Internet of Samples: Creating and Mapping Controlled Vocabularies for Specimen Type, Material Type, and Sampled Feature

Dave Vieglais[1], Quan Gan[2], Yuxuan Zhou[2], Stephen Richard[3], Hong Cui[2], Neil Davies[4], John Deck[4], Eric Kansa[5], Sarah Kansa[5], John Kunze[6], Danny Mandel[2], Chris Meyer[7], Thomas Orrell[8], Sarah Ramdeen[9], Rebecca Snyder[8], Ramona Walls[2], and Kerstin Lehnert[9]

[1]University of Kansas
[2]University of Arizona
[3]U.S. Geoscience Information Network
[4]University of California Berkeley
[5]Open Context, The Alexandria Archive Institute
[6]California Digital Library
[7]Smithsonian National Museum of Natural History
[8]Smithsonian Institution
[9]Columbia University

November 24, 2022

Abstract

Material samples are vital across multiple scientific disciplines with samples collected for one project often proving valuable for additional studies. The Internet of Samples (iSamples) project aims to integrate large, diverse, cross-discipline sample repositories and enable access and discovery of material samples as FAIR data (Findable, Accessible, Interoperable, and Reusable). Here we report our recent progress in controlled vocabulary development and mapping. In addition to a core metadata schema to integrate SESAR, GEOME, Open Context, and Smithsonian natural history collections, three small but important controlled vocabularies (CVs) describing specimen type, material type, and sampled feature were created. The new CVs provide consistent semantics for high-level integration of existing vocabularies used in the source collections. Two methods were used to map source record properties to terms in the new CVs: Keyword-based heuristic rules were manually created where existing terminologies were similar to the new CVs, such as in records from SESAR, GEOME, and Open Context and some aspects of Smithsonian Darwin Core records. For example specimen type =*liquid>aqueous* in SESAR records mapped to specimen type = *liquid or gas sample* and material type = *liquid water*. A machine learning approach was applied to Smithsonian Darwin Core records to infer sampled feature terms from record text describing habitat, locality, higher geography, and higher classification fields. Applying fastText with a 600-billion-token corpus in the general domain, we provided the machine a level of "understanding" of English words. With 200 and 995-record training sets, 87%, 94% precision and 85%, 92% recall were obtained respectively, yielding performance sufficient for production use. Applying these approaches, more than $3x10^6$ records of the four large collections have been mapped successfully to a common core data model facilitating cross-domain discovery and retrieval of the sample records.

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

Dave Vieglais[1], Stephen M. Richard[7], Quan Gan[2], Yuxuan Zhou[2], Hong Cui[2],Danny Mandel[2], Neil Davies[3], John Deck[3], Eric C Kansa[4], Sarah Whitcher Kansa[4], John Kunze[4], Christopher Meyer[5], Thomas Orrell[5], Sarah Ramdeen[6], Rebecca Snyder[5], Ramona L. Walls[2], Kerstin Lehnert[6]
1 University of Kansas; 2 University of Arizona; 3 University of California, Berkeley; 4 The Alexandria Archive Institute; 5 National Museum of Natural History, Smithsonian Institution; 6 Columbia University; 7 independent Contractor

Introduction

Material samples play vital roles in multiple scientific disciplines. A sample initially collected for one project may prove valuable for many more studies. The Internet of Samples (iSamples) project aims to integrate large, diverse, cross-discipline sample repositories and enable access and discovery of material samples as FAIR data (Findable, Accessible, Interoperable, and Reusable). In this poster we report our recent progress in controlled vocabulary development and mapping.

Repository Descriptions

SESAR is a **community platform** that helps make Earth Sciences 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 holds archaeology samples and goes beyond the archive by richly integrating the totality of your analyses, maps, media, and journals together so they can support your interpretations.

Smithsonian Institution is the world's largest museum, education, and research complex and holds natural history of biodiversity.

Controlled Vocabularies (CVs)

Material Type: What kind of material is the specimen?
Specimen Type: What kind of thing is the specimen?
Sampled Feature: What was the sample collected originally to represent?

The vocabularies provide **consistent semantics for high-level integration** of existing vocabularies used in the source collections

Source Terminology to iSamples CVs Mapping

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

Source values are keywords → Keywords-based rules

Source values are free-text → fastText

SESAR System for Earth Sample Registration

GeoMe

OPEN CONTEXT

Smithsonian

Smithsonian

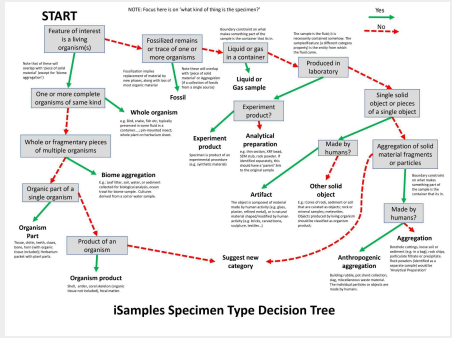
specimen type = liquid>aqueous in SESAR records was mapped to specimen type = liquid or gas sample and material type = liquid water.

Machine learning prediction based on habitat, locality, higher geography, and higher classification fields in biodiversity collections

iSamples
The internet of samples

Applying these approaches, more than 3M records of the four large collections have been mapped successfully to a common core data model facilitating cross-domain discovery and retrieval of the sample records.

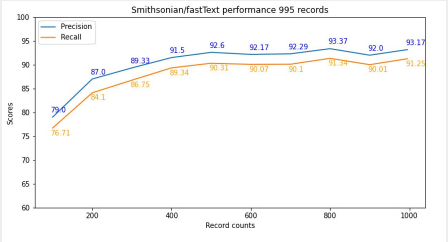
Vocabulary Decision Tree Graphs



Material Type

Sampled Feature

Training Sizes Impact fastText Performance



We trained models with the 100 records to 995 records and found the more training records are used to train an ML model, the higher precision and recall performances.

Acknowledgement

Funded by the US National Science Foundation (CSSI)

Feature Selection for Sampled Feature Prediction

Smithsonian biodiversity collection's 72 fields

Manually select informative text attributes

5 fields related to sampled feature

CollectionCode, habitat, higherGeography, locality, higherClassification

Train a fastText model

Recursively remove the weakest field with insignificant contribution to prediction performance based on T-tests

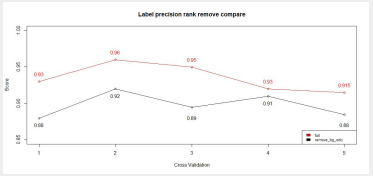
All p-values < 0.05

No

Remove the field associated with the highest p-value

Yes

Best fields reman: higherClassification



The result showed only one attribute mainly influenced performance