PROCEEDINGS OF TDWG

Biodiversity Information Science and Standards 1: e20326
doi: 10.3897/tdwgproceedings.1.20326

OPEN ACCESS

Conference Abstract

A Pipeline for Processing Specimen Images in iDigBio - Applying and Generalizing an Examination of Mercury Use in Preparing Herbarium Specimens

Gaurav Yeole‡, Saniya Sahdev‡, Matthew Collins‡, Alex Thompson‡, Rebecca B Dikow§, Paul Frandsen§, Sylvia Orli§, Renato J Figueiredo‡

‡ University of Florida, Gainesville, United States of America
§ Smithsonian Institution, Washington DC, United States of America

Corresponding author: Gaurav Yeole (gauravyeole@ufl.edu)

Received: 15 Aug 2017 | Published: 15 Aug 2017

Citation: Yeole G, Sahdev S, Collins M, Thompson A, Dikow R, Frandsen P, Orli S, Figueiredo R (2017) A Pipeline for Processing Specimen Images in iDigBio - Applying and Generalizing an Examination of Mercury Use in Preparing Herbarium Specimens. Proceedings of TDWG 1: e20326. https://doi.org/10.3897/tdwgproceedings.1.20326

Abstract

iDigBio currently references over 22 million media files, and stores approximately 120 terabytes worth of those media files co-located with our computing infrastructure (Matsunaga et al. 2013). Using these images for scientific research is a logistical and technical challenge. Transferring large numbers of images requires programming skill, bandwidth, and storage space. While simple image transformations such as resizing and generating histograms are approachable on desktops and laptops, the neural networks commonly used for learning from images require server-based graphical processing units (GPUs) to run effectively.

Using the GUODA (Global Unified Open Data Access) infrastructure, we are building a model pipeline for applying user-defined processing to all or any subset of images stored in iDigBio on servers located in the Advanced Computing and Information Systems lab (ACIS) alongside the iDigBio storage system. This pipeline utilizes Apache Spark, the Hadoop File System (HDFS), and Mesos (Collins et al. 2017). We have placed a Jupyter notebook server in front of this architecture, which provides an easy environment for end users to

write their own Python or R software programs. Users can access the stored data and images and manipulate them per their requirements and make their work publicly available on GitHub.

As an example of how this pipeline can be used in research, we are applying a neural network developed at the Smithsonian Institution to identify herbarium sheets that were prepared with hazardous mercury-containing solutions (Schuettpelz, *in preparation*). The model was trained on Smithsonian servers using their herbarium images and it is being transferred to the GUODA infrastructure hosted at the ACIS lab. All herbarium images in iDigBio are being classified using this model to illustrate the application of these techniques to larger sets of images using a deep convolutional neural network that detects visible mercury crystallization present on digitized herbarium sheets. Such an automated detection process can potentially be used, for instance, to notify other data publishers of any contamination. We are presenting the results of this classification not as a verified research result, but as an example of the collaborative and scalable workflows this pipeline and infrastructure enable.

Keywords

Biocollection Infrastructure, Cloud Computing, Neural Networks, Deep Learning

Presenting author

Matthew Collins

References

- Collins M, Hammock J, Poelen J, Thessen A, Thompson A (2017) http://guoda.bio/about/. Accessed on: 2017-7-17.
- Matsunaga A, Thompson A, Figueiredo R, Germain-Aubrey C, Collins M, Beaman R, MacFadden B, Riccardi G, Soltis P, Page L, Fortes JB (2013) 2013 IEEE 9th International Conference on e-Science. 2013 IEEE 9th International Conference on e-Science https://doi.org/10.1109/escience.2013.48