



Smithsonian

Introduction

Mercuric chloride was commonly used in the past to prevent insect damage to botanical specimens in herbaria. Because this substance is toxic to humans, knowing the number and location of contaminated specimens in a collection is important. Fortunately, the staining is visible upon specimen inspection, because mercuric chloride crystallizes over time. Because the U.S. National Herbarium contains more than 5 million specimens, manual inspection of every specimen is not tractable. In 2017, NMNH and OCIO scientists built a machine learning model to identify mercuric chloride staining on digitized herbarium sheets, but only applied the model to a portion of the digitized herbarium (Schuettpelz et al, 2017). This project seeks to update the model and apply it to all digitized specimens.

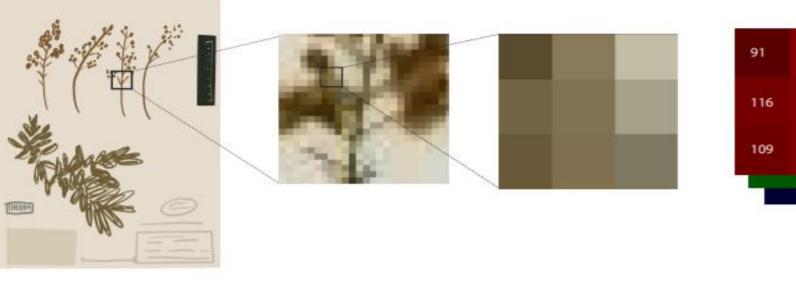


Figure 1: Process of herbarium specimen converted into inputs for machine learning (Borowiec et al., 2021)

Acknowledgments / References

- Elizabeth Cottrell, Gene Hunt, and Virginia Power, for their roles in the NHRE program
- Borowiec, M. L., Frandsen, P., Dikow, R., McKeeken, A., Valentini, G., & White, A. E. (2021,). Deep learning as a tool for ecology and evolution. 10.32942/osf.io/nt3as
- YaleDHLab. (n.d.). YaleDHLab/pix-plot: A WebGL viewer for UMAP or TSNEclustered images. GitHub.
- Schuettpelz, E., Frandsen, P. B., Dikow, R. B., Brown, A., Orli, S., Peters, M., Metallo, A., Funk, V. A., Dorr, L. J. 2017. Applications of deep convolutional neural networks to digitized natural history collections. Biodiversity Data Journal. doi: 10.3897/BDJ.5.e21139
- Funding from NSF OCE-1560088

Using machine learning tools to quantify mercuric **chloride staining across digitized herbarium specimens** Isabella Schrader^{1,2}, Mike Trizna², Ashlyn Powell³, Paul Frandsen^{2,3}, Alex White^{1,2}, Rebecca Dikow², Eric Schuettpelz¹

¹ Smithsonian Institution, National Museum of Natural History, Botany Department ² Smithsonian Institution, Office of the Chief Information Officer, Data Science Lab ³ Brigham Young University, Department of Plant and Wildlife Sciences

Methods

Throughout this project, Jupyter notebooks were used to Pixplot in Python. work First, (https://dhlab.yale.edu/projects/pixplot/) was used to visualize digitized specimens using unsupervised machine learning. Figure 2 shows a PixPlot of 1000 sample images.



Figure 2: 1000 image PixPlot

We used the Smithsonian Institution High Performance Computing Cluster, also known as Hydra, to access files and store data. During the summer, Data Carpentry workshops were attended, which aided in developing skills used to group the data into smaller samples. We Colaboratory Google also used (https://colab.research.google.com) to train a new model to identify mercuric chloride staining, taking advantage of recent advances in machine learning technology.

titl edan_id	e timestam p_unix	lastupdat e_unix	Barcode	specimen _guid	media_c ount	media_g uid	media_g uid_list	media_a ws_id_lis t	aws_me dia_coun t	USNM Number	Other tim Numbers	estam p_dt		
			1619514 189	1.0102e+ ,	http://n2 t.net/ark: /65665/3 886114c e-5e7f- 43c9	1.0	/65665/	t.net/ark: /65665/ m334175 257- 8640-	NMNH-	1.0	1884734	NaN	2021-04- 27 09:03:14	2

Figure 5: Sample DataFrames



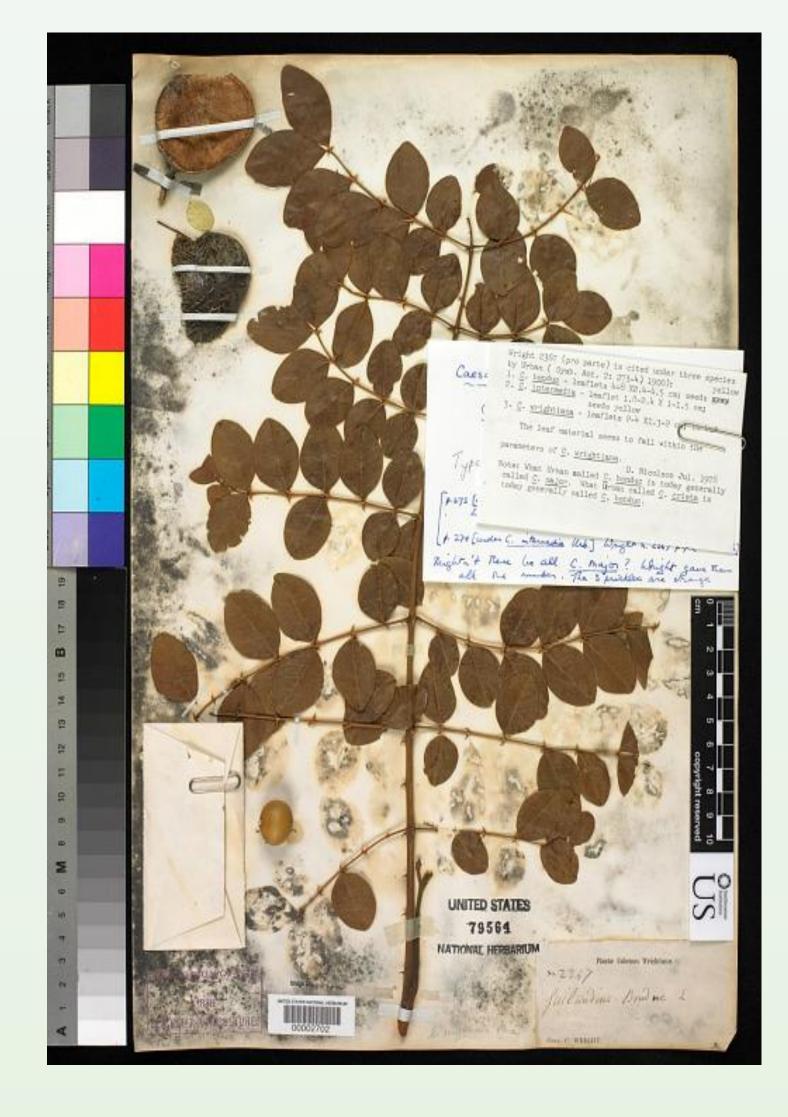


Figure 3: Stained specimen

Results and Discussion

More than two million digitized herbarium specimens are available on the Smithsonian Open Access platform. Python code to download and resize images is available on GitHub (https://github.com/sidatasciencelab/mercury sheets). This code also has the potential to be modified to access Smithsonian Open Access images from other museums so that researchers can more easily assemble datasets for other applications. Sixteen Search the Department of Botany Collections Notany Collections Keyword Search Search by Field Type Register Genetic Samples Plant Photo Archive Botanical Art Help Feedback metadata fields are recorded he plant collections of the Smithsonian Institution began with the acquisition of specimens collecter he United States Exploring Expedition (1838-1842). These formed the foundation of a National for each specimen (Figure 5; erbarium which today numbers over 5 million historical plant records. placing it among the world's argest and most important. Over 4.2 million specimen records (including over 115,000 type specimens with images) are current vailable in this online catalog. two specimens shown). The arch the Botany Collections Select a tab on this page to search by Keyword or Selected Fields. Searches are limited to 10000 records and the results are sorte by taxonomic group. If you need to retrieve a larger record set, contact the Department of Botany's Data Manager. machine learning model we Disclaimer: The US Herbarium's Collections records may contain specimen labels that include offensive or culturally insensitive language. These records do not reflect the US Herbarium's current viewpoint but rather the social attitudes and circumstances of th me period when specimens were collected or cataloge are developing could also be See the Help tab to learn more about searching and then exploring your returned results (sorting, exporting, etc.). applied to specimens from herbaria around the world to identify mercuric chloride NMNH Data Access Policy staining. Figure 6: Open Access





REU Site, OCE-1560088



Figure 4: Unstained specimen

