iSamples (Internet of Samples): Cyberinfrastructure to support transdisciplinary use of material samples

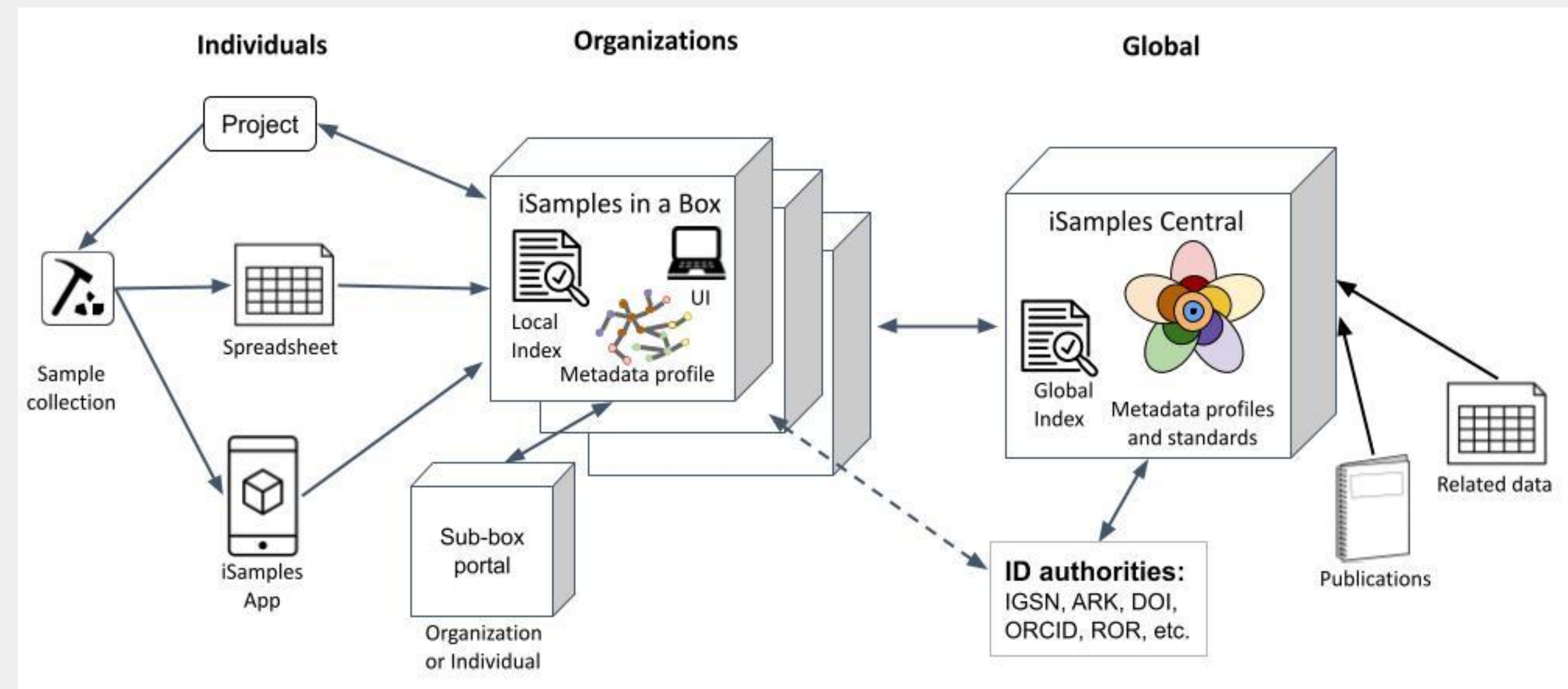Kerstin Lehnert[1], Sarah Ramdeen[1], Hong Cui[2], Neil Davies[3], John Deck[3], Eric Kansa[4], Sarah Kansa[4], John Kunze[5], Christopher Meyer[6], Thomas Orrell[6], Stephen Richard[7], Rebecca Snyder[6], Ramona Walls[2], David Vieglais[8]

1 Columbia University; 2 University of Arizona; 3 University of California, Berkeley; 4 Open Context, The Alexandria Archive Institute; 5 California Digital Library; 6 Smithsonian Institution; 7 Independent Contractor; 8 University of Kansas Biodiversity Institute

iSamples
The internet of samples

Material samples are indispensable data sources in many natural science, social science, and humanity disciplines. But harnessing sample-based data for science is cumbersome and often impractical because data about most material samples are difficult or impossible to Find, Access, Interoperate, and Reuse — they are simply not FAIR. As a consequence, the full value of material samples and the data derived from them is rarely realized, either for basic scientific research or societal applications.

The Internet of Samples (iSamples) project aims to integrate large, diverse, cross-discipline sample data repositories and enable material samples and sample data to be FAIR (Findable, Accessible, Interoperable, and Reusable).

iSamples Architecture

https://doi.org/10.1093/gigascience/giab028

Core Data Model

- Resolvable
  - Identifier uniquely identifies record
  - Record can be retrieved
  - Retrieval from multiple locations yields common properties
- Transformable
  - Machine readable
  - Well formed content; syntax and semantics
- Common fields
  - Core set of properties
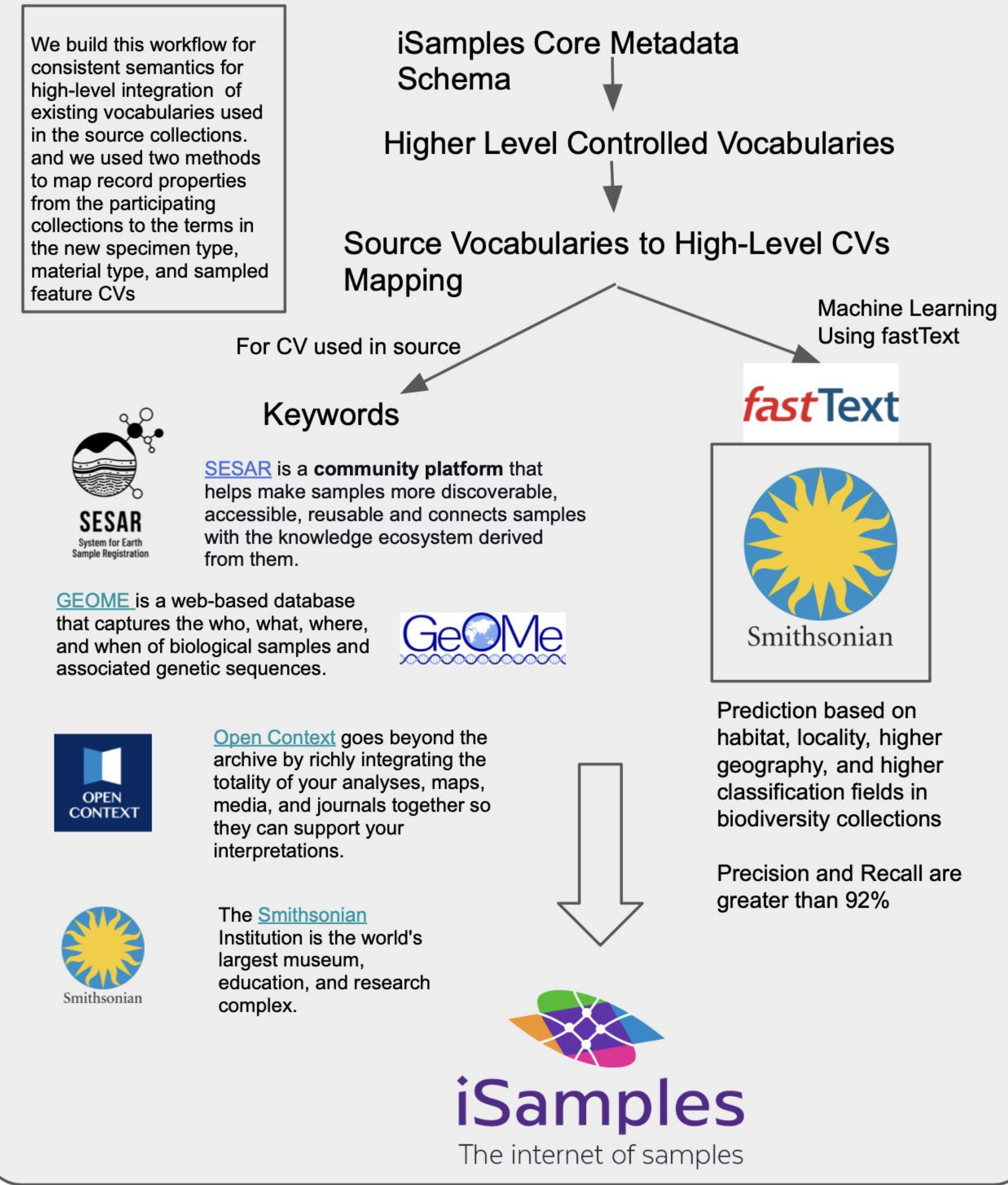 - Present in record or resolvable record lineage

iSamples-in-a-Box (iSB):

- Creates identifiers and associated metadata
- Allows discovery & retrieval of sample information
- Provides tools for updating sample metadata (e.g., augment or correct metadata or append provenance statements),
- Supports different scenarios
  - Reliable community services (e.g. SESAR, Open Context, GEOME)
  - Subset of capabilities ('sub-box') delivered to a smaller community of researchers, (e.g. field station)
    - Create identifiers and collect metadata that are later synchronized with the organization's iSB.

iSamples Central (iSC):

- Gateway between iSB instances and identifier authorities
  - Direct sample registrations to the selected identifier authority
  - Synchronize remote iSB content with relevant authorities
- Mediates metadata standards and data ingest rules
- Central discovery and resolution service
  - Search interface on the web and an API

Source Terminology to iSamples CV Mapping

We build this workflow for consistent semantics for high-level integration of existing vocabularies used in the source collections. and we used two methods to map record properties from the participating collections to the terms in the new specimen type, material type, and sampled feature CVs

iSamples Core Metadata Schema
↓
Higher Level Controlled Vocabularies
↓
Source Vocabularies to High-Level CVs Mapping

For CV used in source → Keywords

Machine Learning Using fastText

fastText

SESAR is a **community platform** that helps make samples more discoverable, accessible, reusable and connects samples with the knowledge ecosystem derived from them.

GEOME is a web-based database that captures the who, what, where, and when of biological samples and associated genetic sequences.

Open Context goes beyond the archive by richly integrating the totality of your analyses, maps, media, and journals together so they can support your interpretations.

The Smithsonian Institution is the world's largest museum, education, and research complex.

Prediction based on habitat, locality, higher geography, and higher classification fields in biodiversity collections

Precision and Recall are greater than 92%

iSamples
The internet of samples

Initial iSamples Implementations

OPEN CONTEXT
GEOME
Smithsonian
SESAR
System for Earth Sample Registration

MVP Target Implementation →

Each partner can implement the iSB stack via Docker

Partners can customize any of the components in the iSB stack to meet their requirements

iSamples Central harvests from the partners via a standard API. Central provides a search endpoint across all partner registries