Improving the discoverability of biodiversity data using the Global Names Finder

Anne E Thessen[‡,§], Dmitry Mozzherin[|], David Peter Shorthouse[¶], David J Patterson[#]

‡ University of Colorado Anschutz Medical Campus, Aurora, CO, United States of America
§ The Ronin Institute for Independent Scholarship, Monclair, NJ, United States of America
| University of Illinois, Champaign, IL, United States of America
¶ Agriculture & Agri-Food Canada, Ottawa, Canada
# University of Sydney, Sydney, New South Wales, Australia

Corresponding author: Anne E Thessen (annethessen@gmail.com)

Academic editor: Elycia Wallis

Abstract

The majority of biodiversity data is not findable, accessible, integratable, or reusable, partially because of a lack of metadata. Taxonomic names as metadata are useful, but not sufficient because these names may be updated as knowledge progresses. There is a great need for tools and services that can scale up to create and maintain metadata for the vast and varied long tail of dark data. Here we examine the use of GNFinder as a tool for creating and maintaining metadata using mentions of taxa in text from publications corresponding to data sets deposited in Dryad. Most studied taxa were mentioned in the publication using a properly formed scientific name, with a few exceptions for studies that only used vernacular names and only mentioned taxa in the corresponding files. GNFinder had a high F1 Score (0.86) representing a balance between precision (0.91) and recall (0.82). GNFinder had lower performance when a name string was an irregular abbreviation, had unexpected capitalization or punctuation, or contained a qualifier (like *aff.* or *cf.*). Approximately 14% of the name strings identified in text published from 1996 to 2012 were outdated and updated to a current, valid name. Automated metadata creation and maintenance at scale using GNFinder can make it easier to find biodiversity publications as demonstrated by the Biodiversity Heritage Library and HathiTrust.

Keywords

taxonomic names, indexing, metadata, named entity recognition

Introduction

Much attention has been given to the “data deluge” (Hey et al. 2009) and the need to render these data computable (Li and Chen 2014) in order to successfully address

pressing societal challenges (Guo et al. 2015). The challenges are especially acute because of the large amounts of “long tail” and “legacy” data that are not findable, accessible, integratable, or reusable (Thessen and Patterson 2011, Heidorn 2008). One of the critical steps to unifying data is the application of data and metadata standards so that information can be effectively discovered, indexed, organized, and made ready for analysis (Schriml et al. 2020). Some subdisciplines within biology have worked to address this problem by developing various types of data and metadata standards (e.g., Field et al. 2008, Wieczorek et al. 2012). Projects like the Monarch Initiative (Shefchek et al. 2020), Planteome (Cooper et al. 2018), Translator (Fecho et al. 2022), the Global Biotic Interactions database (Poelen et al. 2014), the Global Biodiversity Information Facility (Telenius 2011), and many more use a wide variety of community-developed data, metadata, formatting, and exchange standards to integrate and make computable incredibly heterogeneous biology knowledge. The application of these standards has seen a steady increase in some disciplines, but the backlog of non-computable data remains vast (Marshall et al. 2018, Petty et al. 2020). There is a great need for tools and services that can scale up to create and maintain metadata for the quantity and variety of data in the long tail. The absence of computable metadata has severely impaired data discoverability in biodiversity (see Thessen et al. (2012b) for a specific example; also Walls et al. (2014), Mounce (2015)) and slows progress toward data-driven biology.

One unique aspect of biodiversity data is that scientific names can be used as near universal metadata (Patterson et al. 2010). There are rules of nomenclature that govern the use, representation, and modification of scientific names. Despite this standardization, names make poor identifiers because they are not unique or persistent identifiers for taxonomic concepts (such as species) and the continual nature of scientific discovery prevents them from ever being so. Names are not represented consistently in publications and data sources (Patterson et al. 2016, Page 2011). Adequately managing scientific names as metadata requires tools that automatically find names in their various forms in data sources (Gerner et al. 2010, Thessen et al. 2012a, Le Guillarme and Thuiller 2022, Pafilis et al. 2013) and generate computable metadata that allows for resolution of taxonomic concepts over time. This requires more than a standard. This requires a “living” metadata file that can be automatically updated in a transparent, traceable way. Such a file will help to incorporate more biodiversity data into the growing body of computable biological knowledge.

Here we examine the feasibility of living metadata in biodiversity using GNFinder, a tool that can find scientific names in text with a high degree of precision and recall and return the corresponding current, valid name in JSON or CSV format. GNFinder was developed with the goal of processing everything ever published and is currently being used by the Biodiversity Heritage Library (BHL) (Mozzherin and Myltsev 2017), the Encyclopedia of Life (Thessen and Parr 2014), HathiTrust (Mozzherin et al. 2022), and TaxonWorks. For example, GNFinder processes the 60 million pages in BHL in seven hours on a laptop computer and the 6 billion pages in the HathiTrust in less than a day on an HPC cluster (Mozzherin et al. 2022). While GNFinder records the exact location of a taxonomic name in

text, this study examined the utility of GNFinder for adding current taxonomic names as document-level metadata to improve data discoverability.

Materials and Methods

This paper seeks to determine the efficacy of GNFinder for adding taxonomic metadata to the published literature. As a result, annotations are made and results reported at the document level. Multiple instances of the same name string in a document were only counted once.

Description of the Data

Dryad is a repository for ecology and evolution data files that correspond to publications (Vision 2010). We randomly chose 250 data packages from Dryad and retrieved the corresponding publication for analysis in 2012. Each data package consisted of one or more data files (from Dryad) and one publication pdf file (from our institution library). Only some of the manuscript pdf files were machine readable. Others were scanned library copies. The scans were OCRed in 2020 using Adobe Acrobat Pro. Data files were in a variety of formats including txt, nex, docx, and xlsx among others. We were able to analyze the manuscript pdfs and data files from 215 data packages. The txt versions of the manuscript pdfs used in this study are available in GitHub (Mozzherin 2022f).

Description of GNFinder

GNFinder is a web service that uses a combination of naive Bayes, rules, and lists to find scientific names in text (Mozzherin 2022a, Mozzherin et al. 2018). The rules create features that the Bayesian algorithm uses to calculate a score (Fig. 1). There are two types of rules:

1. heuristic, based on a “stop” list containing terms that are likely to appear with, but not be a part of a scientific name (such as “environmental sample”), a “caution” list containing words that are frequently used in European languages that also appear in scientific names, and a “go” list containing terms that are highly likely to only be used in a scientific name as a genus or a specific epithet, and
2. statistical, for example, “are the word endings common in Latin”.

The score (result of naive Bayes) is represented as “odds” instead of a probability. GNFinder output can be configured to show the results from each of the rules and the final Bayesian score (Fig. 1).

Once GNFinder has recognized a name in the text, the name string is parsed into its semantic elements such as genus name, specific epithet, year of publication, authorship, etc. using GNParser (Mozzherin 2022b, Mozzherin et al. 2017). Parsing is essential for matching different variants of taxonomic names (such as *Canis familiaris* Linneaus and *Canis familiaris* L.). This is a process of name string “normalization” that can establish the

canonical form of the name. Normalization is essential for finding the current, valid name according to a user-chosen taxonomic reference using GNVerifier (Mozzherin 2022c).

GNVerifier compares the name string found by GNFinder to names in a list of over 200 reference taxonomies (Mozzherin 2022d). The default setting looks in all available lists and prioritizes results from Catalogue of Life (Hobern et al. 2021). Users can choose to view all available matches or only the best match. GNVerifier compares seven features of the found name and the matched name to calculate a score used to rank the matches (Fig. 2). In this way, if GNFinder finds a name that is no longer in contemporary use, it may be able to return the current name. This process of matching old names to current names is referred to as name resolution.

GNFinder can be accessed directly through the webpage (Mozzherin 2022a) or by using the API (Mozzherin 2022e). Complete documentation of GNFinder and description of the JSON output can be found on GitHub (Mozzherin et al. 2018).

Data Preparation

Two human annotators found every unique name string used to refer to a taxon in every manuscript pdf for the 215 publications (Thessen 2022). Manuscripts were analyzed without the References sections. If a taxon name was used as an adjective, such as in “crocodilian anatomy,” it was not included in the annotator lists. Names of clades that were not taxonomic names, like “deuterostome,” were not included in annotator lists. Vernacular names were collected, but were not part of the GNFinder performance calculations. Mentions of a genus were included as a separate reference to a taxon even if a species within that genus was also mentioned. The results from the annotators were compared to calculate annotator agreement. In order to normalize the different types of pdf files, each pdf was transformed to a text file using Adobe Acrobat Pro before being passed to GNFinder.

Testing Performance

GNFinder returned all of the found name strings and their associated taxon concepts in a CSV file and in a JSON file (Mozzherin 2022f). The results from the annotators and from GNFinder were compared and performance metrics for GNFinder were calculated using a Python script (Thessen 2022). Results from GNVerifier were not used to calculate performance metrics. Results were not filtered using the Bayes odds score. Results from GNFinder and the human annotators were used to calculate precision*[1] (a measure of correctness), recall*[2] (a measure of completeness), and F1 Score*[3] (harmonic mean of precision and recall).

To describe the advances made by GNFinder, we took a subset (17 randomly selected) of the 215 publications and calculated performance metrics using several other published name-finding tools: TaxonFinder (Leary et al. 2007), NetiNeti (Akella et al. 2012), LINNAEUS (Gerner et al. 2010), TaxoNERD (Le Guillarme and Thuiller 2022), ORGANISMS (Pafilis et al. 2013), and Quaesitor (Little 2020) for comparison.

Assessment of Metadata Creation

The subset of publications used to compare GNFinder performance to other, similar tools was also used to explore the utility of GNFinder for creating metadata. To test this, we created a list of taxa represented by all of the name strings recorded by the annotators from the publication and the corresponding data files in Dryad. For each data package (publication and data files) we calculated the total number of taxa present, the taxa only represented in the data files, the taxa only represented by a vernacular name, and the taxa only represented as an improperly formed scientific name. These lists included higher level taxa that appeared in the text or data, or as a vernacular or a scientific name, even when a child taxon was present. Paraphyletic taxa referred to by a vernacular name where counted as being represented by a vernacular name only unless all of the scientific names implied by that vernacular name were also present (e.g., barrel cactus is a paraphyletic group including *Echinocactus* and *Ferocactus*).

To explore the prevalence of outdated names in the literature, we examined the results from GNVerifier. Any names that were found to be exact matches for synonyms (i.e., matchType = Exact and isSynonym = True) were considered, for the purpose of this exercise, as outdated names even though they may reflect different taxonomic preferences of the sources.

Results

Annotator Agreement

To test annotator agreement in recognizing name strings, 27 manuscript pdf files were processed by both annotators. Vernacular names were not included, but abbreviations of scientific names were included. A Cohen's kappa coefficient (Cohen 1960) was calculated for each data file and for the overall dataset. The kappa agreement for individual files ranged from 0.534 to 0.963. The overall kappa agreement was 0.832, which indicates good agreement between the annotators.

GNFinder Performance

GNFinder performance was calculated for 215 manuscripts (Table 1) containing 1,589,065 words and 9,753 name strings. Performance was high (F1 = 0.86). Overall, GNFinder produced 2,559 errors (758 false positives and 1,801 false negatives) out of 9,753 scientific name strings. The annotators recorded 1,939 unique vernacular name strings, but these were not used in the performance metrics.

Error Analysis

GNFinder made 2,559 unique errors, most of which were false negatives (70%) due to GNFinder not being able to read figures, trinomial abbreviations (such as *L. g. confertiflora*), unusual formatting and punctuation used to save room in tables, and parentheses in names (such as *Nanorana* (*Paa*) *bourreti*). GNFinder is not designed to perform well on virus names. Properly formed abbreviations, such as *C. familiaris* were returned and parsed by GNFinder, but were not verified.

Comparison to Other Tools

GNFinder had the highest F1 Score (Table 2), but LINNAEUS had the highest precision and TaxoNERD had the highest recall. The comparison of different tools was very challenging because each tool was designed for a slightly different task. Both LINNAEUS and TaxoNERD were designed to find vernacular names in addition to taxonomic names while GNFinder, Taxon Finder, and NetiNeti were designed to only find Latinized taxonomic names. Vernacular names had to be manually removed from LINNAEUS and TaxoNERD results before calculating performance. Quaesitor conflated finding the name with resolving the name and did not return higher-level taxon names, which artificially lowered the reported performance. For example, because Quaesitor returns the resolved name instead of the found name, it appears to miss abbreviated names (high false negatives) and return names that are not present (high false positives). ORGANISMS returned NCBI Taxon identifiers only, which made comparison impossible.

Metadata Coverage Assessment

For the majority of this subset of the 215 publications, all of the taxa were referenced in the publication, but one data package had 78% of taxa appearing in the data file only (Table 3). One third of the data packages (29%) had 50% or more of the taxon concepts appearing as vernacular names only. The taxa appearing only as anything other than a well formed taxonomic name were few, but not zero.

Of the 8,710 names returned by GNFinder from 215 publications, 1,258 were updated to a current name according to Catalogue of Life (default setting) by GNVerifier (14.4%). The manuscripts containing these names had been published from 1996 to 2012 with most published in 2012 (41%) and 2011 (32%).

Discussion

Data are rendered non-discoverable because of the ways taxonomic names change over time and because of the idiosyncratic ways in which names are expressed. The Global Names project recognizes that names may be expressed in various forms, and the infrastructure has been designed so that we can extend GNFinder to parse additional

variant forms (Patterson et al. 2016). The results can be mapped to the canonical form of a name (i.e., the Latin binomial, genus name capitalized with a single space separating it from the species epithet, no annotations and no authority information), and then track the canonical form to a currently accepted name through an understanding of synonymies. This process is illustrated in this study and with the iPlant TNRS service that uses the GNParser (Boyle et al. 2013).

GNFinder can find scientific names in text and resolve name strings to a current name in a user-chosen list. This is also known as Named Entity Recognition (NER) and is a very active area of research in the Natural Language Processing and Machine Learning fields (Goyal et al. 2018). In addition to GNFinder, there are other algorithms that perform NER for taxa (see results above). LINNAEUS does an excellent job of finding species names in text and resolving those names to concepts in the NCBI taxonomy with very high precision (Gerner et al. 2010). TaxoNERD uses a deep neural network (DNN) to find mentions of taxa in text with very high recall (Le Guillarme and Thuiller 2022). Our intention with GNFinder is to balance precision and recall, as is suggested by these results. Comparing the results of these different algorithms was difficult and can be misleading. The ideal method for comparing algorithm outputs is to compare their results against a publicly-available gold standard corpus. Two potential corpora exist, COPIOUS (Nguyen et al. 2019) and S800 (Pafilis et al. 2013), but none of these algorithms have processed both. Additionally, the algorithms have subtle differences in the scope of their results.

The name-resolution function performed by GNFinder also serves as quality control for resources like BHL, which have used Optical Character Recognition (OCR) as part of the digitization process. OCR can introduce errors in names at rates that depend heavily on the language and typography used (historical texts are particularly vulnerable) (Wei Q et al. 2010). When first checked over a decade ago, approximately 30% of taxonomic name strings in BHL contained an OCR error (Freeland 2009). Since then, OCR errors have been greatly reduced in BHL (Anonymous 2014, Mika 2017), but the need to find and correct misspellings in large corpora at scale has not disappeared. GNFinder can address these errors by resolving novel misspellings through its use of naive Bayes. In this way, a one-of-a-kind, erroneous name string can be recognized and resolved to a current name.

Not all of the 11,692 unique name strings identified by human annotators were properly formed scientific names and their regular abbreviated forms. A properly formed scientific name, for the purposes of this paper, includes a binomial (*Panthera leo*), trinomial (*Felis silvestris lybica*), or higher level taxon name with or without the authority and the regular abbreviation (*P. leo*). This is important because the semi-supervised portion of GNFinder relies on the rules of nomenclature to identify scientific names in text. Out of all of the documented ways a taxonomic name can be represented (Patterson et al. 2016), there were four types of taxon name modification that reduced GNFinder performance.

1. **Irregular abbreviation**. Irregular abbreviations were scientific names shortened by any means except: a) the first one or two letters of the generic name with the first capitalized, b) followed by a full stop and a space, c) followed by the specific epithet. Often these included names with strain designations or location

information. While regular abbreviations were identified by GNFinder, they were not resolved by GNVerifier.

2. **Unusual punctuation or spacing**. Unconventional spacing and punctuation can be used to represent hybrids, species complexes, or unofficial specific epithets such as *Aus bus* × *cus*, *Aus bus*/*cus*, and *Aus* "*bus*".

3. **Improper capitalization**. Publications will sometimes contain a genus name that is not capitalized or a specific epithet that is capitalized, such as *E. Caballo*. GNFinder needs the capitalization to recognize the genus name and specific epithet.

4. **Adding two letter qualifier abbreviations**. Manuscripts often have qualifiers added to the names, such as cf. aff. or sp. When these abbreviations occur within the name string, GNFinder will not recognize the binomial. When they occur after the name string, such as *Bos* sp., GNFinder will include the sp. in the returned name string.

The benefits of including all mentioned taxa as metadata are unclear because a paper may be about one specific taxon, but mention several; so, including all mentioned taxa could lead to less precise document retrieval. Author-supplied keywords and algorithms that can detect keywords from text (Huang et al. 2020) can increase the precision of document search, especially for publications that are about a single taxon; however, keywords are limited to 5-10 and many studies produced data for more than 5-10 taxa. Additionally, keyword extraction algorithms, like NER-RAKE, are not designed for taxonomic name extraction (Huang et al. 2020). GNFinder includes in its results the "mainTaxon" and "mainTaxonRank" for each document, which can partially address this issue by reporting the lowest rank taxon that includes at least 50% of all the mentioned species.

The utility of including mentions of taxa above the rank of genus is also unclear. The argument against this is that parent taxa can be automatically added from an authoritative hierarchy when a taxon is detected; thus, keeping the search criteria broad enough to include them decreases precision of the algorithm and the document search unnecessarily. The arguments for this are the cases where only higher level taxa are mentioned and in the cases where more than one genus has the same name. A path forward is to add both types of higher level taxa (i.e., found and inferred) to the metadata file and label them appropriately.

These results suggest that the majority of relevant taxa are mentioned in the publication and thus searching the publication file will generate most of the needed taxonomic metadata for the accompanying data. This argues that the first priority for future GNFinder development should be improving the extraction of names from published manuscripts, especially proper handling of names in figures. It is known that not all data are published (Heidorn 2008, Shin et al. 2020), but the proportion of digitized data that do not have an accompanying publication is unclear. The data in Table 3 argues that identifying vernacular names and reading files that are not in .txt format should be the next priorities. Name

detection in text written in languages other than English was not discussed here, but is already part of future development plans for GNFinder.

Conclusion

Taxonomic names are useful metadata for finding, accessing, integrating, and reusing data, but only if they can be effectively resolved when there are changes in taxonomy, or when a name represents more than one species concept. GNFinder demonstrates good overall performance on finding name strings that occur in text representing taxa across the tree of life. In this study, approximately 14% of names used in publications 10–20 years old were out-of-date and were mapped to a current, valid name by GNVerifier. Furthermore, the speed of GNFinder makes it possible to apply names as living metadata to the entire body of published literature. Without name-finding algorithms, much biological content cannot be accessed by searches based on the taxon name. The use of GNFinder to tag files with appropriate taxonomic metadata improves discovery on an unprecedented scale. The major advance of GNFinder is the almost unlimited scalability and reliability, while still preserving reasonably high quality of name detection.

Acknowledgements

This work was supported by NSF grants ABI 1062387 Collaborative Research: ABI: Innovation: The Global Names Architecture, an infrastructure for unifying taxonomic databases and services for managers of biological information and ABI 1356347 ABI Development: Global Names Discovery, Indexing and Reconciliation Services. We thank Lakshmi Akella and Patrick Leary for access to software.

Funding program

NSF ABI 1062387 and 1356347

Grant title

Collaborative Research: ABI: Innovation: The Global Names Architecture, an infrastructure for unifying taxonomic databases and services for managers of biological information and ABI Development: Global Names Discovery, Indexing and Reconciliation Services.

Conflicts of interest

Contributions made by David Shorthouse represent work initiated prior to his employment with Agriculture and Agri-Food Canada

References

- Akella LM, Norton CN, Miller H (2012) NetiNeti: discovery of scientific names from text using machine learning methods. BMC Bioinformatics 13: 211. https://doi.org/10.1186/1471-2105-13-211
- Batista-Navarro R, Islam A, W U, et al. (2014) Enriching the legacy literature with OCR corrections and text-mined semantic metadata. TDWG 2014 ANNUAL CONFERENCE PROCEEDINGS. TDWG 2014, Jönköping, Sweden. URL: https://mbgocs.mobot.org/index.php/tdwg/2014/paper/view/588/0
- Boyle B, Hopkins N, Lu Z, Raygoza Garay JA, Mozzherin D, Rees T, et al. (2013) The taxonomic name resolution service: an online tool for automated standardization of plant names. BMC Bioinformatics 14: 16. https://doi.org/10.1186/1471-2105-14-16
- Cohen J (1960) A coefficient of agreement for nominal scales. Educational and Psychological Measurement 2 (1): 37-46. https://doi.org/10.1177/001316446002000104
- Cooper L, Meier A, Laporte M-, Elser JL, Mungall C, Sinn BT, et al. (2018) The Planteome database: an integrated resource for reference ontologies, plant genomics and phenomics. Nucleic Acids Res 46: 1168-1180. https://doi.org/10.1093/nar/gkx1152
- Fecho K, Thessen A, Baranzini S, Bizon C, Hadlock J, Huang S, Roper R, Southall N, Ta C, Watkins P, Williams M, Xu H, Byrd W, Dančík V, Duby M, Dumontier M, Glusman G, Harris N, Hinderer E, Hyde G, Johs A, Su A, Qin G, Zhu Q, Acevedo L, Ahalt S, Alden J, Alkanaq A, Amin N, Avila R, Bada M, Balhoff J, Baumgartner A, Baumgartner W, Belhu B, Brandes MK, Brandon N, Brush M, Bruskiewich R, Burtt N, Callaghan J, Cano MA, Carrell S, Caufield JH, Celebi R, Champion J, Chen Z, Chen M, Chung L, Clemons P, Cohen K, Conlin T, Corkill D, Costanzo M, Cox S, Crouse A, Crowder C, Crumbley M, Dai C, Azevedo RDM, Deutsch E, Dougherty J, Duvvuri V, Edwards S, Emonet V, Fehrmann N, Flannick J, Foksinska A, Gardner V, Gatica E, Glen A, Goel P, Gormley J, Greyber A, Haaland P, Haendel M, Hanspers K, He K, Henrickson J, Hoatlin M, Hoffman A, Huang C, Hubal R, Huellas-Bruskiewicz K, Huls F, Hunter L, Issabekova T, Jarrell M, Jenkins L, Joshi A, Kang J, Kanwar R, Kebede Y, Kim KJ, Kluge A, Knowles M, Koesterer R, Korn D, Koslicki D, Krishnamurthy A, Kvarfordt L, Lee J, Leigh M, Lin J, Liu Z, Liu S, Ma C, Magis A, Mamidi T, Mandal M, Mantilla M, Massung J, Mauldin D, McClelland J, McMurry J, Mease P, Mendoza L, Mersmann M, Mesbah A, Might M, Morton K, Moxon ST, Muller S, Muluka AT, Mungall C, Osborne J, Owen P, Patton M, Peden D, Peene RC, Persaud B, Pfaff E, Pico A, Pollard E, Price G, Putman T, Raj S, Ramsey S, Reilly J, Riutta A, Roach J, Rosenblatt G, Rubin I, Rucka S, Rudavsky-Brody N, Sakaguchi R, Santos E, Schaper K, Schmitt C, Schurman S, Scott E, Seitanakis S, Sharma P, Shefchek K, Shmulevich I, Shrestha M, Shrivastava S, Sinha M, Smith B, Solbrig H, Soman K, Southern N, Stillwell L, Strasser M“M, Thessen A, Tinglin J, Tonstad L, Tran-Nguyen T, Tropsha A, Unni D, Vaidya G, Veenhuis L, Viola A, Grotthuss M, Wang M, Wang P, Weber R, Wei Q, Weng C, Whitlock J, Williams A, Womack F, Wood E, Wu C, Xin JK, Xu C, Yakaboski C, Yao Y, Yi H, Yilmaz A, Zheng M, Zhou X, Zhou E, Zisk T, Translator Consortium (2022) Progress toward a universal biomedical data translator. Clinical and Translational Science https://doi.org/10.1111/cts.13301

- Field D, Garrity G, Gray T, Morrison N, Selengut J, Sterk P, et al. (2008) The minimum information about a genome sequence (MIGS) specification. Nat Biotechnol 26: 541-547. https://doi.org/10.1038/nbt1360
- Freeland C (2009) An evaluation of taxonomic name finding & next steps in Biodiversity Heritage Library (BHL) developments. Nature Precedings1-1. https://doi.org/10.1038/npre.2009.3372.1
- Gerner M, Nenadic G, Bergman CM (2010) LINNAEUS: a species name identification system for biomedical literature. BMC Bioinformatics 11: 85. https://doi.org/10.1186/1471-2105-11-85
- Goyal A, Gupta V, Kumar M (2018) Recent Named Entity Recognition and Classification techniques: A systematic review. Computer Science Review 29: 21-43. https://doi.org/10.1016/j.cosrev.2018.06.001
- Guo HD, Zhang L, Zhu LW (2015) Earth observation big data for climate change research. Advances in Climate Change Research 6: 108-117. https://doi.org/10.1016/j.accre.2015.09.007
- Heidorn P (2008) Shedding Light on the Dark Data in the Long Tail of Science. Libr Trends 57: 280-299. https://doi.org/10.1353/lib.0.0036
- Hey AJ, Tansley S, Tolle KM, et al. (2009) The fourth paradigm: data-intensive scientific discovery. Microsoft Research, Redmond, WA, US.
- Hobern D, Barik SK, Christidis L, Garnett TS, Kirk P, Orrell TM, et al. (2021) Towards a global list of accepted species VI: The Catalogue of Life checklist. Org Divers Evol 21: 677-690. https://doi.org/10.1007/s13127-021-00516-w
- Huang H, Wang X, Wang H (2020) NER-RAKE : An improved rapid automatic keyword extraction method for scientific literatures based on named entity recognition. Proc Assoc Inf Sci Technol 57 https://doi.org/10.1002/pra2.374
- Leary PR, Remsen DP, Norton CN, Patterson DJ, Sarkar IN (2007) uBioRSS: tracking taxonomic literature using RSS. Bioinformatics 23: 1434-1436. https://doi.org/10.1093/bioinformatics/btm109
- Le Guillarme N, Thuiller W (2022) TaxoNERD: deep neural models for the recognition of taxonomic entities in the ecological and evolutionary literature. Methods Ecol Evol https://doi.org/10.1101/2021.06.08.444426
- Little DP (2020) Recognition of Latin scientific names using artificial neural networks. Appl Plant Sci 8: 11378. https://doi.org/10.1002/aps3.11378
- Li Y, Chen L (2014) Big biological data: challenges and opportunities. Genomics Proteomics Bioinformatics 12: 187-189. https://doi.org/10.1016/j.gpb.2014.10.001
- Marshall CR, Finnegan S, Clites EC, Holroyd PA, Bonuso N, Cortez C, et al. (2018) Quantifying the dark data in museum fossil collections as palaeontology undergoes a second digital revolution. Biol Lett 14 https://doi.org/10.1098/rsbl.2018.0431
- Mika K (2017) Crowdsourcing Data Enhancements to Improve Named Entity Recognition in the Biodiversity Heritage Library. Biodiversity Information Science and Standards; Sofia. search.proquest.com https://doi.org/10.3897/tdwgproceedings.1.17354
- Mounce R (2015) Dark Research: information content in many modern research papers is not easily discoverable online. PeerJ PrePrints https://doi.org/10.7287/peerj.preprints.773v1

- Mozzherin D, Myltsev A (2017) A path to continuous reindexing of scientific names appearing in Biodiversity Heritage Library data. Biodiversity Information Science and Standards https://doi.org/10.3897/tdwgproceedings.1.20186
- Mozzherin DY, Myltsev AA, Patterson DJ (2017) “gnparser”: a powerful parser for scientific names based on Parsing Expression Grammar. BMC Bioinformatics 18: 279. https://doi.org/10.1186/s12859-017-1663-3
- Mozzherin DY, Myltsev A, Zalavadiya H (2018) Global Names Finder (GNfinder). 0.19.5. GitHub. Release date: 2022-5-10. URL: https://github.com/gnames/gnfinder
- Mozzherin DY (2022a) Global Names Finder. v0.19.5. URL: https://finder.globalnames.org
- Mozzherin DY (2022b) Global Names Parser. https://parser.globalnames.org/. Accessed on: 2022-3-23.
- Mozzherin DY (2022c) Global Names Verifier. https://verifier.globalnames.org/. Accessed on: 2022-3-23.
- Mozzherin DY (2022d) Global Names Verifier data sources. https://verifier.globalnames.org/data_sources. Accessed on: 2022-5-17.
- Mozzherin DY (2022e) Global Names Finder API Documentation. https://finder.globalnames.org/apidoc. Accessed on: 2022-5-13.
- Mozzherin DY (2022f) GNFinder Dryad Data. GitHub. Release date: 2022-3-07. URL: https://github.com/dimus/dryad-paper
- Mozzherin DY, Yoder MJ, Capitanu B, Dubnicek R (2022) Global Names and the HathiTrust: Towards comprehensive indexing of taxon names in real time. Semantic Scholar. URL: https://www.semanticscholar.org/paper/bae4f41a629679bd695846bc3dc2e929971a6d38
- Nguyen NT, Gabud RS, Ananiadou S (2019) COPIOUS: A gold standard corpus of named entities towards extracting species occurrence from biodiversity literature. Biodivers Data J29626. https://doi.org/10.3897/BDJ.7.e29626
- Pafilis E, Frankild SP, Fanini L, Faulwetter S, Pavloudi C, Vasileiadou A, et al. (2013) The SPECIES and ORGANISMS Resources for Fast and Accurate Identification of Taxonomic Names in Text. PLoS One 8: 65390. https://doi.org/10.1371/journal.pone.0065390
- Page RD (2011) Extracting scientific articles from a large digital archive: BioStor and the Biodiversity Heritage Library. BMC Bioinformatics 12: 187. https://doi.org/10.1186/1471-2105-12-187
- Patterson D, Mozzherin D, Shorthouse DP, Thessen A (2016) Challenges with using names to link digital biodiversity information. Biodiversity Data Journal 4 https://doi.org/10.3897/BDJ.4.e8080
- Patterson DJ, Cooper J, Kirk PM, Pyle RL, Remsen DP (2010) Names are key to the big new biology. Trends Ecol Evol 25: 686-691. https://doi.org/10.1016/j.tree.2010.09.004
- Petty S, Stevenson H, Hadley S (2020) Shining more light on dark data. Sci Editor https://doi.org/10.36591/se-d-4301-7
- Poelen JH, Simons JD, Mungall CJ (2014) Global biotic interactions: An open infrastructure to share and analyze species-interaction datasets. Ecol Inform 24: 148-159. https://doi.org/10.1093/biosci/biu169

- Schriml LM, Chuvochina M, Davies N, Eloe-Fadrosh EA, Finn RD, Hugenholtz P, et al. (2020) COVID-19 pandemic reveals the peril of ignoring metadata standards. Sci Data 7: 188. https://doi.org/10.1038/s41597-020-0524-5
- Shefchek KA, Harris NL, Gargano M, Matentzoglu N, Unni D, Brush M, et al. (2020) The Monarch Initiative in 2019: an integrative data and analytic platform connecting phenotypes to genotypes across species. Nucleic Acids Res 48: 704-715. https://doi.org/10.1093/nar/gkz997
- Shin N, Shibata H, Osawa T, Yamakita T, Nakamura M, Kenta T (2020) Toward more data publication of long-term ecological observations. Ecol Res 35: 700-707. https://doi.org/10.1111/1440-1703.12115
- Telenius A (2011) Biodiversity information goes public: GBIF at your service. Nord J Bot 29: 378-381. https://doi.org/10.1111/j.1756-1051.2011.01167.x
- Thessen AE, Patterson DJ (2011) Data issues in the life sciences. Zookeys 150 https://doi.org/10.3897/zookeys.150.1766
- Thessen AE, Cui H, Mozzherin D (2012a) Applications of natural language processing in biodiversity science. Adv Bioinformatics 2012: 391574. https://doi.org/10.1155/2012/391574
- Thessen AE, Patterson DJ, Murray SA (2012b) The Taxonomic Significance of Species That Have Only Been Observed Once: The Genus *Gymnodinium* (Dinoflagellata) as an Example. PLoS One 7 https://doi.org/10.1371/journal.pone.0044015
- Thessen AE, Parr CS (2014) Knowledge extraction and semantic annotation of text from the encyclopedia of life. PLoS One 9 https://doi.org/10.1371/journal.pone.0089550
- Thessen AE (2022) GNFinder Performance Calculation. GitHub. Release date: 2022-5-05. URL: https://github.com/diatomsRcool/gnrd_performance_test
- Vision T (2010) The Dryad Digital Repository: Published evolutionary data as part of the greater data ecosystem. Nature Precedings https://doi.org/10.1038/npre.2010.4595.1
- Walls RL, Deck J, Guralnick R, Baskauf S, Beaman R, Blum S, et al. (2014) Semantics in support of biodiversity knowledge discovery: an introduction to the biological collections ontology and related ontologies. PLoS One 9: 89606. https://doi.org/10.1371/journal.pone.0089606
- Wei Q, Heidorn PB, C. F (2010) Name matters: Taxonomic Name Recognition (TNR) in Biodiversity Heritage Library (BHL). http://www.ideals.illinois.edu/handle/2142/14919. Accessed on: 2022-5-19.
- Wieczorek J, Bloom D, Guralnick R, Blum S, Doring M, Giovanni R, et al. (2012) Darwin Core: An evolving community-developed biodiversity data standard. PLoS One 7: 29715. https://doi.org/10.1371/journal.pone.0029715

Endnotes

*1 =true positives / (true positives + false positives

*2 =true positives / (true positives + false negatives)

*3 =2 * (precision * recall)/(precision + recall

```
names:
 [{cardinality: 2,
   name: "Canis familiaris",
   oddsLog10: 11.5580405834288852,
   oddsDetails:
      [{feature: "spDect: inSpecies",
        odds: 8904.045433955427
        },
        {
        feature: "uniDict: inGenus",
        odds: 2976.794090112943
        },
        {
        feature: "spEnd3: ris",
        odds: 588.1893458959955
        },
        {
        feature: "uniEnd3: nis",
        odds: 114.12629099474542
        },
        {
        feature: "spLen: 10",
        odds: 9.015350833814008
        },
        {
        feature: "abbr: false",
        odds: 0.8732848865715452
        },
        {
        feature: "uniLen: 5",
        odds: 0.2580257137608618
        },
        {
        feature: "priorOdds: true",
        odds: 0.1
        }
      ],
      start: 0,
      end: 16
 }]
```


Figure 1.

Example output from GNFinder showing the results of the heuristic and statistical rules used by the naive Bayes algorithm to calculate the final score. In this example GNFinder identified the name *Canis familiaris* with high odds of being a taxonomic name based on the following criteria. This name is in two separate “go” lists (A and B). Both the genus and the specific epithet have endings that are common in Latin (C and D). The length of the specific epithet and the genus are within expected values (E and G). The name is not an abbreviation (F). All of these features were used by a naive Bayes algorithm to calculate the final “odds” score, in this case, 11.56. The Bayesian prior was set at 0.1 (H).

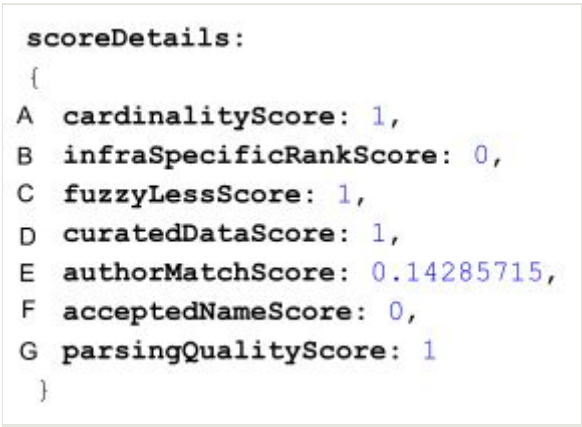

Figure 2.

Example GNVerifier score matching "*Canis familiaris*" found name string to *Canis lupus familiaris* Linnaeus, 1758 in the Catalogue of Life. The final score (G) is calculated based on the following seven attributes and used to sort results: A) Are the names uninomials, binomials, or trinomials? B) Do the names share an infraspecific rank, such as variety or form? C) Do the names match exactly? D) How carefully curated is the source of the matched name? E) Does the author and year information match? F) Is the found name a synonym of the matched name?

Table 1.

GNFinder Performance Metrics.

	All Manuscripts
Precision	0.91
Recall	0.82
F1 Score	0.86
False Positives	758
False Negatives	1801
Total Words	1589065
Total Name Strings	9753

Table 2.

Name-Finding Algorithm Performance Metrics.

Tool	Precision	Recall	F1 Score
Taxon Finder	0.930	0.827	0.875
NetiNeti	0.903	0.803	0.850
GNFinder	0.917	0.838	**0.876**
LINNAEUS	**0.981**	0.166	0.284
TaxoNERD	0.731	**0.879**	0.798
Quaesitor	0.466	0.428	0.446

Table 3.

Location of Taxonomic Names in Data Packages.

Total Number of Taxa	Taxa in manuscript (%)	Taxa in data files only (%)	Taxa as vernaculars only (%)	Taxa as irregular names only (%)
116	98.3	1.72	6.9	0.0
27	100.0	0.0	18.5	0.0
137	99.3	0.7	0.7	0.0
10	100.0	0.0	50.0	0.0
37	100.0	0.0	0.0	0.0
49	100.0	0.0	18.4	2.0
26	100.0	0.0	3.8	3.6
18	100.0	0.0	55.6	0.0
36	100.0	0.0	5.0	0.0
19	100.0	0.0	68.4	0.0
127	21.3	78.7	5.5	0.0
18	100.0	0.0	22.2	0.0
37	100.0	0.0	18.9	0.0
56	100.0	0.0	8.9	8.9
43	100.0	0.0	14.0	2.3
12	100.0	0.0	91.7	0.0
5	100.0	0.0	100.0	0.0