DIGITIZATION TCN: COLLABORATIVE RESEARCH: Digitization, Enrichment, and Quality-control of U.S. Herbarium Data from Tropical Africa to Permit Rigorous Assessment of Seed Plant Biodiversity Patterns

A Data Management Plan created using DMPTool

Creator: Townsend Peterson

Affiliation: University of Kansas (KU)

Funder: National Science Foundation (NSF)

Template: NSF-BIO: Biological Sciences

ORCID iD: 0000-0003-0243-2379

Project abstract:
Proposal to the "Advancing the Digitization of Biological Collections" program, which would involve capture of all African herbarium specimens held in 21 US institutions

Last modified: 08-30-2020

Copyright information:
The above plan creator(s) have agreed that others may use as much of the text of this plan as they would like in their own plans, and customize it as necessary. You do not need to credit the creator(s) as the source of the language used, but using any of the plan's text does not imply that the creator(s) endorse, or have any relationship to, your project or proposal.
DIGITIZATION TCN: COLLABORATIVE RESEARCH: Digitization, Enrichment, and Quality-control of U.S. Herbarium Data from Tropical Africa to Permit Rigorous Assessment of Seed Plant Biodiversity Patterns

Data and Materials Produced

This project will produce data in three forms, as well as programming code for data manipulation. Specifically,

1. Images of herbarium specimens. Specimens will have been applied a unique barcode following standard protocols in each individual herbarium; these barcodes will be used as persistent identifiers throughout the processing of data in this project. We generate images at each of the herbaria that make up the project network. Each herbarium will use DSLR cameras and light boxes (in some cases herbarium scanners) to capture high-resolution digital images of specimens with scale bars and color standards to permit standardization of scale and color. Most herbaria will use existing imaging equipment for this step, but all will capture images at a spatial resolution of >21 megapixels. We will use Lightbox to correct exposure and color balance.

2. Specimen data. All project participant herbaria already have Darwin Core-compliant databases for their specimen data, and already contribute data actively to iDigBio and other biodiversity information portals. As such, each herbarium already has a protocol in place that translates data from herbarium label to institutional database to Darwin Core-compliant format. Participating herbaria host their data on a variety of software platforms, including Specify (CAS, BRU, MIC, PUR, GH, RSA), Symbiota (BRIT, MIN, CM, KSP, LSU, DAV, PH, ARIZ), EMu (NY, US, YU, UT, F), TROPICOS (MOBOT), CollectionSpace (UC), and FileMaker Pro (BYU). We will make no effort to change or standardize these within-institution protocols, but we will require that each institution be able to produce Darwin Core-compliant versions of their data for open sharing.

3. Georeference data. Once specimen data are captured for each institution in the network, we will create a "gazetteer" database, consisting of all unique combinations of key geographic fields, including country, state, county/municipality, and specific locality. We have developed a highly efficient protocol for adding full, best-practices spatial metadata to this information, including harvesting information from Open Street Map, from other specimen data, etc., and involving as little de novo data capture as possible to maximize efficiency. These data will be stored in a master gazetteer database based on a subset of Darwin Core, including those fields relevant to geographic referencing, which will be used as a basis for providing georeference information back to the herbaria that hold the original specimens.

4. Programming code. All programming code produced as part of the quality control and georeferencing portions of this project will be documented carefully, both internally to the code itself, and via detailed metadata. Version-control software (e.g., git) will be used to maintain the full history of all source code revisions (including time of changes, author of changes, and summary messages); version-control code repositories are cloned on several separate computers, and these messages are further protected using cloud-based backups (e.g., CrashPlan).

Standards, Formats and Metadata

The basis for all data to be captured in this project will be Darwin Core, an IOS standard format for biodiversity data. Specimen data will take advantage of all of the full set of data fields documenting taxon, place, and time, as well as ancillary data fields documenting collector, substrate, habitat, etc. Gazetteer data will use only the fields that are related to place. For images, we will follow iDigBio guidelines, with a final pixel density of >21 megapixels in all archival images, which will be captured in DNG/RAW formats that involve no compression or information loss. All images will include size and color standards to permit full image
standardization and color correction. We will also create smaller, JPEG-format images to allow easier access for uses that do not demand maximum resolution.

Roles and Responsibilities

PI/PD Peterson will be responsible ultimately for the progress and advance of data capture and management in this project; data management personnel at each herbarium will be responsible for implementing their individual workflows for translating from image to data; Project Participant Wieczorek will be responsible for collective data management steps related to georeferencing and quality control. At the outset of the project, all data management protocols will be documented in openly accessible documents posted in a shared project directory. All project personnel and institutions will contribute to the design of these documents, and all will similarly subscribe to following their guidelines strictly. Adherence to these protocols will be assured via periodic checks by the PIs, by quality control steps, and by communication among all participants at regular intervals.

Dissemination Methods

Project data will be stored and shared in Darwin Core formats. As regards publication of project results, the University of Kansas has been a national leader in the movement toward assuring open access to its journal-published scholarship. Specifically, the KU faculty voted 10 years ago to grant a license to the University to serve copies of KU-faculty-authored journal papers; this ‘green’ open access is achieved via KU’s institutional repository, KU ScholarWorks. More generally, project participants all adhere to the idea of open access, such that (to every extent possible) project outputs are published in access-friendly journals. As a consequence, we anticipate project outputs appearing in largest part in journals that are universally and openly accessible to any reader.

Policies for Data Sharing and Public Access

Specimen data will be shared via iDigBio and GBIF, both global biodiversity information portals that will integrate project data with the totality of other biodiversity data, such that project data see maximum use and utility. Images will be stored in each institution’s repository, and linked to the Darwin Core data record for the specimen via the <<XXXXX>> field. Gazetteer data will be shared via <<XXXXX>>. All data and programming code will be shared under a Creative Commons (CC BY) license, to assure maximum openness of data and protocols created in this project, including derivative products.

Archiving, Storage and Preservation

Biodiversity data related to physical specimens, such as the herbarium specimens that are the focus of this project, are best stored and managed at the institutions at which the specimens are deposited. As such, each of the herbaria participating in this project will manage its own institutional database and its own repository of images. The institutions that comprise our network all have image repositories (save for CM, for whom PH hosts images), and each will follow its customary protocols. For projectwide datasets, including the specimen data and gazetteer data, we will take advantage of the University of Kansas Research File Storage
service, which provides nightly backup and off-site storage of backups for recovery in the case of disaster. This service is scalable to serve both small and large storage needs. This service is monitored 24/7 by experienced KU Information Technology staff, requires authentication for identity management purposes, and uses Active Directory to manage authorization.