Biodiversity Data Journal 11: e107914
doi: 10.3897/BDJ.11.e107914

Research Article

Ten years and a million links: building a global taxonomic library connecting persistent identifiers for names, publications and people

Roderic Page ‡

‡ University of Glasgow, Glasgow, United Kingdom

Corresponding author: Roderic Page (roderic.page@glasgow.ac.uk)

Academic editor: Lyubomir Penev

Received: 13 Jun 2023 | Accepted: 01 Sep 2023 | Published: 14 Sep 2023

Citation: Page R (2023) Ten years and a million links: building a global taxonomic library connecting persistent identifiers for names, publications and people. Biodiversity Data Journal 11: e107914. https://doi.org/10.3897/BDJ.11.e107914

Abstract

A major gap in the biodiversity knowledge graph is a connection between taxonomic names and the taxonomic literature. While both names and publications often have persistent identifiers (PIDs), such as Life Science Identifiers (LSIDs) or Digital Object Identifiers (DOIs), LSIDs for names are rarely linked to DOIs for publications. This article describes efforts to make those connections across three large taxonomic databases: Index Fungorum, International Plant Names Index (IPNI) and the Index of Organism Names (ION). Over a million names have been matched to DOIs or other persistent identifiers for taxonomic publications. This represents approximately 36% of names for which publication data are available. The mappings between LSIDs and publication PIDs are made available through ChecklistBank. Applications of this mapping are discussed, including a web app to locate the citation of a taxonomic name and a knowledge graph that uses data on researcher ORCID ids to connect taxonomic names and publications to authors of those names.

Keywords

persistent identifiers, linked data, biodiversity knowledge graph, taxonomic databases, DOI, LSID, ORCID

Introduction

One thing the field of biodiversity informatics has been very good at is creating databases. However, this success in database creation has not been matched by an equivalent success in creating deep links between those databases (Thomas 2009). Instead, we create an ever-growing number of silos. An obvious route to “silo-breaking” is the shared use of the same persistent identifiers for the same entities across those databases. For example, rather than mint its own identifier for a publication, a database could reuse the existing Digital Object Identifier (DOI) for that publication. This seemingly trivial step of reusing someone else’s identifier opens up numerous possibilities for interconnection, but comes with some risk: what if that persistent identifier does not, in fact, persist? If we cannot trust that an identifier will continue to be maintained and resolve as we expect, then anything we ourselves build upon that identifier is likely to break. Cross-linkages between databases are more likely to be made between databases that make efforts to maintain their identifiers (Shorthouse 2020).

DOIs are a well known example of a persistent identifier, widely used to identify academic publications and other digital items, including datasets. They have been adopted by publishers, who routinely include DOIs for articles from other publishers in the lists of literature cited in their own publications. Embedding these identifiers in PDFs that are intended to be long-lived versions of records requires a significant degree of trust. In particular, the publishers trust that the persistence of these identifiers will be longer than the typical decade lifespan for web links (Hennessey and Ge 2013). This persistence, coupled with tools to retrieve machine-readable metadata for items with DOIs has led to an ecosystem of services that depend on (or make use of) DOIs, including the citation graph (Peroni and Shotton 2020), measures of attention (e.g. https://www.altmetric.com), populating bibliographies for researchers and machine-learning tools to summarise and interpret article content (Nicholson et al. 2021).

DOIs have gained wide acceptance as identifiers of digital publications and data and have also been adopted for bacterial taxa and their names (Garrity and Lyons 2003) and for species hypotheses for fungi (Nilsson et al. 2019). However, the bulk of the taxonomic community went a different route and adopted Life Science Identifiers (LSIDs) (Clark et al. 2004, Martin et al. 2005) for taxonomic names. These identifiers were attractive for several reasons: they were developed within the life science community, natively supported the Resource Description Format (RDF) and were free. Core taxonomic databases, such as Index Fungorum, International Plant Names Index (IPNI) and the Index of Organism Names (ION) all supported LSIDs, including their novel resolution mechanism. In subsequent years, the ability and/or willingness of databases to support LSIDs has declined until few now do so natively (but work-arounds such as HTTP resolution are still

feasible, for example https://lsid.io). Despite this, LSIDs are still being embedded in taxonomic publications as part of pipelines to register new taxonomic names (Penev et al. 2016).

An additional problem has been the lack of a single, definitive identifier for the same taxonomic name. New plant names typically have LSIDs issued by IPNI. Due to its origins as a combination of three different databases (Croft et al. 1999), IPNI has duplicate names and, hence, often has multiple LSIDs for the same name. Fungal names may have LSIDs issued by Index Fungorum and URLs issued by MycoBank (Robert et al. 2013). These identifiers share the same local identifier (an integer) and so can be regarded as interchangeable.

In zoology, the situation is more complex. Registration of new names is managed by ZooBank, which mints LSIDs for new names and also has LSIDs for some older names. However, the 326,000 records currently in ZooBank represent a small fraction of described animal species. For example, ION has over 5 million names, each with a LSID. Clustering the ION names for duplicates (Page 2013) reduces the total to approximately 4.3 million, still considerably more than in ZooBank or in any other zoological name aggregator. The existence of multiple identifiers for the same name complicates attempts to cross-link databases because it is not obvious which taxonomic name identifier to use. In the absence of a synthesis of these identifiers by the taxonomic community, we may have to rely on third-party identity brokers such as Wikidata (Veen 2019) to manage cross-links between the menagerie of zoological databases.

Linked data needs links

The less than satisfactory history of persistent identifiers for taxonomic names may suggest that the problem was the choice of identifier (i.e. LSID rather than, say, DOI). However, this overlooks the deeper problem that, as implemented, LSIDs offered little of value beyond their persistence. Resolving an LSID typically returns RDF with no external links, that is, no identifiers beyond ones local to the LSID provider. We had, in effect, created yet another data silo, ironically using the data format that was supposed to be a silo-breaker.

Currently, RDF is enjoying something of a renaissance, especially when serialised as JavaScript Object Notation for Linked Data (JSON-LD) which is more readable and developer-friendly than formats such as RDF XML. A growing number of websites relevant to biodiversity are embedding JSON-LD in their pages (see list at https://github.com/rdmpage/wild-json-ld/), including prominent databases, such as the Catalogue of Life (https://www.catalogueoflife.org). Yet, many of these linked data-enhanced web sites are still silos. For example, the JSON-LD for *Cordyceps changchunensis* from the Catalogue of Life shown in Fig. 1 lacks external identifiers for either the taxonomic name *Cordyceps changchunensis* or the publication that includes that name. These identifiers exist (urn:lsid:indexfungorum.org:names:839249 and https://doi.org/10.3897/mycokeys.83.72325, respectively). Including them (Fig. 2) converts a data silo into a record connected to

two other data sources and through those sources potentially connected to an even wider network of information.


```
{
    "@context": "https://schema.org/",
    "@id": "https://www.catalogueoflife.org/data/taxon/B2MC3",
    "name": "Cordyceps changchunensis J.J. Hu, Bo Zhang &amp;amp; Y.
Li",
    "scientificName": {
        "@type": "TaxonName",
        "name": "Cordyceps changchunensis",
        "author": "J.J. Hu, Bo Zhang &amp;amp; Y. Li",
        "taxonRank": "Species",
        "isBasedOn": {
            "@type": "ScholarlyArticle",
            "name": "(2021). In Hu, Dai, Zhao, Guo, Tuo, Rao, Qi,
Zhang, Li &amp;amp; Zhang, IMA Fungus 83: 166."
        }
    }
}
```


Figure 1. doi

Simplified JSON-LD for Catalogue of Life taxon B2MC3, retrieved 16 May 2023. Note the lack of an identifier for either the scientific name or the publication that name is based on.


```
{
    "@context": "https://schema.org/",
    "@id": "https://www.catalogueoflife.org/data/taxon/B2MC3",
    "name": "Cordyceps changchunensis J.J. Hu, Bo Zhang &amp;amp; Y.
Li",
    "scientificName": {
        "@id": "urn:lsid:indexfungorum.org:names:839249",
        "@type": "TaxonName",
        "name": "Cordyceps changchunensis",
        "author": "J.J. Hu, Bo Zhang &amp;amp; Y. Li",
        "taxonRank": "Species",
        "isBasedOn": {
            "@id": "https://doi.org/10.3897/mycokeys.83.72325",
            "@type": "ScholarlyArticle",
            "name": "(2021). In Hu, Dai, Zhao, Guo, Tuo, Rao, Qi,
Zhang, Li &amp;amp; Zhang, IMA Fungus 83: 166."
        }
    }
}
```


Figure 2. doi

The JSON-LD shown in Fig. 1 enhanced by including persistent identifiers for the taxon name (LSID) and its publication (DOI) (highlighted in **bold**).

Putting holes in silos

The goal of the work described here is to make a small hole in taxonomic data silos by linking LSIDs for taxonomic names to DOIs for the works that published those names.

Other bibliographic identifiers are also available and relevant, but the focus in this work will be on DOIs. Imagine that we present an entity, such as a taxonomic name or a publication as a plastic ball and, on that ball, we place strips of velcro, one for each identifier. Imagine that each velcro strip will only connect to the same identifier on another ball. If we think of linked data as a set of balls with velcro strips, then whether those balls stick together depends on how often an identifier is used. I refer to this as the "stickiness" of an identifier. DOIs are a relatively “sticky” identifier often connected to other identifiers, most notably ORCID ids for researchers (Bohannon and Doran 2017). Another reason is the role DOIs play in creating the citation graph, the scholarly network linking works to the works that they either cite or are cited by (Peroni and Shotton 2020). DOIs also make it easier to cite the taxonomic literature. Taxonomists frequently complain about the lack of citations their work receives. Whatever the merits of that complaint, calls for better citation practices (Benichou et al. 2022) are unlikely to improve the situation if the taxonomic literature remains disconnected from taxonomic names. How are we to know what publications should be cited for a name if the links between names and literature are hard to discover?

Storing the mapping

In addition to the challenge of creating these mappings, there is the problem of how to make them available for reuse. Ideally, the source taxonomic databases would incorporate them, on the grounds that they would add value to their users and it would save those databases doing the work themselves. However, this assumes that those databases are willing or have the resources to incorporate this additional data, which rarely seems to be the case. Alternative approaches include developing separate, stand-alone web sites to make the data available or simply putting a data dump in a repository.

I have experimented with various approaches. In 2013, I created a stand-alone database mapping ION LSIDs to DOIs and other identifiers and wrapped this in a user-friendly web site (https://bionames.org) developed with funding from the Encyclopaedia of Life (Page 2013). In 2018, I explored an intermediate approach of using Datasettes to publish a mapping between IPNI names and the literature (Page 2018). This made the data available and queryable, but the interface does not support taxonomic-specific queries. Both these approaches result in stand-alone web sites with little obvious means to integrate the mapping into other databases.

The recent release of ChecklistBank (https://www.checklistbank.org) (Döring et al. 2022) has provided a new way to publish the data so that they complement existing databases. ChecklistBank includes all the taxonomic checklists used to create the Catalogue of Life, as well as taxonomic treatments from Plazi (Agosti and Egloff 2009), but also enables users to upload their own checklists. This means that we can take a taxonomic checklist, add persistent identifiers for the literature, then upload the augmented data to ChecklistBank as a new dataset (with an appropriate citation to the original source database). This augmented checklist can have its own DOI and be citable (hence providing a mechanism to give credit to those making the links). As the augmented dataset uses the same taxon name identifiers as the original database, this also means that, at any point,

the original data publishers could incorporate literature mapping into their own databases. Likewise, any other database that uses those same taxon name identifiers could also use the mapping.

ChecklistBank provides a convenient way to store mappings between names and publications, but this is a single edge in the biodiversity knowledge graph. Storing the deeper links, such as between taxonomic names, publication, people, institutions and funders requires more flexibility. To store these, I follow an approach sketched in Page (2022a) where the mappings are stored as RDF and published to Zenodo. These mappings can then be loaded into a triple store.

Goals

The goal of this work is to make available over a million links between persistent identifiers for taxonomic names and the publications for those names. This paper covers three databases: Index Fungorum (https://www.indexfungorum.org), IPNI (https://www.ipni.org) and ION (http://www.organismnames.com). Each of those uses LSIDs as persistent identifiers for taxonomic names. This gives us substantial coverage of animals, plants and fungi.

In this work, I will focus on "citable" bibliographic identifiers, that is, identifiers that are typically cited by other publications. In practical terms, this means DOIs (Fig. 3). The two advantages of work-level identifiers are that they tend to be persistent (e.g. DOIs) and they are also the basis of measures of scientific activity (e.g. citations) and attention (e.g. altmetrics).

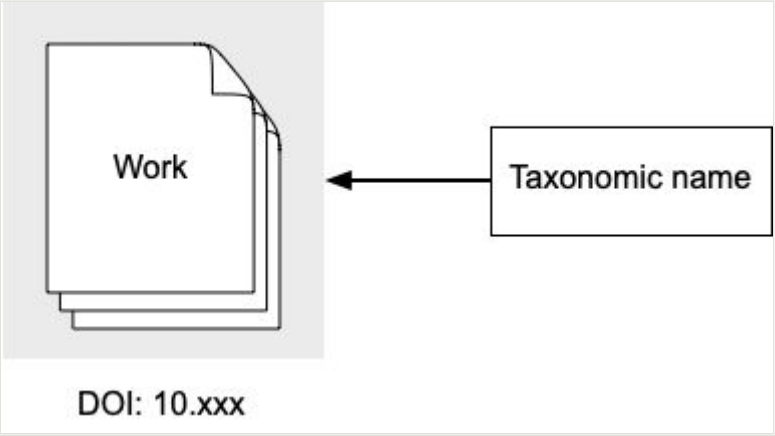


Figure 3. doi

Work-level identifiers. The taxonomic name is linked to a work-level identifier, such as the DOI for the article that published the name.

In contrast, databases such as IPNI and IF, typically store bibliographic information at the level of individual pages or sets of pages. Citations at the page level have been termed

"microcitations" (or "microreferences") and are analogous to what the U.S. legal profession refers to as "point citations" or "pincites". Some bibliographic databases support page-level identifiers. For example, individual pages in the Biodiversity Heritage Library (BHL) have their own unique URL. In cases where there is not an explicit identifier, we can use “fragment identifiers” to identify parts of an entity (Fig. 4). For instance, an individual page in a PDF can be referred to using the fragment #page=*n* where *n* is the position of the page within the PDF, starting from *n* = 1 for the first page (Taft et al. 2004). Blocks of text within a page can be identified using TextQuotes or TextPosition identifiers (Dürst and Wilde 2008). Locations with a HTML or XML document can be referred to using XPath statements. Fragment identifiers enable deep within-document linking, but can be fragile. If the document being linked to changes or has multiple versions, then fragment identifiers may no longer successfully link to the desired content (Brush et al. 2001).

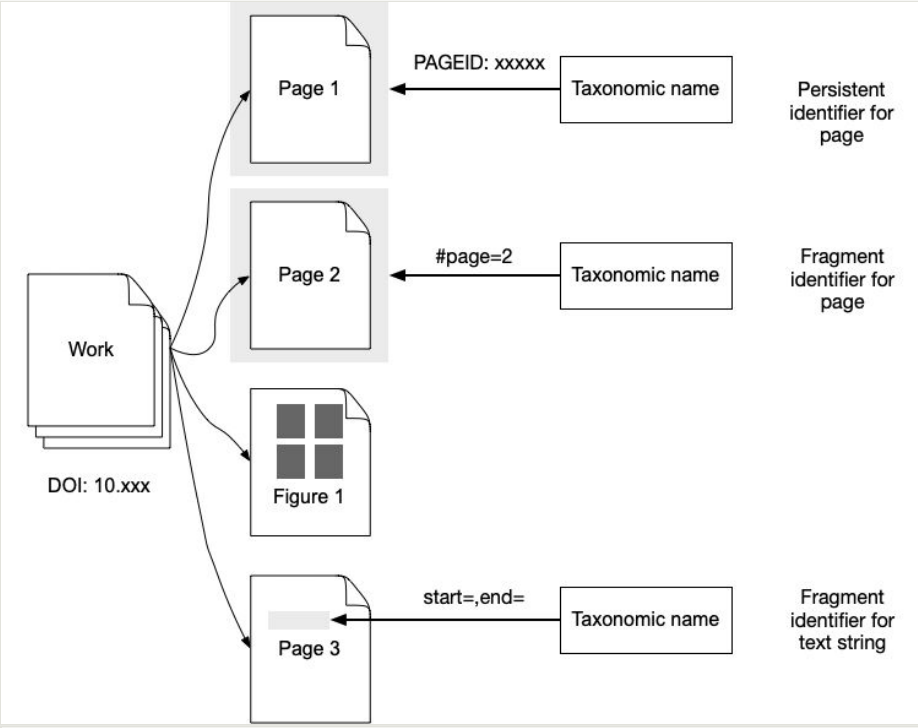


Figure 4. doi

Page and fragment-level identifiers. In contrast to work-level identifiers (Fig. 3), we can use identifiers for pages or parts of pages.

Another approach is to select one or more blocks of text and any associated figures within a publication and treat that collection as a distinct unit (Fig. 5). These can be treated as stand-alone entities or recombined to provide an alternative navigation pathway through a set of papers (Anonymous 2012). The Plazi project (Agosti and Egloff 2009) extracts blocks of text and images as “treatments” and many of these are assigned DOIs. This has

the advantage of creating citable units, although to date, there is little evidence that either taxonomic databases or publications actually cite treatments rather than the entire work.

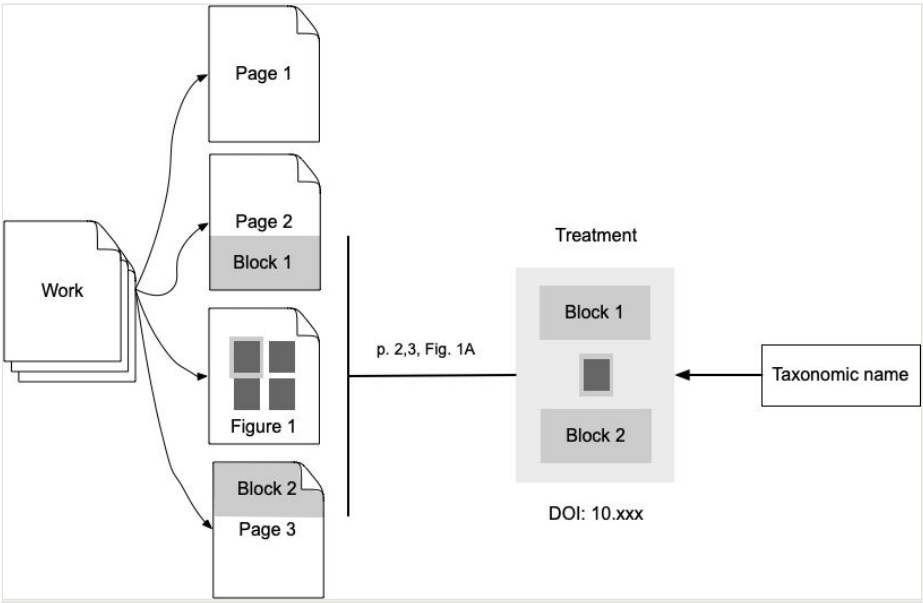


Figure 5. doi

Part of work identifiers. Taxonomic names are linked to one or more parts of a document, such as text and a figure. These parts are packaged into a citable unit, such as a treatment.

Outputs

There are three categories of output from this work. The first is a mapping between LSID for a taxon name and one or more (ideally) persistent identifiers for the publication that established that name. The preference is to use the DOI for the publication, if one is available, but other identifiers may be used, such as Handles and URLs. If the publication has an item in Wikidata, the QID of the Wikidata item is also included. These mappings are published to ChecklistBank.

The second category of output is a version of the mapping in RDF, enabling the mapping to be used in a knowledge graph. In this paper, I combine the mapping with data from ORCID sufficient to construct a simple knowledge graph of taxonomic names and the taxonomists who published those names.

The final category of output is a proof-of-concept web site that uses the mappings stored in ChecklistBank to generate citations for publications of a taxonomic name, as well as locate a PDF for that article.

Methods

Taxonomic names

Taxonomic names and citation data were obtained from Index Fungorum, IPNI and ION at various times over the last decade. For Index Fungorum and IPNI, each taxonomic name represents a nomenclatural act, such as the original description or a new combination (e.g. moving a species from one genus to another and (in almost all cases) each name is associated with the publication in which the name is published for the first time. Taxonomic revisions of plants and fungi that do not result in new names will not be recorded in Index Fungorum and IPNI unless they also result in new names.

ION is the descendant of Zoological Record and has a broader goal than just tracking acts of nomeclature. Hence, it includes duplicate names, spelling variants etc. (Page 2013). The publicly accessible data linked to LSIDs only include bibliographic data for newly-published names; hence, the data do not include publications that, for example, move species from one genus to another. Hence, ION differs from Index Fungorum and IPNI in having somewhat messier data and does not include publication data for names that are the result of taxonomic revision.

Typically, data were retrieved from the source databases by resolving LSIDs for individual names, parsing the resulting RDF into tabular form and storing these data in SQL databases for ease of manipulation. If the database no longer supports LSID resolution, tools such as https://lsid.io can be used to retrieve the RDF. On other occasions, bulk downloads have been made using APIs provided by the databases. Once in a local SQL database, the data have been cleaned, citation strings parsed, any existing bibliographic identifiers extracted, then the citation data are mapped to external bibliographic identifiers.

Mapping citations

For full citations that include data such as authors, title, journal and pagination, there are a number of approaches to mapping these citations to identifiers. These include using search engines, such as CrossRef or ReFindit (https://refindit.org/about.html). Most tools have their own unique search interface, but some support generic search interfaces, such as the Open Refine API.

Matching full citations can be treated as a simple string matching task. However, microcitations (citing a page or a part of a publication) present an additional challenge. The simplest microcitation is a single page within a publication. If we have a database of page ranges for articles (i.e. the start and end page numbers), then matching microcitations to full citations is relatively trivial: find the article in a given volume that has a page range that includes the page in the microcitation. However, given that we lack a freely-accessible database of all taxonomic publications, this can be a challenge. It also assumes that available metadata for articles include page numbers. In some cases, these numbers are not readily available, for example, for the *European Journal of Taxonomy* of 1,201 articles from 2011 - 2023, only 243 had a page range in the CrossRef metadata. Another reason

for the lack of page numbers is the move to online publication where the notion of a “page” becomes problematic. Pagination depends on how the article is rendered and may vary across different representations or be absent altogether.

To facilitate resolving microcitations, I have, for the last decade or more, been building a bibliographic database that includes pagination data. These data come from a variety of sources, such as CrossRef, PubMed, JSTOR, journal websites, article PDFs etc. Managing these data locally is essential as often the metadata available from individual sources are incomplete (e.g. lacking page numbers) and, hence, multiple sources may be required to retrieve sufficient metadata to determine the appropriate persistent identifier for a publication record in a taxonomic database. To make these data more widely available, I am uploading much of it to Wikidata, where it can be further curated and improved (Page 2022b).

Other approaches for mapping citations include using identifiers for articles, or parts thereof, which have also been incorporated into taxonomic databases. For example, the record for the taxonomic name *Neodeightonia mucosa* (urn:lsid:indexfungorum.org:names: 840943) cites “Frontiers in Microbiology, volume 12, issue no. 737541”. This corresponds to the DOI 10.3389/fmicb.2021.737541 (note the shared “737541”). Note that this is an argument against the use of “opaque identifiers” (i.e. an identifier that contains no information about the entity with that identifier). Providing one is aware that information in an identifier might be misinterpreted, non-opaque identifiers (typically based on metadata for the entity being identified) can be a useful aid to making connections between databases. This can be particularly useful in cases where a journal has moved from sequential pagination within a volume to continuous article publication, such that every article starts on page 1 (Anonymous 2014).

One unintended consequence of attempting to map citations is that it can expose errors in the taxonomic databases. A mismatch between journal and volume numbers is often a clue that a record is in error. For example, the citation for urn:lsid:indexfungorum.org:names: 839249 is “Hu, Dai, Zhao, Guo, Tuo, Rao, Qi, Zhang, Li & Zhang, IMA Fungus 83: 166 (2021)”. There is no such volume for IMA Fungus; however, the volume and page number match an article in *MycoKeys* (Hu et al. 2021). Databases inevitably benefit from scrutiny and making links between databases generates a lot of scrutiny.

Data management

The mapping between names and publications is managed in a local SQL database, either SQLite or MySQL. A range of custom scripts manages data import and cleaning and matching bibliographic citations to persistent identifiers. There are also tools to visualise progress, discover gaps and drill down by taxonomic name, publication, date etc. Each mapping project is managed in one or more GitHub repositories, which are listed in Table 2

Table 1.

Numbers of taxonomic names and persistent identifiers for publications. For each database, the Table shows the total number of taxonomic names, how many of those names have publication information and how many of those publications have been mapped to one or more persistent identifiers. The row “Any” records the number of publications that have any identifier.

	Index Fungorum	IPNI	ION	Total
Taxonomic names	507,279	1,721,566	5,309,468	7,538,313
Names with publications	444,235	1,708,187	1,729,338	3,881,760
DOI	75,009	245,846	401,351	722,206
Handle	3	1,157	35,731	36,891
JSTOR	5,578	131,401	28,146	165,125
BioStor	322	52,395	170,250	222,967
BHL	6,818	107,459	5	114,282
URL	32,064	97,396	127,041	256,501
PDF link	12,864	34,319	231,156	278,339
Wikidata	94,192	396,072	515,341	1,005,605
Any	105,886	522,601	769,956	1,398,443

Table 2.

Datasets of names mapped to persistent identifiers for the literature. Each dataset has a DOI for the data, a corresponding ChecklistBank id and the Github repository for the code used to create the mapping.

Dataset name	DOI (all versions)	ChecklistBank ID	Github Repository
Index Fungorum	https://doi.org/10.5281/zenodo.7211134	129659	https://github.com/rdmpage/index-fungorum-coldp
IPNI	https://doi.org/10.5281/zenodo.7208699	164203	https://github.com/rdmpage/ipni-coldp
BioNames (ION)	https://doi.org/10.5281/zenodo.7977714	128415	https://github.com/rdmpage/bionames-coldp

Storing the mapping in ChecklistBank

For each database, a new entry was created in ChecklistBank. A data release in the Catalogue of Life Data Package (CoLDP) format (Döring and Ower 2019) was created and uploaded to Zenodo where it received a DOI. The same data are then uploaded to ChecklistBank.

The CoLDP format requires a unique identifier for each bibliographic reference. This was generated using a trigger in the SQLite database. If the reference had a Wikidata QID, then

that value served as the identifier. In the absence of a Wikidata QID, a new identifier would be generated from one of the persistent identifiers added in the mapping, such as the DOI.

The CoLDP format expects a bibliographic citation string for each reference. LSIDs for the ION database include a citation string, but, in the case of Index Funfgorum and IPNI, complete citations are not available in the original databases as these databases use microcitations. Hence, in this case, complete citations were generated using tools based on the CSL-JSON format (Bennett 2018). Bibliographic metadata in this format were retrieved from Wikidata or via content-negotiation from https://doi.org, then formatted for display.

Outputting the mapping as RDF

In addition to the COLDP format for ChecklistBank, I created linked data files for the names using the N-Triples format. The names were modelled following the draft Bioschemas proposal for taxon names (https://bioschemas.org/TaxonName/), which is also followed by the Catalogue of Life. Rather than output the entire mapping as RDF, I included only those names that have a publication with a DOI. This is because the DOI is likely to be the only bibliographic identifier found in other datasets that we could potentially link to, such as ORCID. For names that have DOIs for their publications, the LSID for that name is linked to the DOI using the schema.org property “isBasedOn” (Fig. 6). For each of Index Fungorum, ION and IPNI, the list of N-Triples was uploaded to Zenodo.

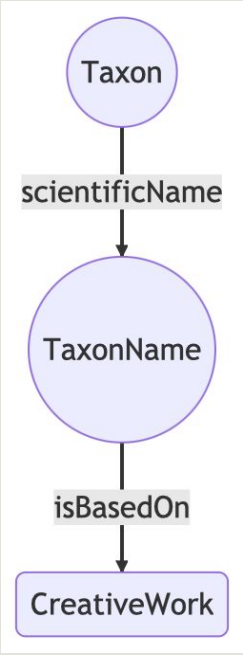


Figure 6. doi

A taxonomic name linked to the publication that makes that name available, expressed using terms from the http://schema.org vocabulary.

Augmenting with ORCIDs

A knowledge graph is only as interesting as its connections, so to augment the simple pairs of taxonomic names and publications, I created an additional RDF file connecting people and their publications. Many researchers have ORCID ids which enable those researchers to uniquely identify themselves (Bohannon and Doran 2017). The ORCID record for an individual may list their publications (and other outputs, such as datasets, peer reviews etc.), many of which have DOIs. It may also link people to other entities, such as organisations where they have studied or worked or funding agencies (Fig. 7).

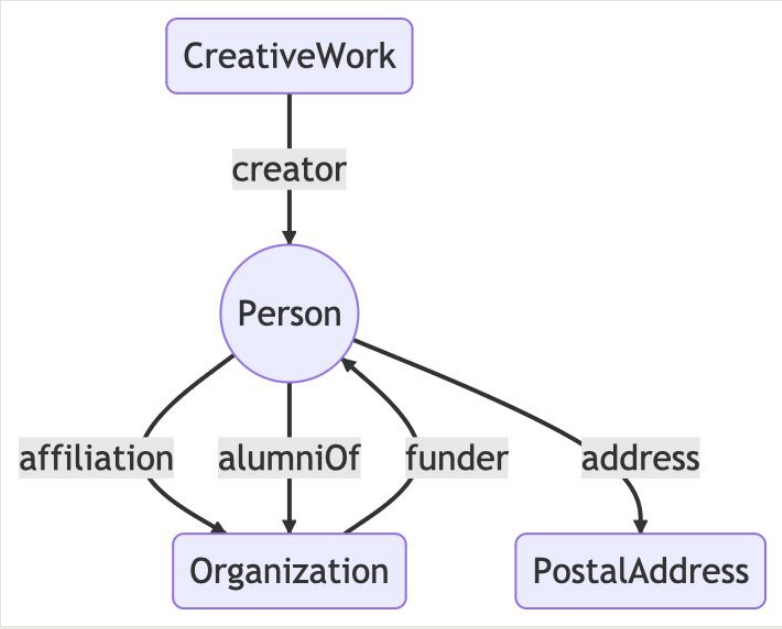


Figure 7. doi

Simplified version of the data model used by ORCID to export data in RDF.

Data in ORCID are available as linked data using content-negotiation. That is, by sending an HTTP request that accepts data in the format “application/ld+json”, ORCID will return structured data about the person with that ORCID id. I retrieved data for a set of ORCIDs associated with DOIs for papers on taxonomy using a tool (https://enchanting-bongo.glitch.me) which queries the ORCID API for ORCID ids associated with DOIs. I also retrieved ORCID ids via queries to Wikidata, for example, for authors in Wikidata that have both an ORCID id and an article on Wikispecies.

Much of the data in ORCID is user-supplied and some of it is messy. A common problem is URLs that are not properly formed. Linked data are in a very unforgiving format when it comes to URLs and these errors cause problems when uploading data to a triple store. Hence, data from ORCID data were run through a series of scripts to clean extraneous characters and eventually output clean RDF in N-Triples format, suitable for uploading into

a triple store. Combining the taxonomic names, literature and ORCIDs yields a small knowledge graph (Fig. 8).

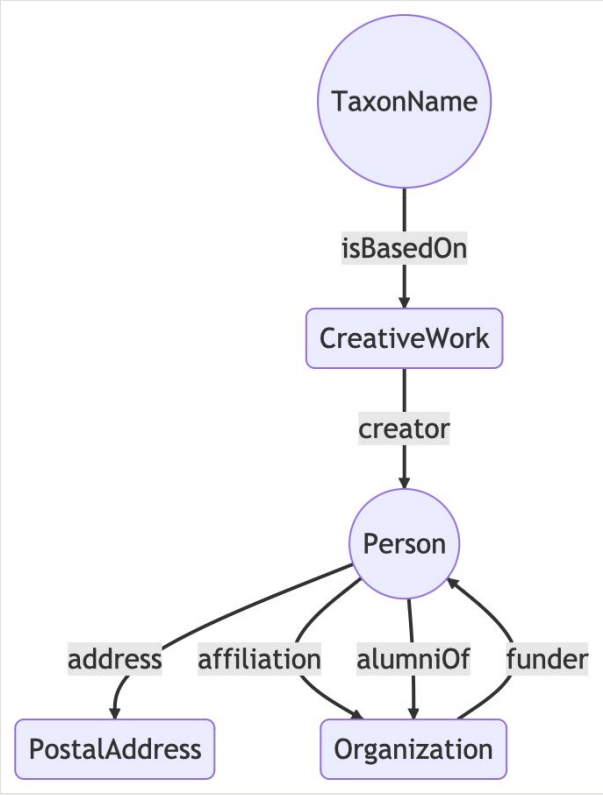


Figure 8. doi

Modelling the relationship between a taxonomic name, its publication and the author of that publication.

Results

Coverage

Table 1 gives basic data on how many names each source database has and how many have been mapped to a persistent identifier for a publication. The number of names with DOIs for the corresponding publication ranges from 75,000 to 722,000 across the three datasets. Not surprisingly, Wikidata is the single largest source of identifiers, with a little over a million names linked to a publication with a Wikidata QID.

To visualise progress on linking names to literature, I computed the number of names per decade from 1750 to 2020 that were published in the 50 publication venues or “containers” (such as journals, monographs and books) that published the most names. The

publications were then sorted by the decade in which they published the most names (their "modal" decade), which enables us to see changes in the fate of publications over time. The diagrams also plot the percentage of works that have any persistent identifier.

Both fungi (Fig. 9) and plants (Fig. 10) show similar patterns of apparent turnover in publications, with the most recent publications having the greater density of persistent identifiers. The diagram for animals (Fig. 11) is truncated relative to the other taxonomic groups with no data prior to the 1860s. This is because Zoological Record, the primary source for ION, started in 1864 (Bridson 1968) approximately a century after the start of zoological nomenclature. Many of the top animal publications are still being published and many have DOIs, which is reflected in the greater density of PIDs in Fig. 11 compared with Fig. 9 and Fig. 10.

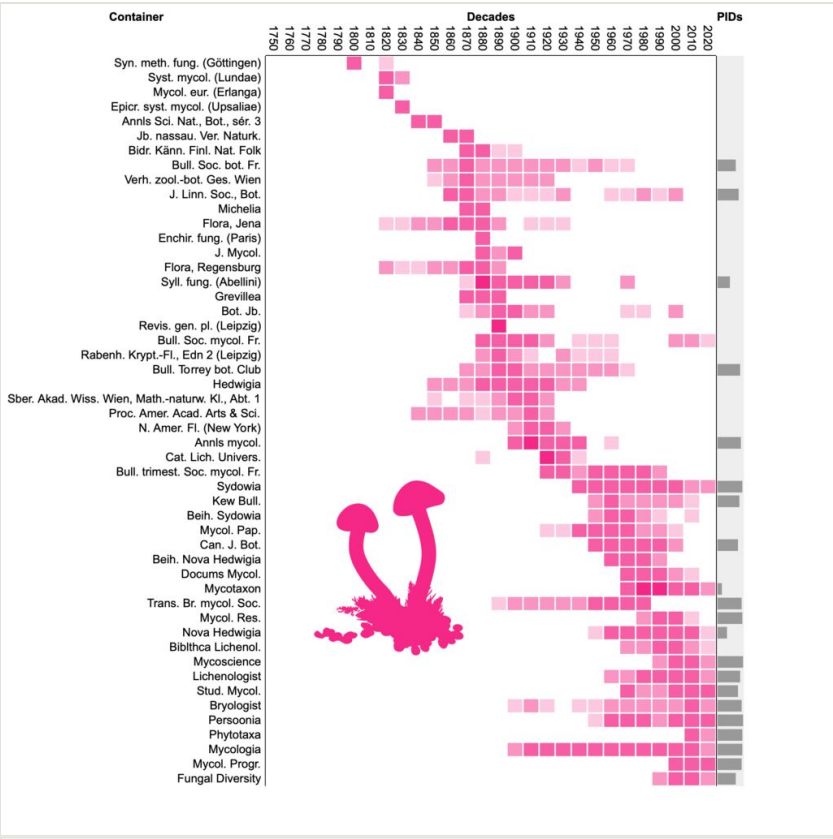


Figure 9. doi

Density distribution of taxonomic names published in the decades from 1750 to 2020 across the 50 publication venues ("containers") that published the most names for fungi. The containers are ordered by the decade with the largest number of names. The column labelled "PIDs" shows bars proportional to the percentage of names in each publication that have been linked to a persistent identifier for that publication.

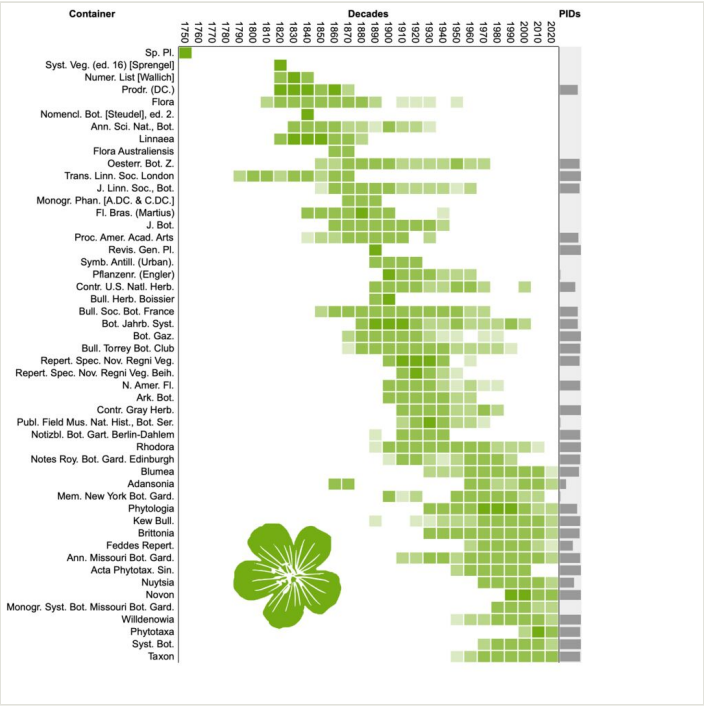


Figure 10. doi

Density distribution of taxonomic names published in the decades from 1750 to 2020 across the 50 publication venues ("containers") that published the most names for plants. See Fig. 9 for further explanation.

Further differences amongst fungal, plant and animal taxonomic publishing can be seen when we plot the total number of names published in each of the top 100 publications (Fig. 12). Animal names are dominated by the journal *Zootaxa* which has published five times as many names as the next publication (*ZooKeys*). The number of names in a publication rapidly declines, such that the top 100 zoological publications combined account for only 30% of all animal names. In contrast, fungal and plant names are dominated by older monographs and 40-50% of names are included in the top 100 publications.

For fungal and plant names, there is not one massively dominant publication and the publications with the most names tend to be old monographic series from the 18th and 19th centuries. Animal taxonomy shows a rather different pattern, with a single journal *Zootaxa* publishing many more names than any other. In contrast to plant and fungi names, many of the top venues for publishing taxonomic names are still currently active journals.

ChecklistBank and demo application

Datasets for Index Fungorum, IPNI and ION have been published to Zenodo and ChecklistBank (Table 2). These datasets are in COLDP format and comprise the subset of names from each data source that have been mapped to one or more persistent identifiers.

Each of these datasets can be queried using the ChecklistBank interface or via the ChecklistBank API. As a proof of concept of what can be done with the mappings, I created a web site called “Species Cite” https://species-cite.herokuapp.com which takes a user-supplied taxonomic name and queries the Index Funforum, IPNI and ION datasets in ChecklistBank for a persistent identifier associated with that name. If it finds either a DOI or a Wikidata item identifier, it displays those, along with a formatted citation of the paper that published the name. It also endeavours to find a PDF of that publication on the web so that the user can read more about that taxon (Fig. 13).

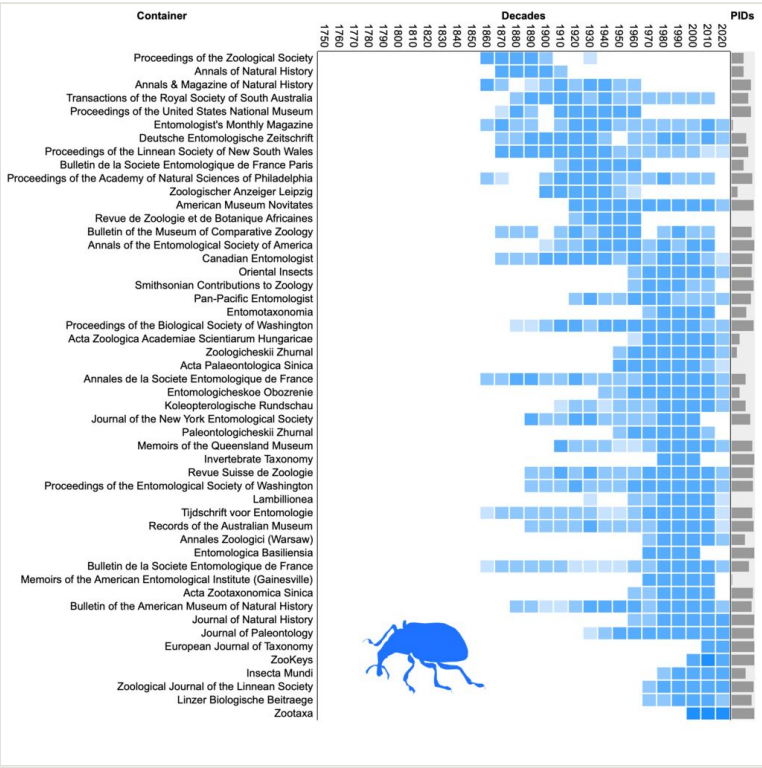


Figure 11. doi

Density distribution of taxonomic names published in the decades from 1750 to 2020 across the 50 publication venues (“containers”) that published the most names for animals. See Fig. 9 for further explanation.

Knowledge graph

The literature mappings for Index Fungorum, IPNI and ION, together with the JSON-LD export from ORCID, are available in Zenodo as RDF in N-Triples format (Table 3). The code to assemble these into a local triple store is available on Github https://github.com/rdmpage/ten-kg.

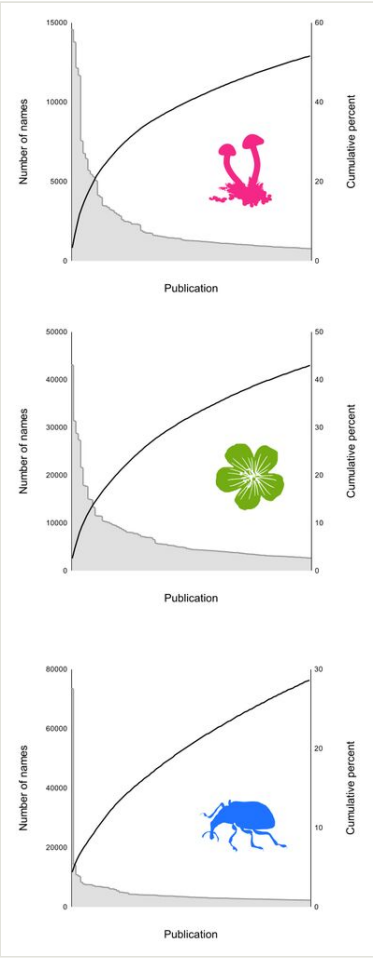


Figure 12. doi

Long tails in Index Fungorum, IPNI and ION. For each database, the top 100 publications are ranked in descending order of the number of names each publishes. The left vertical axis shows the names in each publication, the right vertical axis shows the cumulative percentage of total names in the database that is added by each publication.

Table 3.

The N-Triples datasets for people and taxonomic names.

Dataset name	DOI (all versions)
Linked data for taxonomic authors in ORCID	https://doi.org/10.5281/zenodo.7181180
Index Fungorum	https://doi.org/10.5281/zenodo.7977299
IPNI	https://doi.org/10.5281/zenodo.7977435
ION (BioNames)	https://doi.org/10.5281/zenodo.7977556

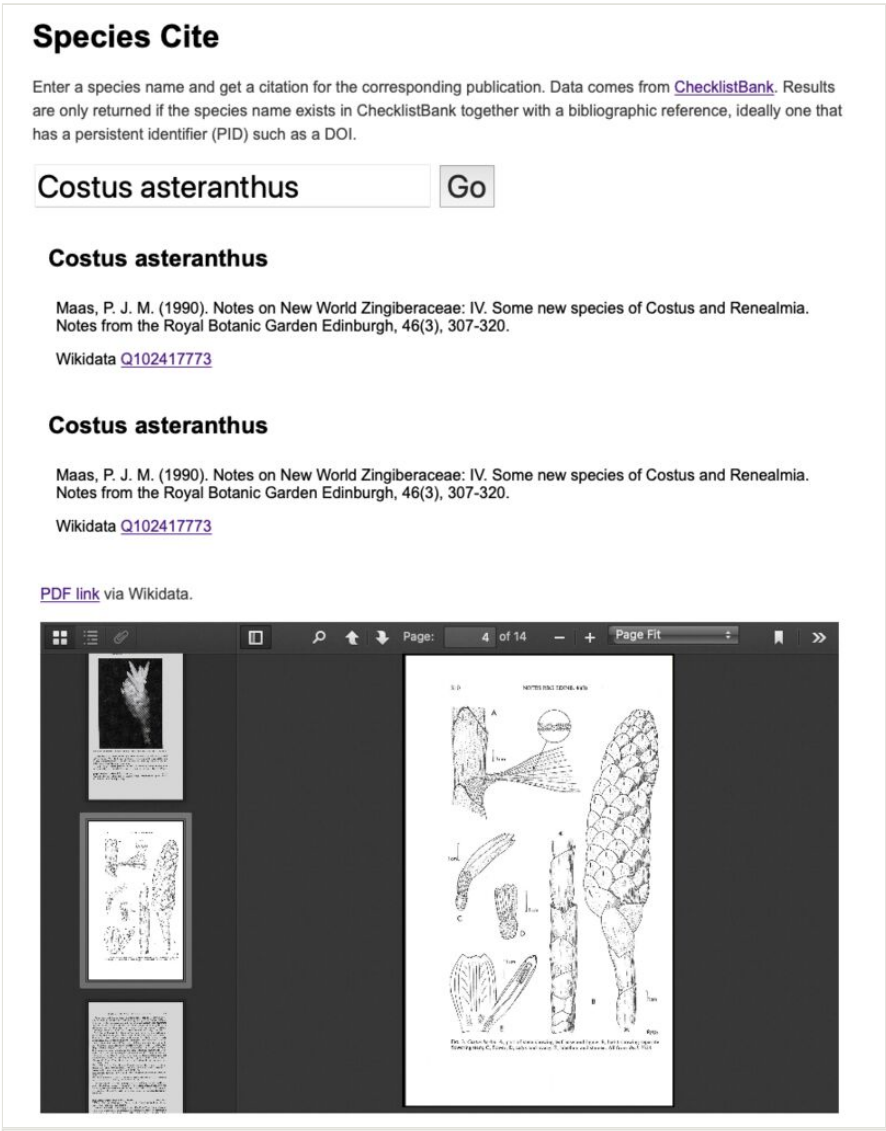
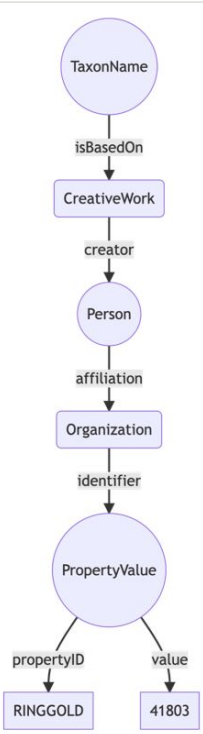


Figure 13. doi

Screenshot of Species Cite showing details for a taxonomic name, with a formatted citation, links to persistent identifiers (e.g. DOI, Wikidata) and a PDF of the publication.

The triples listed in Table 3 are minimalistic in that they lack details on the taxonomic names, other than the name string and the DOI of the associated publication. Likewise, ORCID triples include a bare minimum of information about a publication, typically just the DOI and title. However, the ORCID triples also have links between people and institutions, so we can do queries, such as that shown in Fig. 14 which finds authors affiliated with the Royal Botanic Gardens Edinburgh who have published plant names.



SPARQL query

```
PREFIX schema: <http://schema.org/>
SELECT DISTINCT ?person ?person_name
FROM <https://www.ipni.org>
FROM <https://orcid.org>
WHERE {
 ?id schema:propertyID "RINGGOLD" .
 ?id schema:value "41803" .
 ?org schema:identifier ?id .
 ?person schema:affiliation ?org .
 ?work schema:creator ?person .
 ?taxonName schema:isBasedOn ?work .
 ?taxonName schema:name ?name .
 {
 ?person schema:givenName ?givenName .
 ?person schema:familyName ?familyName .
 BIND(CONCAT(?givenName, " ", ?familyName) AS ?person_name).
 }
 }
ORDER BY ?person_name
```

SPARQL query result

https://orcid.org/0000-0001-9745-3222	Anna Trias-Blasi
https://orcid.org/0000-0002-7619-824X	Colin Pendry
https://orcid.org/0000-0002-0855-4169	Danilo Neves
https://orcid.org/0000-0002-6801-2484	David Harris
https://orcid.org/0000-0001-9144-2848	Elspeth Haston
https://orcid.org/0000-0001-8778-6779	Felicity Anderson
https://orcid.org/0000-0002-3523-1883	Hannah Atkins
https://orcid.org/0000-0002-0235-9506	Laura Forrest
https://orcid.org/0000-0002-2168-0514	Mark Hughes
https://orcid.org/0000-0003-4851-015X	Mark Newman
https://orcid.org/0000-0002-4989-3976	Mark Watson
https://orcid.org/0000-0002-8457-345X	Markus Ruhsam
https://orcid.org/0000-0003-4342-2089	Peter Moonlight
https://orcid.org/0000-0002-3407-8397	Peter Wilkie
https://orcid.org/0000-0002-3410-7258	Sabina Knees
https://orcid.org/0000-0002-8196-288X	Toby Pennington
https://orcid.org/0000-0003-2926-1645	Zoë Goodwin
https://orcid.org/0000-0002-6143-3411	Øystein Lofthus

Figure 14. doi

Example SPARQL query. The graph on the left connects a taxonomic name to an author of the publication of that name and the person, in turn, is connected to an organisation where they worked. This graph can be represented as a SPARQL query, in this case, using the RINGGOLD identifier 41803 which is for the Royal Botanic Gardens Edinburgh. The SPARQL query is restricted to the <https://www.ipni.org> namespace, so the authors listed published papers on plant names.

Discussion

Coverage

The results in Table 1 show that over a million taxonomic names have been linked to a persistent identifier for the associated publication. Overall, this presents almost 20% of the total number of names in the source databases. However, if we include only those names accompanied by publication information, then we have approximately 36% coverage. Clearly much work remains to be done.

The long tails shown in Fig. 12 provide some insight into the problem. Even if we make rapid progress on publications that have DOIs, we quickly end up having to work with a large number of often small, obscure publications that individually contribute little, but in

aggregate hold a great deal of taxonomic information. The “low hanging fruit” are quickly picked, leaving behind a much more challenging harvest.

Readers might be surprised by the very low number of BHL page links in Table 1. This is partly because the BioNames project focuses on work-level identifiers (Fig. 3), such as DOIs. It would be possible to add BHL page links to many more of the names, using tools such as BHLnames (Ower and Mozzherin 2021); however, in this project, I focus on citable, work-level identifiers. Note that BHL is also becoming an important source of work-level identifiers as it mints DOIs for taxonomic publications it has scanned (Kearney and Page 2022).

Integration

Storing the mappings in ChecklistBank enables easy use by the original taxonomic databases or databases that reuse those taxonomic name identifiers. For example, the World Flora Online (Borsch et al. 2020) uses IPNI identifiers for many of its plant names. It could easily add detailed bibliographic data to those names by making use of the name - IPNI mapping. Likewise, UNITE species hypotheses frequently include Index Fungorum identifiers. Once taxonomic databases reuse existing persistent identifiers for names, they will benefit from being able to reuse existing links to the literature.

Using persistent identifiers for the literature offers other benefits (Agosti et al. 2022), such as increasing access to the actual publications. Identifiers, such as DOIs, typically resolve to a publisher’s web site and publication itself may be behind a paywall, potentially inaccessible to a user. Tools, such as Unpaywall, take DOIs and discover whether freely accessible versions of that publication exist. These free versions may exist in institutional repositories or in digital libraries, such as the Biodiversity Heritage Library (Kearney 2020).

Linking names to the literature also opens up possibilities for using summarisation techniques to generate knowledge about a taxon. For example, given the set of names applied to a taxon, we could retrieve abstracts and/or full text for the associated publications, summarise that text and develop query interfaces (e.g. chatbots) that can answer queries about the biology of that taxon.

Missing nodes and edges in the knowledge graph

A knowledge graph consists of nodes (entities) and edges (relationships). To the extent that these are missing, the knowledge graph is incomplete. Although missing nodes is an obvious weakness, we can always expand the scope and utility of a knowledge graph by adding more entities. For example, the knowledge graph described here lacks taxa (it has taxonomic names, but makes no claims about the validity of those names). One reason for this is that taxa are rarely expressed using taxon name identifiers. It is possible to retrieve JSON-LD from web pages for Catalogue of Life taxa, but the corresponding RDF lacks taxon name identifiers (the taxon names are treated as blank nodes). Hence, links between taxa and names would have to be done via matching on name strings, a process that can lead to mistakes (e.g. homonyms). A significant improvement to CoL would be the use of

persistent identifiers for taxonomic names provided by nomenclators, such as Index Fungorum and IPNI (compare Fig. 1 and Fig. 2). Other candidate nodes are type specimens and nucleotide sequences (e.g. DNA barcodes). Most specimens currently lack persistent identifiers - there is considerable folklore about how unstable GBIF occurrence record identifiers are. Hence, maintaining stable links between taxonomic names and type specimens would be a significant challenge.

The role of citations

A potentially very useful class of missing edges are citation links between articles (the “citation graph”). Apart from the obvious, if controversial (Pinto et al. 2021, Loizides et al. 2022), utility in developing metrics for the impact of journals, articles and researchers and the potential for discovering related publications through co-citation, we could potentially use citation patterns as measures of the quality of taxonomic data. Taxonomists make mistakes, in the sense that they partition biodiversity up into sets (e.g. species) that subsequent research may show to be incorrect. This results in taxonomic synonyms, such as having more than one name for the same species. Solow et al. (1995) suggest that it is not uncommon for 50% of taxonomic names to be synonyms and note that a considerable period of time may elapse between a name being published and its eventual discovery to be a synonym. They argue that groups with few synonyms have not necessarily been blessed with very good taxonomists, rather they may suffer from neglect. If these taxa were well-studied, then more synonyms would be discovered. One way to measure taxonomic activity could be citations: if the taxonomic literature of a group has received few citations, especially by other taxonomists, then this could be a clue that a group is neglected and needs more attention. Perhaps citations could be used as a proxy for taxonomic quality. At present, the Catalogue of Life uses an arbitrary “star system” to rate the quality of taxonomic databases, the number of stars being self-assigned by the data provider. Citation-based measures may provide a more objective measure of the current state of knowledge of a taxonomic group.

The notion of citation could be extended to other entities, such as nucleotide sequences, such that we link DNA sequence accession numbers to the publications that cite them. Given the use of DNA sequences to identify species as well as construct phylogenies, it is likely that sequences may be cited by more than just the original publication and, indeed, may link publications that do not have any bibliographic links. That is, a subsequent paper might not cite the original publication of a DNA sequence even if it uses that sequence (Page 2010).

Identifier types

In this work, I have focused on "location based" identifiers, such as DOIs and LSIDs. These identifiers specify a location where one can retrieve information about a digital entity and potentially retrieve that entity itself. Location-based identifiers emphasise the persistence of resolution (for example, through a centralised resolver, such as https://doi.org), but typically make no guarantees that the content returned persists unchanged

over time. For example, academic publishers may update the metadata for an article, but the DOI for that article remains unchanged.

It is worth noting that there is another approach to persistent identifiers, namely using cryptographic hashes of the content as the identifier (Elliott et al. 2020). This has the advantage of ensuring that the data requested have not changed (which we can check by comparing the hash identifier with the hash of the data themselves). Unlike DOIs and similar identifiers, there is typically no centralised mechanism to resolve hash-based identifiers. Some decentralised systems have been developed, but it is unclear if they themselves will persist. To date, there are no widely used hash-based identifiers for publications.

Next steps

This paper has described a small neighbourhood of the biodiversity knowledge graph. It is clear that there is still a considerable amount of taxonomic literature to locate and link to. An increasing fraction of the taxonomic literature is being retrospectively digitised and assigned identifiers (Agosti and Egloff 2009, Kearney and Page 2022). The challenge is now to ensure that this literature is made discoverable, citable and connected to taxonomic names, thus building the bibliography of life (King et al. 2011, Page 2022b). We also need to develop ways to incorporate these links into existing resources, such that a visitor to a biodiversity web site is never faced with the prospect of having to Google a cryptic citation if they want to learn more about a species.

The focus of the work described here has been on bibliographic identifiers at the level of the work, that is, identifiers that are likely to be cited. This is not to discount the value in having deep links below the level of the work, such as to individual pages or to a collection of pages (e.g. treatments). However, identifiers that are cited are more likely to be the basis for new metrics of productivity (McDade et al. 2011). By including persistent identifiers for literature in taxonomic databases, we could explore mechanisms for credit for taxonomists. At present, aggregators, such as ChecklistBank, include ORCIDs for those who contributed to curating individual databases, many of whom are taxonomists. However, the bulk of the taxonomic community does not receive credit for the original work being aggregated (Franz and Sterner 2018). The use of persistent identifiers for names, publications and people means we could start to identify those people who have contributed the most to our taxonomic knowledge. Indeed, we can envisage a case where taxonomic databases and aggregations, such as the Catalogue of Life, give credit directly to the taxonomists whose data they aggregate, based on networks of connected, persistent identifiers.

Acknowledgements

The work described here has benefitted from discussions over the years with numerous people including the late David Remsen, Paddy Patterson, Donat Agosti, David Shorthouse and Rich Pyle. I thank my colleagues in the Biodiversity Heritage Library "Persistent

Identifiers Working Group" for many evenings of Zoom calls discussing the finer points of DOIs, ISSNs and Wikidata. The reviewers Dmitry Mozzherin, Franck Michel and Lyubo Penev provided insightful comments. Images of taxa includied in the figures come from https://www.phylopic.org: *Leotiomycetes* CC0 by Levi Simons, *Geranium maculatum* CC0 by Mason McNair, *Smycronyx* CC0 by Kanako Bessho-Uehara. Support for the publication of this work comes from the BiCIKL project, Grant No 101007492.

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Agosti D, Egloff W (2009) Taxonomic information exchange and copyright: the Plazi approach. BMC Research Notes 2 (1). https://doi.org/10.1186/1756-0500-2-53
- Agosti D, Benichou L, Addink W, Arvanitidis C, Catapano T, Cochrane G, Dillen M, Döring M, Georgiev T, Gérard I, Groom Q, Kishor P, Kroh A, Kvaček J, Mergen P, Mietchen D, Pauperio J, Sautter G, Penev L (2022) Recommendations for use of annotations and persistent identifiers in taxonomy and biodiversity publishing. Research Ideas and Outcomes 8 https://doi.org/10.3897/rio.8.e97374
- Anonymous (2012) Spinning threads. Nature 489 (7414): 5-6. https://doi.org/10.1038/489005b
- Anonymous (2014) Mycological progress converts to continuous article publishing model and monthly publishing frequency. Mycological Progress 13 (4). https://doi.org/10.1007/s11557-014-1007-x
- Benichou L, Buschbom J, Campbell M, Hermann E, Kvaček J, Mergen P, Mitchell L, Rinaldo C, Agosti D (2022) Joint statement on best practices for the citation of authorities of scientific names in taxonomy by CETAF, SPNHC and BHL. Research Ideas and Outcomes 8 https://doi.org/10.3897/rio.8.e94338
- Bennett F (2018) citeproc-js. URL: https://github.com/Juris-M/citeproc-js
- Bohannon J, Doran K (2017) Introducing ORCID. Science 356 (6339): 691-692. https://doi.org/10.1126/science.356.6339.691
- Borsch T, Berendsohn W, Dalcin E, Delmas M, Demissew S, Elliott A, Fritsch P, Fuchs A, Geltman D, Güner A, Haevermans T, Knapp S, le Roux MM, Loizeau P, Miller C, Miller J, Miller J, Palese R, Paton A, Parnell J, Pendry C, Qin H, Sosa V, Sosef M, von Raab-Straube E, Ranwashe F, Raz L, Salimov R, Smets E, Thiers B, Thomas W, Tulig M, Ulate W, Ung V, Watson M, Jackson PW, Zamora N (2020) World Flora Online: Placing taxonomists at the heart of a definitive and comprehensive global resource on the world's plants. Taxon 69 (6): 1311-1341. https://doi.org/10.1002/tax.12373
- Bridson GDR (1968) The Zoological Record—A centenary appraisal. Journal of the Society for the Bibliography of Natural History 5 (1): 23-34. https://doi.org/10.3366/jsbnh.1968.5.1.23
- Brush AJB, Bargeron D, Gupta A, Cadiz JJ (2001) Robust annotation positioning in digital documents. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. [ISBN 978-1-58113-327-1]. https://doi.org/10.1145/365024.365117

- Clark T, Martin S, Liefeld T (2004) Globally distributed object identification for biological knowledgebases. Briefings in Bioinformatics 5 (1): 59-70. https://doi.org/10.1093/bib/5.1.59
- Croft J, Cross N, Hinchcliffe S, Lughadha EN, Stevens PF, West JG, Whitbread G (1999) Plant names for the 21st century: the International Plant Names Index, a distributed data source of general accessibility. TAXON 48 (2): 317-324. https://doi.org/10.2307/1224436
- Döring M, Ower G (2019) The Catalogue of Life Data Package - A new format for exchanging nomenclatural and taxonomic information. Biodiversity Information Science and Standards 3 https://doi.org/10.3897/biss.3.38771
- Döring M, Jeppesen T, Bánki O (2022) Introducing ChecklistBank: An index and repository for taxonomic data. Biodiversity Information Science and Standards 6 https://doi.org/10.3897/biss.6.93938
- Dürst M, Wilde E (2008) URI Fragment Identifiers for the text/plain Media Type. Internet Engineering Task Force. DOI: 10.17487/RFC5147 Num Pages: 17. URL: https://datatracker.ietf.org/doc/rfc5147
- Elliott M, Poelen J, Fortes JB (2020) Toward reliable biodiversity dataset references. Ecological Informatics 59 https://doi.org/10.1016/j.ecoinf.2020.101132
- Franz NM, Sterner BW (2018) To increase trust, change the social design behind aggregated biodiversity data. Database 2018 https://doi.org/10.1093/database/bax100
- Garrity G, Lyons C (2003) Future-Proofing Biological Nomenclature. OMICS: A Journal of Integrative Biology 7 (1): 31-33. https://doi.org/10.1089/153623103322006562
- Hennessey J, Ge SX (2013) A cross disciplinary study of link decay and the effectiveness of mitigation techniques. BMC Bioinformatics 14 (Suppl 14). https://doi.org/10.1186/1471-2105-14-S14-S5
- Hu J, Zhao G, Tuo Y, Dai D, Guo D, Rao G, Qi Z, Zhang Z, Li Y, Zhang B (2021) Morphology and molecular study of three new Cordycipitoid fungi and its related species collected from Jilin Province, northeast China. MycoKeys 83: 161-180. https://doi.org/10.3897/mycokeys.83.72325
- Kearney N (2020) Discovering the platypus: From its scientific description to its DOI. Biodiversity Information Science and Standards 4 https://doi.org/10.3897/biss.4.59089
- Kearney N, Page R (2022) Supplying the Missing Links: Providing immediate access to the taxonomic literature from our taxonomic databases. Biodiversity Information Science and Standards 6 https://doi.org/10.3897/biss.6.91104
- King D, Morse D, Willis A, Dil A (2011) Towards the bibliography of life. ZooKeys 150: 151-166. https://doi.org/10.3897/zookeys.150.2167
- Loizides M, Alvarado P, Moreau P, Assyov B, Halasů V, Stadler M, Rinaldi A, Marques G, Zervakis G, Borovička J, Van Vooren N, Grebenc T, Richard F, Taşkin H, Gube M, Sammut C, Agnello C, Baroni T, Crous P, Fryssouli V, Gonou Z, Guidori U, Gulden G, Hansen K, Kristiansen R, Læssøe T, Mateos J, Miller A, Moreno G, Perić B, Polemis E, Salom JC, Siquier JL, Snabl M, Weholt Ø, Bellanger J (2022) Has taxonomic vandalism gone too far? A case study, the rise of the pay-to-publish model and the pitfalls of Morchella systematics. Mycological Progress 21 (1): 7-38. https://doi.org/10.1007/s11557-021-01755-z
- Martin S, Hohman M, Liefeld T (2005) The impact of Life Science Identifier on informatics data. Drug Discovery Today 10 (22): 1566-1572. https://doi.org/10.1016/S1359-6446(05)03651-2

- McDade L, Maddison D, Guralnick R, Piwowar H, Jameson ML, Helgen K, Herendeen P, Hill A, Vis M (2011) Biology Needs a Modern Assessment System for Professional Productivity. BioScience 61 (8): 619-625. https://doi.org/10.1525/bio.2011.61.8.8
- Nicholson J, Mordaunt M, Lopez P, Uppala A, Rosati D, Rodrigues N, Grabitz P, Rife S (2021) scite: A smart citation index that displays the context of citations and classifies their intent using deep learning. Quantitative Science Studies 2 (3): 882-898. https://doi.org/10.1162/qss_a_00146
- Nilsson RH, Larsson K, Taylor AFS, Bengtsson-Palme J, Jeppesen TS, Schigel D, Kennedy P, Picard K, Glöckner FO, Tedersoo L, Saar I, Kõljalg U, Abarenkov K (2019) The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. Nucleic Acids Research 47 (D1). https://doi.org/10.1093/nar/gky1022
- Ower G, Mozzherin D (2021) Algorithms for connecting scientific names with literature in the Biodiversity Heritage Library via the Global Names Project and Catalogue of Life. Biodiversity Information Science and Standards 5 https://doi.org/10.3897/biss.5.74114
- Page RM (2010) Enhanced display of scientific articles using extended metadata. Web Semantics: Science, Services and Agents on the World Wide Web 8 (2): 190-195. https://doi.org/10.1016/j.websem.2010.03.004
- Page RM (2013) BioNames: linking taxonomy, texts, and trees. PeerJ 1 https://doi.org/10.7717/peerj.190
- Page RM (2018) Liberating links between datasets using lightweight data publishing: an example using plant names and the taxonomic literature. Biodiversity Data Journal 6 https://doi.org/10.3897/BDJ.6.e27539
- Page RM (2022a) Bootstrapping a Biodiversity Knowledge Graph. Biodiversity Information Science and Standards 6 https://doi.org/10.3897/biss.6.91497
- Page RM (2022b) Wikidata and the bibliography of life. PeerJ 10 https://doi.org/10.7717/peerj.13712
- Penev L, Paton A, Nicolson N, Kirk P, Pyle R, Whitton R, Georgiev T, Barker C, Hopkins C, Robert V, Biserkov J, Stoev P (2016) A common registration-to-publication automated pipeline for nomenclatural acts for higher plants (International Plant Names Index, IPNI), fungi (Index Fungorum, MycoBank) and animals (ZooBank). ZooKeys233-246. https://doi.org/10.3897/zookeys.550.9551
- Peroni S, Shotton D (2020) OpenCitations, an infrastructure organization for open scholarship. Quantitative Science Studies 1 (1): 428-444. https://doi.org/10.1162/qss_a_00023
- Pinto ÂP, Mejdalani G, Mounce R, Silveira LF, Marinoni L, Rafael JA (2021) Are publications on zoological taxonomy under attack? Royal Society Open Science 8 (2). https://doi.org/10.1098/rsos.201617
- Robert V, Vu D, Amor ABH, Wiele Nvd, Brouwer C, Jabas B, Szoke S, Dridi A, Triki M, Daoud Sb, Chouchen O, Vaas L, Cock Ad, Stalpers J, Stalpers D, Verkley GM, Groenewald M, Santos FBd, Stegehuis G, Li W, Wu L, Zhang R, Ma J, Zhou M, Gorjón SP, Eurwilaichitr L, Ingsriswang S, Hansen K, Schoch C, Robbertse B, Irinyi L, Meyer W, Cardinali G, Hawksworth D, Taylor J, Crous P (2013) MycoBank gearing up for new horizons. IMA Fungus 4 (2): 371-379. https://doi.org/10.5598/imafungus.2013.04.02.16
- Shorthouse D (2020) Slinging with four giants on a quest to credit natural historians for our museums and collections. Biodiversity Information Science and Standards 4 https://doi.org/10.3897/biss.4.59167

- Solow A, Mound L, Gaston K (1995) Estimating the rate of synonymy. Systematic Biology 44 (1): 93-96. https://doi.org/10.2307/2413485
- Taft EA, Masinter L, Zilles S, Pravetz J (2004) The application/pdf Media Type. Internet Engineering Task Force. https://doi.org/10.17487/RFC3778
- Thomas C (2009) Biodiversity Databases Spread, Prompting Unification Call. Science 324 (5935): 1632-1633. https://doi.org/10.1126/science.324_1632
- Veen Tv (2019) Wikidata:From “an” Identifier to “the” Identifier. Information Technology and Libraries 38 (2): 72-81. https://doi.org/10.6017/ital.v38i2.10886