Internet of Samples: Progress report

Dave Vieglais[‡], Stephen M. Richard[§], Hong Cui[§], Neil Davies[|], John Deck[|], Quan Gan[§], Eric C. Kansa[¶], Sarah Whitcher Kansa[¶], John Kunze[#], Danny Mandel[§], Christopher Meyer[¤], Thomas M. Orrell[¤], Sarah Ramdeen[«], Rebecca Snyder[¤], Ramona L. Walls[§], Yuxuan Zhou[§], Kerstin Lehnert[«]

‡ University of Kansas, Lawrence, United States of America
§ University of Arizona, Tucson, United States of America
| University of California at Berkeley, Berkeley, United States of America
¶ The Alexandria Archive Institute, San Francisco, United States of America
# University of California, Oakland, United States of America
¤ National Museum of Natural History, Smithsonian Institution, Washington, DC, United States of America
« Columbia University, New York City, United States of America

Corresponding author: Hong Cui (hongcui@email.arizona.edu)

Abstract

Material samples form an important portion of the data infrastructure for many disciplines. Here, a material sample is a physical object, representative of some physical thing, on which observations can be made. Material samples may be collected for one project initially, but can also be valuable resources for other studies in other disciplines. Collecting and curating material samples can be a costly process. Integrating institutionally managed sample collections, along with those sitting in individual offices or labs, is necessary to faciliate large-scale evidence-based scientific research. Many have recognized the problems and are working to make data related to material samples FAIR: findable, accessible, interoperable, and reusable.

The Internet of Samples (i.e., iSamples) is one of these projects. iSamples was funded by the United States National Science Foundation in 2020 with the following aims:

1. enable previously impossible connections between diverse and disparate sample-based observations;
2. support existing research programs and facilities that collect and manage diverse sample types;
3. facilitate new interdisciplinary collaborations; and
4. provide an efficient solution for FAIR samples, avoiding duplicate efforts in different domains (Davies et al. 2021)

The initial sample collections that will make up the internet of samples include those from the System for Earth Sample Registration (SESAR), Open Context**,** the Genomic Observatories Meta-Database (GEOME), and Smithsonian Institution Museum of Natural History (NMNH), representing the disciplines of geoscience, archaeology/anthropology, and biology.

To achieve these aims, the proposed iSamples infrastructure (Fig. 1) has two key components: iSamples in a Box (iSB) and iSamples Central (iSC). The iSC component will be a permanent Internet service that preserves, indexes, and provides access to sample metadata aggregated from iSBs. It will also ensure that persistent identifiers and sample descriptions assigned and used by individual iSBs are synchronized with the records in iSC and with identifier authorities like International Geo Sample Number (IGSN) or Archival Resource Key (ARK). The iSBs create and maintain identifiers and metadata for their respective collection of samples. While providing access to the samples held locally, an iSB also allows iSC to harvest its metadata records.

The metadata modeling strategy adopted by the iSamples project is a metadata profile-based approach, where core metadata fields that are applicable to all samples, form the core metadata schema for iSamples. Each individual participating collectionis free to include additional metadata in their records, which will also be harvested by iSC and are discoverable through the iSC user interface or APIs (Application Programming Interfaces), just like the core. In-depth analysis of metadata profiles used by participating collections, including Darwin Core, has resulted in an iSamples core schema currently being tested and refined through use. See the current version of the iSamples core schema.

A number of properties require a controlled vocabulary. Controlled vocabularies used by existing records are kept, while new vocabularies are also being developed to support high-level grouping with consistent semantics across collection types. Examples include vocabularies for Context Category, Material Category, and Specimen Type (Table 1). These vocabularies were also developed in a bottom-up manner, based on the terms used in the existing collections. For each vocabulary, a decision tree graph was created to illustrate relations among the terms, and a card sorting exercise was conducted within the project team to collect feedback. Domain experts are invited to take part in this exercise here, here , and here. These terms will be used as upper-level terms to the existing category terms used in the participating collections and hence create connections among individual participating collections.

iSample project members are also active in the TDWG Material Sample Task Group and the global consultation on Digital Extended Specimens. Many members of the iSamples project also lead or participate in a sister research coordination network (RCN), Sampling Nature. The goal of this RCN is to develop and refine metadata standards and controlled vocabularies for the iSamples and other projects focusing on material samples. We cordially invite you to participate in the Sampling Nature RCN and help shape the future standards for material samples. Contact Sarah Ramdeen (sramdeen@ideo.columbia.edu) to engage with the RCN.

Keywords

iSamples system design, material samples, Sampling Nature RCN, sample material, specimen type, sampled feature, metadata profiles

Presenting author

Dave Vieglais

Presented at

TDWG 2021

Conflicts of interest

References

- Davies N, Deck J, Kansa EC, Kansa SW, Kunze J, Meyer C, Orrell T, Ramdeen S, Snyder R, Vieglais D, Walls RL, Lehnert K (2021) Internet of Samples (iSamples): Toward an interdisciplinary cyberinfrastructure for material samples. GigaScience 10 (5). https://doi.org/10.1093/gigascience/giab028

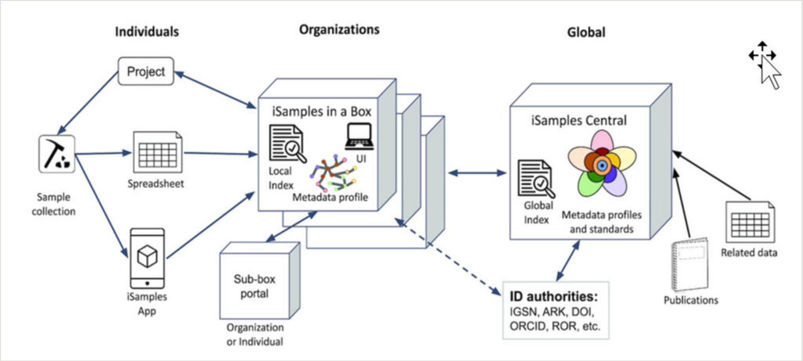

Figure 1.

Design of the iSamples system (used with permission from Davies et al. 2021).

Table 1.

Current controlled vocabulary terms for hasSpecimenCategories, hasMaterialCategories, and hasContextCategories for the iSamples project.

Specimen Type v2	Material Category v.3	ContextCategory v.20210703
decision graph of specimen type terms	decision graph of material terms	decision graph of sampled feature terms
Aggregation Analytical Preparation Anthropogenic aggregation Artifact Biome aggregation Experiment product Fossil Liquid or gas sample Organism part Organism product Other solid object Whole organism	Anthropogenic material Anthropogenic metal Biogenic non-organic material Dispersed media Gaseous material Liquid water Mineral Mixed soil sediment rock Non-aqueous liquid material Organic material Particulate Rock Sediment Soil Water ice	Active human occupation site Atmosphere Earth interior Experiment setting Extraterrestrial environment Glacier environment Laboratory or curation environment Lake river or stream bottom Marine water body Marine water body bottom Site of past human activities Subaerial surface environment Subsurface fluid reservoir Terrestrial water body