GlobalNames and the HathiTrust

ACS awardees: Dmitry Mozzherin and Matt Yoder (Species File Group, Illinois Natural History Survey, Prairie Research Institute, University of Illinois)

Our team of researchers, the Species File Group, develop and use digital tools for biodiversity informaticians, those scientists who study the Earth's species. One of the things we focus on is locating information about the Earth's species via their scientific names, a project called GlobalNames. The idea is straightforward, find a biological name like *Homo sapiens* (humans), *Apis mellifera* (the Western honey bee), or *Anopheles gambiae* (a mosquito that transmits Malaria), and you may discover information important to scientists "nearby". In the context of the GlobalNames project finding a name means parsing digitized literature or datasets, small or large. Thanks to funding from the National Science Foundation (NSF ABI 1645959, 2015) initial tools developed by Dmitry Mozzherin and Alex Myltsev were developed and hardened against the large, free corpus of scientific publications in the Biodiversity Heritage Library (BHL). Within the BHL the diversity of data (e.g. different languages, publication types, general quality of parsed text), and its structure therein let us find and resolve many edge cases in the name detecting algorithms. While finding specially formatted latinized names is challenging, the results of this work are fairly simple: at their core, they are an index indicating that "*this* name was found *there*". From these simple data many downstream features and explorations emerge, for example the list of names found on any given page of the BHL (e.g. Scientific Names on this Page), is derived from our tools.

With over 17 million items of the perhaps over 130 million known, the HaithiTrust corpus represents a wealth of knowledge waiting to be explored in novel ways. It seemed obvious to us that this would be an interesting target for GlobalNames, and when our attention was drawn to the Advanced Collaborative Support program by the HathiTrust by Prairie Research Institute Librarian Susan Braxton, we applied. While some in our group raised eyebrows about a grant that paid nothing, the potential access to huge body of copyrighted (and open) text, and the technical and project management support offered in the award, seemed worthwhile. In retrospect, the "free" grant was anything-but, supercomputing access facilitated by the time of highly skilled system administrators (Boris Capitanu) and supporting staff (Ryan Dubnicek) who handled the thorny issues of copyright was invaluable. The award process and project management has also been rewarding rather than burdensome, further adding value relative to the time it might have taken us had we used more traditional granting agencies.

Stepping back, our goal in GlobalNames is to do one "simple" thing, very (very) quickly: detect the presence of taxon names and generate an index that links the found name to the page it was located in. Keeping the project hyper-focused on this initial goal minimizes the chances that the diversity inherent in huge datasets will cause unforeseen issues. For example, pages in the HathiTrust all follow an indexing system, if this index is slightly off such that we can not trace a path from it to the processed text it points to, then our algorithm needs to make a decision as to what to do, for example skip it and move on, log an error, or die. These decisions alone are not particularly troublesome, but when processing across 50 machines, each using hundreds of threads (executing instructions in parallel to speed processing), small problems like this can have a crippling effect on the overall process. A second process, validating the results that are found (is this scientific name still used and current?), and cross linking them (is this scientific name a synonym of that scientific name?) is also part of the pipeline, as you can imagine this is an entirely separate challenge. From a software architecture standpoint we keep the tools that operate at both stages isolated so that they do their specific jobs without complications from one-another.

Our initial work, ongoing discussion, and outcomes raised a number of interesting questions that we are actively working on. Ultimately our goal is to index everything ever published in "real-time", i.e. as fast as possible, as often as possible. If the HathiTrust is around 10% of everything ever published, and it currently takes us 9 hrs on 50 multiprocessor computers to do an initial parse, then we can estimate 90 hours as an outer bound (given unlimited access to computing power, etc.). Thinking about the scope of the knowledge at this size this is remarkably little time. Fast iterations times are important as the algorithms that find names are expected to continue to improve. Each time there is a minor change to the name finding code it's desirable to run the whole process again. While small changes may at first glance seem "locally" important (tested against a much-smaller subset of the data) they may ultimately turn out to be globally misleading, for example returning more false-positives once more diverse data are encountered. Tweaking the algorithm then running the "experiment" (name finding) in an iterative fashion is facilitated by the concept of "continuous integration," a method that drives processes from stage to stage via automatic triggers in the software. How a continuous integration process might return useful information back into the resources and services provided by the HathiTrust is an open question.

Other open questions focus on when to share our results, which are essentially very large (> 1 TB) text files. Since results from any given indexing run may change (sometimes drastically, in unforeseen ways), anticipating when a new index will be useful to other researchers is difficult. Ultimately specific results will be tagged, and, ideally, offered as "gold standards" to encourage others to a) repeat the work as closely as possible in novel ways, or b) use the result for new purposes, with the derivative results themselves potentially becoming new science. We want to share as often as we can, but also try to minimize the inherent error or fuzziness captured by the algorithms.

We're excited about the downstream possibilities that might emerge from this work. Some copyrighted work, particularly that outside the domain of biologists, is not often considered as a source of research data, as it's simply too hard to know where to start to look. With even the simplest of assumptions (e.g. this is a plant name, that is an animal name) overlaid on top of the core name index, we can imagine, and seek new, previously hidden correlations that could be of interest. For example, by locating proxy relationships (i.e. found close together in the literature) between species, perhaps we can identify and explore historical periods in which pests (animals), and their crop hosts (plants) drove an increase in policy, research, or societal impact. These data could then lead to models or profiles that ultimately predict or detect similar newly emerging patterns via the process of continuously re-processing new additions to the HathiTrust or other electronic corpa. Visualization of our vast results, tools for subsequent human-assisted refinement, and software that feeds specific results to specific downstream consumers are all potential avenues of research.

While not everyone can access the supercomputing resources as we were able to as part of our ACS project, we worked hard to make as much as we could open to others. The SFG are strong advocates of open-source software and open data. All the code behind this project is readily available through GlobalNames repositories. We're also actively working on the best way to share result sets. If you have any questions or ideas as to where to go next don't hesitate to contact us. We look forward to ongoing work and to seeing what others will do with what hopefully becomes a new way to engage and explore the HathiTrust.