

Wikidata and the bibliography of life

Roderic D. M. Page

Abstract

[to do]

Introduction

Much of the primary data about the planet's biodiversity is contained in the taxonomic literature, a corpus that dates back to the eighteenth century. Whereas other biological disciplines have created substantial bibliographic databases, such as PubMed (<https://pubmed.ncbi.nlm.nih.gov/>), or open access repositories for work funded by specific funding agencies and charities agencies, such as EuroPMC (The Europe PMC Consortium, 2015), the taxonomic literature mostly lingers in obscurity (Page, 2016b). There are several projects trying to redress this problem by digitising the taxonomic literature, ranging from global initiatives such as the Biodiversity Heritage Library (BHL) (<https://www.biodiversitylibrary.org/>) to extensive, regional repositories such as ZOBODAT (Gusenleitner & Malicky, 2017). While the bulk of BHL content comprises legacy works that are out of copyright, recently this has been supplemented by an influx of more content so that BHL includes content less than a year old. Another initiative, the Biodiversity Literature Repository (BLR) (<https://www.biolitrepo.org/>) is focussed on recently published "born digital" content and its component parts, such as figures and taxonomic treatments (Egloff et al., 2017). Taxonomy also benefits from digitising initiatives that don't specifically target the taxonomic literature, but which include taxonomic journals, such as E-Periodica cite (Wanger & Ehrismann, 2016).

Digitisation greatly increases the accessibility, but not necessarily the discoverability of content. The BHL has scanned volumes for many journals, but unless articles contained within those volumes are indexed those articles will be difficult to find. This is the motivation for my BioStor project (Page, 2011), which to date has extracted over 200,000 articles from content scanned by BHL. Another impediment to discoverability is the widespread taxonomic practice of using "micro citations", that is, citing a page or set of pages within a work, rather than the work itself (Page, 2009). Experts in a particular group are usually familiar with these micro citations, but non-experts may find them challenging to interpret.

Discoverability of the taxonomic literature would be greatly improved if we had a single, easily accessible database of all taxonomic publications (King et al., 2011). While the field has some highly visible journals, there is a long tale of taxonomic publication in small, often obscure journals (Page, 2016b). Not only does discoverability hamper taxonomic research, it also hampers recognition of the value of that research. Taxonomists have long complained that standard measures of academic impact do not work well for taxonomists (Garfield,

2001), and the ranking of major taxonomic journals by commercial organisations such as Clarivate can undergo dramatic and seemingly capricious changes (Hamilton et al., 2021). A commonly proposed remedy is increased citation of taxonomic work, such as original descriptions of new species. Regardless of the merits of these proposals, they founder when confronted with the practical issue that we don't have citable references for many, if not most, species descriptions.

The challenge of discoverability is not unique to taxonomic literature. There have been long standing calls for what Cameron ([Cameron, 1997](#)) described as a “universal citation database”. Recent developments such as the Open Citations Corpus ([Shotton, 2013](#)) and the WikiCite project (<http://wikicite.org>) have brought us considerably closer to this goal. Indeed, in the last few years there has been a growing effort to add bibliographic details for the entire academic corpus to Wikidata, an open database of structured information. Bibliographic metadata is at the heart of measures of academic performance and impact, and these measures are typically provided from closed data held by commercial organisations ([Aspesi & Brand, 2020](#)). Having an open bibliographic database for taxonomy opens the possibility of more transparent analytics for the discipline.

In this paper I make the case for Wikidata as the logical venue for a “bibliography of life”. I begin by providing background on Wikidata, describing how it models bibliographic data, and how it is populated with data. I then summarise some analyses that assess the extent to which the Wikidata community curates bibliographic data, and estimate the “density” of the Wikidata knowledge graph for bibliographic data. I also describe a simple web interface for navigating bibliographic data in Wikidata.

Wikidata

Wikidata is a store of structured information or “statements” about things or concepts (“items”) (<https://www.mediawiki.org/wiki/Wikibase/DataModel>). Each statement comprises a key-value pair where the key is a community-defined property, and the value is editable by any Wikidata user. Each Wikidata item has a unique identifier of the form “Q n ” (where n is an integer), each property has an identifier in the form “P n ” (in this article I often refer to Wikidata properties by their P number). A given key-value pair can have one or more qualifiers (Vrandečić & Krötzsch, 2014), that is, a statement about that value. For example, a multi-author publication will have multiple values of the property “author” (P50). Adding the qualifier “series ordinal” (P1545) to each value enables us to express order of authorship, i.e., the first author has a series ordinal qualifier of “1”.

Ideally values in Wikidata are accompanied by one or more references to the sources of those values. Typically references are links to external sources (such as a web site or database), but they can also be links to another item in Wikidata (for example the item corresponding to a publication that is the source of that value). Among the strengths of Wikidata is its support for multiple languages, and for multiple values for the same property. Hence Wikidata can accommodate cases where there is legitimate disagreement about the value a property should take (for example, the date of publication of a work). While any user can edit values, properties are added by community consensus. A property is proposed, discussed, and if it receives community support it becomes available for editors to add to an

item. The information stored in Wikidata can be expressed as Resource Description Framework (RDF) triples (Erxleben et al., 2014) and there is a SPARQL endpoint that enables anyone to query the data.

Wikicite

The original scope of Wikidata was to provide structured data to underpin the different Wikipedia projects. Hence notionally each item in Wikidata had a corresponding entity in at least one of the various Wikipedias. However, as Wikidata has grown the potential of having a single, queryable, community-edited database of structured information has become increasingly clear. Hence many items being added to Wikidata might not have a Wikipedia page, but are relevant to the content and goals of Wikipedia. A good example of this are bibliographic citations, which are a key source of support for factual statements made on Wikipedia.

The Wikicite project started out with the goal to provide structured bibliographic data for citations across the different Wikipedia projects. Given that the scope of Wikipedia includes taxonomy, many of the publications cited in Wikipedia (and hence destined to be in Wikidata) are relevant to taxonomy. Furthermore, there is a wiki devoted entirely to taxonomy (Wikispecies), which includes pages for taxa, taxonomists, and taxonomic publications. Many of these pages have corresponding items in Wikidata. Hence a considerable amount of taxonomic literature has already been added by contributors to the WikiCite project.

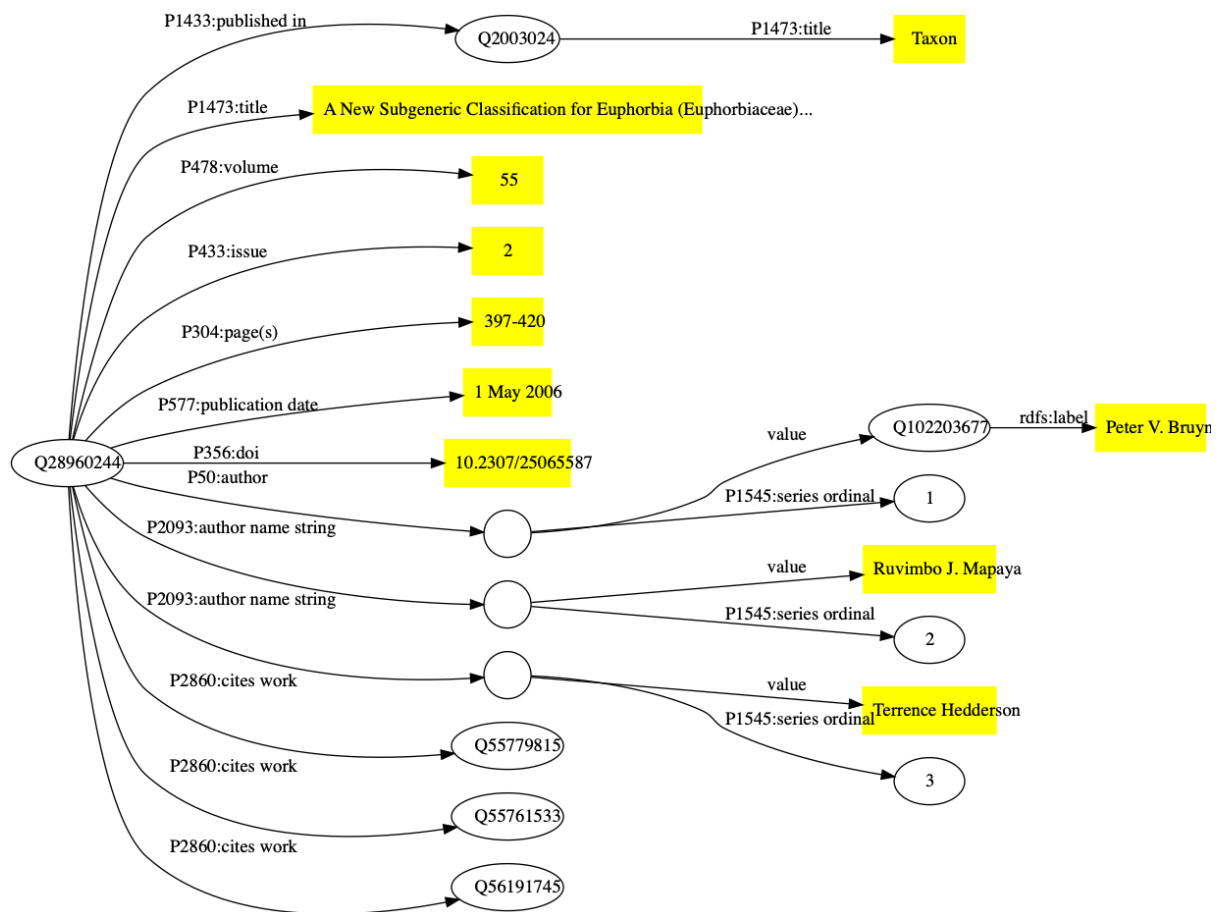
Data contributions to Wikidata typically come in two forms, manual edits by people or automated edits by software (“bots”). A number of bots add bibliographic metadata sourced from databases such as PubMed and CrossRef. For example, given a CrossRef DOI for an article CrossRef provides an API to retrieve the metadata for the corresponding article. If one wanted to include only publications cited by Wikipedia, one would then need a list of DOIs cited on Wikipedia pages. Alternatively one could proactively include DOIs even if they aren’t currently cited, on the assumption that as Wikipedia grows it is likely that more and more articles will be cited. This means it is a short step to expanding the scope of WikiCite to include most, if not all of the academic corpus in Wikidata. One motivation for this is to have openly accessible bibliographic data which can be used to enable freely accessible measures of the activity and impact of researchers (Nielsen, Mitchen & Willighagen, 2017).

As a consequence of work done by the WikiCite community, and the prominence of taxonomy in Wikipedia and Wikispecies, Wikidata already contains a considerable number of publications relevant to taxonomy. This, coupled with the sophistication of the data model, powerful query language, and the existence of an enthusiastic community of editors makes a strong case for Wikidata being a promising platform for a “bibliography of life”.

Bibliographic data in Wikidata

The Wikidata model for a publication has evolved over time as the community adds properties and recommendations for their use. Fig. BIBMODEL shows a basic model for a scientific publication (Bruyns, Mapaya & Hedderson, 2006).

BIBMODEL for Q28960244



Wikidata items are given one or more “types” using Wikidata property P31 (instance), such as Q13442814 for a scholarly article, and Q571 for a book. There are properties for the typical metadata associated with an article, such as title, journal that contains the article, volume, pagination, and date of publication. Wikidata supports values in multiple languages, so that articles with titles in multiple languages can have all those titles represented. Authorship is handled in two distinct but complementary ways. If an author of a publication is known to have a Wikidata entry then the author property (P50) links the item for the publication to the item for that author. If it is not known whether the author exists in Wikidata their name can be stored as a simple string value (P1545). In Fig. BIBMODEL there are examples of both authors. There are tools available to subsequently map those name strings to the corresponding Wikidata items.

External identifiers, such as ones provided by the publishing industry (e.g., DOIs), archiving services (e.g., Handles), and domain-specific databases (such as PubMed, ZooBank, etc.) can also be added to the Wikidata item. Wikidata items are being decorated with an

increasing number of diverse identifiers, hence Wikidata is increasingly playing a role as an “identity broker” enabling cross-links between identifiers (Veen, 2019).

Links between publications

Publications rarely exist in isolation from each other, hence we can connect them using a range of properties. The most obvious relationship is citation, where one publication cites another. Adding this information helps flesh out the citation graph, and enables us to track the provenance of an idea, as well as discover potentially related publications through co-citation (Marshakova-Shaikovich, 1973; Small, 1973).

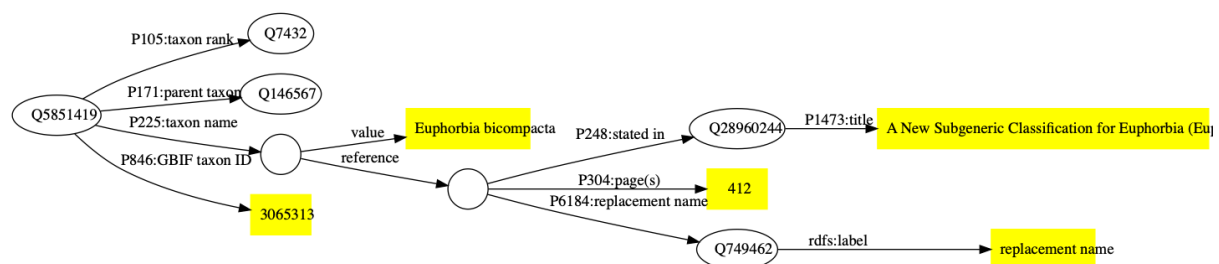
Other relationships supported by Wikidata include errata where one publication corrects errors or mistakes in a previous publication, and translations, where a publication may exist in more than one language. For example, the paper (Korotyaev, 2018) is an English translation of (Коротяев, 2018), the corresponding items in Wikidata can be connected by properties reflecting that relationship.

Links to facts

A key motivation for including publications in Wikidata is as a source of references for statements made in that database. For example, statements about the birth and death dates for a person, the exact date of publication of a work, the date at which a journal changed its name, or the publication of a taxonomic name can all be supported by adding references to the relevant source.

As an example, the taxonomic name *Euphorbia bicompecta* Bruyns was published in (Bruyns, Mapaya & Hedderson, 2006) as a replacement for the name *Synadenium compactum* N.E.Br. This publication (Q28960244) is the one discussed above in Fig. BIBMODEL. The Wikidata item for *Euphorbia bicompecta* (Q5851419) has a property “taxon name” (P225) with the value “Euphorbia bicompecta” and Wikidata item Q28960244 as a reference for that value (see Fig TAXONMODEL).

Fig. TAXONMODEL



Populating Wikidata

Creating a bibliography of life would only be conceivable if much of it could be automated, and if freely accessible sources of data were available. Bibliographic metadata from CrossRef and PubMed are being added by automated tools (“bots”). This means that publications that have a CrossRef Digital Object Identifier (DOI), or have an entry in PubMed are likely to be already in Wikidata. If they aren’t, then it is straightforward to add them. Data from these sources are typically of high quality, although this is not always the case. Sometimes the data is limited, for example, in not including lists of literature cited, or there may be typographic or character encoding errors in the data. An advantage of a community-editable resource is that these can be found and corrected by anyone.

While much of the biomedical literature, and an increasing fraction of modern taxonomic literature has CrossRef DOIs, much of the taxonomic corpus either lacks a DOI, or may have a DOI issued by a different registration agency. The DOI foundation has several members that issue DOIs, and these differ in the support they provide for resolving DOIs to machine readable data. CrossRef DOIs can return extensive metadata about an article in CiteProc JSON, a default standard for bibliographic metadata that originates in the Zotero project (<https://www.zotero.org/>). Some DOI agencies support CiteProc (albeit not as fully populated as CrossRef), however agencies such as ITISC - which is issuing DOIs for many Chinese articles (Wang et al., 2018) - do not support machine readability at all.

There are also publications with persistent identifiers that are not DOIs (such as Handles), publications which lack persistent identifiers but are online, and publications which may not be online at all. There are various strategies we can use to gather bibliographic data for these publications.

Scrape metadata from the web

Many journal web sites embed machine readable metadata about publications in their web pages to enhance discoverability by search engines such as Google Scholar. These tags also enable software tools (e.g., reference managers) to easily extract bibliographic metadata to be stored by users of those tools. Although typically there are journal and publisher-specific idiosyncrasies in how the metadata is marked up, it is relatively straightforward to write software to fetch these web pages and extract the metadata.

Lists of literature cited

Most articles will have a list of literature cited, so when taxonomists publish their work they are also continually publishing bibliographic metadata. These lists are becoming increasingly accessible to machines. CrossRef is encouraging publishers to include lists of references cited in their submissions to CrossRef. If both a cited article has a DOI, then this citation link may ultimately find its way into Open Citations (Peroni & Shotton, 2020). While this helps grow the citation network, it overlooks all those publications that lack DOIs (or which lacked them at the time the citing article was published). However, the metadata for references cited which lack DOIs can still be used to help populate Wikidata.

Some publishers provide article text in machine-readable formats such as XML where the references are identified and can be easily extracted. Other publishers may provide lists of references in the web view of an article, sometimes with embedded markup. Hence we can regard taxonomic researchers as, in effect, “crowd sourcing” the taxonomic literature simply by the act of publishing their research. For example, articles published in the journal *Zootaxa* together contain over a million citation strings (Roderic D. M. Page, 2020).

Taxonomic databases

The numerous taxonomic databases being developed by the community, often focussed on a particular taxonomic group, are yet another source of bibliographic data. Regrettably, in many cases taxonomic databases do not treat the taxonomic literature as a first class citizen, and hence the data may be stored in an abbreviated form (such as the micro-citations mentioned earlier). But some databases do provide high-quality curated literature, such as MycoBank (Robert et al., 2013).

Databases of researchers

Yet another potential source of data are the collections of articles created by researchers as part of an online profile or identity, such as ORCID or ResearchGate. Using a combination of manual input and web services, ORCID (<https://orcid.org/>) assembles a list of publications (and other outputs) linked to a researcher’s unique identifier (their ORCID id). This data is openly available via an API. In contrast, ResearchGate (<https://www.researchgate.net/>) is a commercial website where members can upload lists of their publications, and provide access to the publications themselves (on the understanding that their members have the legal right to do so). Although ResearchGate is closed in that it lacks a publicly available API, they do embed structured markup in their web pages which links authors to their publications using terms from the schema.org vocabulary.

Wikis

The sources which perhaps most closely match the notion of “crowd sourcing” are Wikidata itself, and other wikis of the Wikipedia Foundation, such as Wikipedia and Wikispecies. Indeed, in much the same way that we can regard Wikipedia as an Encyclopaedia of Life (Page, 2010), Wikispecies can be regarded as a version of the “bibliography of life” where volunteers are assembling a wiki with one page per taxon, often including extensive lists of references cited. However, these references are often entered as simple text strings with little or no structured markup, making it challenging to extract structured metadata.

Full-text

Wikidata stores metadata rather than full-text content, that is, it stores information about a publication not the contents of the publication itself. A growing proportion of the taxonomic literature is being digitised, such that articles may be available in formats such as PDF or sets of images (e.g., scans of printed works). Given the alarming ease with which links to

content can break (Laakso, Matthias & Jahn, 2020) a convention on Wikidata is to include not only a link to a freely available PDF but also a link to an archived version, e.g. on the Internet's Wayback Machine (<https://archive.org/web/>). Another strategy (one that I have regularly used) is to store a copy of the PDF on Internet Archive itself and include the Internet Archive identifier as a property of the publication on Wikidata.

There are other ways to access content. There are tools that take a DOI and return a PDF if one is available online, either freely available Unpaywall (<https://unpaywall.org/>) or “pirated” (Bohannon Apr. 28, 2016 & Pm, 2016). Some publishers such as the China National Knowledge Infrastructure (CNKI) (<https://cnki.net/>) have mobile phone apps that provide access to their content through that app.

Being able to access the content of the articles themselves not only means that we can read the article, but it also provides a way to augment existing metadata. In my own experience for a number of journals page numbers were not recorded in the available metadata. This can make it harder to link publications to taxonomic names using “microcitation”, where the only information we have is a journal, a volume, and a page number. However, if we have access to a digital version of the article we can extract the page numbers. This need not be a manual process, for instance the Internet Archive generates a file for each PDF that contains a best-guess of the page numbers in the PDF. We can use those to add missing pagination values to the corresponding Wikidata items.

Exploring Bibliographic Data in Wikidata

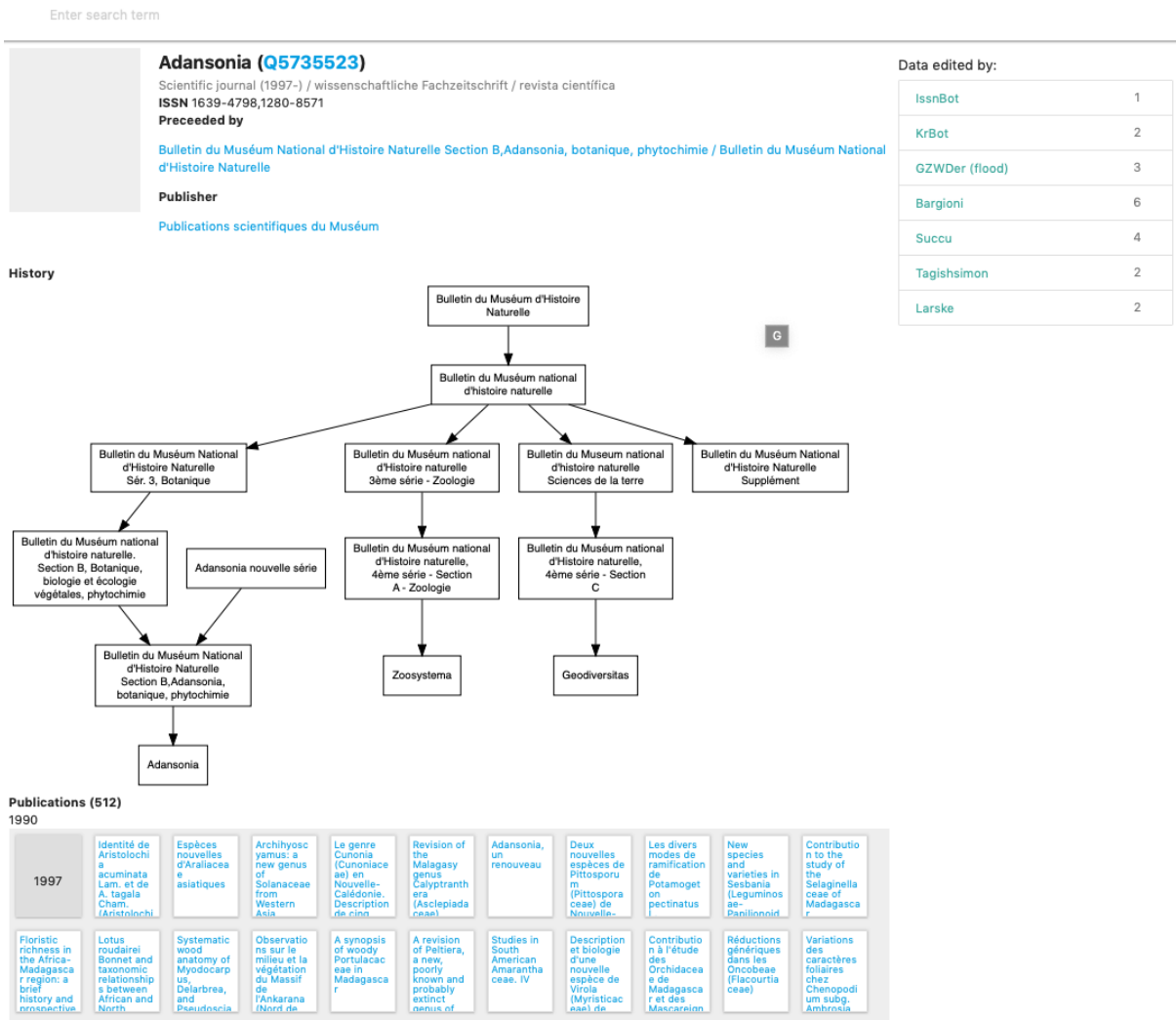
Having discussed sources of bibliographic data and how we can get that data into Wikidata, I now turn to exploring that data. First I describe a tool I developed to navigate through bibliographic (and related) data, then I present some results exploring the editing activity of the Wikidata community and density of the knowledge graph the community is building through those edits. I then look at the coverage of taxonomic literature and taxonomic authors.

User interface


The user interface of Wikidata is heavily focussed on data entry, and hence is not particularly friendly to anyone wanting to explore the knowledge accumulated in Wikidata. The underlying data can be queried using SPARQL, which is a powerful but challenging language to use. Given this, a number of more accessible tools have emerged, including generic tools such as resonator (<https://reasonator.toolforge.org>), and more focussed tools such as Scholia (Nielsen, Mietchen & Willighagen, 2017). The latter provides a wealth of visualisations for a publication, including its citation network, major topics, and the network of connections amongst a publication's authors.

To complement existing tools I have developed a simple website called ALEC (<https://alec-demo.herokuapp.com>) to make it easy to find and view publications, links






between publications, and links between people, publications, and other entities such as taxa. The primary motivation was a tool that could display gaps in coverage in Wikidata, for example, by displaying a list of all articles in a journal.



Enter search term



Victoria Ann Funk (Q19060876)
US-amerikanische Botanikerin (1947-2019) / botánica y brióloga estadounidense / American botanist (1947-2019)
26 Nov 1947 - 22 Oct 2019

Publications by (110)
1970

1978

Cladistics for the Practicing Plant Taxonomist

1980

1980

Polyploidy in Montanoa Cerv. (Compositae, Heliantheae)

1983

Cladistic Analysis of Complex Natural Products: Developing Transformation Series from Sesquiterpenes

Advances in Cladistics

Sesquiterpene Lactones as Taxonomic Characters in the Asteraceae

1982. A Monograph of the Fern Genus Platycerium (Polypodiaceae)

1984

RECIENTES CROMOSOMOS EN COMPOSITAE DE COLOMBIA

Advances in Cladistics, Volume 2

The Systematics of Montanoa (Asteraceae, Heliantheae)

1985

Preface

Cladistics: Perspectives on the Reconstruction of Evolutionary History

Phylogenetic Patterns and Hybridization

Cladistics and Generic Concepts in the Compositae

1986

A phylogenetic analysis of the Orchidaceae

Cladistics. A Reply

A phylogenetic analysis of the Orchidaceae

1988

Griffiths' "Cacti Glass Negatives" Collection

1989

A bibliography of plant collectors in Bolivia


Data edited by:

Local	1
Liridon	1
GeertivpBot	9
Vojtěch Dostál	1
Uncommon fritillary	7
Dcflyer	1


Vicki Ann Funk (November 26, 1947 – October 22, 2019) was a Senior Research Botanist and Curator at the Smithsonian's National Museum of Natural History, known for her work on members of the composite family (Asteraceae) including collecting plants in many parts of the world, as well as her synthetic work on phylogenetics and biogeography.[1][2][3] (from Wikipedia)

Fig. Screenshot of how ALEC displays information for a person, in this case the late Vicki Funk. In addition to data from Wikidata, such as external identifiers and publications authors the web page includes a text summary retrieved from DBpedia.

Enter search term



Sulawesimetopus henryi (Q79381426)
 species of insect



Data edited by:

YoaR	1
SuccuBot	15
Edoderoobot	1
BotNinja	1
Myrmoteras	2

Sulawesimetopus > Sulawesimetopus henryi

Reference (1)

Sulawesimetopus henryi, a new genus and species of Isometopinae (Hemiptera, Heteroptera, Miridae) from Sulawesi


1 Jan 2018

DOI: 10.3897/ZOOKEYS.796.21273

Fig. Screenshot of a taxon in Wikidata that is linked to the publication that originally described the species.

If an article has an Internet Archive identifier ALEC embeds the Internet Archive viewer in the web page for that article. If the article has a PDF that has been archived by the Wayback Machine then ALEC displays a link to open that PDF using the PDF.js viewer (<https://mozilla.github.io/pdf.js/>). If the article has a CNKI identifier then ALEC displays a QR code that can be opened by CNKI mobile phone apps.


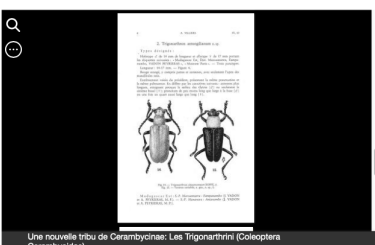
Enter search term




Une nouvelle tribu de Cerambycinae: Les Trigonarthrini (Coleoptera Cerambycidae). (Q105118008)
 A Villiers

Published 1 Jan 1983 in *Bulletin de l'Institut Royal des Sciences Naturelles de Belgique, entomologie* 55 (10) pages 1-8

[VIEW](#)



Enter search term



悼念福建省昆虫分类区系研究开拓者马骏超先生 / IN MEMORY OF THE LATE PROF. MAA TSING-CHAO (T.C. MAA), THE PIONEERING EXPLORER OF FUJIAN INSECT FAUNA (Q89665527)
 Hsu-tu Chao

Published 1 May 1993 in *Wuyi xueshu / 武夷科学 / Wuyi Science Journal* 10 pages 1-3

[CNKI:WYKX199301001](#)

Subject(s)

Tsing-Chao Maa

Fig. Screenshots of two articles displayed in ALEC. (a) Item Q105118008 shows a link to a copy of the PDF in the Wayback Machine, and an embedded Internet Archive viewer displaying the article. (b) Item Q89665527 displays a QR code that can be read by an app from CNKI that displays full text content.

Source code for ALEC is available on GitHub. The implementation makes use of the SPARQL interface to Wikidata. CONSTRUCT queries are used to retrieve information about specific entities, such as a publication or a person. Information about these entities is expressed using terms from the <http://schema.org> vocabulary. Lists of entities, such as articles in a journal or articles cited by a publication, are retrieved as schema.org DataFeeds (analogous to a RSS feed). These RDF documents are expressed as JSON-LD and are then converted into HTML using simple Javascript templates. If images are available these are displayed as thumbnails. To find Wikidata items ALEC supports basic search via Wikidata's API. If the search term looks like an identifier (such as a DOI or an ISSN then ALEC will use a SPARQL query in an attempt to find the corresponding item. In future more sophisticated search tools could be added that are more specific to the task of searching bibliographic data.

A community of editors

One of the challenges in community-based editing of scientific data is assembling that community. One approach is to create a domain-specific database and hope a community coalesces around that database. Another is to take the data to where a community already exists. This is the approach taken by projects such as Gene Wiki (Good et al., 2012).

Hence we can ask “does the community edit taxonomic publications?” To assess this I looked at the edit history of a sample of Wikidata items for papers that have published scientific names. For each item I retrieved the number of edits made since the record was created, when those edits were made, and what properties were edited.

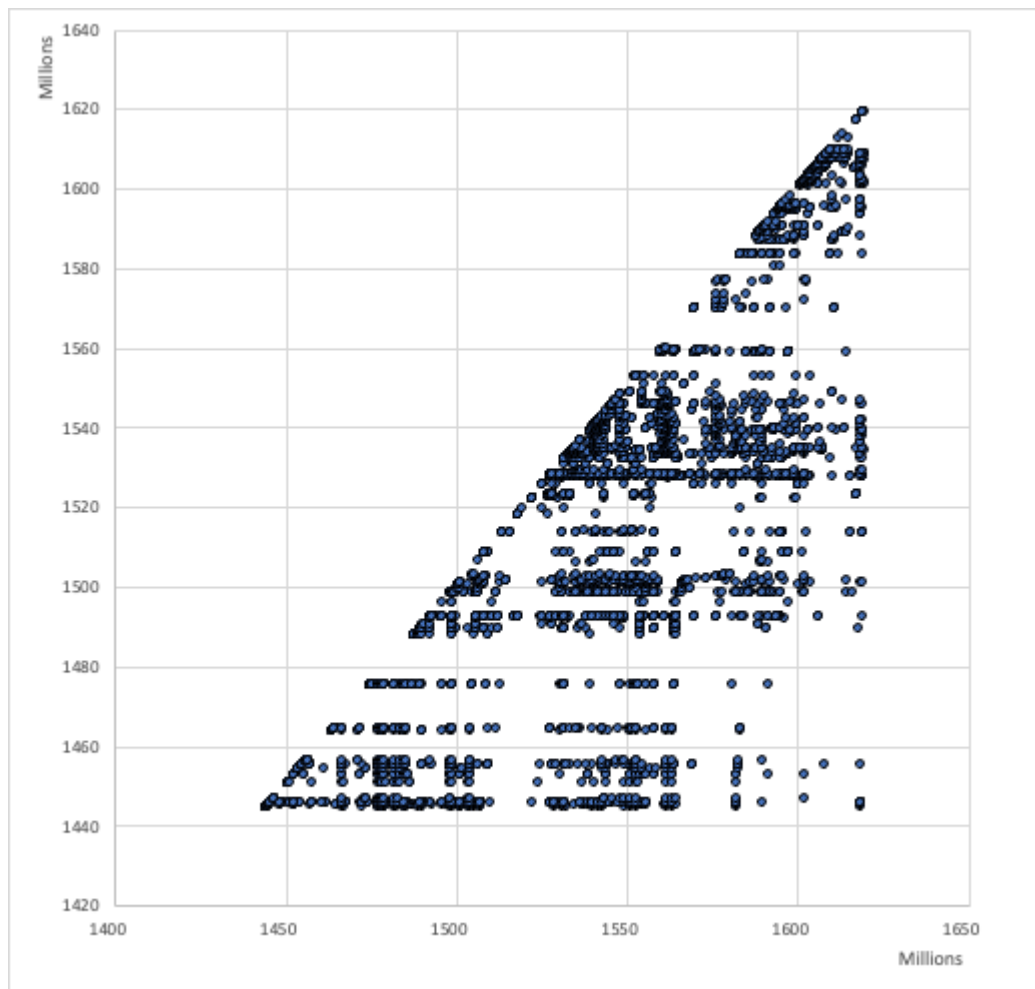


FIG TRIANGLE visualises the edit history for a sample of 1000 publications in Wikidata as a scatter plot of creation timestamp against edit timestamp. If an item was only edited at the time it was created then all points would fall along the diagonal. This diagonal continues to go up and to the right as time goes on. Any edit to an item appears as a dot to the right of the dot on the diagonal representing the item's creation. If there are no dots to the right then an item has not been edited since its creation. What we see is that most items undergo a series of sporadic edits over time. Some of these edits occur shortly after item creation. For example there are bots whose function is to add a description for a new item in a specific language. Other edits may happen much later, for example if a user associates a publication with its author, or links a publication to its main subject. Or there may be a bulk update of many items, for example when a bot that edits a specific property for many items. Of the properties that are edited the top ten properties of publication items in Wikidata are shown in TABLE PROP.

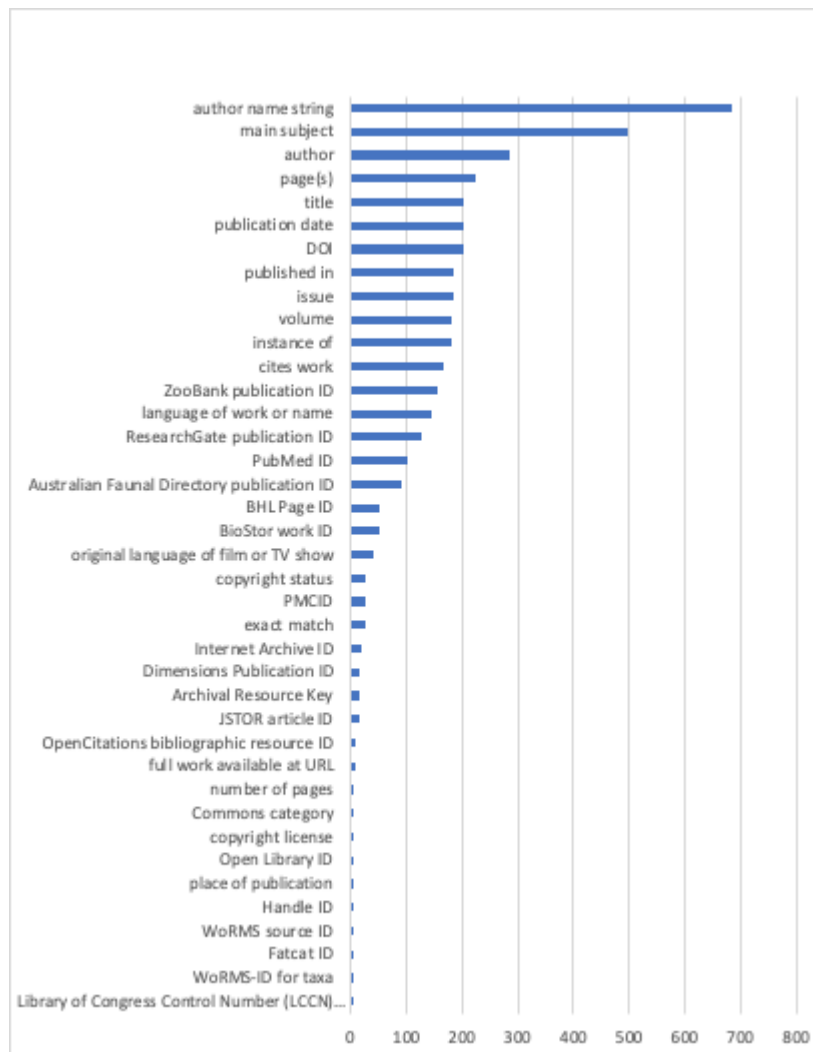


FIG Number of edits for each property in a sample of bibliography items in Wikidata.

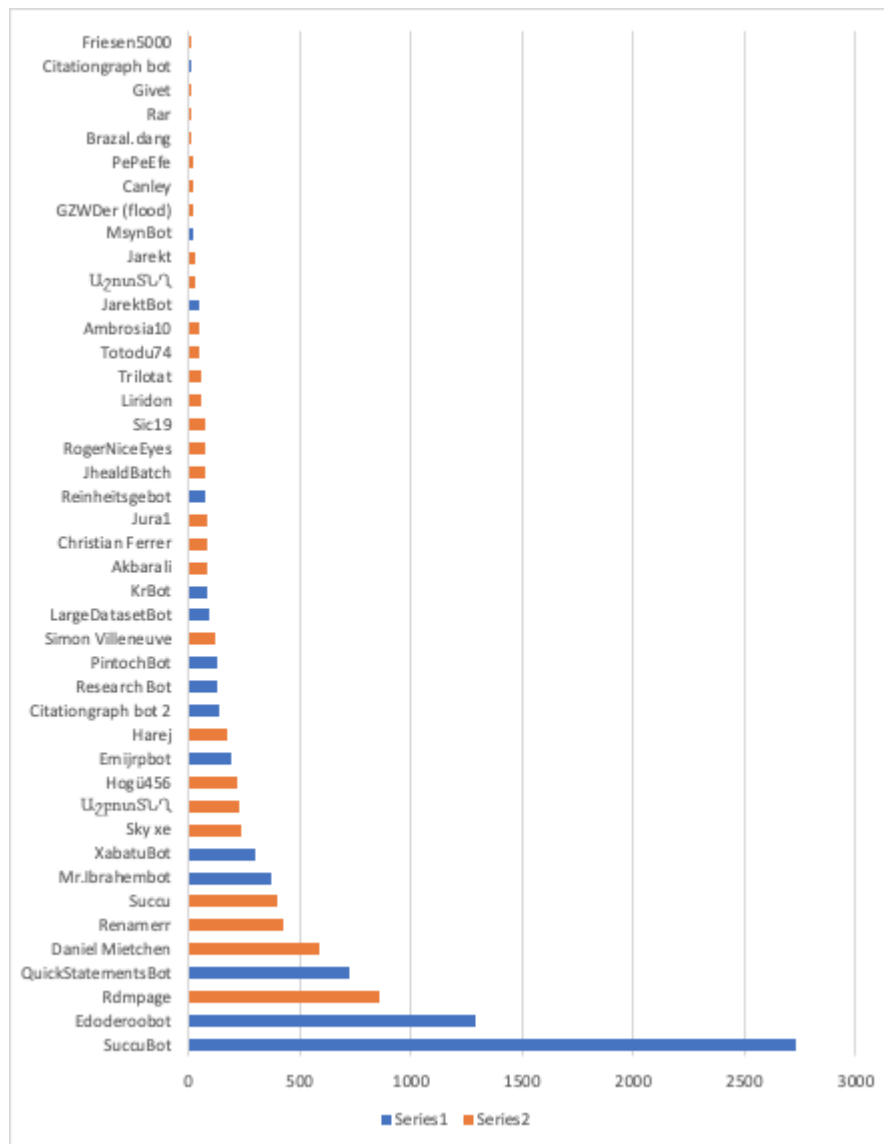


FIG Plot of number of edits of bibliographic items by user. For the sample of 1000 bibliographic items the number of edits by each user is shown. Users are colour-coded by whether they are bots or people.

Edits in Wikidata can be made by people, either directly by editing a record, or using bulk tools such as Quickstatements (<https://quickstatements.toolforge.org/>). Edits can also be made by automated programs (“bots”). Of the top ten editors of publications, half are bots.

This approach to measuring edit activity assumes that only edits made to an item itself are relevant, however edits may be made to other items that link those items to the current item. For example, adding a “cites work” statement to an item does not result in any changes to the item being cited, but does affect the Wikidata knowledge graph.

Knowledge graph density

Conceptually a knowledge graph comprises entities (nodes) that are connected by facts (edges). The number of facts for an entity is a measure of the knowledge graph's density, which for many graphs is low, often averaging less than two facts per entity (Hegde & Talukdar, 2015). Note that this definition of "facts" ignores simple statements associated with an entity (e.g., the number of pages in an article). These are also facts, but we don't need a knowledge graph to store them. The true power of a knowledge graph comes from the density of the connections between entities.

To assess the density of bibliographic data in Wikidata I counted the number of links between bibliographic items and other Wikidata entities. In counting connections some entities, such as those for language, were not counted so as to avoid inflating the estimate of knowledge density based on what are essentially administrative metadata.

The average link density for the sample of publications was 4.24, with the modal number of connections being one. Hence this part of the knowledge graph is sparse with most publications having just the connection to a parent publication (typically a journal). Some publication items are connected to other items via citation relationships, either as the source or the target of that relationship (i.e., citing or cited by). Less than a third of the publications were connected to an author.

Table: Properties that represented links between Wikidata items in the sample of taxonomic publications.

Property	Frequency
published in	980
cites work	2844
author	288
place of publication	3
stated in	80
sponsor	20
part of	2
publication in which this taxon name was established	2

Taxonomic coverage

The primary motivation for this project is to link every taxonomic name for eukaryotes to its original description using a unique identifier (e.g., a DOI) and ideally a link to a digitised version of that publication. The scale of the challenge was discussed in (Page, 2016a), and an attempt to do this for animal names led to BioNames (Page, 2013). Similar work for plants is mostly unpublished, although preliminary data has been released as Plant Names GBIF, IPNI datasets (Page, 2018) and raw data dumps (“IPNI plant names linked to bibliographic identifiers,” 2020). My work on Index Fungorum is unpublished. The bulk of the articles I have added to Wikidata using the approaches described above are taxonomic publications (HOW DO I GET THIS COUNT?).

Database	Number of taxonomic names with publications	Number of names linked to publication in Wikidata	Coverage (%)
ION/BioNames	1693166	341333	20.1
IPNI	1667909	312822	18.8
Index Fungorum	436039	69186	15.9

```
SELECT COUNT(id) FROM names WHERE publication IS NOT NULL;
SELECT COUNT(id) FROM names WHERE wikidata IS NOT NULL;
```

```
SELECT COUNT(Id) FROM names WHERE Publication IS NOT NULL;
SELECT COUNT(Id) FROM names WHERE wikidata IS NOT NULL;
```

```
SELECT COUNT(Id) FROM names_indexfungorum WHERE Title IS NOT NULL;
SELECT COUNT(Id) FROM names_indexfungorum WHERE wikidata IS NOT NULL;
```

Author coverage

The bulk of publications added to Wikidata treat authors as “strings” not “things”, that is, most authors are listed as names using the P2093 “authors name string” property, rather than as Wikidata items using the P50 “author” property (see Fig. BIBMODEL). Ideally all authors of publications would be Wikidata items not strings. Realising that goal requires that all authors of taxonomic publications have items in Wikidata.

There are several databases of taxonomists that have representation in Wikidata, although coverage is variable. For example, the International Plant Names Index (IPNI) contained approximately 43,000 authors in 2013 (Lindon et al., 2015), currently some 53,073 Wikidata items have IPNI author ids. At the time of writing (2021) ZooBank (Pyle & Michel, 2008) contains some 79,000 authors, of which 15,833 are in Wikidata. The BHL has 26,371 authors in Wikidata, while Wikispecies contributes 56,994 authors to Wikidata. Overlap between these sources is variable. For example, almost all of the ZooBank authors that are in Wikidata are also in Wikispecies, whereas the majority of authors sourced from IPNI are

unique to IPNI (Fig. CLUSTERS). What is unclear is how much of the lack of overlap between authors in the different sources databases is real, versus a lack of mapping between identifiers. In other words, it is possible that the same person exists in BHL and IPNI, but has two records in Wikidata, one for each identifier. There is considerable scope for reconciling authors between these databases, as well as other sources of information on people, such as ORCID and ResearchGate.

It is not enough to merely have authors represented in Wikidata, we also need to link them to their publications. This is a major undertaking (cross-reference the fact that it is being done). The source databases (BHL, IPNI, Wikispecies, and ZooBank) all contain links between authors and their publications, and much more use could be made of these sources to add P50 author links (Page, 2019).

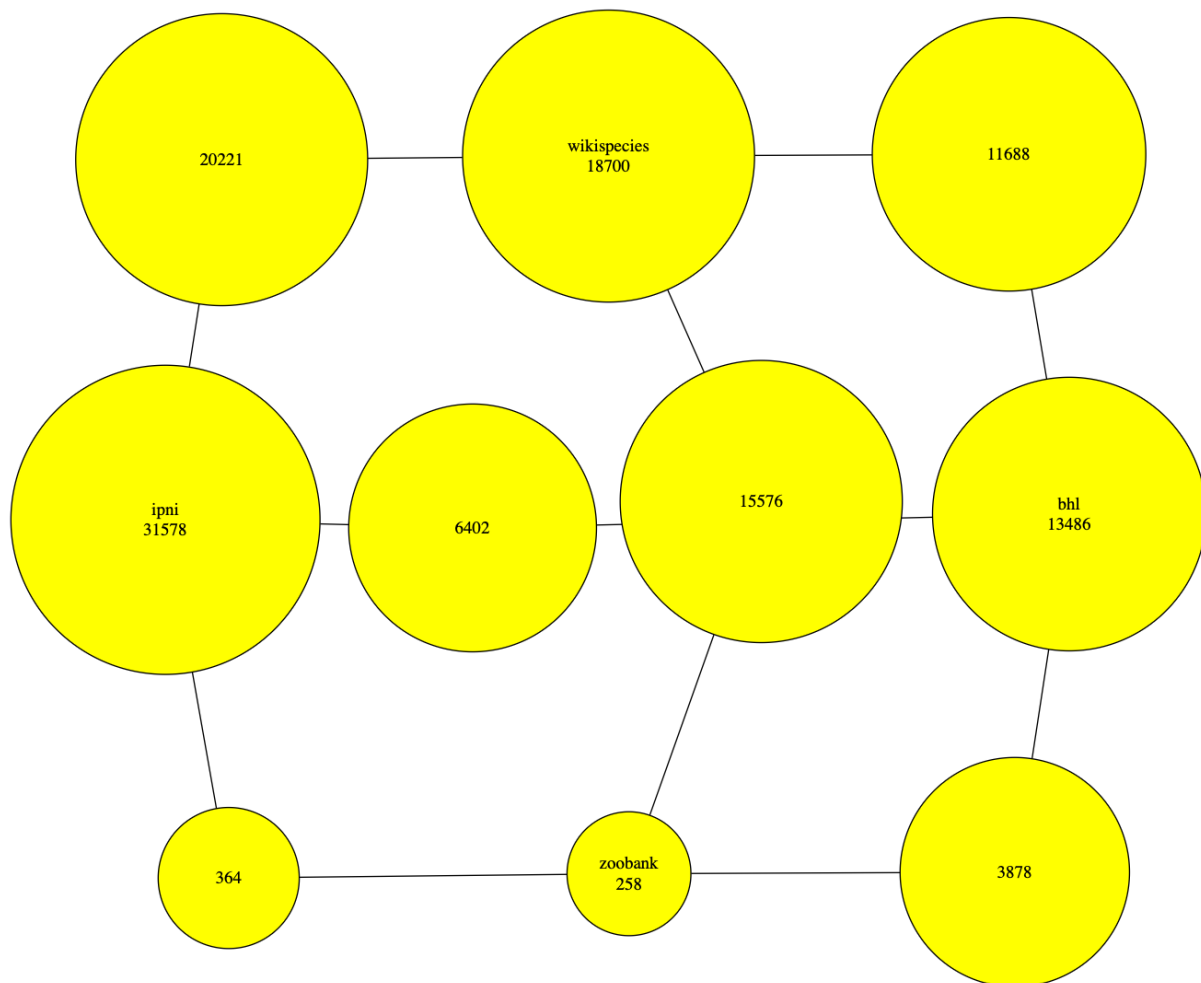


Fig. CLUSTERS Contribution of four different data sources (BHL, IPNI, Wikispecies, and ZooBank) to authors in Wikidata. The nodes labelled by source comprise authors unique to that database. The authors shared by each pair of data sources are represented by the node on the path between the pair of sources.

Discussion

By providing a robust, open platform for community editing of structured data, Wikidata seems like the ideal platform for the “bibliography of life”. It not only benefits from a community of active editors, it piggy backs on the remarkable fact that taxonomy is the only discipline to have its own Wikimedia Foundation project (Wikispecies). This means that a large number of taxonomic works and their authors already exist in Wikidata. As more and more taxonomic publications acquire DOIs, and as more taxonomists acquire ORCID ids the taxonomic literature component of Wikidata will automatically grow because content linked to these identifiers is routinely harvested by Wikidata bots. This leaves a large fraction of the taxonomic literature to be added by other means, but as discussed here there are numerous ways to do that. It is not unreasonable to expect that the bulk of the taxonomic literature will find its way into Wikidata in the next few years.

Being a community project Wikidata has a number of quirks. It is possible for people to create multiple Wikidata items for the same thing, although there is a simple mechanism for merging such duplicates. The way Wikidata models a given class of entities (such as “taxa” or “books”) is determined on an ad hoc basis by a self-assembling community of interested people. This can lead to multiple ways to do the same thing, which presents challenges to both editing and querying the data. Tools such as ALEC described here often need quite elaborate SPARQL queries to accommodate this multiplicity of ways to represent essentially the same information.

In the context of biological taxonomy perhaps the greatest limitation of Wikidata is the way it models taxa and their names. Ideally these would be separate entities, but in Wikidata (in common with many taxonomic databases) names and taxa are conflated. This makes it difficult to adequately model the relationship between taxa, names, and publications. Whether the existing model can be improved will have a major impact on the broader taxonomic utility of Wikidata.

Acknowledgements

References cited

- Aspesi C, Brand A. 2020. In pursuit of open science, open access is not enough. *Science* 368:574–577. DOI: 10.1126/science.aba3763.
- BohannonApr. 28 J, 2016, Pm 2:00. 2016. Who’s downloading pirated papers? Everyone. *Science* | AAAS. DOI: 10.1126/science.aaf5664.

- Bruyns PV, Mapaya RJ, Hedderson TJ. 2006. A new subgeneric classification for Euphorbia (Euphorbiaceae) in southern Africa based on ITS and psbA-trnH sequence data. *TAXON* 55:397–420. DOI: <https://doi.org/10.2307/25065587>.
- Cameron RD. 1997. A Universal Citation Database. *First Monday*. DOI: 10.5210/fm.v2i4.522.
- Egloff W, Agosti D, Kishor P, Patterson D, Miller J. 2017. Copyright and the Use of Images as Biodiversity Data. *Research Ideas and Outcomes* 3:e12502. DOI: 10.3897/rio.3.e12502.
- Erxleben F, Günther M, Krötzsch M, Mendez J, Vrandečić D. 2014. Introducing Wikidata to the Linked Data Web. In: Mika P, Tudorache T, Bernstein A, Welty C, Knoblock C, Vrandečić D, Groth P, Noy N, Janowicz K, Goble C eds. *The Semantic Web – ISWC 2014*. Lecture Notes in Computer Science. Cham: Springer International Publishing, 50–65. DOI: 10.1007/978-3-319-11964-9_4.
- Garfield E. 2001. Taxonomy is small, but it has its citation classics. *Nature* 413:107–107. DOI: 10.1038/35093267.
- Good BM, Clarke EL, de Alfaro L, Su AI. 2012. The Gene Wiki in 2011: community intelligence applied to human gene annotation. *Nucleic Acids Research* 40:D1255–D1261. DOI: 10.1093/nar/gkr925.
- Gusenleitner FJ [Friedrich], Malicky M. 2017. Die Datenbank www.ZOBODAT.at als Recherchewerkzeug für biologische und erdwissenschaftliche Forschung. *Linzer Biologische Beiträge* 49:1163–1208.
- Hamilton CA, Shockley FW, Simmons R, Smith A, Ware J, Zaspel JM. 2021. The Future for a Prominent Taxonomy. *Insect Systematics and Diversity* 5. DOI: 10.1093/isd/ixaa020.
- Hegde M, Talukdar PP. 2015. An Entity-centric Approach for Overcoming Knowledge Graph Sparsity. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, 530–535. DOI: 10.18653/v1/D15-1061.

IPNI plant names linked to bibliographic identifiers. 2020. DOI:

10.6084/m9.figshare.13055804.v1.

King D, Morse D, Willis A, Dil A. 2011. Towards the bibliography of life. *ZooKeys*

150:151–166. DOI: 10.3897/zookeys.150.2167.

Korotyaev BA. 2018. Two New Species of the Weevil Genus *Mecysmoderes* Schoenherr,

1837 (Coleoptera, Curculionidae: Ceutorhynchinae) from Vietnam. *Entomological*

Review 98:899–906. DOI: 10.1134/S0013873818070114.

Laakso M, Matthias L, Jahn N. 2020. Open is not forever: a study of vanished open access

journals. *arXiv:2008.11933 [cs]*.

Lindon HL, Gardiner LM, Brady A, Vorontsova MS. 2015. Fewer than three percent of land

plant species named by women: Author gender over 260 years. *TAXON* 64:209–215.

DOI: <https://doi.org/10.12705/642.4>.

Marshakova-Shaikovich I. 1973. System of Document Connections Based on References.

Scientific and Technical Information Serial of VINITI 6:3–8.

Nielsen FÅ, Mietchen D, Willighagen E. 2017. Scholia, Scientometrics and Wikidata. *Lecture*

Notes in Computer Science:237–259. DOI: 10.1007/978-3-319-70407-4_36.

Page RD. 2009. bioGUID: resolving, discovering, and minting identifiers for biodiversity

informatics. *BMC Bioinformatics* 10:S5. DOI: 10.1186/1471-2105-10-s14-s5.

Page RDM. 2010. Wikipedia as an encyclopaedia of life. *Organisms Diversity & Evolution*

10:343–349. DOI: 10.1007/s13127-010-0028-9.

Page RD. 2011. Extracting scientific articles from a large digital archive: BioStor and the

Biodiversity Heritage Library. *BMC Bioinformatics* 12:187. DOI:

10.1186/1471-2105-12-187.

Page RDM. 2013. BioNames: linking taxonomy, texts, and trees. *PeerJ* 1:e190. DOI:

10.7717/peerj.190.

Page RD. 2016a. Surfacing the deep data of taxonomy. *ZooKeys*:247.

Page RDM. 2016b. DNA barcoding and taxonomy: dark taxa and dark texts. *Phil. Trans. R.*

Soc. B 371:20150334. DOI: 10.1098/rstb.2015.0334.

Page R. 2018. Liberating links between datasets using lightweight data publishing: an example using plant names and the taxonomic literature. *Biodiversity Data Journal*. DOI: 10.3897/BDJ.6.e27539.

Page RDM. 2019. Reconciling Author Names in Taxonomic and Publication Databases. In: *Proceedings of the 12th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences (SWAT4HCLS)*. CEUR Workshop Proceedings. Edinburgh, Scotland, 36–43.

Peroni S, Shotton D. 2020. OpenCitations, an infrastructure organization for open scholarship. *Quantitative Science Studies* 1:428–444. DOI: 10.1162/qss_a_00023.

Pyle RL, Michel E. 2008. ZooBank: Developing a nomenclatural tool for unifying 250 years of biological information. *Zootaxa* 1950:39–50. DOI: 10.11646/zootaxa.1950.1.6.

Robert V, Vu D, Amor ABH, Wiele N van de, Brouwer C, Jabas B, Szoke S, Dridi A, Triki M, Daoud S ben, Chouchen O, Vaas L, Cock A de, Stalpers JA, Stalpers D, Verkley GJM, Groenewald M, Santos FB dos, 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 DL, Taylor JW, Crous PW. 2013. MycoBank gearing up for new horizons. *IMA Fungus* 4:371–379. DOI: 10.5598/imafungus.2013.04.02.16.

Roderic D. M. Page. 2020. Zootaxa articles cited as TSV file. DOI: 10.6084/m9.figshare.12630773.

Shotton D. 2013. Publishing: Open citations. *Nature News* 502:295. DOI: 10.1038/502295a.

Small H. 1973. Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science* 24:265–269. DOI: <https://doi.org/10.1002/asi.4630240406>.

The Europe PMC Consortium. 2015. Europe PMC: a full-text literature database for the life sciences and platform for innovation. *Nucleic Acids Research* 43:D1042–D1048. DOI: 10.1093/nar/gku1061.

Veen T van. 2019. Wikidata: From “an” Identifier to “the” Identifier. *Information Technology*

and Libraries 38:72–81. DOI: 10.6017/ital.v38i2.10886.

Vrandečić D, Krötzsch M. 2014. Wikidata. *Communications of the ACM* 57:78–85. DOI: 10.1145/2629489.

Wang W, Deng L, You B, Zhang P, Chen Y. 2018. Digital object identifier and its use in core Chinese academic journals: A Chinese perspective. *Learned Publishing* 31:149–154. DOI: 10.1002/leap.1137.

Wanger R, Ehrismann M. 2016. *E-Periodica: die Plattform für digitalisierte Schweizer Zeitschriften*. Jülich: Forschungszentrum Jülich Zentralbibliothek, Verlag.

Коротяев БА. 2018. Два новых вида долгоносиков рода *tecysmoderes* schoenherr, 1837 (coleoptera, curculionidae: ceutorhynchinae) из Вьетнама. *Энтомологическое Обозрение* 97. DOI: 10.1134/S0367144518030115.