PROCEEDINGS OF TDWG

Biodiversity Information Science and Standards 1: e20186
doi: 10.3897/tdwgproceedings.1.20186

OPEN ACCESS

Conference Abstract

# A path to continuous reindexing of scientific names appearing in Biodiversity Heritage Library data

Dmitry Mozzherin[‡], Alexander A Myltsev[§]

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

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

Received: 10 Aug 2017 | Published: 11 Aug 2017

Citation: Mozzherin D, Myltsev A (2017) A path to continuous reindexing of scientific names appearing in Biodiversity Heritage Library data. Proceedings of TDWG 1: e20186.
https://doi.org/10.3897/tdwgproceedings.1.20186

## Abstract

Biodiversity Heritage Library (BHL) is a massive, constantly expanding repository of Open Access biodiversity literature. It currently serves over 50 million pages of biological texts to scientific community. Metadata attached to this textual data dramatically enhances its usefulness. One of the most important categories of metadata is an index of scientific names that binds names to pages in BHL. This index helps researches to find information about species or higher taxons they study. Finding scientific names in 50 million pages is a challenging task not only because of the scale, but also because of inevitable mistakes made during the process. Optical character recognition (OCR) mistakes, historical changes in common practices of nomenclature, name abbreviations, variations in formats of scientific names etc. - all create potential problems for name indexing. Ability to cope with such problems mainly depends on the quality of applied computer algorithms. With time improvements in OCR, error reports from BHL users, improvements in name-finding algorithms make it beneficial to re-index the whole data set. If we are able to regularly re-index BHL data, we should be able to systematically improve the quality of the index. However with tools existed so far such task was impractical. In 2012 our team at Global Names Architecture (GNA) project accomplished a complete index of BHL. The process took more than 40 days, during which the whole toolset of name-finding, name-parsing, name-verification programs was unavailable for public use. Repeating such extensive exercise was not feasible since then. We think it is important to constantly improve indexing

quality, and for that we need much faster approach that would allow to re-index BHL data on a regular basis. Global Names Architecture team is developing a system that should be able to scan 50 million pages and extract scientific names from there in a matter of one-two days. Our approach includes dramatic increase in speed and scalability of the tools at every stage of the process. On the hardware level we created a computer cluster of more than 20 servers. The cluster uses Kubernetes -- an Open Source-based cloud operating system that allow us flexibly increase or decrease power of every component by scaling services up or down at will. This allows us to fine tune amount of resources for every component and use computer power to its maximum. On the software level we have to solve several problems: scientific name recognition, name verification, name parsing. We developed new faster tools for each of the steps -- gnparser allows to break names into its elements, gnindex analyses the quality of found names, Global Names Recognition and Detection service (GNRD) detects names in BHL texts. For this talk we are presenting the architecture of the system for massive indexing of biodiversity information. The tools and system administration approaches are easily transferable to cloud services that support Kubernetes and it allows with to scale the system even further with relative ease. After successful scanning of the BHL pages the system should be fast enough to to massive name finding in all available scientific literature.

## Keywords

scientific name, Biodiversity Heritage library, name-finding, parser, reconciliation, resolution

## Presenting author

Dmitry Mozzherin