Using standard keywords in publications to facilitate updates of new fungal taxonomic names

Conrad L. Schoch[1], M. Catherine Aime[2], Wilhelm de Beer[3], Pedro W. Crous[4], Kevin D. Hyde[5], Lyubomir Penev[6], Keith A. Seifert[7], Marc Stadler[8], Ning Zhang[9], and Andrew N. Miller[10]

[1] National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 45 Center Drive, Bethesda, MD 20892, USA; corresponding author e-mail: schoch2@ncbi.nlm.nih.gov

[2] Department of Botany & Plant Pathology, Purdue University, West Lafayette, IN 47907, USA

[3] Department of Microbiology and Plant Pathology, Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, Pretoria, South Africa

[4] Westerdijk Fungal Biodiversity Institute, P.O. Box 85167, 3508 AD Utrecht, The Netherlands

[5] Center of Excellence in Fungal Research, Mae Fah Luang University, Chiang Rai 57100, Thailand

[6] Institute for Biodiversity and Ecosystem Research, Sofia, Bulgaria

[7] Ottawa Research and Development Centre, Biodiversity (Mycology and Microbiology), Agriculture and Agri-Food Canada, 960 Carling Avenue, Ottawa, Ontario K1A 0C6, Canada

[8] Department Microbial Drugs, Helmholtz Centre for Infection Research GmbH (HZI), Braunschweig, Germany

[9] Department of Plant Biology, Rutgers University, New Brunswick, NJ 08901, USA

[10] Illinois Natural History Survey, University of Illinois, Champaign, IL 61820, USA

Abstract: The combination of manual curation and the reliance on updates from submitters to the public sequence databases is currently inefficient and impedes the comprehensive and timely release of records with new taxonomic names. This should be improved by making several steps during data release more efficient. This article focuses on one such step by proposing a standard way for publications to flag papers with novel taxonomic information. As a result, the potential for automated searches of publication aggregators are improved, as well as the accurate curation of taxonomic information.

Key words: Data curation, NCBI Taxonomy, Novel taxa, Publication standards.

An abundance of online sources and databases are available to the modern mycologist. Nomenclature and taxonomic information is easily accessed *via* resources such as MycoBank and Index Fungorum, in addition to numerous smaller databases focused on functional or taxonomic groups (Yahr *et al*. 2016). Computational analyses can be performed by accessing molecular data stored by the International Sequence Database Consortium (INSDC) that includes GenBank at the National Center for Biotechnology Information (NCBI), DNA database for Japan (DDBJ) and the European Nucleotide Archive (ENA). However, the accuracy of taxonomic names associated with these records has remained a concern for many biologists, and mycologists have long been vocal in this regard (Nilsson *et al.* 2006, Bidartondo *et al*. 2008). Because the NCBI Taxonomy database acts as a central organizing hub of the databases of the INSDC it fulfils a crucial role in taxonomic labelling of sequence records (Federhen 2012). Recent improvements such as the ability to track type material, ability to link out to third party databases with additional information (Federhen 2015) and the addition of curated markers (Schoch *et al*. 2014) have been steps toward resolving this problem. Additionally, the Biocollection Database at NCBI (https://www.ncbi.nlm.nih.gov/biocollections/) now provides the ability to structure submissions to account for standardized repository information. However, many molecular sequence submissions to the public databases still have inadequate and incomplete metadata relating to samples (isolate, strain, bio material, culture collection or specimen voucher). The result is additional complications for any attempts at improved curation.

The current system of NCBI curation requires submitters to be diligent in updating their information after publication. When a provisional name is evident it is added with a temporary label and an "unpublished name" property in the NCBI Taxonomy database – meaning that it will not be displayed publicly but its associated records can be found in a direct text search. The eventual release of these names upon valid publication are primarily the responsibility of the author. Authors are therefore asked to update the database when their sequence data is published and their newly proposed taxonomic names validated. However, this step is often neglected. Taxonomic curators do make updates independently but this is inefficient and a comprehensive scanning of all taxonomic literature is still impossible. This increasing problem is illustrated in Fig. 1. All unpublished names in the NCBI Taxonomy since 2009 are indicated (when the unpublished name type was first introduced for provisional names). Over recent years, there appears to be an acceleration in the number of all unreleased taxonomic names. This effect is mirrored in fungal names with a small increase in the percentage from 2009 (25–30 %).

It remains a concern that despite having multiple electronic resources available, several new names that were proposed with sequence data still elude the public domain. This is not unique to Fungi (Uetz & Garg 2017), but with judicious use of several important sources of available information the process of the public release and propagation of new fungal taxonomic names can be much improved. This can be done by improving information along all the required author submission steps for publication and propagation, to whit: taxonomic name registration, public sequence deposit and publication. Firstly, we urge submitters to revisit their data after publication to ensure the publication details are updated correctly. Secondly, information sharing between databases should be

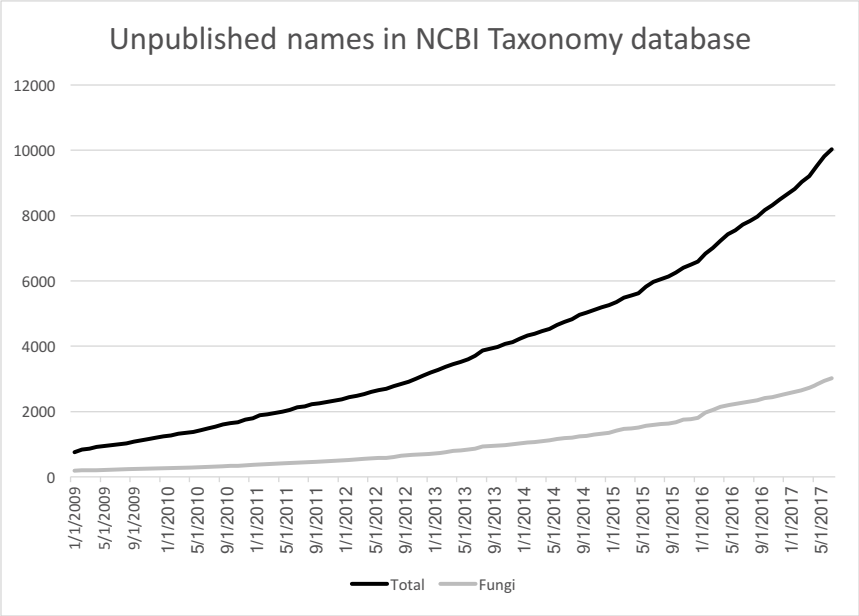

Fig. 1. Increase in unpublished names in NCBI Taxonomy for each month from first application of this name type over the period 1 January 2009 to 1 June 2017. The total number of all taxonomic names and fungal names only are indicated.

improved. This includes the synchronization between the name registries as well as the use of third party managed links, such as LinkOut in the NCBI databases (Federhen 2012). Finally, newly published taxonomic information should be more easily discoverable, to lower the burden of manual curation.

The 25 journals with the most taxonomy identification numbers (taxids) added to INSDC over the last five years are indicated in Table 1. These taxids are used to track unique names in the NCBI Taxonomy database and are also used by INSDC partners and other databases. In the second column, the total number of new species described in each journal since 2012 is indicated. The disparity between these numbers indicate that either a species was not annotated correctly in NCBI Taxonomy, or no sequence data was associated with it. This set of 25 journals represents 88 % of all deposited taxids from newly described species and 80% of all new species published over the last five years. If one only considers journals with a PubMed presence the equivalent numbers are 34 % and 55 %. Another important number in this table is the species still tagged with an unpublished name type which represents 6 % of the total published names over this period. While these unpublished names can now be verified and publicized as part of the NCBI taxonomic curation, an unknown number of species were not flagged with unpublished name types before publication.

Many fungal journals already provide extensive information on new taxa in the abstract either as text or under specific heading, such as "Taxonomic novelties". Often new names are documented in the title of a paper. This is commendable and very effective. However, this is not done consistently across journals. In order to expand the ability to detect taxonomic novelties automatically and reliably across many different journals we propose a simplified, agreed upon standard keyword that could clearly be used to flag fungal publications with novel taxonomic information. This will only improve taxonomic curation of the public sequence databases and provide benefits to third party users and external resources. This standard should be improved and expanded as the need arises.

This simple proposal arose out of discussions in a working group set up by the International Commission on the Taxonomy of Fungi (ICTF; http://www.fungaltaxonomy.org/) and the contributors to this paper include several current or former editors of major fungal taxonomic journals. A small set of keywords are proposed that can be flagged in PubMed and other aggregators of literature:
x new taxa *(With x denoting the number of taxa as a digit and "taxa" including all taxa: species, genera, etc.) OR in the case of one new taxon:* **1 new taxon.**

In the case of additional typifications, not related to new species: **x new typifications / 1 new typification.**

The keywords can be associated with optional footnotes (e.g. new taxa[1]). In the footnote, all the actual new taxa can be spelled out separated by semi colons. This can provide a way to remain within word limitations in an abstract; and with a single keyword instead of multiples that indicate rank, key word limits would not be squandered.
Example:
Key words: 3 new taxa[1], 1 new typification[2]
[1] *Exemplum* gen. nov.; *Exemplum secundum* sp. nov.; *Exemplum unum* comb. nov.
[2] Epitype proposed for *Exemplum unum*

The participating journals commit to add the following text on a visible place in their Instructions to Author guidelines: **"This journal requires that, in case the manuscript contains descriptions of new taxa of any taxonomic rank, to put a keyword "X new taxa" where the X is a digit indicating the number of new taxa in the manuscript. In the case of a single taxon it should read "1 new taxon". We strongly recommend to add the list of new taxon names according to journal specifications. This feature will help timely recording of your new taxa in INSDC and other relevant aggregators of taxonomic information". Where additional typifications (lectotypes, neotypes, epitypes etc.) are proposed we propose listing those under a separate keyword, "X new typifications" under the same conditions as above.**

In addition, participating journals are urged to promote this new feature to their users and to introduce the respective changes in their routine editorial policies and workflows by January 2018.

CONCLUSIONS

The challenge discussed here exists within the larger context of improving the timely public release of all published data. Recently the development of Wide-Open, a programmatic approach using text mining to detect published but unreleased data was described (Grechkin *et al.* 2017). The first run of this approach focused on records in the Gene Expression Omnibus (GEO) repository and the Sequence Read Archive (SRA) at NCBI. The process of scanning PubMed articles for unique identifiers related to these resources found several

Table 1. The top 25 mycological journals publishing new species ranked by newly assigned NCBI taxids for the five years 1 January 2012 to 1 June 2017.

Publication	Species with new taxids	Total new species published	Unpublished names in NCBI Taxonomy	h5-index (GOOGLE)	h5-median (GOOGLE)	PubMed ID (PMID)	PubMed Status
Studies in Mycology	1301	1679	29	N/A	N/A	8411984	Indexed
Persoonia	1236	1662	54	24	35	19540520R	Indexed
Fungal Diversity	1126	1853	175	45	61	100955518	Selected citations only
Mycologia	841	1263	126	32	44	400764	Indexed
Mycological Progress	632	999	104	20	23	101136371	Selected citations only
Phytotaxa	417	1077	84	19	28	101517955	Not currently indexed
Mycotaxon	307	1032	44	17	27	9876348	Not currently indexed
Fungal Biology	268	356	30	28	37	101524465	Indexed
The Lichenologist	234	938	45	18	24	100955368	Selected citations only
IMA Fungus	206	671	14	19	37	101557546	Indexed
Nordic Journal of Botany	171	227	2	14	25	9886922	Not currently indexed
Mycoscience	159	309	38	17	22	9890476	Selected citations only
Index Fungorum	156	359	4	N/A	N/A	101615729	Not currently indexed
Cryptogamie Mycologie	149	304	14	12	17	100961513	Not currently indexed
IJSEM	129	208	12	40	58	100899600	Indexed
Mycosphere	120	429	21	13	21	101534483	Not currently indexed
Antonie van Leeuwenhoek	116	170	10	29	41	372625	Indexed
PloS ONE	106	183	2	166	215	101285081	Indexed
Ferrantia	104	326	104	N/A	N/A	N/A	Not currently indexed
Acta Botanica Hungarica	85	250	1	9	13	101582241	Not currently indexed
Sydowia	84	186	5	8	9	100955200	Not currently indexed
Nova Hedwigia	72	250	20	14	23	101317353	Not currently indexed
CBS Biodiversity Series	63	266	2	N/A	N/A	N/A	Not currently indexed
The Bryologist	55	189	4	13	14	100955480	Selected citations only
Mycokeys	48	140	13	N/A	N/A	101569696	Indexed

overdue datasets that could be released. This elicited a positive response from curators at NCBI and resulted in the accelerated release of several records (Williams 2017), but this can and should be expanded to other data sets. Another promising new development in this direction is automated workflows for text mining of taxonomic journals providing alert services on new taxa and other semantically recognizable sub-article elements (e.g. taxon treatments, images, occurrence records, identification keys) on the day of publication. Such a service is currently being developed through the RDF-based Open Biodiversity Knowledge Management System (OpenBiodiv; Senderov *et al.* 2017). This system will provide automated machine-readable information in RDF to be harvested by aggregators such as GBIF, Catalogue of Life, NCBI, and others.

We focused on a very particular problem in this paper: the validation of unpublished names attached to the newly released sequence data. The curation process continues to depend on inefficient workflows that requires submitters to provide updates after publication. While improvements can and will be made, important opportunities exist to utilize information within abstracts in PubMed and other database aggregators. Fungal biology has become reliant on a public collection of nucleotide sequence data to compare and improve the cataloguing of diversity. This simple proposal, if widely applied, can significantly improve the synchronization of the release of new taxonomic data in publications and public databases, aiding discovery and data integration in biology

ACKNOWLEDGEMENTS

This paper resulted from discussions in Group 5 under the auspices of the International Commission on the Taxonomy of Fungi (ICTF). The Intramural Research Program of National Institutes of Health, National Library of Medicine is also acknowledged. We are grateful to Stacy Ciufo (NCBI) for her help in preparing the figure and table.

REFERENCES

Bidartondo MI, Bruns TD, Blackwell, M, Edwards I, Taylor AFS, *et al.* (2008) Preserving accuracy in GenBank. *Science* **319**: 1616.

Federhen S (2012) The NCBI Taxonomy database. *Nucleic Acids Research* **40**: D136–D143.

Federhen S (2015) Type material in the NCBI Taxonomy Database. *Nucleic Acids Research* **43**: D1086-D1098.

Grechkin M, Poon H, Howe B (2017) Wide-Open: accelerating public data release by automating detection of overdue datasets. *PLoS Biology* **15**: e2002477.

Nilsson RH, Ryberg M, Kristiansson E, Abarenkov K, Larsson KH, *et al.* (2006) Taxonomic reliability of DNA sequences in public sequence databases: a fungal perspective. *PLoS ONE* **1**: e59.

Schoch CL, Robbertse B, Robert V, Vu D, Cardinali G, *et al.* (2014) Finding needles in haystacks: linking scientific names, reference specimens and molecular data for Fungi. *Database*: bau061.

Senderov V, Georgiev T, Agosti D, Catapano T, Sautter G, *et al.* (2017) OpenBiodiv: an implementaion of a semantic system running on top of the biodiversity knowledge graph. *Proceedings of TDWG* **1**: e20084.

Uetz P, Garg A (2017) Molecular taxonomy: species disconnected from DNA sequences. *Nature* **545**: 412.

Williams R (2017) Making public data public. *The Scientist Magazine* **June 8** https://www.the-scientist.com/?articles.view/articleNo/49629/title/Making-Public-Data-Public/.

Yahr R, Schoch CL, Dentinger BT (2016) Scaling up discovery of hidden diversity in fungi: impacts of barcoding approaches. *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences* **371**: 20150336.