Natural History collections are essential resources for taxonomy, systematics, ecological, and climate change research. These collections also reflect the biases and colonial past of our collecting institutions and the science that results from using these collections also reflects these biases. Mass digitization of the herbarium at the National Museum of Natural History has provided researchers with the opportunity to study botanical specimens and their associated metadata at a scale that was previously impossible. In this talk, I describe two research projects using mass-digitized natural history data. The first is using computer vision to understand the evolution and diversification of ferns by considering all available digitized fern specimen images, more than 1000 species. The second project, part of the American Women’s History Initiative, is using natural language processing to correct the scientific record, particularly for women, often volunteers or spouses of Smithsonian employees. These women were historically critical to building collections and enabling science at the Smithsonian, but their contributions were often missed or erased. We use specimen metadata, including the specimen collector and identifier fields, as well as Smithsonian Annual Reports, which date back to the founding of the Smithsonian in 1846, to put into context the work these women did for the Smithsonian. Both projects highlight the opportunities and challenges of machine learning approaches for digitized museum collections.

Biography

Dr. Rebecca Dikow is a Research Data Scientist and leads the Smithsonian Institution Data Science Lab, part of the Office of the Chief Information Officer. Since its start in 2016, members of the Data Science Lab have been conducting biodiversity research using genomics, informatics, and machine learning tools. While the Data Science Lab research focus was initially on biodiversity, as the pace of digitization of collections from across the Smithsonian has increased, the Data Science Lab has expanded its areas of inquiry to...
include biodiversity informatics and digital humanities in collaborations with more than a dozen Smithsonian units. Data Science Lab members now work with researchers studying digitized collections and archives data outside the biodiversity sphere and strive to collaborate with scholars all across the Smithsonian. The Data Science Lab also provides support for researchers using the High-Performance Computing Cluster and training in data science and bioinformatics tools. Rebecca is also an affiliated faculty member in the George Mason University School of Systems Biology and the Smithsonian-Mason School of Conservation.