Automated Metadata Enhancement for Physical Sample Record Aggregation in the iSamples Project

Song, Hyunju University of Arizona, USA | hyunjusong@arizona.edu
Cui, Hong University of Arizona, USA | hongcui@arizona.edu
Vieglais, Dave University of Kansas, USA | vieglais@ku.edu
Mandel, Danny University of Arizona, USA | dmandel@arizona.edu
Thomer, Andrea K. University of Arizona, USA | athomer@arizona.edu

ABSTRACT
Large amounts of samples have been collected and stored by different institutions and collections across the world. However, even the most carefully curated collections can appear incomplete when aggregated. To solve this problem and support the increasing multidisciplinary science conducted on these samples, we propose a method to support the FAIRness of the aggregation by augmenting the metadata of source records. Using a pipeline that is a combination of rule-based and machine learning-based procedures, we predict the missing values of the metadata fields of 4,388,514 samples. We use these inferred fields in our user interface to improve the reusability.

KEYWORDS
digital archive; samples; metadata; automated metadata generation

INTRODUCTION
Physical samples – for example, a chip from a geologic outcrop – play a key part in scientific research and form a fundamental unit of data that represents nature and the environment. Just as research data must be made findable, accessible, interoperable, and reusable (FAIR; Wilkinson et al., 2016), so must physical samples and their metadata.

This poster describes work from the Internet of Samples (iSamples) project, which aims to aggregate diverse physical sample databases from the earth sciences, biosciences, and archaeology via consistent metadata and links (Davies et al., 2021). iSamples developed a high-level metadata schema that can be applied to multiple sample-collecting domains. However, when aggregating records into a single digital infrastructure, we found many incomplete sample records with missing fields such as “material” and “specimen type”. Missing metadata is a crucial problem as it impacts the quality of the aggregation and limits the usability of individual records. To address this issue, we developed a procedure that uses rules and machine learning-based predictions.

METHOD: COMBINING RULES-BASED AND MACHINE LEARNING-BASED APPROACHES
Given an incomplete record that has a missing value in the material or specimen type field, we first check if it falls into any of the rules that were curated based on expert domain knowledge (**Figure 1)**. If it does not correspond to any rules, we apply the machine learning model to predict the missing metadata field.

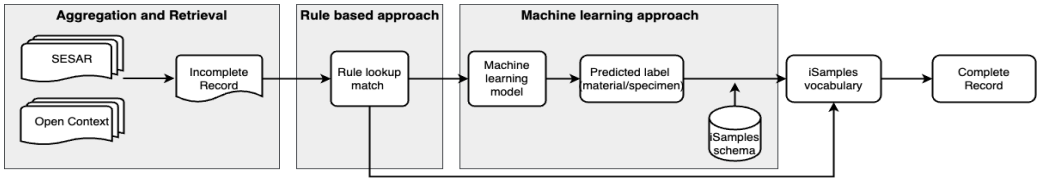

Figure 1. Pipeline of procedure

Rule-based approach
After test runs and expert evaluations on the machine learning model’s prediction results, we observed that some of the existing metadata fields of a record could be used to predict the missing values. Dedicating part of the prediction task to rule-based approach can embed the collection-specific knowledge that was not picked up by the machine learning model and help improve system efficiency. The rules were created by the collection curator after reviewing a random subsample of the records **(Figure 2)**.

```
IF collection= 'SESAR' and metadata field=’sample type’ and value=’CTP’, THEN material = ‘Liquid’
IF collection= 'Open Context' and metadata field=’item category’ and value=’Animal bone’, THEN material = ‘Biogenic non-organic material’
```

Figure 2. Example of curated rules. Based on the metadata field value, determines the missing material value

Machine learning approach
As the sample records have rich textual descriptions, we used BERT, a transformer-based architecture that was developed to solve natural language processing tasks. After experiments, we decided to use BERT-E (Koirala et al.,

86th Annual Meeting of the Association for Information Science & Technology | Oct. 27 – 31, 2023 | London, United Kingdom.

2021), an earth science-focused language model that is trained on a similar domain of dataset for our task. We developed a multiclass model that is fine-tuned on the records we have for each collection.

We integrated the models into our curation pipeline. SESAR had 3,565,478 (78%) earth sample records that were missing the material field. 3,534,384 records were rule-assigned, and 31,094 records were machine assigned. Open Context had 790,375 (96%) records missing the material field, and 793,751 (97%) records missing the specimen type. The material field of 520,632 records were rule-assigned, and 269,743 records were machine assigned. The specimen type field of 349,556 records were rule-assigned and 444,195 were machine assigned.

Collection-Field	Label	Precisio	Recall	F1	Count
SESAR-Material	Biology	0.998	0.999	0.999	198192
	Earth Material	1.000	0.494	0.662	400
	Gas	0.662	0.891	0.760	546
	Liquid	0.996	0.989	0.992	25131
	Mineral	0.993	0.995	0.994	244422
	Other	0.971	0.892	0.930	16932
	Particulate	1.000	0.95	0.974	123
	Rock	0.995	0.998	0.996	408599
	Sediment	0.998	0.995	0.997	78976
	Soil	0.990	9.995	0.993	14790
	Experimental Material	0.943	0.985	0.964	300
OpenContext-Material	Anthropogenic	0.980	0.942	0.962	21945
	Anthropogenic Metal	0.754	0.834	0.784	3297
	Biogenic Non-Organic	0.680	0.712	0.680	2294
	Mineral	0.802	0.746	0.700	433
	Natural Solid Material	0.418	0.340	0.300	1326
	Organic Material	0.396	0.498	0.426	306
OpenContext-Specimen	Artifact	1.000	0.992	0.996	23164
	Organism Part	0.858	1.000	0.908	613
	Organism Product	0.650	1.000	0.746	19

Table 1. Performance by class on test set. Evaluation results by average scores of stratified 5-fold cross validation. Macro F1 score was used to evaluate the imbalanced data. Count represents the number of records. We do not predict the specimen type of the SESAR as there were none that had missing values in this field.

Application

Using this augmented metadata, we developed a faceted search interface for iSamples (**Figure 3**). This allows users to locate samples via the augmented metadata fields, which we hope will improve the discoverability of samples.

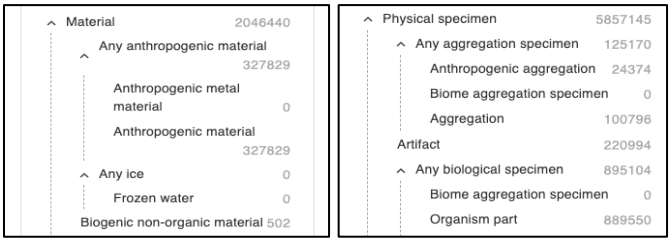

Figure 3. Subsets of "material" (left) and "specimen type" (right) used as a search field in search interface

CONCLUSION

In this poster, we presented an automated metadata enhancement procedure to create a more complete, higher-quality metadata aggregation that improves the FAIRness of the physical samples in the iSamples project. Our approach is broadly applicable to numerous other domains and collections that grapple with "lossy" metadata aggregation resulting in incomplete records. We utilize the large amount of textual data that is available in the earth science and archaeological records for the fine-tuning process. We find that developing and utilizing domain-specific language models may be a solution for automatic metadata generation of digital libraries even beyond the area of natural science. In our future research, we plan to conduct a comprehensive user study to both test the accuracy of our predicted labels, and to understand how best to flag predicted metadata values (vs verbatim legacy values) in records, and thereby increase trust in and usability of the aggregated collection.

REFERENCES

Davies, N., Deck, J., Kansa, E. C., Kansa, S. W., Kunze, J., Meyer, C., Orrell, T., Ramdeen, S., Snyder, R., Vieglais, D., Walls, R. L., and Lehnert, K. (2021). Internet of Samples (iSamples): Toward an interdisciplinary cyberinfrastructure for material samples. *GigaScience*, 10(5). giab028.

Koirala, P., Ramasubramanian, M., Gurung, I., Maskey, M., and Ramachandran, R. (2021). BERT-E: An Earth Science Specific Language Model for Domain-Specific Downstream Tasks. *In AGU Fall Meeting Abstracts, volume 2021*, pages IN15B–06.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., et al. (2016). The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9.