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

This paper is a non-peer reviewed preprint submitted to EarthArXiv. It has been submitted to *Nature Scientific Data* for publication and is under review as of May 30, 2024.

Joan Damerow[1], Natalie Raia[2], Val Stanley[3], Saebyul Choe[4], Mikayla A. Borton[5], Neil Byers[6], Ellen R. Cassidy[7], Shreyas Cholia[8], Rorie Edmunds[9], Brieanne Forbes[10], Kathleen Forrest[11], Amy Goldman[10], John Kunze[11], Sara Lafia[12], Kerstin Lehnert[4], Marcella McIntyre-Redden[13], Richard Naples[14], Dylan O'Ryan[1], Charles Parker[6], Esther Plomp[15], Beck Powers-McCormack[10], Sarah Ramdeen[16], Stephen Richard, Anne Thessen[17], Cody W. Thompson[18], Dave Vieglais[19], Kristina Vrouwenvelder[20], Elisha M Wood-Charlson[21], Lesley Wyborn[22], T.B.K. Reddy[6], Andrea Thomer[2]

1) Earth and Environmental Sciences Area, Lawrence Berkeley National Laboratory, Berkeley, CA, 94720, USA
2) College of Information Science, University of Arizona, Tucson, AZ, 85721, USA
3) Marine and Geology Repository, Oregon State University, Corvallis, OR, 97331, USA
4) Lamont Doherty Earth Observatory, Columbia University, Palisades, NY, 10964, USA
5) Colorado State University, Fort Collins, CO, 80526, USA
6) Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA, 94720, USA
7) University of Michigan, Museum of Zoology and Department of Ecology & Evolutionary Biology, Ann Arbor, MI, 48108, USA
8) Computing Sciences Area, Lawrence Berkeley National Laboratory, Berkeley, CA, 94720, USA
9) DataCite, Welfengarten 1B, 30167 Hannover, Germany
10) Pacific Northwest National Laboratory, Richland, WA, 99354, USA
11) Drexel University, 3141 Chestnut St, Philadelphia, PA, 19104, USA
12) NORC at the University of Chicago, 1155 East 60th St, Chicago, IL 60637
13) Geological Survey of Alabama, Tuscaloosa, AL, USA
14) Smithsonian Libraries and Archives, Washington, DC, USA
15) Delft University of Technology, Lorentzweb 1, 2628 CJ Delft
16) Ronin Institute for Independent Scholarship
17) University of Colorado Anschutz Medical Campus, USA
18) University of Michigan, Museum of Zoology and Department of Ecology & Evolutionary Biology, Ann Arbor, MI, 48108
19) Biodiversity Institute and Natural History Museum, University of Kansas, Lawrence, KS, 66045
20) American Geophysical Union, 2000 Florida Ave NW, 20009, Washington, DC, USA
21) Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, CA, 94720, USA
22) Australian National University, Canberra ACT 2600, Australia

Email correspondence to: Joan Damerow (JoanDamerow@lbl.gov), Andrea Thomer (athomer@arizona.edu)

Abstract

Physical samples and their associated (meta)data underpin scientific discoveries across disciplines, and can enable new science when appropriately archived. However, there are significant gaps in community practices and infrastructure that currently prevent accurate provenance tracking, reproducibility, and attribution. For the vast majority of samples, descriptive metadata is often sparse, inaccessible, or absent. Samples and associated (meta)data may also be scattered across numerous physical collections, data repositories, laboratories, data files, and papers with no clear linkages or provenance tracking as new information is generated over time. The Physical Samples Curation Cluster has therefore developed 'A Scientific Author Guide for Publishing Open Research Using Physical Samples.' This involved synthesizing existing practices, community feedback, and assessing real-world examples to identify community and infrastructure needs. We identified areas of work needed to enable authors to efficiently reference samples and related data, link related samples and data, and track their use. Our goal is to help improve the discoverability, interoperability, use of physical samples and associated (meta)data into the future.

Introduction

Physical samples and their associated (meta)data are primary building blocks across a wide range of research. They represent features of interest or living things [1,2], underpin discoveries across disciplines, and are critical to the scientific process. This may include, for example, soil or water samples collected to represent environmental conditions at a given site and depth, a rock from a geologic outcrop, or a preserved organism, such as a plant or animal specimen. When samples and associated (meta)data are findable, accessible, interoperable, and reusable [FAIR; 3] and as 'open as possible' [4–6], *new science* becomes possible [7,8]. For example, species occurrence records published and aggregated globally through the Global Biodiversity Information Facility (GBIF) are now cited in more than two publications per day [9]. GBIF infrastructure enables studies that integrate from hundreds to billions of records to answer questions on conservation, species distribution, climate change impacts, macroecological patterns, and more [9,10]. However, for many data types and disciplines, widespread adoption of community practices and useful tools for sample and (meta)data discovery, integration, and use are in much earlier stages [11], or do not yet exist.

Progress in funding policies, community standards, and infrastructure for samples continue to improve the discovery and reuse of samples and associated (meta)data [9,11]. Recent updates to the US National Science Foundation, Division of Earth Sciences (EAR) Data and Sample Policy require that:

> *"All data and sample metadata underlying peer-reviewed scholarly publications resulting from EAR support must now be made publicly accessible at or before the time*

of publication, and no later than two (2) years after completion of data collection or generation, via appropriate long-lived FAIR-aligned repositories" [12].

However, there are significant gaps that prevent accurate provenance tracking [13], reproducibility [7,14], and attribution [15,16]. For the vast majority of samples, descriptive metadata is often sparse, inaccessible, or absent [17–19]. Samples and associated (meta)data may also be scattered across numerous physical collections, data repositories, laboratories, data files, and papers with no clear linkages or provenance tracking as new information is generated over time [13,20]. There is a growing need to connect related interdisciplinary sample-associated (meta)data spanning diverse fields and data systems [21].

There is also a need for researchers to respect Indigenous Data Sovereignty and Indigenous Data Governance for samples collected on lands and waters belonging to Indigenous peoples. While beyond the scope of the current work, researchers should be aware of and comply with the CARE Principles for Indigenous Data Governance (Collective Benefit, Authority to Control, Responsibility, and Ethics) [22] for both the collection and long-term management of samples and any derivative data on those samples.

Practices for publishing and citing sample-associated data have been inconsistent and there is a lack of clear guidelines across disciplines. This has led to many consequences, such as: 1) research that uses samples may not be reproducible [7]; 2) it can be time-consuming or even impossible to track related data and information about samples [23]; 3) samples can be difficult to find and reuse, minimizing the repeatability of the science; and 4) sample collection managers are less able to show the impact of their collections and curatorial work [16].

To address these challenges, the Earth Science Information Partners (ESIP) Physical Samples Curation Cluster sought to develop recommended practices for publishing and citing physical samples in scientific research. This involved synthesizing existing practices, extensive community feedback, and use case review. We assessed community-identified use cases in which sample metadata sharing, citation, and tracking need to be improved, focusing on how to:

1. Efficiently publish and cite large number of samples and associated (meta)data;
2. Provide credit for those involved in physical sample collection and curation to demonstrate value of investing in collections;
3. Track use of sample data generated by analysts and laboratories; and
4. Connect related interdisciplinary sample (meta)data and other research outputs.

Existing Community Practices and Infrastructure

Community practices and tools for assigning persistent identifiers (PIDs) or accession numbers to samples have been in place for decades and enable access, integration, and reuse of high-value (meta)data. International Geo Sample Number, now International Generic Sample Number (IGSN IDs), Archival Resource Keys [ARKs; 24], and other sample PIDs are globally unique, associated with standardized human- and machine-readable metadata that are accessible online, and resolve to a landing page where users can link and exchange sample information [25,26].

The IGSN ID was established in 2004 for earth science samples and has since expanded to include a wide range of interdisciplinary samples [20,25,27–29]. At the time of writing, there are 5,070,114 (mostly earth and environmental science) samples in SESAR, and >12.5 million IGSN IDs across allocating agents [30]. Major organizations such as the geological surveys of the US, UK, Australia, Korea, and Germany use the IGSN IDs for their collections.

Sample PIDs are particularly useful to track information about long-term samples, or samples used in multiple analyses (including subsamples sent to multiple laboratories) or publications [26]: they can link to datasets, images and other information derived from them. There are non-resolvable identifiers, such as Universally Unique IDs (UUIDs) and Darwin Core Triplets, that are commonly used for biological specimens in natural history collections. While they both generate strings that are effectively globally unique, they are often not associated with standard metadata, nor are they readily web-accessible unless they are modified to be Uniform Resource Locators (URLs) and maintained over time by an institution committed to long-term preservation. In addition, analyses have found Darwin Core Triplets (which include institution and collection codes followed by a numeric string, such as MVZ:Mamm:165861) commonly used in natural history collections to often contain errors and duplicates, and to be ineffective for linking related data [23]. As such, only resolvable PIDs, maintained by a long-term institution, and associated with standard metadata are suitable for sample identification and tracking use of samples.

Metadata templates and standards to describe physical samples are available for biodiversity records [31], 'omics (such as genomics, metagenomics) material [32], earth and environmental science samples [33,34], and ecosystem sciences [35]. For example, the Biosample database maintained by the National Center for Biotechnology Information (NCBI) contains records with information and metadata describing the physical materials from which the sequence information stored in other NCBI databases like GenBank are derived [31]. Implementation of standard metadata practices has enabled search and access to genetic sequence data in Genbank since 1979 [36] and aggregated species occurrence records in GBIF starting in 2001 [37]. GenBank contains over 2.9 billion nucleotide sequences for 504,000 formally described species [38], and there are now close to 2.7 billion species occurrence records in GBIF that enable a wide variety of synthesis studies. SESAR [33,34] contains metadata records for >5 million samples including rock, mineral, sediment, and soil samples; rock, sediment, and ice cores; as well as samples of volcanic gas, different types of fluids (seawater, river water, hydrothermal fluids), biological specimens collected as part of earth, planetary and environmental sciences research.

Current Sample Citation Recommendations and Practices

For publications that include physical samples and associated data, there are no central recommendations across disciplines for citing samples and associated data, and attribution practices vary. Clear and consistent recommendations across institutions and disciplines are greatly needed and must be clearly communicated and enforced by publishers.

While many physical sample repositories and natural history collections request acknowledgement when their samples are used, there is no standard citation practice that enables tracking of sample use [39]. Each institution sets their own recommended practice, which often includes museum catalog numbers and the institution name; PIDs may or may not be required.

The Field Museum in Chicago, for example, recommends that specimens or objects be cited in their preferred format: [occurrenceID].[catalogNumber].[data publisher] (The Field Museum, 2024). The citation formats for museum collections at the Smithsonian National Museum of Natural History and the American Museum of Natural History (AMNH) are dependent on the division or department under their loan policies. For example, the Smithsonian Mineral Sciences collection requires users to cite their collections based on what is available: catalog number, ARK, and/or IGSN ID. The AMNH Paleontology department requires a copy of the manuscript for records and catalog citation.

This variation in recommended citations results in even greater variations in how authors actually acknowledge long-term collections and samples used, if they do so at all. For example, authors will often mention sample repositories in the Acknowledgements section, and list individual samples on a map or table shared as supplemental materials [see 40]. In both cases, the identifiers may be inconsistently abbreviated, with no information about current archives where the physical samples are held, which is important if the samples need to be accessed and re-analyzed. This reduces the reproducibility of the study, and makes automated tools to identify citations very difficult or even impossible.

Samples destined for genetic sequencing have more infrastructure in place that connects them to the resulting data or publication than some other fields [41]. However, similar variation in guidance and interpretation of genomic sample citation requirements by authors, editors, and publishers have contributed to inconsistent practices in this field of research as well. Guidance on citing accession numbers for genetic data is provided for NCBI, but different/additional requirements may be requested by specific laboratories conducting the genomic analyses; for example, the Joint Genome Institute (JGI) requests that authors cite JGI proposal DOIs in their publication [42].

In the microbiology literature, the standard practice for citing samples is to refer to an isolate by a strain identifier (e.g., “Kra1”) or by culture collection accession number with a prefix that indicates a Biological Resource Center (BRC) and a numeric or alphanumeric catalog number uniquely identifying the sample within the collection (e.g., ATCC 35583, DSM 2078, JCM 9277) [43]. A strain identifier is not typically unique, and as it is a free-form string, it can lead to ambiguity in the literature and public databases. Culture collection accession numbers, while not resolvable, are more easily identifiable due to their typical use of a three or four letter repository prefix coupled with an integer that uniquely connects the sample with a specific collection. However, the lack of standards (for accession format or metadata retrieval) among repositories limits the utility of these accession numbers in searching 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 accession numbers [44]. The metadata associated with these identifiers can be rich, providing information about the source of the isolate (such as location, host, organization, personnel), as well as references to associated projects, genome assemblies, and DNA sequence data. Additional identifier classes exist for biological projects, analyses, and sequence data. These accession numbers 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 to prevent ambiguity (for example, a project may be associated with multiple biological samples). The same strain or culture collection may end up with multiple entries in the BioSample repository as different researchers submit the same strain as different BioSamples. However, journals and indexers do not currently recognize BioSample or BioProject accession numbers as a formal related identifier or citation.

Ethics, the Nagoya Protocol and the CARE Principles for Samples and Associated (Meta)data

We recognize that not all samples and data derived from these samples can be fully open. Samples that are sensitive or restricted must be protected through appropriate access controls and have any restrictions documented (such as permits, ethics agreements, access moratoriums). Samples and derivative data should be as open as possible and as closed as necessary. The decision as to whether the samples and derived data can be made public is not necessarily that of the researcher. For example, the Nagoya Protocol addresses 'Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from their Utilization to the Convention on Biological Diversity' [87].

When collecting and managing samples related to Indigenous Peoples and lands and waters, authors should consult the CARE Principles for Indigenous Data Governance (Collective Benefit, Authority to Control, Responsibility, and Ethics) [22]. The CARE Principles were developed by Indigenous Peoples, scholars, non-profit organizations, and governments to address concerns about secondary use of their data and samples. The CARE Principles are designed to 1) respect Indigenous data sovereignty), and 2) support Open data, including secondary use [22,88]; they are designed to complement the FAIR Principles [89]. For future sample acquisition, it is essential that the relevant Indigenous communities are engaged prior to any samples being collected, and that wherever possible, local knowledge is included in the collection process to avoid incidents such as the unauthorized sampling of the Bishop Tuff in California and other cases elsewhere [90].

Operationalizing the CARE Principles by both repositories and researchers is just beginning, and there are several communities currently working to make progress. For example, the Indigenous Metadata Bundle Communique [91] provides guidance on the "Collective Benefit" and "Authority to Control" Principles, and has identified five categories for metadata elements: governance, provenance [92], lands and waters, protocols and Local Contexts Notices and Labels [93]. Publishers are also increasingly concerned with how adherence to the CARE Principles can be documented in publications, with several publishers developing position statements outlining their commitment and intention to recognise Indigenous Data Sovereignty and Indigenous Data Governance (for example, Data Science Journal, [94]. The ESIP Physical Samples Curation Cluster is monitoring these efforts and will incorporate recommendations as they develop in the future.

Results

Use Case Review: Needs for Tracking Sample Use

We identified several use cases that illustrate common needs for tracking sample use. These real-world examples informed the guidelines presented in the next section.

Use Case 1: Efficiently publish and cite large number of samples and associated (meta)data

Many studies that involve physical samples include dozens, hundreds, or even thousands of samples and subsamples. There is no widely-adopted method to efficiently cite large numbers of datasets [45], let alone the physical samples linked to them. However, we can include metadata indicating all the samples used in a given dataset. Several data repositories currently include sample PIDs as related identifiers in dataset metadata (such as EarthChem Library (ECL), Pangaea, GFZ Data Services).

A Real-world Example: Connecting Samples and Data in the Interdisciplinary Earth Data Alliance

The Interdisciplinary Earth Data Alliance (IEDA2) is a collaborative, NSF-funded data infrastructure that consists of several complementary data systems— ECL, the EarthChem Synthesis, the Library of Experimental Phase Relations and Trace Element Distribution Experimental Database (LEPR/TraceDs), and the System for Earth and Extraterrestrial Sample Registration (SESAR). These systems provide services for publishing sample-based analytical data, primarily from laboratories, that ensure use of consistent sample metadata and PIDs for samples (IGSN IDs) used to unambiguously connect samples to data in EarthChem and LEPR/TraceDs.

SESAR offers IGSN ID registration services for researchers and collection curators, enabling them to permanently store and update sample metadata—as well as images and links to related datasets and publications—on a persistent and publicly accessible digital sample landing page (e.g., doi:10.58052/IENHR006K; Figure 1). Researchers may register IGSN IDs by entering metadata for a single sample in a web form, uploading a standardized spreadsheet template with metadata for one or more samples in their MySESAR account (batch registration process, or sending XML-encoded sample metadata from their local sample metadata management systems to SESAR through an Application Programming Interface (API). SESAR also enables linking of related samples (collection sites, parent–child samples, and/or sibling samples) based on metadata provided by researchers, making sample metadata more discoverable.

EarthChem provides two distinct, but complementary services. Firstly, it gives access to large volumes of published laboratory analytical data for terrestrial samples (ca. 50 million analytical data points); these are aggregated and harmonized into synthesis databases (PetDB, EarthChem Portal) with human and machine-actionable interfaces for search, access, and retrieval of analysis-ready data [46]. Secondly, EarthChem enables archiving and publication of datasets in

ECL, a data repository recommended by funders and publishers, where researchers can publish their datasets in compliance with the FAIR Principles, using standardized or discipline/method-specific data templates developed with community input, such as sample and data templates for volcanic Tephra samples [47]. Researchers contributing data to ECL have the ability to provide IGSN IDs within a designated column in the data templates and in a distinct metadata field during dataset submission. Upon publication, IGSN IDs in this metadata field become fully resolvable, linking to their respective SESAR IGSN ID metadata landing pages (Figure 1).

As of November 2023, >25% (360) of EarthChem's 1,336 published datasets included links to IGSN IDs, with 31,506 unique IGSN IDs recorded. Within SESAR, 25,000 publicly available samples had been linked to EarthChem datasets. These numbers reflect strong community interest and buy-in for a future where these systems have automated links for sample and data discovery. The IEDA2 Geosamples Data Nexus is currently under development [48] to automate the linking of data and samples across IEDA2's sample data systems and to provide an open, central discovery point for samples and related data in which other sample-based data systems can also participate.

An example of an ECL dataset with linked IGSN IDs is represented in Figure 1. This example consists of three distinct relationships: samples linked as a parent and child relationship, a sample IGSN ID linked to a published dataset in ECL, and an ECL dataset linked to a manuscript.

Summary of Needs

To support efficiently publishing and citing large numbers of samples and associated (meta)data, we need the following to happen:

1) Authors should use PIDs for their samples, and include them as a column in data files and/or dataset metadata.
2) Data repositories need to provide metadata field(s) that support related PIDs and include these in the metadata records registered with DOI registries such as DataCite.
3) Sample (meta)data and data repositories should enable automatic updates to sample metadata profiles and dataset landing pages as new (meta)data is published. For example, when samples are included in a dataset, the sample landing page should automatically be updated with a link to that dataset. PIDs must therefore be processed through an indexer or other functional links must exist between pertinent repositories and sample ID landing pages.
4) When a dataset is cited, samples included in that dataset should also be automatically recognized and tracked in metrics.
5) Users should be able to easily access sample PIDs and metadata on dataset landing pages; for example, through a weblink or the option to download.

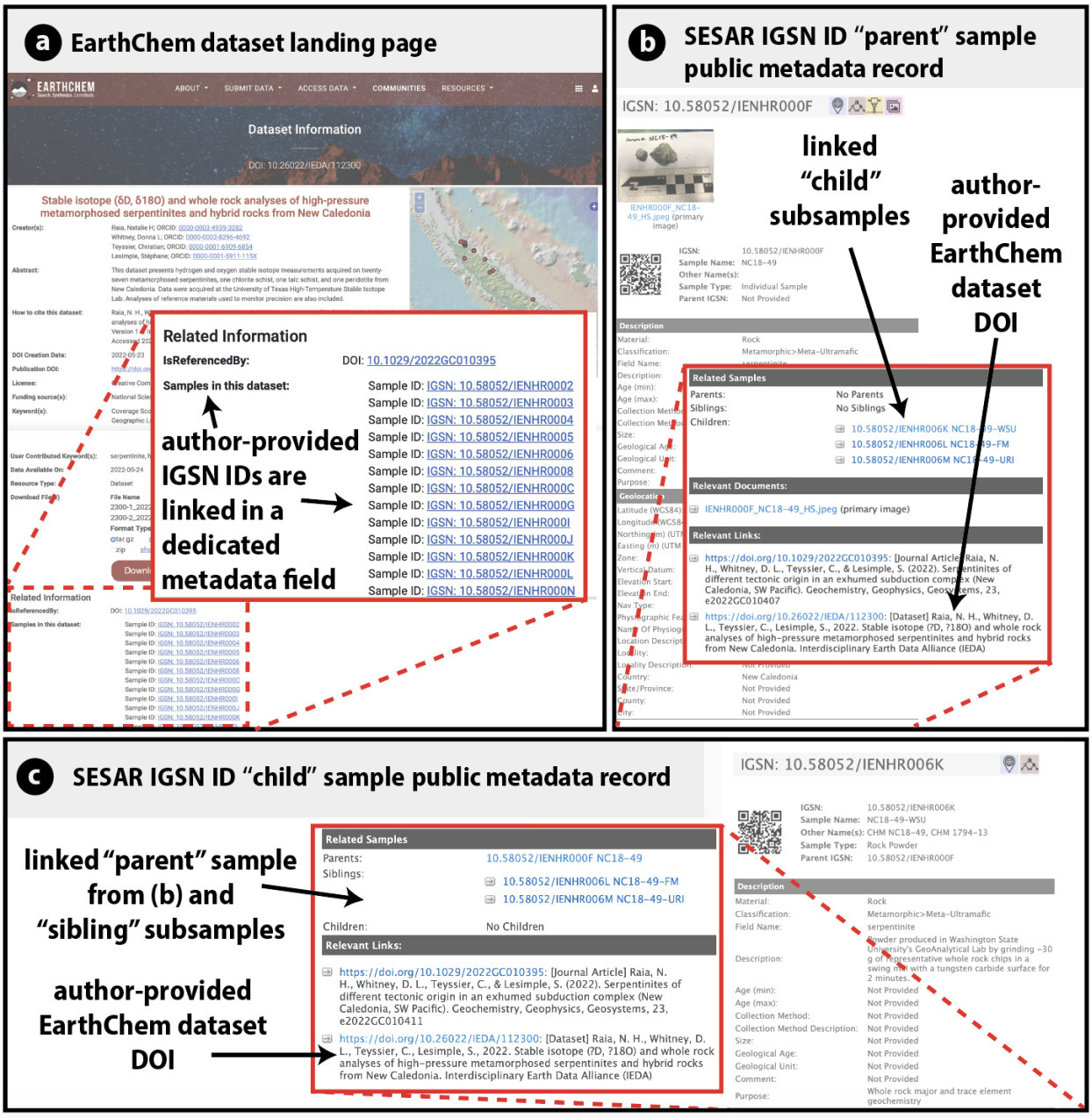

Figure 1. Diagram depicting linkages between EarthChem (doi:10.26022/IEDA/112300) and SESAR. a) During dataset submission, authors are provided with a dedicated PID metadata field to provide persistent identifiers for samples. Once the dataset is submitted, the system verifies and hyperlinks PIDs; in this case, linked IGSN IDs are shown. b) Linked IGSN IDs lead to a permanent, publicly-available metadata record page. For the sample shown, additional subsamples ("child") IGSN IDs have been registered and are linked. The IGSN ID registrant has provided the DOI for the dataset shown in (a) in a dedicated metadata field for related URLs or DOIs. c) A "child" subsample metadata record links back to the "parent" sample IGSN ID (b) and to other subsamples ("siblings"). The IGSN ID registrant has again manually provided the DOI for the dataset shown in (a).

Use Case 2: Provide credit for those involved in physical sample collection and curation to demonstrate value of investing in collections

Citing samples and tracking sample provenance is crucial for giving credit to individuals and organizations involved in sample collection and curation over time, including sample collectors, the repositories and collection managers who curate and manage samples, and funders evaluating impact. For example, physical sample repositories must regularly show the impact of their collections to justify their work and continue to acquire funding. When samples are not cited, collection managers are less able to demonstrate how these collections are used to advance research and quantitatively demonstrate impact of that research, which in turn threatens the sustainability of these valuable scientific assets. Furthermore, individual collection managers, repository managers, analysts, and other data stewards are unable to fully document their contributions to science and scholarship [15].

A Real-world Example: Showing the Impact of the University of Michigan Museum of Zoology (UMMZ)

The UMMZ mammal division manages over 150,000 specimens that are used in a broad range of scientific studies. Each of these specimens has a catalog number—a unique identifier *within the UMMZ* that is associated with both the physical sample and its metadata—but not a PID. To track the use of their collections, mammal division collections staff (led by author CWT) 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. CWT 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.

While the Google Scholar page shows one form of impact—the papers in this bibliography have received over 91,000 citations—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, authors [SL, ERC, KF, RN, CWT, AT [49]] employed multiple types of text mining pipelines to extract catalog numbers and generate metrics to show their use. The results were somewhat underwhelming: of the 1,297 papers analyzed, only 245 included catalog numbers. This was much lower than expected; while Lafia et al. (2022) expected the corpus to include papers that excluded catalog numbers, they did not expect that it would be over 80% of the papers. After reviewing the corpus, they concluded that many of the papers using specimens from UMMZ simply did not cite them in their papers. Researchers typically thanked the collection in the acknowledgements section without citing specimens, listed specimens in supplementary material that could not be effectively identified and mined, or listed other identifiers that were not used by the UMMZ, and thus limited the repeatability of the science and the recognition of the collection.

Summary of Needs

To provide credit for those involved in physical sample collection and curation to demonstrate value of investing in collections, we need the following:

1) Managers of physical collections should explore assigning PIDs to their specimens. While this takes considerable time and effort, the pay off is potentially high in terms of showing the impact of collections over time. PIDs are much easier to "mine" and aggregate than catalog numbers because they are consistently formatted and resolvable to an online metadata catalog.
2) Paper authors should reference individual samples/specimens using a PID that is managed and assigned by the sample's curators. Depending on the study and number of samples used, this could be done by listing the PID in the text of the paper, by formally citing a sample in the references section, or by including sample PIDs in a dataset cited by the paper.
3) Publishers, indexers, and data repositories need to work together to make it possible to aggregate and track use of all PID types. This might mean that publishers expose PID metadata in a way that makes it easier to index, indexers build new tools to harvest PIDs from papers and datasets, or data repositories take steps to expose sample PIDs to indexers.
4) Subsamples taken from a parent sample should be clearly linked to the parent through related identifiers. For instance, in the example above, some authors included GenBank PIDs in their papers rather than UMMZ catalog numbers. However, long-term collections such as museums must be able to easily traverse a network linking GenBank PIDs to their original source/parent sample.

Use Case 3: Track Use of Sample Data Generated from Laboratories

Similar to the sample collectors and physical collections described in Use Case 1, laboratories conducting analyses on samples need to be able to demonstrate the value of their work to funders. Understanding how data are reused is also essential for identifying service improvements that can benefit the laboratories themselves and the communities they serve; for example, focusing on thematic areas that are heavily cited, improving the efficiency of laboratory processes, or allocating resources towards products and services with a high-impact potential. However, a laboratory that publishes data or provides samples loses control over provenance information (records of how the sample and data are used) as soon as it ends up in the hands of a third party. Approaches that preserve provenance information for samples and data, and that accumulate metadata in a consistent manner across systems, are greatly needed.

Real-world Example: Citations for Data Generated by the Joint Genome Institute (JGI)

The JGI provides integrated high-throughput sequencing of samples, DNA design and synthesis; metabolomics; and computational analysis. To track its impact on scientific research, JGI developed the Data Citation Explorer [50] [preprint, manuscript in review], a web service that identifies use of genomic data products in published literature even in instances where those products are not properly cited. The service employs heuristics to discover occurrences of unique identifiers associated with genomic data in the text and reconstructs graphs that restore many of the missing connections among these related classes of identifiers. The Data Citation Explorer has been able to identify ca. 4,000 publications citing JGI data using NCBI identifiers or other

standard identifier types. However, concurrent manual expert-analyses identified that most researchers cite publications associated with datasets produced from samples, if they cite anything at all. The authors estimate that there are tens of thousands of such "nonstandard" references to JGI data that cannot yet be identified using automated tools [50].

Summary of Needs

The following would facilitate tracking use of sample data generated by laboratories:

1. Researchers should follow consistent guidelines on how samples and associated (meta)data should be cited. Particularly with a rise in interdisciplinary work, it would be beneficial to use and enforce similar practices across disciplines, journals, and institutions.
2. Scholars, laboratory managers, and others that register sample identifiers should use PIDs that can be identified and indexed using automated tools (unique, no white space, easily used as part of a Uniform Resource Identifier).
3. Sample metadata and data repositories should use consistent methods of search and retrieval of sample (meta)data (for example, URL formats, API standards, metadata formats), and implement standards to unambiguously link and exchange information for related PIDs [51]. Provenance information must be propagated when laboratory and/or sample PIDs are used.
4. Sample metadata publishers should include a contributor role type (for example, using the Contributor Roles Taxonomy [CRedIT] taxonomy) that indicates the form of credit that should be attributed to an author, institution, or funder, ideally using a PID for the practitioner.

Use Case 4: Connect Interdisciplinary Sample (Meta)data and Other Research Outputs

Interdisciplinary studies that connect diverse data to understand multiscale processes often involve sample data. This highly related data may be analyzed and published separately on multiple data systems, creating a challenge to connect different data types from the same samples. Future researchers attempting to find and reuse such data often have no way of knowing sample provenance without contacting the authors, which makes data synthesis involving interdisciplinary samples nearly impossible. For example, it is currently difficult to integrate and reuse existing sample data across multiple studies to validate land and earth system models [52].

Real-world Example: Biogeochemical Samples from Projects of the United States Department of Energy's Biological and Environmental Research Program (U.S. DOE BER)

The U.S. DOE BER program is highly interdisciplinary, and samples from its projects are often used to enhance models and predictions of ecological processes and biogeochemical responses to contamination, warming, and other disturbances. We reviewed citation practices for 30 publications from projects funded by this program, including both environmental data and associated 'omics data. Over 60% of these papers included some form of sample-related ID, mostly citing biological 'omics data at a collection level because there are existing

recommendations to do this as described above (e.g., NCBI Bioproject, JGI GOLD Study ID, IGSN ID, NCBI SRA Accession). However, less than half of the publications provided associated environmental samples and data (Table 1).

Table 1. Analysis of citation and publication practices for 30 publications from projects funded by the U.S. DOE BER program, with both environmental and 'omics data.

Citation and Publication Practice	Percentage of Papers	Example Publication(s)
Sample Identifiers within Data Availability	53%	Ward et al. 2017 McGivern et al. 2021
Sample Identifiers within Paper Text (e.g., tables, methods)	20%	Henske et al. 2018 Shaw et al. 2020 Yee et al. 2023
Sample Identifiers within Supplemental File	13%	Ward et al. 2018
Sample Identifiers at Collection Level	87.5%	Evans et al. 2021 Liao et al. 2021
Sample Identifiers at Sample Level	13%	Lynes et al. 2023 Vigneaud et al. 2023
Associated Environmental Data within Supplemental Files	40%	Reichart et al. 2021 Hestrin et al. 2022
Associated Environmental Data Published within Repository	20%	Woodcroft et al. 2018 Matheus Carnevali et al. 2021

Scientists on these projects have faced sample tracking challenges due to inefficiencies in the processes of submitting samples to different data systems and laboratories and then compiling the resulting data. One such project, the River Corridor and Watershed Biogeochemistry Scientific Focus Area, studies hydrologic, biogeochemical, and microbial function within river corridors [53]. In one study, researchers collected a series of individual surface water samples (e.g., igsn:10.58052/IEWDR00RT), sediment samples (e.g., igsn:10.58052/IEWDR0149), and filter samples (e.g., igsn:10.58052/IEWDR00UI) at almost 100 global sites (e.g., igsn:10.58052/IEWDR00P4). DNA and RNA material were extracted from the filter and sediment samples (subsamples/child samples; e.g., igsn:10.58052/IEWDR00UI), and sent to JGI for metagenomic and metatranscriptomic sequencing. Water and sediment samples were also sent to the Environmental Molecular Sciences Laboratory (EMSL) for metabolomics analyses (Figure 2). The researchers obtained raw data from these respective online systems, generated additional data, and conducted further processing, analysis, and visualization across all the data types. They created sample sets and documented their workflows in the DOE Systems Biology Knowledgebase as a part of the Genome Resolved Open Watersheds effort [KBase; 54]. Analysis and visualizations from the sample set were incorporated into formally published datasets for long-term preservation and documentation in the Environmental System Science Data Infrastructure for a Virtual Ecosystem (ESS-DIVE) data repository [55,56]. These datasets were then referenced in the final journal publications associated with the data [57–60].

The process of submitting the associated data to multiple systems, and adding links and other

information over time as new (meta)data is generated, is currently inefficient. The relevant BER data systems are therefore working towards developing a more deeply integrated data ecosystem that automatically links and connects related data [manuscript in progress].

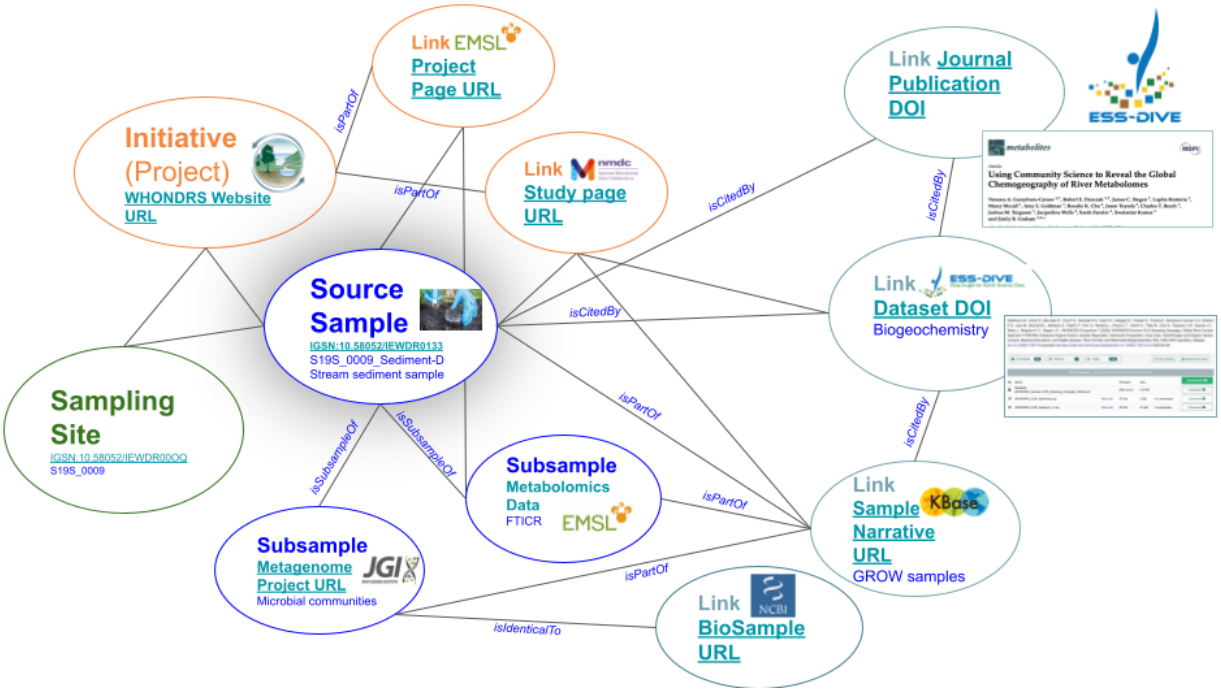

Figure 2. Tracking and linking a source material sample from the River Corridor and Watershed Biogeochemistry Scientific Focus Area project (based on the iSample relational data model, which links related samples based on entities such as project, sampling site, subsamples, as well as other related links: narrative workflow, dataset, analytical results for specific data types, and journal publication).

Summary of Needs

To efficiently connect interdisciplinary sample (meta)data and other research outputs, we need the following:

1. Researchers should use sample PIDs for environmental source samples and subsamples sent to laboratories.
2. (Meta)data repositories and laboratories should promote or provide field apps and other tools for automated registration of sample PIDs with standard metadata at the time of collection/creation of the sample, or soon after, and upon sending subsamples to different laboratories and user facilities (automatically creating resource maps that specify and display sample relationships). These tools should enable more efficient metadata curation using controlled vocabularies, systematic names for organisms, standardized microbiome names for environmental samples.
3. Laboratories and data systems should provide tools that map varying, but similar, metadata requirements across different systems [61].

4. (Meta)data repositories and data systems should develop APIs to automatically connect and exchange (meta)data, for example, automatically crosslink across data systems as new (meta)data is generated; for example, APIs for metadata exchange and registration of BioProjects and BioSamples at NCBI to facilitate data submission (Mukherjee et. al. 2022 and Mukherjee et. al. 2023).
5. The ability to search and integrate samples across projects to support global sample search (such as searching by sample type, environmental context, analysis type, location, date).
6. Methods and systems for tracking sample use and citations as new (meta)data are published and (re)used over time [62].

Recommended Practices for Scientists Publishing Sample-Based Research

The ESIP Physical Sample Curation Cluster synthesized community feedback and the experiences from the above use cases to develop "A Scientific Author Guide for Publishing Open Research Using Physical Samples" [63]. This author guide includes foundational elements to make samples and associated (meta)data Open and FAIR, and will be updated as technologies and practices evolve and become commonly adopted. Implementing this guidance helps to enable usage tracking of samples over time, which in turn supports reproducible research, data integration, reuse, and credit. The full guidance document includes links to specific examples, and additional information on why each step is needed. We have also condensed the guidelines into a flyer and postcard for community distribution within the earth sciences (Figure 3) [Figure X; 64]. These guidelines can be used directly by individual researchers, journal publishers, or data repositories, and can be modified to provide more targeted instructions for specific communities.

Figure 3. Condensed postcard and flyer [64] illustrating the author guidelines developed by the ESIP Physical Sample Curation Cluster, and targeted towards the earth science community.

A summary of key elements of the full scientific author guide for publishing Open research using physical samples [63] include:

Step 1. Describe Samples with Rich Metadata

Describe key characteristics and collection details of the samples used for the paper; for example, by including a sample metadata file or table. This can be a csv file with sample PIDs 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). In line with the FAIR Principles (Wilkinson et al 2016), use a domain-specific standard or community reporting format relevant for your sample type [31–34,65].

Step 2. Assign and/or Use Identifiers for Samples

Assign and/or use sample identifiers, ideally PIDs, to track samples and associated data; some institutions or data systems that you use may assign sample PIDs for you. Identifiers and specific steps may vary depending on your use case, with specific recommendations for long-term physical collections, samples used in multiple analyses of publications, subsamples sent to multiple laboratories for analysis, and samples used only once.

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 (such as column header "IGSN" or "Sample PID") within all data files containing sample data. If your samples do not have PIDs associated with standard metadata, also include a sample metadata file that clearly describes all sample collection details (Step 1) as part of your dataset. Then cite the dataset in the reference section of your paper and include it in your data availability statement.

Step 4. Reference 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. When referencing PIDs in text, files/tables, and data availability statements, include a prefix identifying what kind of PID you are using before writing the number, and a hyperlink to the sample landing page (e.g., igsn:10.58052/IEGRW002B) or the full url, depending on journal requirements. This will make your PID findable by both humans and computers.

Note that for valuable samples archived in collections, you should cite sample PIDs in the text or references section where possible. However, when using large numbers of samples, you can cite a dataset that in turn cites the individual samples included (Step 3).

Applying Recommended Practices in the ESS-DIVE Data Repository

Through work completed by author JED and collaborators, we illustrate how a community not previously exposed to standard practices for publishing sample-based research can implement the recommendations resulting from the above guidance. The ESS-DIVE data repository has 26 datasets (as of March 2024) compiled into a data portal collection, each of which includes IGSN IDs and standard metadata for associated samples. This includes seven datasets with detailed links to related samples and other research outputs (Table 2). We outline below how we were able to follow the recommendations using existing infrastructure.

Table 2. DOE Environmental System Science Projects following the ESIP Physical Sample Curation Cluster's author guidelines, including linking source samples, subsamples, datasets, and published papers to the extent possible with current infrastructure.

Project	Sample Source Material IDs (IGSN)	ESS-DIVE Dataset(s)
LBNL Watershed Function Scientific Focus Area (SFA)	53 Soil Samples	doi:10.21952/WTR/1573029 doi:10.15485/1577267
LBNL Belowground Biogeochemistry SFA	60 Soil Samples	doi:10.15485/1830417

LLNL SBR SFA - Biogeochemistry of Actinides	5 Sites,195 Sediment Cores,19 Sediment Samples, 83 Water Samples	doi:10.15485/1910298 doi:10.15485/1910299
River Corridor and Watershed Biogeochemistry SFA	97 Sites, 97 Water Samples, 97 Filters, 290 Sediment Samples	doi:10.15485/1603775 doi:10.15485/1729719

Author Checklist Step 1 & 2: For each sample, we documented standard metadata following SESAR requirements, extended for Environmental System Science samples [33–35,65]. We then submitted the standard metadata to SESAR to obtain IGSN IDs for each sample. All sample metadata is now readily accessible through SESAR's API or by visiting the sample landing page(s). We recorded the relationships among samples and subsamples in the following ways:

- Sample relationships were recorded by documenting Parent IGSN ID. SESAR landing pages then populate links to all related parent and sibling samples.
- IGSN IDs were listed as the source material sample IDs sent to JGI and EMSL for 'omics analyses.
- We updated sample metadata for individual IGSN IDs when analyses were completed and published on JGI and ESS-DIVE, 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: We published seven datasets [55,56,Table 2; 66–70] that included samples with PIDs/IGSN IDs, along with links to related data. Each dataset includes IGSN ID 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 IGSN IDs 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 ID.

- The ESS-DIVE data repository currently does not have a specific metadata field for samples or related identifiers. Implementing related identifiers, including samples, will help support sample tracking. We need a user-friendly way to list and display a large number of samples associated with the dataset; for example, by automatically extracting the IGSN IDs from the sample metadata and/or data files, and linking to sample landing pages.
- The data repository could enforce IGSN IDs in sample data files, or provide an automated way to connect sample names to IGSN IDs as we develop tools for advanced search within data files.
- IGSN IDs should be provided within a sample identifier field, and SESAR API should be used to harvest relevant metadata for the samples that would support sample search.
- Small data repositories may struggle with resources needed to build and maintain new tools to support these new functionalities.
- The sample datasets in the ESS-DIVE data repository were cited in associated journal publications.

Author Checklist Step 4: We did not reference individual sample PIDs in associated papers, because papers involved hundreds of samples. Instead, we cited the sample datasets (Step 3),

which in turn reference the individual samples.

- Following paper publication, we added links on the sample landing pages to associated publications, which is highly inefficient. This should be done automatically when datasets and associated samples are cited in a publication.

Discussion

The author guide for publishing sample-based research (summarized above) is one step towards enabling physical sample discovery, tracking, and attribution. However, author guidelines alone are not enough; there are multiple ways in which scientists, repositories, PID organizations, publishers, and citation indexers need to further develop the physical sample research ecosystem (Figure 4). Additionally, there are notable ongoing infrastructural developments and efforts that can support adaptation and use of these guidelines.

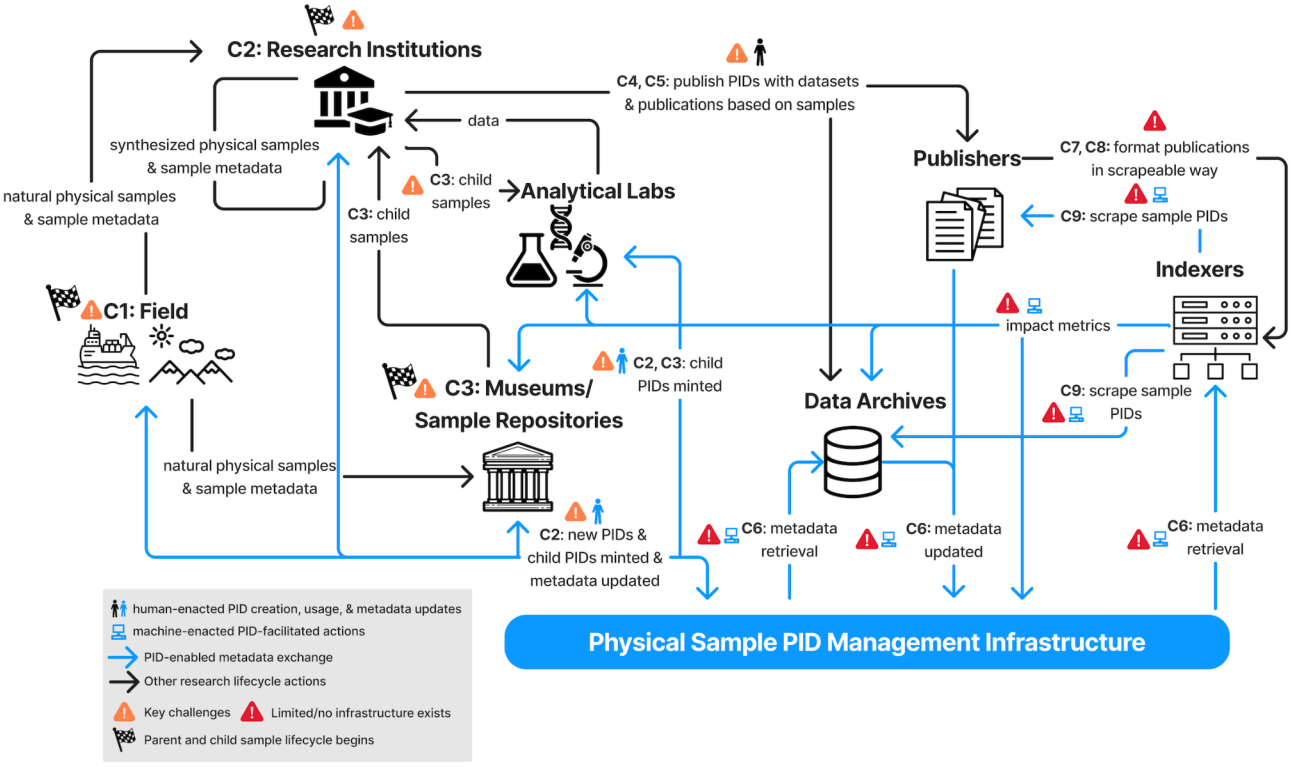

Figure 4. Diagram of the sample-based research ecosystem, with emphasis on the role of physical sample PIDs (blue arrows and icons) and human vs. machine-enacted actions (blue and black icons). Parts of the ecosystem that need development are called out with symbols **C*N***, including key challenges (orange symbols) and areas where limited or no infrastructure currently exists (red symbols). **C1:** Minimal technology for assigning PIDs in the field. **C2:** Few efficient ways to maintain links between child and parent samples and mechanisms for automated metadata transfer. **C3:** Missing protocols for where and when to assign PIDs (including child PIDs). **C4:** Limited to no guidance for sample citation. **C5:** Publications lacking space and/or section structures for samples citation. **C6:** Few technical links between PID Management Infrastructure and repositories or indexers, with some exceptions such as EarthChem. **C7:** No

editorial checks for appropriate PID usage. **C8:** PIDs not exposed in repositories and journals in scrapeable ways. **C9:** Samples (and some data) not indexed.

Adoption of Standard Practices

One of the biggest obstacles to supporting sample tracking and citation is cultural; it is simply not the norm for most scientists to precisely cite their samples at this time and many researchers do not see the value in taking additional time to deploy PIDs when they are not required by journals and funders. Scientists may not be aware of the possibility and benefits of citing PIDs, and lacking incentives otherwise, they follow disciplinary traditions. We need to promote widespread adoption of standard practices for publishing sample-based research. This involves advancing the incentives that encourage researchers to follow recommended practices, and tools that make this process easier and rewarding. Such incentives include having citation counts and records of where and how samples and associated (meta)data are used [50,71].

Clearly citing samples used also improves the ability of future researchers to find, access, integrate, and reuse existing relevant physical samples for new work (which may include access to materials that no longer exist or are no longer available to be sampled). This saves time and money, and enables science that would not be possible otherwise [72]. Perhaps most importantly, new fields of research become possible when communities publish data using standard practices for identifiers and (meta)data [see 9,10]. Such standard practices enable useful tools for data discovery, integration, analysis, and/or visualization. GBIF now supports new publications every day, addressing topics such as conservation, species distribution, climate change impacts, macroecological patterns, and more [9,10]. Reuse of omics datasets has contributed to research with diverse applications, for example in the industrial biotechnology [73] and biomedical [74] sectors, and has enabled researchers to better understand biological effects of climate change[75].

We can promote a culture of sample citation and PID use through mentorship and training of early career researchers, as well as through funding and journal requirements. Some funders now include sample PIDs as a recommendation or requirement in data management plans, which is an important step. For example, the US National Science Foundation (NSF) GEO data and sample policy requests IGSN ID registration through SESAR [12]. supplement already requests IGSN registration, and NSF GEO data policies request registration of samples with SESAR. And journals can include guidance for samples in their publication requirements (for example, AGU includes IGSN IDs in their guidance for authors). We can also provide guidance and support for retiring researchers and their collections.

Widespread adoption of standards for consistent sample metadata across different institutions may present further challenges due to differences in operating procedures and technical or institutional barriers to change, such as mapping lab-internal metadata (for example, from Laboratory Information Management Systems [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. This will help both funders and institutions better track the impact of their research.

Recommendations for Research Institutions, Physical Collections, and Laboratories

The research institutions, physical collections, and laboratories that manage physical samples have a major role to play in facilitating sample citation and tracking. To encourage consistent sample citation, the managers of physical sample collections should register their samples with PIDs that researchers can use in their publications. For many existing (legacy) curated or long-term collections, sample PIDs are more effectively assigned and managed by the physical repository that holds them or by institutional (meta)data repositories. For example, PID registration can be more readily incorporated into required (and ideally automated) workflows throughout the sample collection and management life cycle. Once PIDs are assigned, research institutions, physical sample repositories, and laboratories analyzing samples may additionally facilitate sample tracking and citation by working with researchers to mint child PIDs for any subsamples taken from their collections.

However, we recognize that there are many barriers to PID adoption by physical repositories. First, the time to assign PIDs to collections is not trivial; it requires gathering, organizing, and registering sample metadata and including the PID in digital sample catalogs and could involve physically re-labelling samples, which is a laborious task that many collections do not have the staff to support. However, another option can be to include PIDs in the repository or museum digital catalog to enable linking and updates over time. Second, it requires access to sample PID allocation services and infrastructure to support sample and data management. While there are some PID allocation services, like EZID, that are free, these require significant time and technology investments to adopt. Other services, like SESAR provide curation services and a user-friendly sample registration interface for no fee.

For newly collected samples, PIDs and standard metadata can be effectively assigned in the field at the time of sample collection using automated "Dirt to Desktop" field apps. GPS enabled field apps automate the capture of precise geographical coordinates at time of collection, and can be preset to collect a consistent set of metadata attributes for a major field sampling campaign. Not all sampling locations have internet access, but the information can be stored off line and automatically loaded to the home database when an internet connection becomes available [76]. These apps also remove the chance of transcription errors and save time and money [77] and repositories and long-term collections should encourage researchers that store their samples with them to use these apps when feasible. The PID and its associated metadata can then be shared with the long-term repository for the physical sample for archival, sharing and curation. There are a number of apps that already support this work, such as Dirt to Desktop [76] and StraboSpot [78], and there are more in development. By using these apps, the IGSN ID, the geolocation, time of collection, and other critical metadata are consistently captured in the field and then the sample is ready for submission on return.

Recommendations for Data Repositories, Journal Publishers, and Indexers

Clear Guidelines for Publishing and Citing Sample-Based Research

Journal publishers must recognize the role of citations beyond their current focus on research articles, and require citations for datasets, physical samples, and beyond. While our author guidelines provide high level guidance on the kind of metadata needed for sample tracking, journals need to provide complementary guidance for their specific publications. Some journal publishers already provide data and software citation guidance [79]; similar author instructions are needed on where and how to cite samples in publications and/or associated datasets. This includes information about how to encode sample PIDs so that they become linked in the publication process (for example, Elsevier provides this guidance for authors of Palaeogeography, Paleoclimatology, Palaeoecology, Earth and Planetary Science Letters). This guidance should outline procedures for all components of a paper (how to cite sample PIDs in line in text, in tables, and how they should appear in Data Availability statements or reference sections) or dataset. We hope that our author guidelines provide journal and data publishers with a starting point and are eager to collaborate with publishers in this important work.

Editors, reviewers, and research authors can help advocate for these policy changes in their interactions with publishers. During the review process, journal and data publishers should ensure that PIDs are formatted in a way that they can be easily harvested or indexed and are reliably linked to related metadata records. Editor and reviewer guidelines can play a supporting role in encouraging authors to use sample PIDs where appropriate. Development, training, and uptake of editorial checks for ensuring consistent linking of sample PIDs is a significant undertaking.

Sample repositories can contribute to this guidance by including a "how to cite this sample PID" section within the sample landing page that gives direct reference authors can use in their manuscript. The citation guide for each repository should match the current IGSN ID citation guidelines from DataCite, which allow for multiple formats.

Coordination and Integration across Systems

Sample (meta)data and data repositories are often siloed and need better integration across one another, as well as connections to journal publishers. For example, many Allocating Agents of IGSN IDs are specific to a country, discipline, or organization. This is beneficial to researchers when community-relevant and user-friendly tools are provided for data management. However, these distributed services mean that researchers must search multiple systems to find sample data. Additionally, these distributed services are often not connected to other key systems where associated metadata and data are added over time, such as laboratories, data repositories, and journal publishers.

Data repositories and publishers must coordinate and implement community practices and technical solutions that enable automated linking and information exchange described below. Groups such as the ESIP Physical Sample Curation Cluster (including many of the authors of this paper), the RDA Physical Samples and Collections in the Research Data Ecosystem, RDA Coordinating Earth, Space, and Environmental Science Data Preservation and Scholarly Publication Processes Working Group, and the Coalition for Publishing Data in the Earth and Space Sciences (COPDESS) [80] can help promote and facilitate such coordination.

There are also emergent infrastructure development projects that aim to bridge these silos. For example, the iSamples project builds connections across distributed sample metadata catalogs by aggregating sample metadata into iSamples Central [11]. This aggregation means that researchers would only need to search for samples in one place, rather than in multiple repositories. Additionally, several US federal agencies have plans to develop federated systems allowing discovery and access to federally-funded data and articles [for example, 81]. There is generally a push towards 'open research commons' within geoscience and more broadly. These important efforts should include samples and associated (meta)data as a major component.

Related Identifiers and Connection Metadata

Sample PIDs and standard metadata are the foundational elements necessary to track and update provenance information. Yet, many data archives do not have dataset metadata fields specifically for samples and other related identifiers. We recommend that data repositories implement related identifiers as part of dataset metadata, particularly sample PIDs, but also ORCiDs for people and RORs for organizations. Specifically, data repositories serving sample-based research should further provide functionality to recognize sample PIDs as related entities associated with and cited by the dataset. For example, EarthChem automatically extracts IGSN IDs from data files, and clearly displays links to the samples on the dataset landing page (for example, doi:0.26022/IEDA/112300).

Sample PIDs should be linked to other identifiers using defined relationship types, such as DataCite related identifiers and relation types [82]. 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 "related identifier+relationship type" enables DataCite to automatically create and track directional linkages. Furthermore, we need to make these related identifiers agnostic to identifier type, going beyond DOIs to include the range of identifiers in use, such as Archival Resource Keys (ARKs), BioSample Accession numbers [44,83], and more.

Tracking Use of Samples and Complex Citations

We have found that scientists and sample managers face similar challenges with regards to sample tracking across specific use-cases and disciplines. And believe that a limited set of community practices and improved infrastructure designed to track sample use can solve many current challenges and enhance sample discovery, integration, and use. A key element of these recommendations is the wide implementation and adoption of the Sample PID, which provides a powerful way to link and exchange relevant scientific information across facilities and data systems.

All institutions involved in sample collection and (meta)data life cycle can contribute to a network of related identifiers that links (meta)data across PID registries and related research outputs. If sample PIDs and related identifiers are captured in parent-child sample records and dataset metadata, we can design APIs to efficiently cross-link and exchange information where needed across sample repositories, data repositories, journals, and more when samples PIDs are referenced. This will make it possible to track the use of samples and attribute appropriate credit to those involved in sample collection, management, and analysis, as well as document provenance and relationships that make samples and associated data more useful. Tracking sample use will often require traversing multiple links in a graph of related PIDs. For example, this may involve a paper citing a dataset, the dataset citing analyses done on subsamples, and subsamples citing the original source sample collected in the field and/or archived in a museum. Currently, there are few effective ways of doing this traversal, making it challenging to track sample usage *en masse*.

Citation and usage metrics work fairly well for journal publications and researchers, improvements are needed [84], including better support for data and sample citation. Indexers that currently provide paper and data citation metrics, such as CrossRef and DataCite, need to recognize samples as an entity in tracking metrics. Further, at the present time, metrics and usage tracking are only available for DOIs. We need metrics and usage tracking to be implemented for a range of identifiers in order to make sample-based research truly open and FAIR. Existing initiatives, such as the Make Data Count effort, are working towards making data citation work more consistently [85,86]. Perhaps there is a need for a parallel "Making Samples Count" initiative.

We encourage indexers to explore the work of the RDA Complex Citation Working Group, which has outlined needs across multiple use cases to enable citing large numbers of objects (that may originate across multiple data systems) in a single container citation. One of the key use cases for complex citations is to make it possible for authors to cite as many samples as needed in a paper or dataset in a machine-readable way, with the goal to enable both provenance tracking and credit. Indexers then need to actually harvest those citations accurately from datasets and journal articles.

New fields of research become possible when communities publish data using standard practices for identifiers and (meta)data. Adoption of these practices is needed to create useful tools for sample and (meta)data discovery, integration and use. In this paper, we have described the need for citation guidelines, cultural changes, and infrastructure development to better facilitate physical sample discovery, citation and tracking. Through years of iterative development, we created author guidelines for sample citation as one step towards this vision. Other actors, such as data repositories and publishing entities, can use and adapt the author guidelines developed by the ESIP Physical Samples Curation Cluster to provide clear guidelines for authors submitting data and journal publications [63]. These guidelines would enable future development of automated tools to track sample use over time, while making samples and associated data open and FAIR.

Methods

The Earth Science Information Partners (ESIP) is a 501(c)(3) nonprofit supported by NASA, NOAA, USGS and 130+ member organizations, providing leadership in promoting the collection, stewardship and use of Earth science data, information and knowledge. This includes about 30 collaboration areas where members meet regularly to work together on common data challenges. The Physical Samples Curation Cluster is one such group that we organized in January 2021 to promote discovery, access, and use of physical samples and associated data. Members and our target community includes researchers who collect/identify/analyze/use samples and related data products, professionals who manage samples in physical collections, data repository managers and other cyber infrastructure providers who support tools and services for physical samples. This includes subject-matter experts from universities, federal organizations such as the U.S. Geological Survey, NASA, NOAA, US Department of Energy, major U.S. scientific sample repositories such as the USGS Core Research Center and the Oregon State University Marine and Geology Repository, data repositories such the Interdisciplinary Earth Data Alliance and the ESS-DIVE data repository, and the international IGSN e.V.

The working group started with a goal of addressing social and technical needs for tracking and publishing sample-related research across scientific disciplines.

Use Case Needs for Tracking Sample Use

In community discussions, we identified specific use cases that demonstrate common needs, across disciplines, for tracking samples to better support sample and data management, data synthesis, and appropriate credit for researchers and institutions. Here we present a) background information, b) example use case(s), and c) a summary of needs and challenges for the following priorities identified by the sample data community:

1. Efficiently publish and cite large number of samples and associated (meta)data;
2. Provide credit for those involved in physical sample collection and curation to demonstrate value of investing in collections;
3. Track use of sample data generated by analysts and laboratories; and
4. Connect related interdisciplinary sample (meta)data and other research outputs.

These real use-cases encountered in our work as sample-data experts provide additional background and testing to inform the final recommendations for scientific authors, journal publishers, data repositories, and indexers presented in the results section. We then tested the recommendations by applying them to samples and associated datasets that illustrate connecting related interdisciplinary sample data.

Drafting Guidelines Through Community Feedback and Review

We developed author guidelines for researchers submitting scientific publications involving physical samples. These guidelines are based on existing best practices [26,29,65], use cases that illustrate current challenges and needs, and extensive community feedback and review.

We gathered community feedback in regular working meetings and conference sessions. During monthly meetings we held discussions, working sessions, and relevant talks from stakeholders on challenges, needs, and visions for publishing and tracking scientific samples and associated data. The group engaged the broader community by convening seven conference sessions at bi-annual ESIP meetings, the 2022 American Geophysical Union meeting, and the Society for the Preservation of Natural History Collections conference. We designed ESIP sessions in particular to collect specific feedback through individual reflection (via digital collaborative documents and whiteboards), community discussion, and anonymous poll/survey questions [97,98]. We gathered input from community presentations, and feedback during community meetings, which informed drafts of the guidelines and improved later versions.

To further refine the guidelines, we coordinated with several related projects and international efforts in relevant communities. This included the Sampling Nature Research Coordination Network, Internet of Samples (iSamples) project [11], Australian Research Data Commons (ARDC) Information Management for Physical Samples Community of Practice, Research Data Alliance (RDA) Complex Citations Working Group, and the RDA Physical Samples and Collections in the Research Data Ecosystem Interest Group.

Example of applying recommended practices

We worked through several examples where projects from a specific community not previously using standard sample identifiers and metadata applied recommended practices. The ESS-DIVE data repository worked in-depth with four scientific projects from the Department of Energy's (DOE's) Biological and Environmental Research (BER) program program to apply recommended practices to the extent possible for interdisciplinary data across multiple data systems, using current infrastructure. This involved interdisciplinary data to test how we could track analysis and use of Samples sent to multiple labs and facilities and then published data and associated papers.

Each project assigned IGSN IDs and standard metadata to their samples [65], and 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. We worked across five laboratories and data systems to ensure that the IGSN ID 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.

We used this experience to refine the guidelines, and identify additional community and infrastructure needs to make this kind of cross-linking and provenance tracking more feasible, accurate, and useful.

Acronyms

Table 3. List of acronyms used throughout the paper, with links to more information.

Abbreviation	Name and link to more information
AMNH	American Museum of Natural History
API	Application Programming Interface
ARDC	Australian Research Data Commons
ARK	Archival Resource Key identifier
BER	Office of Biological and Environmental Research
BRC	Biological Resource Center
CARE	Collective Action, Authority to Control, Responsibility, and Ethics
COPDESS	Coalition for Publishing Data in the Earth and Space Sciences
U.S. DOE	United States Department of Energy
DOI	Digital Object Identifier
EAR	NSF Division of Earth Sciences
ECL	Earthchem Library
EMSL	Environmental Molecular Sciences Laboratory
ESIP	Earth Science Information Partners
ESS	Environmental System Science Program
ESS-DIVE	Environmental System Science Data Infrastructure for a Virtual Ecosystem
EZID	University of California Identifiers Service
FAIR	Finability, Accessibility, Interoperability, and Reusability
GBIF	Global Biodiversity Information Facility
GFZ	German Research Centre for Geosciences (GeoForschungsZentrum)
ID	Identifier
IEDA2	Interdisciplinary Earth Data Alliance
IGSN e.V.	IGSN Implementation Organization
IGSN ID	International Generic Sample Number
iSamples	Internet of Samples

JGI	Joint Genome Institute
JGI GOLD	Joint Genome Institute Genomes OnLine Database
Kbase	Department of Energy Systems Biology Knowledgebase
LBNL	Lawrence Berkeley National Laboratory
LIMS	Laboratory Information Management System
LLNL	Lawrence Livermore National Laboratory
ORCiD	Open Researcher and Contributor ID
NASA	National Aeronautics and Space Administration
NCBI	National Center for Biotechnology Information
NMDC	National Microbiome Data Collaborative
NOAA	National Oceanic and Atmospheric Administration
NSF	National Science Foundation
PID	Persistent Identifier
PNNL	Pacific Northwest National Laboratory
RDA	Research Data Alliance
SESAR	System for Earth and Extraterrestrial Registration
UMMZ	University of Michigan Museum of Zoology
URL	Uniform Resource Locator
USGS	United States Geological Survey
UUID	Universally Unique Identifier

Data Availability

The resulting "Scientific Author Guide for Publishing Open Research Using Physical Samples," as well as relevant community meeting presentations are available in the ESIP Figshare research repository [63,64,97,98].

The ESS-DIVE data repository has 26 datasets (as of March 2024) compiled into an data portal collection for Environmental System Science samples, each of which includes IGSN IDs and standard metadata for associated samples. This includes seven datasets with detailed links to related samples and other research outputs [55,56,66–70].

Code Availability

No new code was generated in this work.

References

1 Haller A, Janowicz K, Cox SJD, Lefrançois M, Taylor K, Le Phuoc D *et al.* The modular SSN ontology: A joint W3C and OGC standard specifying the semantics of sensors, observations, sampling, and actuation. *Semantic Web* 2018. doi:10.3233/SW-180320.

2 Janowicz K, Haller A, Cox S, Phuoc DL, Lefrancois M. SOSA: A Lightweight Ontology for Sensors, Observations, Samples, and Actuators. 2018. doi:10.2139/ssrn.3248499.

3 Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 2016; **3**: 160018.

4 National Academies of Sciences, Engineering, and Medicine, Policy and Global Affairs, Board on Research Data and Information, Committee on Toward an Open Science Enterprise. *Open Science by Design: Realizing a Vision for 21st Century Research*. National Academies Press, 2018.

5 UNESCO Recommendation on Open Science. 2023.https://www.unesco.org/en/open-science/about (accessed 20 Mar2024).

6 Mons B, Neylon C, Velterop J, Dumontier M, da Silva Santos LOB, Wilkinson MD. Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. *Inf Serv Use* 2017; **37**: 49–56.

7 McNutt M, Lehnert K, Hanson B, Nosek BA, Ellison AM, King JL. Liberating field science samples and data. *Science* 2016; **351**: 1024–1026.

8 Sidlauskas B, Ganapathy G, Hazkani-Covo E, Jenkins KP, Lapp H, McCall LW *et al.* Linking big: the continuing promise of evolutionary synthesis. *Evolution* 2010; **64**: 871–880.

9 Heberling JM, Miller JT, Noesgaard D, Weingart SB, Schigel D. Data integration enables global biodiversity synthesis. *Proc Natl Acad Sci U S A* 2021; **118**. doi:10.1073/pnas.2018093118.

10 Ball-Damerow JE, Brenskelle L, Barve N, Soltis PS, Sierwald P, Bieler R *et al.* Research

applications of primary biodiversity databases in the digital age. *PLoS One* 2019; **14**: e0215794.

11 Davies N, Deck J, Kansa EC, Kansa SW, Kunze J, Meyer C *et al.* Internet of Samples (iSamples): Toward an interdisciplinary cyberinfrastructure for material samples. *Gigascience* 2021; **10**. doi:10.1093/gigascience/giab028.

12 US National Science Foundation. Division of Earth Sciences (EAR) Data and Sample Policy Division of Earth Sciences National Science Foundation. 2023https://www.nsf.gov/geo/geo-data-policies/ear/ear-data-policy-jul2023.pdf.

13 Troudet J, Vignes-Lebbe R, Grandcolas P, Legendre F. The Increasing Disconnection of Primary Biodiversity Data from Specimens: How Does It Happen and How to Handle It? *Syst Biol* 2018; **67**: 1110–1119.

14 Shiffrin RM, Börner K, Stigler SM. Scientific progress despite irreproducibility: A seeming paradox. *Proc Natl Acad Sci U S A* 2018; **115**: 2632–2639.

15 Thessen AE, Woodburn M, Koureas D, Paul D, Conlon M, Shorthouse DP *et al.* Proper attribution for curation and maintenance of research collections: Metadata recommendations of the RDA/TDWG working group. *Data Sci J* 2019; **18**: 54.

16 Rouhan G, Dorr LJ, Gautier L, Clerc P, Muller S, Gaudeul M. The time has come for Natural History Collections to claim co-authorship of research articles. *Taxon* 2017; **66**: 1014–1016.

17 Deck J, Gaither MR, Ewing R, Bird CE, Davies N, Meyer C *et al.* The Genomic Observatories Metadatabase (GeOMe): A new repository for field and sampling event metadata associated with genetic samples. *PLoS Biol* 2017; **15**: e2002925.

18 Pope LC, Liggins L, Keyse J, Carvalho SB, Riginos C. Not the time or the place: the missing spatio-temporal link in publicly available genetic data. *Mol Ecol* 2015; **24**: 3802–3809.

19 Roche DG, Kruuk LEB, Lanfear R, Binning SA. Public Data Archiving in Ecology and Evolution: How Well Are We Doing? *PLoS Biol* 2015; **13**: e1002295.

20 Klump J, Lehnert K, Ulbricht D, Devaraju A, Elger K, Fleischer D *et al.* Towards globally unique identification of physical samples: Governance and technical implementation of the IGSN global sample number. *Data Sci J* 2021; **20**. doi:10.5334/dsj-2021-033.

21 Schindel DE, Cook JA. The next generation of natural history collections. *PLoS Biol* 2018; **16**: e2006125.

22 Carroll SR, Garba I, Figueroa-Rodríguez OL, Holbrook J, Lovett R, Materechera S *et al.* The CARE principles for indigenous data governance. *Data Sci J* 2020; **19**. doi:10.5334/dsj-2020-043.

23 Guralnick R, Conlin T, Deck J, Stucky BJ, Cellinese N. The Trouble with Triplets in

Biodiversity Informatics: A Data-Driven Case against Current Identifier Practices. *PLoS One* 2014; **9**: e114069.

24 Peyrard S, Tramoni J-P, Kunze J. The ARK Identifier Scheme: Lessons Learnt at the BnF and Questions Yet Unanswered. 2014.https://escholarship.org/uc/item/58d52295 (accessed 20 Nov2019).

25 Klump J, Huber R. 20 Years of Persistent Identifiers – Which Systems are Here to Stay? *Data Science Journal* 2017; **16**: 9.

26 McMurry JA, Juty N, Blomberg N, Burdett T, Conlin T, Conte N *et al.* Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data. *PLoS Biol* 2017; **15**. doi:10.1371/journal.pbio.2001414.

27 Lehnert KA, Goldstein SL, Lenhardt C, Vinayagamoorthy S. SESAR: Addressing the need for unique sample identification in the Solid Earth Sciences. 2004, p SF32A–06.

28 Lehnert KA, Klump J, Arko RA, Bristol S, Buczkowski B, Chan C *et al.* IGSN e.V.: Registration and Identification Services for Physical Samples in the Digital Universe. *AGU Fall Meeting Abstracts* 2011; **13**: IN13B–1324.

29 Lehnert K, Klump J, Wyborn L, Ramdeen S. Persistent, Global, Unique: The three key requirements for a trusted identifier system for physical samples. *Biodiversity Information Science and Standards* 2019; **3**: e37334.

30 Lehnert K, Klump J, Ramdeen S, Wyborn L, Haak L. IGSN 2040 Summary Report: Defining the Future of the IGSN as a Global Persistent Identifier for Material Samples. 2021 doi:10.5281/zenodo.5118289.

31 Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R *et al.* Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. *PLoS One* 2012; **7**: e29715.

32 Yilmaz P, Kottmann R, Field D, Knight R, Cole JR, Amaral-Zettler L *et al.* Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nat Biotechnol* 2011; **29**: 415–420.

33 System for Earth Sample Registration (SESAR). SESAR XML Schema for samples. 2020 doi:10.5281/zenodo.3875531.

34 System for Earth Sample Registration (SESAR). SESAR Batch Registration Quick Guide. 2020 doi:10.5281/zenodo.3874923.

35 Damerow J, Varadharajan C, Boye K, Brodie E, Chadwick D, Cholia S *et al.* Sample Identifiers and Metadata Reporting Format for Environmental Systems Science. Environmental Systems Science Data Infrastructure for a Virtual Ecosystem (ESS-DIVE). 2020. doi:10.15485/1660470.

36 Strasser BJ. Genetics. GenBank--Natural history in the 21st Century? *Science* 2008; **322**: 537–538.

37 Robertson T, Gonzalez ML, Høfft M, Grosjean M. Documenting Natural History Collections in GBIF. *Biodiversity Information Science and Standards* 2019; **3**. doi:10.3897/biss.3.37216.

38 Sayers EW, Cavanaugh M, Clark K, Pruitt KD, Sherry ST, Yankie L *et al.* GenBank 2023 update. *Nucleic Acids Res* 2023; **51**: D141–D144.

39 Miller SE, Barrow LN, Ehlman SM, Goodheart JA, Greiman SE, Lutz HL *et al.* Building Natural History Collections for the Twenty-First Century and Beyond. *Bioscience* 2020; **70**: 674–687.

40 Cui X, Mucci A, Bianchi TS, He D, Vaughn D, Williams EK *et al.* Global fjords as transitory reservoirs of labile organic carbon modulated by organo-mineral interactions. *Sci Adv* 2022; **8**: eadd0610.

41 Whitlock MC. Data archiving in ecology and evolution: best practices. *Trends Ecol Evol* 2011; **26**: 61–65.

42 JGI Publication Policy. DOE Joint Genome Institute. 2024.https://jgi.doe.gov/user-programs/pmo-overview/policies/ (accessed 22 Mar2024).

43 Smith D. Culture Collections and Biological Resource Centres (BRCs). Encyclopedia of Industrial Biotechnology. 2009. doi:10.1002/9780470054581.eib246.

44 Barrett T, Clark K, Gevorgyan R, Gorelenkov V, Gribov E, Karsch-Mizrachi I *et al.* BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. *Nucleic Acids Res* 2012-1; **40**: D57–D63.

45 Agarwal D, Damerow J, Varadharajan C, Christianson D, Pastorello G, Cheah Y-W *et al.* Balancing the needs of consumers and producers for scientific data collections. *Ecol Inform* 2021; : 101251.

46 Lehnert K. EarthChem - FAIR data for geochemistry, volcanology, and petrology. 2023. doi:10.5281/zenodo.10737711.

47 Wallace KL, Bursik MI, Kuehn S, Kurbatov AV, Abbott P, Bonadonna C *et al.* Community established best practice recommendations for tephra studies-from collection through analysis. *Sci Data* 2022; **9**: 447.

48 Profeta L, Lehnert K, Ramdeen S, Ji P, Nielsen RL, Ustunisik GK *et al.* The IEDA2 Facility - Harmonizing FAIR Sample (Meta)Data for VGP Research. 2022, p V42A–05.

49 Lafia S, Thomer A, Thompson C, Cassidy E, Polasek K. Surfacing Specimen Citations: Machine Learning, Manual Annotation, and Impact Metrics for Natural History Collections. American Geophysical Union (AGU), 2022, p IN55A–01.

50 Byers N, Parker C, Beecroft C, Reddy TBK, Salamon H, Garrity G *et al.* Identifying genomic data use with the Data Citation Explorer. bioRxiv. 2024; : 2024.01.26.577091.

51 Cross-Domain Interoperability Framework (CDIF) Working Group, Richard S, Gregory A, Hodson S, Fils D, Kanjala C *et al.* Cross Domain Interoperability Framework (CDIF): Discovery Module (v01 draft for public consultation). 2023 doi:10.5281/zenodo.10252564.

52 Bouskill NJ, Riley WJ, Tang JY. Meta-analysis of high-latitude nitrogen-addition and warming studies implies ecological mechanisms overlooked by land models. *Biogeosciences* 2014; **11**: 6969–6983.

53 Stegen JC, Goldman AE. WHONDRS: a Community Resource for Studying Dynamic River Corridors. *mSystems* 2018; **3**: e00151–18.

54 Borton K. KBase Narrative - GROWdb US River Systems - Samples. 2022. doi:10.25982/109073.30/1895615.

55 Toyoda JG, Goldman AE, Chu RK, Danczak RE, Daly RA. WHONDRS Summer 2019 Sampling Campaign: Global River Corridor Surface Water FTICR-MS, NPOC, and Stable Isotopes. 2020.https://data.ess-dive.lbl.gov/view/doi:10.15485/1603775 (accessed 16 Nov2020).

56 Goldman AE, Arnon S, Bar-Zeev E, Chu RK, Danczak RE, Daly RA *et al.* 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). 2020. doi:10.15485/1729719.

57 Borton MA, Collins SM, Graham EB, Garayburu-Caruso VA, Goldman AE, de Melo M *et al.* It Takes a Village: Using a Crowdsourced Approach to Investigate Organic Matter Composition in Global Rivers Through the Lens of Ecological Theory. *Frontiers in Water* 2022; **4**. doi:10.3389/frwa.2022.870453.

58 Garayburu-Caruso VA, Danczak RE, Stegen JC, Renteria L, Mccall M, Goldman AE *et al.* Using Community Science to Reveal the Global Chemogeography of River Metabolomes. *Metabolites* 2020; **10**. doi:10.3390/metabo10120518.

59 Stadler M, Barnard MA, Bice K, de Melo ML, Dwivedi D, Freeman EC *et al.* Applying the core-satellite species concept: Characteristics of rare and common riverine dissolved organic matter. *Frontiers in Water* 2023; **5**. doi:10.3389/frwa.2023.1156042.

60 Buser-Young JZ, Garcia PE, Schrenk MO, Regier PJ, Ward ND, Biçe K *et al.* Determining the biogeochemical transformations of organic matter composition in rivers using molecular signatures. *Frontiers in Water* 2023; **5**. doi:10.3389/frwa.2023.1005792.

61 Gill IS, Griffiths EJ, Dooley D, Cameron R, Savić Kallesøe S, John NS *et al.* The DataHarmonizer: a tool for faster data harmonization, validation, aggregation and analysis of pathogen genomics contextual information. *Microb Genom* 2023; **9**. doi:10.1099/mgen.0.000908.

62 Wood-Charlson EM, Crockett Z, Erdmann C, Arkin AP, Robinson CB. Ten simple rules for getting and giving credit for data. *PLoS Comput Biol* 2022; **18**: e1010476.

63 Damerow J, Raia N, Stanley V, Byers N, Choe S, Edmunds R *et al.* A Scientific Author Guide for Publishing Open Research Using Physical Samples. 2024. doi:10.6084/m9.figshare.24669057.v1.

64 Raia N, Damerow J, Stanley V, Lehnert K, O'Ryan D, Plomp E *et al.* 4 Steps to Publish Open Earth Science Samples. 2023. doi:10.6084/m9.figshare.24291148.v1.

65 Damerow JE, Varadharajan C, Boye K, Brodie EL, Burrus M, Chadwick KD *et al.* Sample identifiers and metadata to support data management and reuse in multidisciplinary ecosystem sciences. *Data Sci J* 2021; **20**: 11.

66 Sorensen P, Brodie E, Beller H, Wang S, Bill M, Bouskill N. Soil nitrogen, water content, microbial biomass, and Archaeal, bacterial and fungal communities from the East River Watershed, Colorado collected in 2016-2017. 2019. doi:10.15485/1577267.

67 Sorensen P, Brodie E, Beller H, Wang S, Bill M, Bouskill N. Sample collection metadata for soil cores from the East River Watershed, Colorado collected in 2017. 2019. doi:10.21952/WTR/1573029.

68 Alves RJE, Callejas IA, Marschmann GL, Mooshammer M, Singh HW, Whitney B *et al.* Kinetic and temperature sensitivity properties of soil exoenzymes through the soil profile down to one-meter depth at a temperate coniferous forest (Blodgett, CA). 2021. doi:10.15485/1830417.

69 Coutelot F, Powell B. Biogeochemistry of pond B (Savannah River Site, South Carolina, USA): Sediment core, total extraction data, pond B Savannah River Site July 2019. Subsurface biogeochemistry of actinides SFA. 2023. doi:10.15485/1910299.

70 Merino N, Powell B, Coutelot F, Zavarin M, Kersting A, Jiao Y *et al.* Biogeochemistry of Pond B (Savannah River Site, South Carolina, USA): Water column and Sediments. Environmental System Science Data Infrastructure for a Virtual Ecosystem …, 2021https://data.ess-dive.lbl.gov/view/ess-dive-a2da471f864e297-20230808T205831128.

71 Colavizza G, Hrynaszkiewicz I, Staden I, Whitaker K, McGillivray B. The citation advantage of linking publications to research data. *PLoS One* 2020; **15**: e0230416.

72 Brown J, Jones P, Meadows A, Murphy F. Revised cost-benefit analysis for the UK PID Support Network. 2022 doi:10.5281/zenodo.7356219.

73 Stewart RD, Auffret MD, Warr A, Wiser AH, Press MO, Langford KW *et al.* Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. *Nat Commun* 2018; **9**: 870.

74 Bernheim A, Millman A, Ofir G, Meitav G, Avraham C, Shomar H *et al.* Prokaryotic viperins produce diverse antiviral molecules. *Nature* 2021; **589**: 120–124.

75 Hu R, Li X, Hu Y, Zhang R, Lv Q, Zhang M *et al.* Adaptive evolution of the enigmatic Takakia now facing climate change in Tibet. *Cell* 2023; **186**: 3558–3576.e17.

76 Ross S, Ballsun-Stanton B, Cassidy S, Crook P, Klump J, Sobotkova A. FAIRer Data through Digital Recording: The FAIMS Mobile Experience. 2022; **5**: 271–285.

77 Noble R, Reid N, Klump J, Robertson J, Cole D, Fox D *et al.* Testing a rapid sampling and analysis workflow in the remote Nullarbor Plain, Australia. *Newsletter for the Association of Applied Geochemists* 2020; **186**: 6–18.

78 Walker DJ, Tikoff B, Newman J, Clark R, Ash J, Good J *et al.* StraboSpot data system for structural geology. *Geosphere* 2019; **15**: 533–547.

79 Fox P, Erdmann C, Stall S, Griffies SM, Beal LM, Pinardi N *et al.* Data and Software Sharing Guidance for Authors Submitting to AGU Journals. 2021 doi:10.5281/zenodo.5124741.

80 Lehnert K, Hanson B, Sallans A, Elger K. COPDESS (Coalition for Publishing Data in the Earth & Space Sciences): An Update on Progress and Next Steps. 2016, pp EPSC2016–16120.

81 NSF National Center for Atmospheric Research. Computational and Information Systems Lab: Innovations in Open Science (IOS) Planning Workshop: Community Expectations for a Geoscience Data Commons. 2024.https://www2.cisl.ucar.edu/events/innovations-open-science-ios-planning-workshop-community-expectations-geoscience-data (accessed 23 Mar2024).

82 DataCite Metadata Working Group. DataCite Metadata Schema for the Publication and Citation of Research Data and Other Research Outputs. DataCite e.V., 2024 doi:10.14454/znvd-6q68.

83 Barrett T. *BioSample*. National Center for Biotechnology Information (US), 2013.

84 Stall S, Bilder G, Cannon M, Chue Hong N, Edmunds S, Erdmann CC *et al.* Journal Production Guidance for Software and Data Citations. *Sci Data* 2023; **10**: 656.

85 Kratz JE, Strasser C. Making data count. *Scientific Data* 2015; **2**: 1–5.

86 Cousijn H, Feeney P, Lowenberg D, Presani E, Simons N. Bringing Citations and Usage Metrics Together to Make Data Count. *CODATA* 2019; **18**: 9–9.

87 Buck M, Hamilton C. The Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from their Utilization to the Convention on Biological Diversity. *Rev Eur Community Int Environ Law* 2011; **20**: 47–61.

88 Williamson B, Provost S, Price C. Operationalising Indigenous data sovereignty in environmental research and governance. *Environment and Planning F* 2023; **2**: 281–304.

89 Carroll SR, Herczog E, Hudson M, Russell K, Stall S. Operationalizing the CARE and FAIR Principles for Indigenous data futures. *Sci Data* 2021; **8**: 108.

90 Sahagún L. Caltech says it regrets drilling holes in sacred Native American petroglyph site.

Los Angeles Times. 2021.https://www.latimes.com/environment/story/2021-07-19/caltech-fined-for-damaging-native-american-cultural-site (accessed 24 Mar2024).

91 Taitingfong R, Martinez A, Carroll SR, Hudson M, and Anderson J. Indigenous Metadata Bundle Communique. Collaboratory for Indigenous Data Governance, ENRICH: Equity for Indigenous Research and Innovation Coordinating Hub, and Tikanga in Technology. 2023https://indigenousdatalab.org/3006-2/.

92 Golan J, Riddle K, Hudson M, Anderson J, Kusabs N, Coltman T. Benefit sharing: Why inclusive provenance metadata matter. *Front Genet* 2022; **13**: 1014044.

93 Liggins L, Hudson M, Anderson J. Creating space for Indigenous perspectives on access and benefit-sharing: Encouraging researcher use of the Local Contexts Notices. *Mol Ecol* 2021; **30**: 2477–2482.

94 Lock M, McMillan F, Bennett B, Martire JL, Warne D, Kidd J *et al.* Position statement: Research and reconciliation with Indigenous peoples in rural health journals. *Aust J Rural Health* 2022; **30**: 6–7.

95 Hudson M, Garrison NA, Sterling R, Caron NR, Fox K, Yracheta J *et al.* Rights, interests and expectations: Indigenous perspectives on unrestricted access to genomic data. *Nat Rev Genet* 2020; **21**: 377–384.

96 Sheridan C. Kenyan dispute illuminates bioprospecting difficulties. *Nat Biotechnol* 2004; **22**: 1337.

97 Damerow J, Thomer A, Stanley V. How can we connect and track use of physical samples and associated data? 2023. doi:10.6084/m9.figshare.25483765.v1.

98 Damerow J, Thomer A, Stanley V. Community and Technical Needs to Facilitate Sample Citation. 2024. doi:10.6084/m9.figshare.25483771.v1.

Author Information

Authors and Affiliations

Earth and Environmental Sciences Area, Lawrence Berkeley National Laboratory, Berkeley, CA, 94720, USA
Joan Damerow, Dylan O'Ryan

College of Information Science, University of Arizona, Tucson, AZ, 85721, USA
Andrea Thomer, Natalie Raia

Marine and Geology Repository, Oregon State University, Corvallis, OR, 97331, USA
Val Stanley

Pacific Northwest National Laboratory, Richland, WA, 99354, USA
Amy Goldman, Brieanne Forbes, Beck Powers-McCormack

Lamont Doherty Earth Observatory, Columbia University, Palisades, NY, 10964, USA
Kerstin Lehnert, Sae Choe

Australian National University, Canberra ACT 2600, Australia
Lesley Wyborn

Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA, 94720, USA
Charles Parker, Neil Byers, T.B.K. Reddy

Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, CA, 94720, USA
Elisha Wood-Charlson

Colorado State University, Fort Collins, CO, 80526, USA
Mikayla Borton

Delft University of Technology, Lorentzweb 1, 2628 CJ Delft
Esther Plomp

Ronin Institute for Independent Scholarship
Sarah Ramdeen, John Kunze

University of Colorado Anschutz Medical Campus, USA
Anne Thessen

Geological Survey of Alabama, Tuscaloosa, AL, USA
Marcella McIntyre-Redden

Biodiversity Institute and Natural History Museum, University of Kansas, Lawrence, KS, 66045, USA
Dave Vieglais

DataCite, Welfengarten 1B, 30167 Hannover, Germany
Rorie Edmonds

American Geophysical Union, 2000 Florida Ave NW, 20009, Washington, DC, USA

Kristina Vrouwenvelder

University of Michigan, Museum of Zoology and Department of Ecology & Evolutionary Biology, Ann Arbor, MI, 48108, USA
Cody W. Thompson, Ellen R. Cassidy

Smithsonian Libraries and Archives, Washington, DC, USA
Richard Naples

Contributions

Conceptualization: J.E. Damerow, A. Thomer, V. Stanley, S. Ramdeen. Write section drafts: J.E. Damerow, N. Raia, S. Choe, A. Thomer, C. Parker, N. Byers. Contribute to use-case assessment: J.E. Damerow, A. Thomer, N. Raia, S. Choe, K. Lehnert, D. O'Ryan, N.Byers, M.A. Borton, C. Parker, E. Wood-Charlson, B. Forbes, A. Goldman, C.W. Thompson, S. Lafia, K. Forrest, R. Naples, E.R. Cassidy, T.B.K. Reddy, B. Powers-McCormack, S. Cholia. Review and edit: all co-authors.

Corresponding Author

Correspondence to Joan Damerow.

Competing Interests

The authors declare no competing interests.

Acknowledgements

J.E. Damerow, D. O'Ryan, and S. Cholia were supported through the ESS-DIVE repository by the U.S. DOE's Office of Science Biological and Environmental Research Program under contract number DE-AC02-05CH11231. A.Thomer, N. Raia, S. Choe, and K. Lehnert were supported through the iSamples Project NSF grant number 2004562 and SESAR NSF grant number 2148939. The work conducted by the U.S. Department of Energy Joint Genome Institute (https://ror.org/04xm1d337), a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy operated under Contract No. DE-AC02-05CH11231.

The PNNL River Corridor Science Focus Area portion of this work was supported by the United States Department of Energy, Office of Biological and Environmental Research (BER), Environmental System Science (ESS) Program. PNNL is operated by Battelle Memorial Institute for the United States Department of Energy under contract no. DE-AC05-76RL01830. In addition, RNA and DNA samples were processed through the "Creating the GROW (Genome

Resolved Open Watershed) Database: Leveraging Distributed Research Networks to Understand Watershed Systems" award (doi:10.46936/10.25585/60001289).

We thank those who contributed to the ESIP Physical Samples Curation Cluster and the feedback that all participants provided in developing the author guide described in this work. We also thank those who contributed to the DOE sample interoperability working group, and their feedback/work on approaches to link related samples and data across DOE BER data systems.