Biodiversity Heritage Library and Global Names: Successes, opportunities and the challenges for the future collaboration

Dmitry Mozzherin ‡

‡ University of Illinois, Champaign, United States of America

Corresponding author: Dmitry Mozzherin (dmozzherin@gmail.com)

Abstract

The Biodiversity Heritage Library (BHL) is a major aggregator of biodiversity literature with more than 200,000 volumes. The Global Names Architecture (GNA) strives to develop and provide tools for finding, parsing and verifying scientific names. GNA and BHL have enjoyed 10 years of collaboration in the creation of a scientific names index for BHL. Such an index provides researchers with a means for finding data about more than a million species.

Recently, BHL and GNA developed a workflow for the creation of an index that covers more than 50 million pages of BHL, and finds and verifies scientific names in less than a day. The unprecedented speed of the index creation opens an opportunity to dramatically increase its quality and reach. The following challenges can now be addressed.

1. Abbreviated names reconciliation.

From 20% to 25% of all scientific names in BHL are abbreviated. It is much harder to reconcile and verify abbreviated names, because their specific epithets are not unique. We plan to reconcile the vast majority of such names via a statistical approach.

2. Linking of biodiversity publication titles with actual pages in BHL.

Scientific names are closely connected to publications of original description, taxonomic treatments, and other usages. We plan to build algorithms for finding out how different lexical variants of the same publication reference can be disambiguated and connected to corresponding BHL pages.

3. Using taxonomic intelligence for finding information about species.

According to our estimation, on average, there are three scientific names (historical and current) per taxon. Names of species often change over time as a result of misspellings, and homotypic or heterotypic synonymy. We plan to link outdated names with currently

accepted names of taxa. This functionality provides all information about a taxon in BHL, no matter what names were used to reference the taxon at the time of publication.

4. Finding information about original descriptions of genera and species.

For every species there is a publication with the original description. We want to create an index of species that are described in the publications aggregated by BHL.

5. Detection of species names in spite of "incorrect" capitalization.

Previously, or in horticultural sources, specific epithets were often capitalized (e.g., *Bulbophyllum Nocturnum*), or for patronyms in which the species was named in honor of someone (e.g., *Notiospathius Johnlennoni*). We plan to detect names with non-standard capitalization of this sort.

6. Removal of false positives.

Texts in Latin language, names of people, and geographical entities often create false positives that look like scientific names. Using machine learning techniques will allow us to detect and remove most of these errors from the names index.

7. Detection of the names of biodiversity scientists and geographical entities in texts.

Finding names of biologists and geographical places in addition to scientific names would allow us to draw connections between these data and to create metadata demonstrating these links. We plan to add tools and algorithms for indexing person names and geographical names.

In this talk I will present plans for a dramatic quality increase in the scientific name-finding algorithms, as well as an introduction of other elements that would enhance usability of BHL for its patrons.

Keywords

nomenclature, taxonomy, scientific name

Presenting author

Dmitry Mozzherin

Presented at

TDWG 2021

Hosting institution

University of Illinois at Urbana-Champaign

Conflicts of interest