

## Conference Abstract

# Digitization of US Herbaria - How close did we get to the 2020 goal?

David Peter Shorthouse<sup>‡</sup>, Jocelyn Pender<sup>‡</sup>, Richard Rabeler<sup>§</sup>, James A Macklin<sup>‡</sup>

<sup>‡</sup> Agriculture & Agri-Food Canada, Ottawa, Canada

<sup>§</sup> University of Michigan, Ann Arbor, United States of America

Corresponding author: David Peter Shorthouse ([davidpshorthouse@gmail.com](mailto:davidpshorthouse@gmail.com))

Received: 30 Sep 2020 | Published: 09 Oct 2020

Citation: Shorthouse DP, Pender J, Rabeler R, Macklin JA (2020) Digitization of US Herbaria - How close did we get to the 2020 goal? Biodiversity Information Science and Standards 4: e59166.

<https://doi.org/10.3897/biss.4.59166>

## Abstract

A discussion session held at a National Science Foundation-sponsored Herbarium Networks Workshop at Michigan State University in September of 2004 resulted in a rallying objective: *make all botanical specimen information in United States collections available online by 2020*. Rabeler and Macklin 2006 outlined a toolkit for realizing this ambitious goal, which included:

1. a review of relevant and state-of-the-art web resources, data exchange standards and,
2. mechanisms to maximize efficiencies while minimizing costs.

Given that we are now in the year 2020, it seems appropriate to examine the progress towards the objective of making all US botanical specimen collections data available online. Our presentation will attempt to answer several questions:

- How close have we come to meeting the original objective?
- What fraction of “digitized” specimens are minimally represented by a catalog number, a determination, and/or a photograph? What fraction has been thoroughly transcribed?

- How close have we come to attaining a seamlessly integrated, comprehensive, and national view of botanical specimen data that guides a stakeholder to appropriate resources regardless of their entry point?
- What “holes” in this effort still exist and what might be required to fill them?

Given our interest in the success of both the Global Biodiversity Information Facility ([GBIF](#)) and the Integrated Digitized Biocollections ([iDigBio](#)), as well as the overwhelming likelihood that either one of these initiatives is the usual entry point for someone seeking US-based botanical data, we approached the answers to the above questions by first crafting a repeatable data download and processing workflow in early July 2020. This resulted in 25.6M records of plant, fungi, and Chromista from 216 datasets available through GBIF and 32.8M comparable records available through iDigBio from 525 recordsets. We attempted to align these seemingly discordant sets of records and also chose [Darwin Core](#) terms that were best suited to match the four hierarchical levels of digitization defined in the Minimal Information for Digital Specimens (MIDS) (van Egmond et al. 2019).

During the analysis/comparison of the datasets, we found several examples where the number of data records from an institution seemed much lower than expected. From a combination of analyzing record content in GBIF/iDigBio and consulting regional/taxonomic portals, it became evident that, besides datasets only being included in either GBIF or iDigBio, there was a significant number of records in regional/taxonomic portals that were not yet made available through either GBIF or iDigBio.

Progress on digitization has benefited greatly from the US National Science Foundation's creation of the Advancing Digitization of Biodiversity Collections ([ADBC](#)) program, and funding of the 15 Thematic Collection Networks (TCN). The launching of new projects and the ensuing digitization of herbarium collections have led to a multitude of new specimen portals and the enhancement of existing software like [Symbiota](#) (Gries et al. 2014). But, it has also led to insufficient data sharing among projects and inadequately aligned data synchronization practices between aggregators. Consistency in terms of data availability and quality between GBIF and iDigBio is low, and the chronic lack of record-level identifiers consistently restricts the flow of enhancements made to records. We conclude that there remains substantial work to be done on the national infrastructure and on international best practices to help facilitate collaboration and to realize the original objective of making all US botanical specimen collections data available online.

## Keywords

digitization, aggregator, MIDS

## Presenting author

David Peter Shorthouse

## Presented at

TDWG 2020

## References

- Gries C, Gilbert E, Franz N (2014) Symbiota – A virtual platform for creating voucher-based biodiversity information communities. *Biodiversity Data Journal* 2 <https://doi.org/10.3897/bdj.2.e1114>
- Rabeler R, Macklin J (2006) Herbarium networks in the United States: Towards creating a toolkit to advance specimen data capture. *Collection Forum* 21 (1-2): 223-231.
- van Egmond E, Willemse L, Paul D, Woodburn M, Casino A, Gödderz K, Vermeersch X, Bloothoofd J, Wijers A, Raes N (2019) Design of a Collection Digitisation Dashboard, ICEDIG Deliverable D2.3. <https://doi.org/10.5281/zenodo.2621055>