Wikidata and the bibliography of life

Deleted:

Roderic D. M. Page[1]
[1]IBAHCM, MVLS, University of Glasgow, Glasgow G12 8QQ, United Kingdom
https://orcid.org/0000-0002-7101-9767

Corresponding author
Roderic D. M. Page
IBAHCM, MVLS, University of Glasgow, Glasgow G12 8QQ, United Kingdom

Abstract

Biological taxonomy rests on a long tail of publications spanning nearly three centuries. Not only is this literature vital to resolving disputes about taxonomy and nomenclature, for many species it represents a key source - indeed sometimes the only source - of information about that species. Unlike other disciplines such as biomedicine, the taxonomic community lacks a centralised, curated literature database (the "bibliography of life"). This paper argues that Wikidata can be that database as it has flexible and sophisticated models of bibliographic information, and an active community of people and programs ("bots") adding, editing, and curating that information.

Deleted:

Deleted: The paper also describes a tool to visualise and explore bibliography information in Wikidata and how it links to both taxa and taxonomists.

Introduction

Much of the primary data about the planet's biodiversity is contained in the taxonomic literature, a corpus that dates to the eighteenth century. Whereas other biological disciplines have created substantial bibliographic databases, such as PubMed ("PubMed"), and open access repositories for work sponsored by specific funding agencies and charities agencies, such as Europe PubMed Central (The Europe PMC Consortium, 2015), the taxonomic literature mostly lingers in relative obscurity (Page, 2016a) . 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) (Gwinn & Rinaldo, 2009) to extensive, regional repositories such as the Zoological-Botanical Database (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 recent content so that BHL is no longer "legacy only". A complementary initiative, the Biodiversity Literature Repository (BLR) 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 (Wanger & Ehrismann, 2016) .

Deleted: dates back to

Deleted:

Deleted: PMC

Deleted: ZOBODAT

Deleted: cite

Digitisation greatly increases the accessibility, but not necessarily the discoverability of content. The Biodiversity Heritage Library has scanned volumes for many journals, but unless articles contained within those volumes are indexed those articles will be difficult to

Deleted: BHL

find. This was 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.

Deleted: is

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 taxonomy has some highly visible journals, there is a long tail of taxonomic publication in small, often obscure journals (Page, 2016c). Not only does lack of 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 (Werner, 2006), 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.

Deleted: ¶
¶

Deleted: the field

The challenge of discoverability is not unique to taxonomic literature. There have been long standing calls for what Cameron (1997) described as a “universal citation database”. Recent developments such as the OpenCitations infrastructure (Peroni & Shotton, 2020) and the WikiCite project (“WikiCite”) 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 (“Wikidata”), an open database of structured information (Waagmeester et al., 2020). 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 leads to the possibility of more transparent analytics for the discipline.

Deleted: Open Citations Corpus (Shotton, 2013)

Deleted: opens

In this paper I make the case for Wikidata as the logical venue for a global database of taxonomic literature, the so-called “bibliography of life” (King et al., 2011). Given that this

Deleted:

may exclude much of the literature on medicine, agriculture, genomics, etc. this may seem an overly narrow definition of what constitutes a bibliography of life. But I argue that the term is justified given the taxonomic breadth of such a bibliography. The effort devoted to studying different taxa is very uneven, such that in species-rich groups such as Coleoptera (beetles) an individual species may be the subject of a publication only once every 100 years (May, 1988). This uneven coverage is only likely to increase with the growing importance of citizen science (Troudet et al., 2017) and the increasing dominance of research on model organisms (Farris, 2020). For many species the taxonomic literature will be the best (possibly the only) source of published information on that species, hence arguably the only database that could claim to be a bibliography of life is one that includes all taxa, that is, it includes all the taxonomic literature.

Deleted: (May, 1988; Troudet et al., 2017)(May, 1988)

In this paper I make the case for constructing the bibliography of life using Wikidata. I begin by providing background on Wikidata, describing how it models bibliographic data, and how it can be 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.

Deleted: . 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"). 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 particular 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", the second author has the value "2", and so on.

Formatted: Font: Italic

Formatted: Font: Italic

Formatted: Font: Italic

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 (SPARQL Protocol and RDF Query Language) endpoint that enables anyone to query the data.

Formatted: Font: Not Bold

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 themselves 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.

Deleted: ,

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 also 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 individual 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 the CrossRef API can be used 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 add articles with DOIs to Wikidata even if they aren't currently cited on Wikipedia, 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 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, Mietchen & 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. Figure 1 shows how a scientific article can be modelled in Wikidata.

[Fig 1 here]

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 (P2093). In Fig. 1 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 from different databases (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, enables us to track the provenance of an idea, and also discover potentially related publications through co-citation (Marshakova-Shaikevich, 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 to provide trustworthy sources of references for statements made in Wikidata. 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 bicompacta* Bruyns was published in Bruyns et al. (2006) as a replacement for the name *Synadenium compactum* N.E.Br. This publication (Q28960244) is the one discussed above in Fig. 1. The Wikidata item for *Euphorbia* *bicompacta* (Q5851419) has a property “taxon name” (P225) with the value “Euphorbia bicompacta” and Wikidata item Q28960244 as a reference for that value (see Fig 2).

[Fig. 2 here]

Populating Wikidata

Creating a bibliography of life would only be conceivable if much of the work of populating it could be automated, and if freely accessible sources of data were available. Bibliographic metadata from CrossRef and PubMed are constantly being added by automated tools (“bots”). This means that many 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 sometimes the data is limited or incorrect, 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 subsequently corrected by the community.

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 registration agency other than CrossRef. 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 (Willighagen, 2019; Bennett, 2021). 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. Hence not all DOIs are equally easy to work with.

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

Formatted: Font: Italic

Deleted:), or

publications. Below I describe some of these strategies. Source code for some of these approaches is available at https://github.com/rdmpage/wikidata-bibliographic-data .

Scrape metadata from the web

Web sites for some journals contain embedded 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 such as Zotero) 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. Furthermore, CrossRef is encouraging publishers to include lists of references cited in their submissions to CrossRef. If both the citing article and the cited article have Crossref DOIs, then this citation link may ultimately find its way into COCI, the OpenCitations index of open Crossref DOI-to-DOI citations (Heibi, Peroni & Shotton, 2019). 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.

Deleted: If both a cited article has a DOI

Deleted: Open Citations (Peroni & Shotton, 2020)

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 taxonomists 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 references cited (Page, 2020a).

Deleted: Hence

Formatted: Font: Italic

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 above). But some databases do provide high-quality curated literature which can be used to help populate Wikidata.

Deleted: first class

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 (“ORCID”) or ResearchGate (“ResearchGate”). Using a combination of manual input and web services, ORCID 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 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.

Deleted: (“ORCID”)

Deleted: (“ResearchGate”)

Wikis

Formatted: Font: Arial, 14 pt

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 crowd sourced “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, and hence limiting the utility of Wikispecies.

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 online 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 ("Wayback Machine"). 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.

Other ways to access content include tools that take a DOI and return a PDF if one is available online, either freely available, e.g. Unpaywall ("Unpaywall") or "pirated" (Bohannon, 2016). Some publishers such as the China National Knowledge Infrastructure (CNKI) 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 key data such as page numbers were often not recorded in the available metadata for an article. This can make it harder to link publications to taxonomic names using "microcitations", 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

A key goal for the bibliography of life is to be able to link every taxonomic name for eukaryote species 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 this challenge was discussed in (Page, 2016a) , and an attempt to do this for animal names led to my BioNames project (Page, 2013). I have done similar work for plants and fungi based on the International Plant Name Index (IPNI) ("International Plant Names Index") and Index Fungorum ("Index Fungorum Home Page"), a subset of which has been released on GBIF (Page, 2016b), and published as a both a "datasette" (Page, 2018) and raw data dumps (Page, 2020). Based on this work across animals, plants, and fungi, a little under 4 million taxonomic names have

Deleted: .

Deleted: There are o

Deleted: . There are

Deleted: for a number of journals

Formatted: Heading 2, Line spacing: 1.5 lines

Deleted: ¶
Taxonomic coverage¶

Moved (insertion) [1]

Deleted: The primary motivation for this project is to be able to link every taxonomic name for eukaryote species 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 this challenge was discussed in (Page, 2016a), and an attempt to do this for animal names led to my BioNames project (Page, 2013). I have done similar work for plants, although this is mostly unpublished. Preliminary data has been released on GBIF (Page, 2016), as a "datasette" (Page, 2018), and raw data dumps (Page, 2020b). My work on Index Fungorum is currently unpublished. Typically 20-40% of publications have been mapped to one or more identifiers, but only 15-20% of the publications currently exist in Wikidata.¶
¶
[Table 2 here]¶

associated bibliographic metadata (Table 1), such as a citation to a publication or a page in a publication. Depending on taxonomic group, anywhere between 20-40% of those citations have been mapped to an external identifier such as a DOI, and some 16-25% of taxonomic names have their associated publication in Wikidata. The 880,000 links between names and Wikidata publication items correspond to just under 200,000 distinct publications. A random sample of 10,000 of these publications was used in the analyses described below.

[Table 1 here]

A community of editors

One of the challenges in community-based editing of scientific data is assembling that community. We could create a domain-specific database and hope a community coalesces around that database. Alternatively, we take the data to where an active community already exists. This is the approach taken by projects such as Gene Wiki (Good et al., 2012). If Wikidata is going to be the place to assemble the bibliography of life, a natural question is "does the community actually edit taxonomic publications?" To assess this, I looked at the edit history of the random sample of 10,000 Wikidata items generated above. For each of these items I retrieved the number of edits made since the Wikidata item was created, when those edits were made, and what properties were edited.

[Fig. 3 here]

Figure 3 visualises the edit history for the sample of 10,000 publications 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 and the lower right triangle in Figure 3 would be empty. 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 that represent the item's creation. If there are no dots to the right of the diagonal, then an item has not been edited since its creation. Figure 3 shows that many items undergo a series of sporadic edits over time. Some of these edits occur shortly after item creation. For example, there are Wikidata bots whose function is to add a description for a new item in a specific language. Other edits may happen later in the life cycle of an item, for example if a user associates a publication

Deleted: 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 somewhat challenging language to use. Hence, a number of more accessible tools have emerged, including generic tools such as resonator ("Reasonator"), and more focussed tools such as Scholia (Nielsen, Mietchen & Willighagen, 2017). The latter provides a wealth of visualisations for a publication and its authors, including its citation network, major topics, and the network of connections amongst a publication's authors.

To complement these tools I have developed a simple website called ALEC (All Literature Electronically Catalogued) (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 (Figs. 3-5).

[Fig. 3 here]

[Fig. 4 here]

[Fig. 5 here]

If an article has an Internet Archive identifier then ALEC embeds the Internet Archive viewer in the web page for that article (Fig. 6). 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 ("PDF.js"). If the arti ... [1]

Deleted: Alternatively

Deleted: this

Deleted: a

Deleted: Wikidata items

Deleted: for papers that have published scientific names.

Deleted: record

Deleted: 7

Deleted: 7

Deleted: a

Deleted: sample of

Deleted: publications in Wikidata

Deleted: diagonal

Deleted: 7

Deleted: example

with its author or links a publication to its main subject. Or there may be a bulk update of many items by a bot that edits a specific property.

[Fig 4 here]

[Fig. 5 here]

The most common edits observed in the publications involved the authors of those publications (properties P2093 and P50), as well as adding values for P921 “main subject” (a form of tagging an item) (Fig. 4). Edits in Wikidata can be made by people, either directly by editing a record in Wikidata, or using bulk tools such as Quickstatements (“QuickStatements”). Edits can also be made by automated programs (“bots”). Of the top ten editors of publications, half are bots (Fig. 5).

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 (i.e., the target of the “cites work” statement).

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 in the sense of being statements about an entity, 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 connection density of bibliographic entities in Wikidata, I counted the number of links between bibliographic items and other Wikidata entities in the sample of 10,000 bibliographic items. In counting these connections some entities, such as those for language, were not counted to avoid inflating the estimate of knowledge density based on what are essentially administrative metadata. The properties that were counted are shown in Table 2.

Deleted: author, or

Moved (insertion) [2]

Deleted: 8

Deleted: 9

Deleted: sample of 1000

Deleted: for P

Moved up [2]: ¶
[Fig.8 here]¶
¶
[Fig. 9 here]¶

Deleted: ¶

Deleted: ¶

Deleted: 9

Deleted: , h

Deleted: To assess the density of bibliographic data in Wikidata …

Deleted: so as to

Deleted: 1

The average link density for the sample of publications was 4.17, with the modal number of connections being one. Hence this part of the knowledge graph is relatively 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).

[Table 2 here]

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. 1). Ideally all authors of publications would be Wikidata items, not simply text strings, and indeed making that conversion is among the most commonly made edits (Fig. 4). Realising this goal requires that all authors of taxonomic publications have items in Wikidata, which in turn is part of a broader goal of having a Wikidata item for everyone involved in taxonomic research (Groom et al., 2020).

[Fig. 6 here]

There are several databases of taxonomists that have representation in Wikidata, although their coverage in Wikidata is variable. For example, the International Plant Names Index (IPNI) contained approximately 43,000 authors in 2013 (Lindon et al., 2015), and currently some 53,000 Wikidata items have IPNI author ids. At the time of writing (2021) ZooBank (Pyle & Michel, 2008) contains some 87,000 authors, of which 17,000 are in Wikidata. The Biodiversity Heritage Library has 28,500 authors in Wikidata, while Wikispecies contributes 61,000 authors to Wikidata. There is overlap among these sources. 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. 6). What is unclear is how much of the lack of overlap between authors in the different sources databases is real (do they represent

Deleted: 24

Deleted: Less than a third of the publications were connected to an item for an author (hence most authors were "strings" rather than "things").

Deleted: 1

Moved up [1]: Taxonomic coverage
The primary motivation for this project is to be able to link every taxonomic name for eukaryote species 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 this challenge was discussed in (Page, 2016a), and an attempt to do this for animal names led to my BioNames project (Page, 2013). I have done similar work for plants, although this is mostly unpublished. Preliminary data has been released on GBIF (Page, 2016), as a "datasette" (Page, 2018), and raw data dumps (Page, 2020b). My work on Index Fungorum is currently unpublished. Typically 20-40% of publications have been mapped to one or more identifiers, but only 15-20% of the publications currently exist in Wikidata.

[Table 2 here]

Deleted: .

Deleted: that

Deleted: 073

Deleted: 79

Deleted: 15,833

Deleted: 26,371

Deleted: 56,994

Deleted: 10

different sets of authors?), versus a lack of mapping between identifiers (how many records are for the same people, just using different identifiers?). 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. 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).

Discussion

By providing a robust, open platform for community editing of structured data, Wikidata seems an 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). Consequently, a large number of taxonomic works and their authors already exist in Wikidata. As more and more taxonomic publications acquire DOIs, and as more working taxonomists acquire ORCID ids, the taxonomic literature component of Wikidata will automatically grow as 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 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.

Wikidata has a higher density that most knowledge graphs (Hegde & Talukdar, 2015), highlighting the importance of having an active community of editors. However, being a community project Wikidata has a number of quirks. It is possible for people working independently 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. While these quirks would be less likely in a

Deleted: [Fig. 10 here]¶
¶
unique to that source. The authors shared by each pair of data sources are represented by the nodes on the paths between each pair of sources, these nodes are labelled by the number of authors the two sources share.¶

Formatted: Heading 2, Line spacing: 1.5 lines

Deleted: ¶

Deleted: {Citation}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). Consequently a large number of taxonomic works and their authors already exist in Wikidata. As more and more taxonomic publications acquire DOIs, and as more working taxonomists acquire ORCID ids, the taxonomic literature component of Wikidata will automatically grow as 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.

Deleted: (2015)

Deleted: ¶
¶

Deleted: B

Deleted:

Formatted: Font: Italic

Deleted: Tools such as ALEC described here often need quite elaborate SPARQL queries to accommodate this multiplicity of ways to represent essentially the same information.

domain-specific database, it is unlikely that such a database would have the level of community engagement we see in Wikidata.

Deleted: (Troudet et al., 2017)(Troudet et al., 2017)

Given that the "bibliography of life" is of little use unless it has content, I have focussed here on where that content comes from, and to what extent the Wikidata community contributes to the curation and improvement of that content. There is considerable scope for analysing gaps in coverage in geography and language (Miquel-Ribé & Laniado, 2021) as well as taxonomy. Wikidata's user interface is aimed at data entry and editing rather than search and visualisation. Creating engaging, user-friendly interfaces (Whitelaw, 2015) to navigate the bibliography of life is major challenge which will be addressed elsewhere.

Deleted: (Whitelaw, 2015)

Deleted: ¶

Deleted: ¶
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. However, Wikidata's utility as a bibliography of life seems clear.¶

Acknowledgements

I thank the numerous Wikidata contributors (some known only by their usernames) who have helped me learn the ropes and navigate the active and opinionated Wikidata community. Among these contributors are Christian Ferrer, Siobhan Leachman, Andy Mabbett, Daniel Mietchen, Succu, Andra Waagmeester, and Egon Willighagen. Lars Willighagen provided helpful feedback on an earlier draft. I'm grateful to David Shotton, John Mittermeier, and an anonymous reviewer for their helpful critiques of the manuscript.

Deleted: and

References

Aspesi C, Brand A. 2020. In pursuit of open science, open access is not enough. *Science* 368:574–577. DOI: 10.1126/science.aba3763.

Bennett F. 2021. *Juris-M/citeproc-js*. Juris-M Project.

Bohannon J. 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.

Farris SM. 2020. The rise to dominance of genetic model organisms and the decline of curiosity-driven organismal research. *PLOS ONE* 15:e0243088. DOI: 10.1371/journal.pone.0243088.

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.

Gwinn NE, Rinaldo C. 2009. The Biodiversity Heritage Library: sharing biodiversity literature with the world. *IFLA Journal* 35:25–34. DOI: 10.1177/0340035208102032.

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.

Heibi I, Peroni S, Shotton D. 2019. Software review: COCI, the OpenCitations Index of Crossref open DOI-to-DOI citations. *Scientometrics* 121:1213–1228. DOI: 10.1007/s11192-019-03217-6.

Index Fungorum Home Page. *Available at http://www.indexfungorum.org/* (accessed March 18, 2022).

International Plant Names Index. *Available at https://ipni.org/* (accessed March 18, 2022).

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-Shaikevich I. 1973. System of Document Connections Based on References. *Scientific and Technical Information Serial of VINITI* 6:3–8.

May RM. 1988. How Many Species Are There on Earth? *Science* 241:1441–1449. DOI: 10.1126/science.241.4872.1441.

Miquel-Ribé M, Laniado D. 2021. The Wikipedia Diversity Observatory: helping communities to bridge content gaps through interactive interfaces. *Journal of Internet Services and Applications* 12:10. DOI: 10.1186/s13174-021-00141-y.

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.

ORCID. *Available at https://orcid.org/* (accessed November 27, 2018).

Page RD. 2009. bioGUID: resolving, discovering, and minting identifiers for biodiversity informatics. *BMC bioinformatics* 10: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. The Plant List with literature. DOI: 10.15468/btkum2.

Page RDM. 2016c. 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.

Page RDM. 2020. IPNI plant names linked to bibliographic identifiers. DOI: 10.6084/m9.figshare.13055804.v1.

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

PubMed. *Available at https://pubmed.ncbi.nlm.nih.gov/* (accessed April 27, 2021).

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.

QuickStatements. *Available at https://quickstatements.toolforge.org/#/* (accessed May 14, 2021).

ResearchGate. *Available at https://www.researchgate.net/profile/Stefan_Wanke/research* (accessed December 26, 2019).

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.

Troudet J, Grandcolas P, Blin A, Vignes-Lebbe R, Legendre F. 2017. Taxonomic bias in biodiversity data and societal preferences. *Scientific Reports* 7:9132. DOI: 10.1038/s41598-017-09084-6.

Unpaywall. *Available at http://unpaywall.org/* (accessed April 16, 2021).

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.

Waagmeester A, Stupp G, Burgstaller-Muehlbacher S, Good BM, Griffith M, Griffith OL, Hanspers K, Hermjakob H, Hudson TS, Hybiske K, Keating SM, Manske M, Mayers M, Mietchen D, Mitraka E, Pico AR, Putman T, Riutta A, Queralt-Rosinach N, Schriml LM, Shafee T, Slenter D, Stephan R, Thornton K, Tsueng G, Tu R, Ul-Hasan S, Willighagen E, Wu C, Su AI. 2020. Science Forum: Wikidata as a knowledge graph for the life sciences. *eLife*. DOI: 10.7554/eLife.52614.

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.

Wayback Machine. *Available at https://web.archive.org/* (accessed April 15, 2021).

Werner YL. 2006. The case of impact factor versus taxonomy: a proposal. *Journal of Natural History* 40:1285–1286. DOI: 10.1080/00222930600903660.

Whitelaw M. 2015. Generous Interfaces for Digital Cultural Collections. *Digital Humanities Quarterly* 9.

Willighagen LG. 2019. Citation.js: a format-independent, modular bibliography tool for the browser and command line. *PeerJ Computer Science* 5:e214. DOI: 10.7717/peerj-cs.214.

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

Deleted: Aspesi C, Brand A. 2020. In pursuit of open science, open access is not enough. Science 368:574–577. DOI: 10.1126/science.aba3763.
Bennett F. 2021. Juris-M/citeproc-js. Juris-M Project.
Biodiversity Heritage Library. Available at https://www.biodiversitylibrary.org/ (accessed May 14, 2021).
Biodiversity Literature Repository. Available at http://plazi.org/resources/bibliography-of-life-bol/biodiversity-literature-repository-blr/ (accessed November 28, 2018).
Bohannon J. 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.
Groom Q, Güntsch A, Huybrechts P, Kearney N, Leachman S, Nicolson N, Page RDM, Shorthouse DP, Thessen AE, Haston E. 2020. People are essential to linking biodiversity data. Database 2020. DOI: 10.1093/database/baaa072.
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.
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].
... [2]

Page 12: [1] Deleted **Roderic Page** **18/03/2022 16:34:00**

Page 20: [2] Deleted **Roderic Page** **16/03/2022 18:15:00**