Plan Overview

A Data Management Plan created using DMPTool

Title: 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

Creator: Townsend Peterson

Affiliation: University of Kansas

Principal Investigator: A. Townsend Peterson

Data Manager: John Wieczorek

Funder: National Science Foundation (nsf.gov)

Funding opportunity number: xxxxxx

Grant: xxxxxxx

Template: NSF-BIO: Biological Sciences

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

Start date: 08-01-2021

End date: 07-31-2025

Last modified: 09-01-2020

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

This project will produce three categories of data as well as programming code for data manipulation and access. 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 and beyond the processing of data in this project. We will 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 and color bars. Most herbaria will use existing imaging equipment for this step, but all will capture images in raw format 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 specimen databases capable of exporting data as Darwin Core, and already contribute data in this format to iDigBio, GBIF, and other biodiversity information portals. Participating herbaria host their data on a variety of software platforms, including Specify (CAS, BRU, MICH, PUL, GH, RSA), Symbiota (BRIT, MIN, CM, KSP, LSU, DAV, PH, ARIZ), EMu (NY, US, YU, UT, F), TROPICOS (MO), 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 continue 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 master reference gazetteer consisting of distinct combinations of terms of the Darwin Core Location class (geographic fields, including country, state, county/municipality, specific locality, coordinates, uncertainty, and process metadata) from among all of the aggregated specimen records. These Location records will be the basis of georeferencing, adding spatial information in the form of coordinates, uncertainty, and process metadata following the recently updated "Georeferencing Best Practices" (Chapman and Wieczorek 2020) and Georeferencing Quick Reference Guide (Zermoglio et al. 2020). The master gazetteer database will be used as the staging area from which to return georeference data back to the herbaria that hold the original specimens.

4. **Programming code**. The project will generate code used to construct the gazetteer from original specimen records and to provide resulting georeferences for the contributing institutions to incorporate into their working databases. Code will be documented carefully, both within the code and via external explanatory documents.

The basis for all data to be captured in this project will be Darwin Core, a standard format for biodiversity data exchange maintained by Biodiversity Information Standards (TDWG). Specimen data will take advantage of the full set of terms documenting taxon, place, and time, as well as ancillary data fields documenting collector, substrate, habitat, etc. with persistent unique identifiers.

Gazetteer data will use the subset of Darwin Core terms related to location.

For images, we will follow iDigBio guidelines, using the Audubon Media Description extension to Darwin Core for the metadata and capturing images in DNG/RAW formats with a final pixel density of >21 megapixels, with no compression or information loss. All images will include size and color standards to permit full image standardization and color correction. We will also extract images of lower resolution in JPEG format to allow easier access for uses that do not demand maximum resolution.

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 related to georeferencing and quality control. At the outset of the project, all data management protocols will be captured 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.

Digitized specimens records will be captured and secured in the native database platforms used by each participating institution. Data will be shared to data aggregators (e.g., iDigBio, GBIF) via Darwin Core archives using existing sharing infrastructures (e.g., Symbiota, Integrated Publishing Toolkit instances) as soon as they are incorporated in the native databases, via protocols already established at each herbarium.

Gazetteer data will be managed in cloud-based Google BigQuery and accessible via the BigQuery Console and an Application Programming Interface (API). By the end of the project, georeferenced locations will be incorporated into the Biodiversity Enhanced Location Services (BELS) being developed under the recently recommended NSF Proposal (DBI-2027241) entitled "Collaborative Research: CIBR: Leaping the Specimen Digitization Gap: Connecting Novel Tools, Machine Learning and Public Participation to Label Digitization Efforts". This gazetteer is an aggregation of location information for major biodiversity data initiatives, including iDigBio and GBIF. A snapshot of finished best-practice georeferences from this project will be made at the end of the project as a reference and made freely accessible and citable via Zenodo.

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.

Specimen data digitized in this project will be shared openly under one of three open Creative Commons licenses (CC0, CC-BY, or CC-BY-NC), as chosen by each contributing institution; these three licenses are the only ones allowed for integration in and sharing via GBIF. Data of a sensitive nature (such as the specific location of endangered species) can be withheld or generalized at the discretion of the data publishing institution. Darwin Core provides fields (informationWithheld and dataGeneralizations) to support this manner of data publishing while alerting the potential *bona fide* user that further pertinent information may be available upon request.

Images will be stored in each institution's repository, and metadata using the Audubon Core Extension to Darwin Core will accompany and be linked to the specimen record via the Darwin Core occurrenceID. Images also will be shared openly under CC0, CC-BY, or CC-BY-NC licenses.

Gazetteer data will be shared under the open CC0 waiver via web services and periodic snapshots of CSV files. 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.

Version-control software (git) will be used to maintain the full history of all source code revisions (including time of changes, author of changes, and summary messages) in free and open cloud-based GitHub repositories with a

GNU General Public License to assure maximum openness and reusability; version-control code repositories are cloned on several separate computers, and are further protected with cloud-based backups (e.g., CrashPlan).

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 project-wide datasets, including the specimen data and gazetteer, we will take advantage of the University of Kansas Research File Storage service to store before- and after-georeference snapshots. This service 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.