# Liberating links between datasets using lightweight data publishing: an example using plant names and the taxonomic literature

## Abstract

Constructing a biodiversity knowledge graph will require making millions of cross links between diversity entities in different datasets. Researchers trying to bootstrap the growth of the biodiversity knowledge graph by constructing databases that comprises links between these entities lack obvious ways to publish these datasets. One appealing and lightweight approach is to create a “datasette”, a database that is wrapped together with a simple web server that enables users to query the data. Datasettes can be packaged into Docker containers and hosted online with minimal effort. This approach is illustrated using a dataset of links between globally unique identifiers for plant taxonomic names and identifiers for the taxonomic articles that published those names.

## Introduction

A venerable tradition in taxonomy is compiling and publishing lists of scientific names, whether in printed form or as online databases. If the lists include bibliographic information these lists can also serve as search indices to the taxonomic literature. However, this functionality is often

hampered by the use of cryptic citations, and a mismatch between the granularity sought by taxonomists (often page-level) and that used by researchers when citing sources. A practical consequence is that a biologist seeking information about a species may struggle to locate the original taxonomic description, which for many species may be the best (or, indeed, only) source of basic biological data for that species [Page 2016a].

In an ideal world each taxonomic name would be linked to a detailed bibliographic record of where that name was published, and that publication would be available in digital form, as would any subsequent taxonomic revisions [Agosti & Egloff,2009, Page 2016a]. There are notable examples of resources like this for particular taxonomic groups (e.g., world spiders)[Gloor et al. 2017], but there is no freely accessible resource that covers, for example, all animals, or all plants. There are detailed databases of names that also cite the primary literature, but typically these citations are simply text strings, not actionable digital identifiers.

Motivated by this lack of links I have spent the last few years obsessively collecting digital identifiers for taxonomic publications and linking them to taxonomic names. This project is far from complete, nor is it likely to be in the near future given the continuing discovery of new species, and the increasing number of taxonomic works that are becoming available online. One consequence of this Sisyphean task is that it becomes tempting to simply continue to accumulate links in a local database, forever postponing actually publishing them. This is not a particularly successful career strategy, nor is it helpful to people who might make use of these links. However, publishing sets of links is not necessarily a straightforward task.

One option for publication is to create a custom interface to the links, to make them both discoverable and interesting. Examples include links

between the NCBI taxonomy [Federhen 2011] and Wikipedia [Page 2011], or BioNames [Page 2013], which comprises links between animal names and the primary literature. These may be user friendly, but they provide limited functionality, especially if a user wants programmatic access to the underlying data.

Rather than expend effort on developing idiosyncratic solutions, one could simply publish the data to an existing platform. I adopted this approach for the names in the Plant List [http://www.theplantlist.org/], for which I linked a subset of names to publications with Digital Object Identifiers (DOIs) or with a link to JSTOR. This dataset was uploaded to the Global Biodiversity Information Facility (GBIF) [Page 2016c]. GBIF uses the Darwin Core Archive [Wieczorek et al. 2012] as its data format, which does not handle literature particularly well, and literature is not a first class citizen of the GBIF portal. Consequently, uploading this data to GBIF does not make the most of the effort that went into creating the links.

GBIF is a domain-specific data publisher. An alternative may be to publish in a venue with broader scope, such as Wikidata [Vrandečić & Krötzsch 2014, https://www.wikidata.org]. Wikidata is rapidly becoming a useful platform for cross linking scientific data [Burgstaller-Muehlbacher 2016], and has enormous potential. However, because the data is published at the level of individual statements, it becomes difficult to point to who did what in a simple way. In contrast, GBIF treats datasets as both a bundle of data that can be identified as a specific contribution (with a dataset-specific DOI), as well as "unbundling" the data and merging it into a single index.

A particularly appealing route for publishing links would be to treat each link as a "nanopublication" [Groth et al. 2010], which is minimally a single linked data "triple". Nanopublications have built in mechanisms for provenance and attribution [Kuhn et al. 2017], and have been used to

publish large datasets [Queralt-Rosinach et al. 2016]. Because nanopublications are grounded in linked data they will be of most use in communities where linked data has been widely adopted. To date the biodiversity informatics community has shown lukewarm enthusiasm for linked data, despite various calls to exploring its use [Page 2016b], and some working implementations [Michel et al., 2017; Senderov et al. 2018], we have nothing on the scale, say, of Uniprot [The UniProt Consortium 2016]. Hence, despite their attractiveness, publishing the links as nanopublications does not seem to be a way to encourage their reuse, although this may well change in the future.

If we find the three options discussed so far (custom web site, existing data publisher, and nanopublications) unsatisfactory, then it seems that the only remaining approach is simply to deposit the dataset as a "dumb" file in a repository such as Datadryad or Zenodo, minting a DOI to make it citable, and then hope that somebody makes use of it. But multi-megabyte data files are often not the easiest for users to work with, and it might not be obvious to a potential user why the data would be worth them investing time in discovering whether it was useful.

However, other possibilities are emerging. For example, Simon Willison [2017] has recently proposed a lightweight approach to data publishing called "datasettes". A datasette comprises one or more comma separated value (CSV) files which are merged into a SQLite database. The database and a simple API are bundled together with a web server that can be queried interactively by the user via a web interface. The datasette can be run on the user's local machine, or easily pushed to a server in the "cloud", for example using Docker containers. Datasettes and the other approaches listed above are not, of course, mutually exclusive. But the attraction of the datasette is that it makes it easy to publish data that might otherwise either not be published, or might be published as a large "blob" of data whose utility is opaque to its potential users.

In this paper I describe the creation of a datasette for a longstanding but mostly unpublished project on linking plant names in The Index of Plant Names Index (IPNI) to the taxonomic literature.

## Materials and methods

The Index of Plant Names Index (IPNI, http://www.ipni.org) is an international register of published plant names based at the Royal Botanic Gardens, Kew but which has contributions from the Harvard Gray Index and the Australian Plant Name Index. Both new taxonomic names (e.g., for newly described species) and new combinations (e.g., reflecting transfers of species from one genus to another) are recorded in IPNI, together with a citation to the scientific work that published that name. These citations typically comprise an abbreviation for the publication (such as a a journal or a book), a description of the location of the name within that publication, such as a combination of volume number and page number, and the year of publication. One or more of these items may be missing, different journal abbreviations may be used in data sourced from different datasets, and the volume and pagination may be in either Roman or Arabic numerals. For some records the IPNI curators have added a link to the corresponding page in the Biodiversity Heritage Library (BHL), and for some recently added records the IPNI web site may give the DOI for a publication, but the majority of IPNI records are not linked to a digital identifier for the publication associated with each name.

In much the same way as for BioNames [Page 2013], I have harvested the IPNI database via its API and have developed software for matching the text string citations to digital identifiers such as DOIs, Handles, JSTOR links, etc. Whereas the source data for BioNames comprised citations at the level of a work (e.g., an article, chapter, or a book) which are relatively easy to match, the citations in IPNI are at the level of one or more pages

within a work. Hence a big part of the challenge is to map page-level citations to work-level bibliographic data (Page 2009). Given a complete bibliography of the taxonomic literature, this would be a relatively straightforward task, in that we could treat each work as comprising a set of pages, and we simply ask which works include the page in the IPNI citation. However, as yet we don't have a comprehensive bibliography of life [King et al. 2011], hence much of the work in making the mapping involves scouring the web for sources of bibliographic information in the hope that these will include works containing the IPNI citations (the bibliographic database being assembled as a consequence of this work will be described in more detail elsewhere). I manage the mapping between IPNI names and the literature in a local MySQL database, and the results are periodically uploaded to a GitHub repository https://github.com/rdmpage/ipni-names, which also has code for a custom interface to that mapping.

## Datasette

A CSV file containing basic metadata for a plant name, such as IPNI LSID, scientific name, bibliographic details, and any identifiers found was generated from the current IPNI LSID to literature identifier mapping. This CSV file was then converted into a SQLite database using csv-to-sqlite, and the resulting database (ipni.db) was wrapped in a web server using the command "datasette serve ipni.db". This datasets runs on the user's local machine. We can also package the datasets into a Docker container using the following command:

```
datasette package -t <username>/ipni ipni.db
```

Where is your username at docker.com. The container can be run locally, or can be pushed to a repository where others can access it. To push to Docker's Hub the commands are:

```
docker login -u <username> -p <password>
docker push <username>/ipni
```

A container for this project is available at https://hub.docker.com/r/rdmpage/ipni/

# Results

The datasette generated here can be seen online at https://ipni.sloppy.zone. If this demo is offline, the reader can simply deploy a copy of the container from the Docker repository https://hub.docker.com/r/rdmpage/ipni/. The interface (Fig. 1) is simple and generic, but enables the user to browse the data as well as perform some straightforward queries. It should be noted that the interface can be customised to add more features, for this example I have stuck with the defaults.

Figure 1: Screenshot of datasette of IPNI names.

Some simple queries include finding the DOIs for publications of new names in a given genus, such as *Begonia*:

```
select Id, Full_name_without_family_and_authors, doi
from ipni where Genus="Begonia" and doi is not null;
```

JSTOR has digitised many botanical journals, so for some taxa such as the genus *Tiquilia* it is an excellent source of taxonomic literature:

```
select Id, Full_name_without_family_and_authors, doi
```

```
from ipni where genus='Tiquilia' and jstor is not
null;
```

Although the primary goal of the name-to-literature mapping is to find digital versions of the descriptions for each species, the datasette enables queries that might address other questions. For example, the database includes information on the agency that registers the DOI for a publication. For most publications this is CrossRef, but there are other agencies, such as DataCite, the multilingual European Registration Agency (mEDRA), and the Airiti Incorporation (華藝數位). Table 1 summarises the relative importance of these agencies. Different DOI agencies expose metadata for their DOIs in different ways, so the existence of multiple DOI agencies has implications for any researcher or tool that attempts to harvest bibliographic metadata. It could also can be used to investigate the pace of digitisation of legacy literature in different parts of the world. For example, a growing number of articles from journals published in China, Taiwan, and Japan now have DOIs assigned by local DOI agencies.

Table 1: Number of DOIs for articles linked to IPNI names, grouped by registration agency.

| **Agency** | **Number of DOIs** |
|---|---:|
| crossref | 197525 |
| 10.SERV/JALC | 2876 |
| 10.SERV/ISTIC | 832 |
| 10.SERV/AIRITI | 576 |
| | |

| | |
|---|---|
| 10.SERV/ETH | 107 |
| medra | 23 |
| 10.SERV/KISTI | 4 |
| datacite | 3 |

# Discussion

Links between taxonomic names and the scientific literature have many possible uses. One is simply to be able to read the description of a new species, or discover the reasoning behind subsequent changes in name. Given that many of these sources are available in machine-readable text, the links could be used to generate a corpus for text mining to extract information on the species being described [e.g., Cui 2012].

The use of global bibliographic identifiers also enables queries that can span multiple databases. For example, knowing the DOI for a paper that changes the taxonomy of a plant genus, we could ask whether the evidence for that is supported by phylogenetic analysis by seeing whether that DOI also occurs in TreeBASE [https://treebase.org]. We could ask to what extent the discovery of new plants species is being driven by molecular data by seeing whether the DOI for the species description also occurs in sequence databases such as GenBank. However, these examples all require the existence of links between these databases, which are often incomplete [Miller et al. 2009], and hence represent further instances where the kind of mapping described here would be worthwhile.

In the absence of an existing knowledge graph, and the lack of a centralised infrastructure supporting its development, datasettes provide

an easy mechanism for publishing links that places minimal burden on the researcher or curator doing the mapping, but also provides an interface that is potentially useful to users, even as we wait for the knowledge graph itself to coalesce.

## Acknowledgements

## References

Agosti, D., & Egloff, W. (2009). Taxonomic information exchange and copyright: the Plazi approach. BMC Research Notes, 2(1), 53. doi:10.1186/1756–0500–2–53

Burgstaller-Muehlbacher, S., Waagmeester, A., Mitraka, E., Turner, J., Putman, T., Leong, J., … Su, A. I. (2016). Wikidata as a semantic framework for the Gene Wiki initiative. Database, 2016, baw015. doi:10.1093/database/baw015

Cui, H. (2012). CharaParser for fine-grained semantic annotation of organism morphological descriptions. Journal of the American Society for Information Science and Technology, 63(4), 738–754. doi:10.1002/asi.22618

Federhen, S. (2011). The NCBI Taxonomy database. Nucleic Acids Research, 40(D1), D136–D143. doi:10.1093/nar/gkr1178

Gloor, D., Nentwig, W., Blick, T., & Kropf, C. (2017). World Spider Catalog. Natural History Museum Bern. https://doi.org/10.24436/2

Groth, P., Gibson, A., & Velterop, J. (2010). The anatomy of a nanopublication. Information Services & Use, 30(1–2), 51–56. doi:10.3233/isu–2010–0613

King, D., Morse, D., Willis, A., & Dil, A. (2011). Towards the bibliography of life. ZooKeys, 150, 151–166. doi:10.3897/zookeys.150.2167

Kuhn, T., Chichester, C., Krauthammer, M., Queralt-Rosinach, N., Verborgh, R., Giannakopoulos, G., … Dumontier, M. (2016). Decentralized provenance-aware publishing with nanopublications. PeerJ Computer Science, 2, e78. doi:10.7717/peerj-cs.78

Franck Michel, Olivier Gargominy, Sandrine Tercerie, Catherine Faron Zucker. A Model to Represent Nomenclatural and Taxonomic Information as Linked Data. Application to the French Taxonomic Register, TAXREF. S4Biodiv 2017 - 2nd International Workshop on Semantics for Biodiversity co-located with ISWC 2017, Oct 2017, Vienna, Austria. CEUR Vol. 1933, pp.1–12, 2017, https://hal.archives-ouvertes.fr/hal–01617708

Miller, H., Norton, C. N., & Sarkar, I. N. (2009). GenBank and PubMed: How connected are they? BMC Research Notes, 2(1), 101. doi:10.1186/1756–0500–2–101

Page, R. D. (2009). bioGUID: resolving, discovering, and minting identifiers for biodiversity informatics. BMC Bioinformatics, 10(Suppl 14), S5. doi:10.1186/1471–2105–10-s14-s5

Page, R. D. M. (2011). Linking NCBI to Wikipedia: a wiki-based approach. PLoS Currents, 3, RRN1228. doi:10.1371/currents.rrn1228

Page, R. D. M. (2013). BioNames: linking taxonomy, texts, and trees. PeerJ, 1, e190. doi:10.7717/peerj.190

Page, R. D. M. (2016). Surfacing the deep data of taxonomy. ZooKeys, 550, 247–260. doi:10.3897/zookeys.550.9293

Page, R. D. M. (2016). Towards a biodiversity knowledge graph. Research Ideas and Outcomes, 2, e8767. doi:10.3897/rio.2.e8767

Page, R D M. (2016) The Plant List with literature. Institute of Biodiversity, Animal Health and Comparative Medicine, College of Medical, Veterinary and Life Sciences, University of Glasgow. Checklist Dataset https://doi.org/10.15468/btkum2.

The Plant List (2013). Version 1.1. Published on the Internet; http://www.theplantlist.org/ (accessed 1st January).

'The International Plant Names Index (2012). Published on the Internet http://www.ipni.org [accessed 1 July 2012]*.

Queralt-Rosinach, N., Kuhn, T., Chichester, C., Dumontier, M., Sanz, F., & Furlong, L. I. (2016). Publishing DisGeNET as nanopublications. Semantic Web, 7(5), 519–528. doi:10.3233/sw–150189

Senderov, V., Simov, K., Franz, N., Stoev, P., Catapano, T., Agosti, D., … Penev, L. (2018). OpenBiodiv-O: ontology of the OpenBiodiv knowledge management system. Journal of Biomedical Semantics, 9(1). doi:10.1186/s13326–017–0174–5

The UniProt Consortium (2016). UniProt: the universal protein knowledgebase. Nucleic Acids Research, 45(D1), D158–D169. doi:10.1093/nar/gkw1099

Vrandečić, D., & Krötzsch, M. (2014). Wikidata. Communications of the ACM, 57(10), 78–85. doi:10.1145/2629489

Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., … Vieglais, D. (2012). Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLoS ONE, 7(1), e29715. doi:10.1371/journal.pone.0029715

Wikicite https://meta.wikimedia.org/wiki/WikiCite

Simon Willison 2017 Datasette: instantly create and publish an API for your SQLite databases
https://simonwillison.net/2017/Nov/13/datasette/
https://github.com/simonw/datasette