Using machine learning tools to quantify mercuric chloride staining across digitized herbarium specimens

Isabella Schrader[1,2], Mike Trizna[2], Ashlyn Powell[3], Paul Frandsen[2,3], Alex White[1,2], Rebecca Dikow[2], Eric Schuettpelz[1]

[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

REU Site, OCE-1560088

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.

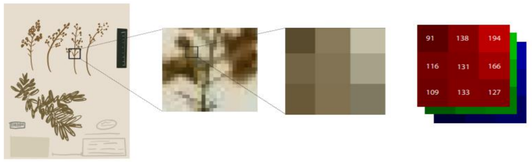

Figure 1: Process of herbarium specimen converted into inputs for machine learning (Borowiec et al., 2021)

Methods

Throughout this project, Jupyter notebooks were used to work in Python. First, Pixplot (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 also used Google Colaboratory (https://colab.research.google.com) to train a new model to identify mercuric chloride staining, taking advantage of recent advances in machine learning technology.

edan_id	title	timestamp_unix	lastupdate_unix	Barcode	specimen_guid	media_count	media_guid	media_guid_list	media_aws_id_list	aws_media_count	USNM Number	Other Numbers	timestamp_dt	lastupdate_dt
edanmdm-nmnhbotany_10489462	Peperomia hernandiifolia (Vahl) A. Dietr.	1619514194	1619514189		http://n2t.net/ark:/65665/3886114ce-5e7f-43c9-...	1.0	http://n2t.net/ark:/65665/m334175257-8640-4e70...	http://n2t.net/ark:/65665/m334175257-8640-4e70...	NMNH-01010204	1.0	1884734	NaN	2021-04-27 09:03:14	2021-04-27 09:03:09

Figure 5: Sample DataFrames

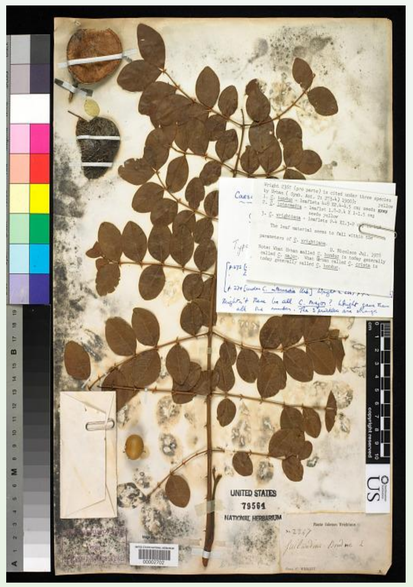

Figure 3: Stained specimen

Figure 4: Unstained 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 metadata fields are recorded for each specimen (Figure 5; two specimens shown). The machine learning model we are developing could also be applied to specimens from herbaria around the world to identify mercuric chloride staining.

Figure 6: Open Access

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 TSNE-clustered 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