BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 3: e35353
doi: 10.3897/biss.3.35353

OPEN ACCESS

Conference Abstract

Finding scientific names in Biodiversity Heritage Library, or how to shrink Big Data

Dmitry Mozzherin[‡], Alexander A Myltsev[§], David Patterson[|]

‡ University of Illinois, Champaign, United States of America
§ IP Myltsev, Moscow, Russia
| University of Sydney, Sydney, Australia

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

Received: 10 Apr 2019 | Published: 13 Jun 2019

Citation: Mozzherin D, Myltsev A, Patterson D (2019) Finding scientific names in Biodiversity Heritage Library, or how to shrink Big Data. Biodiversity Information Science and Standards 3: e35353. https://doi.org/10.3897/biss.3.35353

Abstract

The Biodiversity Heritage Library contains 57 million pages of biological information. The majority of this information is a scanned and digitized non-structured text. This "raw" text is hard to access by computers or humans, without the addition of rich metadata. Recent improvements in natural language processing (NLP) and machine learning (ML) promise to facilitate the creation of such metadata.

One obvious approach to improve BHL usability is to extract and provide an index of scientific names thereby enabling biologists to find useful information easier and faster. The Global Names Architecture (GNA) detects, verifies, collects, and indexes scientific names from many sources. Six years ago GNA developers created an index of the scientific names in the BHL by parsing every page one by one. This took 45 days to accomplish. Almost immediately BHL users began to find problems in the index and suggest improvements. However, the cost of repeating such a gigantic job was insurmountable and as a result the index remained nearly unchanged for 6 years.

Two problems were at the heart of dealing with the “Big Data” of the BHL, the time it took to transfer the raw data *prior* to processing, and the computational time it took to detect the names themselves.To solve these problems we could either throw more hardware resources into the problem (expensive), or find ways to dramatically improve performance of the tasks (cheaper). We decided to achieve our goal by utilizing hardware more effectively, and by using fast, scalable programming languages.

We wrote several Open Source applications in Go and Scala to detect candidate scientific names then verify them as names by comparing them to 27 million scientific name-strings aggregated by GNA. We were able to speed up data mobilization from 24 hours to 11 minutes, and decrease the time for name detection from 35 days to 5 hours. Name-verification time decreased from 10 days to 9 hours. Overall our computing requirements shrank from 4 high-end servers to one modern laptop. As a result we achieved our goal and indexed BHL in only 14 hours and unlocked the reality of iterative improvements to the scientific name index.

We also wanted to make it possible to study BHL data in its entirety remotely, in real-time. We created an HTTP2 service that is able to stream gigantic amounts of BHL textual data together with scientific names to a researcher. Sending the text of 50 million pages with an associated 250 million name occurrences takes ~5 hours. For comparison, simply copying BHL text data from Smithsonian Institute to University of Illinois using more traditional methods took us 10 days.

What do we hope to achieve with these tools as next steps? To make it possible for everyone to make new discoveries by computing in real-time across the complete BHL text. For example 20% of all names in BHL are abbreviated, and, as a result, very poorly searchable given their existing full-text indexing. We plan to develop algorithms to expand abbreviated genera reliably. Digitized texts contain huge amounts of character recognition mistakes. The tools might help to detect badly digitized pages and mark them for re-digitization. Tools can help to extract scientific names that are identical to "normal" words, such as "Atlanta", or "America", to find common names in texts, and to localize information on locations, adding new search contexts. Finally, we are exploring tools that allow researchers to stream such results back to source thereby growing the “Big Data” and ultimately improving the BHL’s end-user experience.

Keywords

Biodiversity Heritage Library, BHL, Global Names Architecture, GNA, nomenclature, natural language processing, NLP, scientific names

Presenting author

Dmitry Mozzherin

Funding program

NSF: ABI development, Species File Group at University of Illinois

Grant title

Award Abstract #1645959 ABI Development: Global Names Discovery, Indexing and Reconciliation Services

Hosting institution

University of Illinois at Urbana/Champaign, Illinois Natural History Survey

Author contributions

Dmitry Mozzherin is the principle investigator of the NSF grant. He developed scientific names finding, parsing, and verification tools, participated in development of Global Names Architecture concepts.

Alexander Myltsev developed scientific names parsing and verification tools.

David Patterson is one of the original authors of Global Names Architecture project. He had been a leader of GNA, a mentor of the group and a consultant at various times.