Consistent formatting is critical for both human and machine readability; the latter in particular is difficult without the context created by consistent formatting. Referencing sample PIDs in publications is important for compliance with the FAIR Data Principles, and citations in publications with a standard structure and format align with those principles. Taking this final step will increase the impact of your science for years to come.

Example 1: Provide sample PIDs in their standard compact format with hyperlink (, igsn:10.58052/IEGRW002B), or the full url (https://doi.org/10.58052/IEGRW002B).

Example 2: Provide sample PIDs in a table alongside laboratory identifiers and summary of geochemical results (see Table 2, Appendix C, Appendix D1, and Appendix D2 in https://doi.org/10.1016/j.gca.2013.08.001).

Do not use variations (such as abbreviations or wildcards) of an identifier to refer to a sample. In some domain-specific cases where dozens or hundreds of identifiers need to be referenced, it may be appropriate to refer to a range of identifiers (such as referencing a range of nucleotide identifiers as “MN40555-602”).

If possible, reference individual samples in the data availability statement to make it easier to track use of samples, particularly when referencing a small number of samples (<10).

Example 3: Sample PIDs included in the Data Availability Statement: https://doi.org/10.1029/2021GL094036.