Global Names and the HathiTrust: Towards comprehensive indexing of taxon names in real time

Dmitry Mozzherin, Matthew Yoder, INHS; Boris Capitanu, Ryan Dubnicek, HTRC.

Abstract

Biology is organized by reference to the species it pertains to. This is done via taxonomic names that are formalized and governed by several international bodies, or ad-hoc produced for studies in progress. The Global Names (http://globalnames.org/) project ultimately seeks to index all the earth's biological names in real-time, (re)processing all digital documents ever produced in a timespan of hours. We document progress towards this goal in the context of an one-year award from the HathiTrust's (HT) Advanced Collaborative Support (ACS) program. Here we adapted software previously used to index the Biodiversity Heritage Library to run on an HPC cluster ("Big Red 3" at Indiana University). Over 110 million metadata-delimited items (> 6 billion pages of text) in the HT were processed. A multi-step processing strategy was used to create a base-index of raw data, and then several enriched, derivative datasets. First, raw name-detection using heuristic algorithms employing natural language processing in combination with allow, deny, and uncertain lists was used to create an index that included the name-string, reference to the source document, and a probabilistic estimate of name-detection confidence for the record. This step took ~9 hours. Next, this index was then verified against more than 100 biodiversity datasets. It detected 200 million unique name-candidates. These names were verified against > 100 "authoritative" datasets to produce 30 million verified, partially verified, and fuzzy matched names over ~7 days of processing. This list was cross-referenced to the original index to produce a 375G index linking names to HT documents. A final step trimmed this index. We observed that some of the HT corpus predicted to have biological names did in fact not and excluded these, i.e. results were excluded because of their occurrence in a set of a-posteriori excluded HT items. The final result is a 231G index linking canonical (authoritative) names to HT documents. The core indexing software, and instructions for their use are available at https://github.com/gnames/htindex and https://github.com/gnames/gnfinder. We identify shortcomings and logical next steps to improving speed and quality of the indexing process including the application of iterative windowed-based enrichment (if verified names are nearby, uncertain names are more probably certain), source metadata (e.g. language, or page-range), and improvements to fuzzy matching.

Introduction

The Earth's species are referenced in scientific literature by their taxonomic names. These names are governed by various international codes, for example the International Code of Zoological Nomenclature (https://www.iczn.org/) and the Botanical Code of Nomenclature (https://www.iapt-taxon.org/nomen/main.php). The rules encoded in these standards can be exploited in algorithms that crawl digitized literature and index the location of these names. The Global Names initiative (https://globalnames.org) seeks to index all taxonomic names ever published. In order to keep up with the massive, and continuously growing, corpus of new and old literature, it is desirable to plan for near "real-time" processing speeds, i.e. we should minimize the gap between when a name is born-digital, and when it is indexed. Speed is also important, as the underlying algorithms which detect, then ultimately refine lists of names are continuously improving and faster approaches can iterate more often. This automated indexing will augment manually-driven efforts (e.g. http://plazi.org/resources/treatmentbank/) to indexing species concepts, and publishing standards that integrate born-digital indexing that are emerging (e.g. efforts like those of the publisher Pensoft, http://plazi.org/).

Previously, the Global Names project focused on the corpus of literature in the Biodiversity Heritage Library (BHL, https://www.biodiversitylibrary.org/), an archive that seeks to contain all literature pertaining to the description of biological species. This corpus is clearly particularly relevant to the algorithms developed in Global Names, as many publications therein contain scientific names. Results of the indexing of the BHL are accessible through widgets on their pages, and in the GlobalNames "GN Index" application programming interface (API, https://bit.ly/2AZJngf). Processing the BHL corpus (around 50 million pages) currently takes around 4 hours on a stock Intel Core i7 or similar processor and 12 more hours for verification on a kubernetes cluster (48 cores).

This paper extends the work done in conjunction with the BHL to the digitized corpus managed by the HathiTrust (HT, https://www.hathitrust.org). While the BHL corpus is mostly out of copyright or has a permissive license, the HT corpus is around 60% copyrighted material and, as such, can typically only be accessed through certain "non-consumptive" (see https://www.hathitrust.org/htrc_ncup) technologies or data products, or through special collaboration.This project was undertaken via special access through the HathiTrust Research Center's Advanced Collaborative Support Program (https://www.hathitrust.org/htrc-advanced-collaborative-support-program). The HT corpus is, by its nature, also not restricted to biological literature. These two factors result in a corpus that is around 100 times larger than the BHL, equating to over 17 million items with around 6.2 billion pages. Published estimates of the total number of digitized documents are rare, but some suggest over 130 million exist (https://booksearch.blogspot.com/2010/08/books-of-world-stand-up-and-be-counted.html). At an order of magnitude scale we can roughly assume the HT to contain around 10% of everything ever published.

Here we report the details of our efforts to extend the Gnfinder tools to compute across the HaithTrust corpus. The key novel contribution is modification of existing indexing software to perform on a high-performance computing cluster (HPC) while referencing the HathiFiles metadata (https://www.hathitrust.org/hathifiles) to drive the indexing. Materials and methods

Prior work

The Global Names architecture, collectively open source and available at https://github.com/gnames, has evolved over around 10 years. Major speed increases were seen the past several years, thanks to NSF support (NSF ABI 1645959), such that finding names within the whole of the BHL (50 million pages of optical character recognition, OCR, derived text) was reduced in time from 40+ days to 12 hours on a circa 2018, 12-core laptop. There are two core components of the software: *gnfinder*, the code that finds names and *gnresolve* that reads names and resolves them against known data-sources to “verify” them. These are supported by several smaller libraries, including *gndict*, largely a list of names to allow/deny. The APIof all components can be used independently of the other tools, for example you can feed gnfinder a directory of OCRed text documents with a wrapping script. To use them in concert, a wrapping library for the BHL, *bhlindex* (https://github.com/gnames/bhlindex), was created. Most libraries now use Go (https://golang.org/), though Rust (https://www.rust-lang.org/) is also being explored for potential speed benefits.

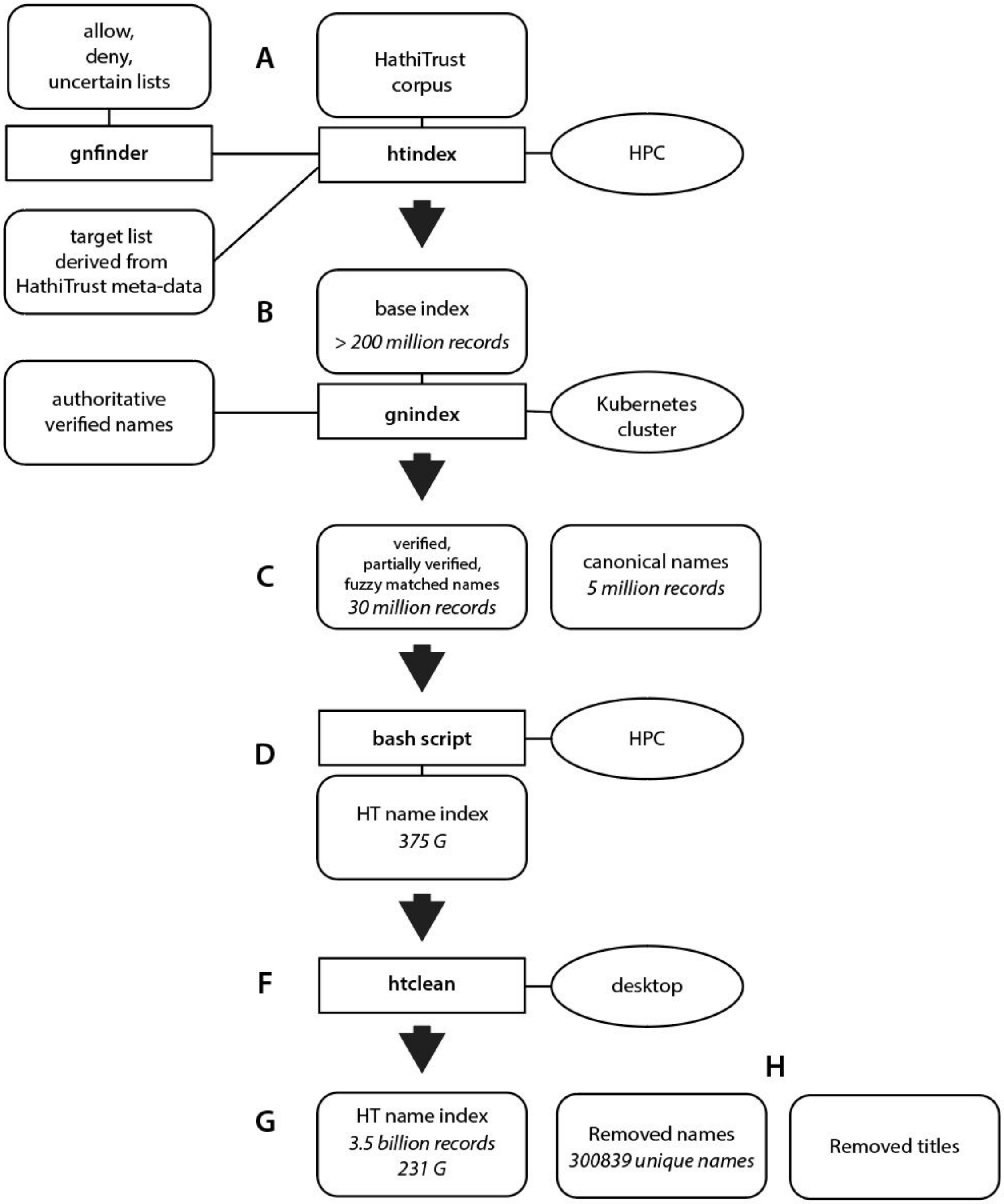
Adaptations for HPC

The primary roadblock to deploying the existing Global Names codebase within a wrapper around the HT corpus was the requirement to have all code execute on HTRC-trusted computing resources due to copyright and policy restrictions on accessing the underlying data. The 17 million volumes to be processed were stored on infrastructure managed by Indiana University, on a high-performance parallel file system (Lustre, https://kb.iu.edu/d/aqnk) attached to several high-performance computing systems (HPCs). Due to the size of the dataset (over 6 billion pages) and the technical constraints arising from running on the HPCs (https://kb.iu.edu/d/bcqt and https://kb.iu.edu/d/aoku), the existing codebase could not be used unmodified.

To address these roadblocks, Mozzherin coded a new executable, *htindex*, to drive the indexing. In brief, this executable wraps *gnfinder*, feeding it documents drawn from pre-processed lists generated by examining the directory structure housing HT documents. Technical execution details are available in the htindex repository. This architecture was successfully deployed after a series of iterative fixes to the *htindex* code (see closed issues, https://github.com/gnames/htindex/issues?q=is%3Aissue+is%3Aclosed).

Processing pipeline

Figure 1. A-G major processing steps and outcomes, see text for description. Rounded boxes: text or CSV files; ovals: hardware used; rectangles: executable/script.



The pipeline is summarized in Figure 1. We iterated our initial *htindex* run (Fig. 1A) 2 full times, with many partial runs that did not complete (e.g. due to memory leak bugs) but lead to improvements in the process and code. The underlying engine in *htindex, gnfinder,* produces a predicted list of names. Known and unknown limitations to *gnfinder* algorithms result in errors (e.g. false positives). This leads to the need for the second step (Fig. 1B), verification and normalization of the initial name-string index (a name-string is a candidate scientific name) via *gnindex. Gnindex* (Fig. 1C) takes a name-string and runs it through a lexical normalization step. The result is mapped via exact, partial, or fuzzy matching against over 100 authoritative nomenclatural datasets. Two lists are then produced, one contains the various matches, the second is a reference that isn't further used in the pipeline- a list of canonical forms, that is how a name *should* appear according to the authoritative lists. The results in the match list are then cross-referenced (Fig. 1D) back into the original index, mapping the results from C back to B. At this step we have a very large file indexing authoritative names to documents in the HathiTrust. A final step, *htclean* (Fig. 1F) was taken based on examining the results of step C. Here we observed that some HT documents (Supplementary Material, *removed-titles.txt*) identified as containing biological names in fact did not, for example the partial match algorithm combined an element from a good name with an element that was not a name. Knowing this, we created a "deny" list of some titles, and a corresponding list of names (Fig. 1H). These names were subtracted from the initial index (results of Fig 1A) to produce the final, cleaned result (Fig. 1G).

Results

Libraries

Htindex is available as open source at https://github.com/gnames/htindex. The post processing pipeline code, *htclean* is at https://github.com/gnames/htclean.

Runtime

Initial runs of the code were focused on testing out the code and workflow, which continued to evolve over the course of the project. Though the workflow and code became more complex over the course of the project, runtimes were mostly consistent due to improvements in the code and the ability to utilize improved HPC systems at Indiana University. The final processing of the corpus took 9 hours on 50 machines with a total of 2,400 CPU cores, with a subsequent verification step that took 7 days.

Names

The final product (Fig. 1G), a 231G text file resulting from the processing pipeline is available at http://opendata.globalnames.org/ht/ (see also Supplementary Material) under a CC0 license. This is the result of cleaning the result of step G (Fig. 1). The cleaning process removed over

140G of data. Each record is an instance of a name in a document. Column descriptors are available in the README.md (Supplementary Material). Our ultimate run returned 3.5 billion name instances, for 5.5 million canonical names (names recognized by an authority). For reference, there are an estimated 2.1 million valid taxon names, with ~5 million synonyms known (http://www.catalogueoflife.org/).

Discussion

Our core result, an index of biological names across over 10% of everything ever published is a starting point. There are several logical next steps focused around exploring our result. Some of these are obvious, for example examining a small subset of results (Supplementary Material, Table 1) we can see that the very common Spanish word “La” was detected as a scientific name candidate over 26 million times (Supplementary Material, Table 3, and “*La cerveza*” as a species epithet 7,121 times, Supplementary Material, Table 2). There is indeed a genus *La* from Crambidae (grass moths) family. To distinguish biological and non-biological use of such a word would require state-of-the-art machine learning algorithms that are able to analyse the context of the word usage. Another example *Habeas corpus*, an unfortunate name for species from Diplommatinidae, a small land snail family, was detected over 260,000 times, mostly apparently in legal documents. This is a logical candidate for the “Deny” list. Setting up a pipeline for post processing of the core result will enable a relatively huge number of instances to be eliminated.

We observed that some documents in the HT corpus contained strings predicted in the first steps (Fig. 1A) to be names, but on human inspection they appeared not to be. Based on this we created a list of volumes to exclude (Fig. 1H). Because this excluded list was not completely verified as containing no biological names, some biological names may have been erroneously excluded. Refining a master deny/allow list of those works that have biological names that won’t will ultimately greatly speed the indexing processes. Creating a master list that can be applied beyond reference to the HT corpus is a major challenge that requires cross-referencing metadata, for example between the BHL and the HT.

Making the raw data accessible via more user friendly APIs was beyond the scope of this project, but is another logical next step. Once examined and cleaned, further integrating it into the native HT metadata should enable a broad range of questions to be addressed. Adding the result to the core gnresolver api (https://index.globalnames.org/) will expose a vast array of new literature to a wide range of biologists already using this resource.

The project crystallized several concepts of interest to the ultimate goal of real-time processing. Given that the process of indexing as implemented now is relatively fast, major improvements will likely be advanced by exploiting 1) a-priori knowledge of the documents being processed, for example their asserted language and 2) a-posteriori knowledge from prior runs, for example the detected absence of few-to-no biological names in some works may be exploited to narrow or eliminate names as false positives. The project considered indexing the words surrounding each name result, for example 3 on either side, to build a windowed profile to exploit for accuracy estimates. Copyright restrictions, however, limited our ability to both access and release such contextual information.

From a broader perspective, the HT may accession in excess of 40,000 documents *per day*, meaning that processes which take 7 days to complete are “obsolete” before they finish. While 280,000 documents is a fraction of the total, it suggests that if we are to index in real time then additional layers that track differences to both the algorithms that detect and clean names, and the quality of the target documents (e.g. improvements to OCR) will be necessary. Continuous integration approaches, i.e. those that trigger processing based on changes to code-bases or data, will almost certainly be key elements of this pipeline.

Acknowledgements

We thank the Species File Group (https://speciesfilegroup.org) for technical resources and the opportunity to explore these types of questions. ACS staff supported the process throughout.

Funding

This work was supported by a grant from the Advanced Collaborative Support program by the HathiTrust
(https://wiki.htrc.illinois.edu/display/COM/Advanced+Collaborative+Support+%28ACS%29+Awards.

Supplementary material

Table 1. All supplementary/addendum files are available at http://opendata.globalnames.org/ht/

File	Description	sha512
ht-sci-names-2020-06-30.tar.gz	CSV of the instance data (Fig 1. “”).	b715ba74387fc37716c4e445b10235a87c36bbbc5e9c9834080889461ff22e2638bdd92599535a9d5328e24d7a2fcda420f83d7eeefb1979021d00d1504573c8
ht-sci-names-addendum-only-2020-06-30.tar.gz	Compressed collection of addendum/supplementary material. *Includes files below.*	
README.md	Describes addendum-only files, and columns in ht-sci-names-2020-06-30.tar.gz.	
ht-sci-names-grouped.csv	See README.md	
removed-names-grouped.csv	See README.md	
removed-titles.txt	See README.md	

Table 2. Fifty name-string hits of note and the number of times they were encountered. Most strings are also real biological names, most for common species, but not all.

Habeas corpus	269766	*Canis minor*	1970
Oedipus complex	100784	*Cornu copiae*	1866
Homo sapiens	73338	*Canis major*	1831
La paloma	19920	*Venus bella*	1729
La cerveza	7121	*Bacterium coli*	1678
Cannabis sativa	7018	*Solanum tuberosum*	1628
Staphylococcus aureus	5658	*Aspergillus niger*	1547
Homo erectus	5578	*Dalla zona*	1507
Aedes aegypti	5272	*Sequoia sempervirens*	1400
Drosophila melanogaster	4985	*Bos primigenius*	1352
Pero otra	4825	*Venus aurea*	1256
La cucaracha	4537	*Clostridium difficile*	1238
Oedipus rex	4509	*Ginkgo biloba*	1224
Viola tricolor	3405	*Datura fastuosa*	1076
Sequoia gigantea	3019	*Teredo navalis*	1037
Unda maris	2882	*Cassia fistula*	1028
Victoria regia	2872	*Mus musculus*	1009
Primula veris	2802	*Rosa divina*	888
Cannabis indica	2716	*Mimosa pudica*	860
Villa una	2361	*Gemma purpurea*	855
Zea mays	2315	*Rosa bella*	849
Homo habilis	2247	*Questa media*	815

Tuber cinereum	2106	*Giardia lamblia*	813
Bacillus coli	2051	*Amanita muscaria*	807
Lapsus calami	2028	*Musca domestica*	781

Table 3. Fifty strings that look like biological names, and were detected as names, but are not, and the number of times they were encountered.

La	26690395	Rita	4260892
Emma	11234801	Alain	4247787
Felix	9512313	Lisa	4218182
Paulus	8696243	Lucia	4187554
Petrus	8358640	Camara	4047476
Patricia	8226764	Amelia	3996307
Eleanor	6861684	Sylvia	3980769
Arte	6070153	Elena	3880278
Justicia	5992744	Nora	3799740
Castilla	5975838	Faber	3676111
Platon	5352145	Romana	3591381
Nantes	5308507	Lucius	3583782
Dana	5289051	Angela	3577263
Leonardo	5064615	Osaka	3573184
Pereira	5042899	Marta	3572422
Paula	4937446	Casas	3508934
Sancho	4876805	Ortega	3508169
Moreno	4776598	Jana	3459484
Tulsa	4711918	Enrico	3441098
Louisa	4668040	Alexis	3440813
Mater	4556914	Claudius	3436524

Regis	4497534	Memoria	3432920
Guillermo	4477448	Una	3342326
Mercedes	4442144	Marietta	3333753
Roche	4387542	Quijote	3323771