Organization: Smithsonian Libraries and Archives
Primary mentor: Richard Naples, Data Manager, Smithsonian Libraries and Archives
Secondary mentor(s): Martin Kalfatovic, Associate Director, Digital Programs and Initiatives Division | Program Director, Biodiversity Heritage Library

Project Title	Enhancing the Museum Data Ecosystem through Surfacing Specimen Data from Research Publications
Description	Smithsonian Research Online, a program of the Smithsonian Libraries and Archives, tracks the research output of the Smithsonian Institution's 21 museums, 9 research centers, and the National Zoo. Integral to much of this research are the collections of specimens and museum objects from across the institution. Typically, specimens used by taxonomists and other researchers are cited in the body of an article rather than the list of references, and is done so using a collection code and a number. Beyond specimens, there are other occurrences such as the mention of new species, geographic localities, research equipment, and more that could be extracted with similar techniques. This project will continue to explore and refine methods to identify and extract specimen citations from Smithsonian-authored research publications from Smithsonian Research Online (SRO), possibly expanding to the Biodiversity Heritage Library (BHL).
Problems/ Research Questions	1. How are specimens (and possibly other relevant topics) mentioned in the literature? 2. Can mentions of specimens and/or other data from research articles be consistently recognized and extracted, through pattern recognition or machine learning techniques. Also, how do these techniques compare? Can they be evaluated and/or visualized? 3. Can a linkage be created between the publication and the museum data systems that track specimens, whether using existing systems or not? If not, what are viable methods of data sharing?

	4. How can such techniques be scaled up to encompass more data and more collections? How about other useful topics like identifying new species or geographic locations?
Techniques	Text extraction techniques will be used to structure article content into JSON, then using machine learning and/or regular expressions, specimen citations will be identified and extracted.
Tools/ Languages used	Python (GROBID, SpaCy, Prodigy) OpenRefine, Command Line, Zotero (As this is a continuation of a project, there is existing code available for review and further revision.)
Data	**Description:** The data consist of journal article pdfs covering calendar year 2022. This includes both articles with and without specimens cited, authored by researchers at the Smithsonian Institution. **Data Type:** PDFs **Data Size**: under 10GB Further data for analysis may be identified and gathered during this fellowship.
Outcome	Outcomes may include: • Comparison of machine learning model vs. pattern recognition. • Refined method/software for extracting specimen catalog numbers from publications. • Exploration of and possibly method for linking specimen catalogs from publications to their respective collections databases and persistent identifiers. • Research Output (e.g. Github repository, report, article, and/or presentation)
Milestone Timeline	• M1: Literature review and exploration of ways specimens are currently referenced to develop model. • M1-2: Begin processing data set of pdf files to prepare for implementing information extraction techniques for specimen IDs, species data, and/or other key terms. • M3-5: Development and refinement of models for extracting specimens. • M5-6: Scope storage options in order to make associations openly available to data managers. • M5-6: Scope expanding the application of tools to the Biodiversity

	Heritage Library, and to include other museum collections information systems. • M6: Create a report, presentation, or other output.
References	https://research.si.edu/ (https://repository.si.edu/ and https://profiles.si.edu/ as well), https://www.biodiversitylibrary.org/, https://collections.si.edu/search/, https://naturalhistory.si.edu/research/nmnh-collections/museum-collections-policies