Ozymandias: A biodiversity knowledge graph

Roderic D. M. Page
https://orcid.org/0000-0002-7101-9767
Institute of Biodiversity, Animal Health and Comparative Medicine, College of Medical, Veterinary and Life Sciences, Graham Kerr Building, University of Glasgow, Glasgow, UK
Email address: Roderic.page@glasgow.ac.uk

Abstract
Enormous quantities of biodiversity data are being made available online, but much of this data remains isolated in their own silos. One approach to breaking these silos is to map local, often database-specific identifiers to shared global identifiers. This mapping can then be used to construct a knowledge graph, where entities such as taxa, publications, people, places, specimens, sequences, and institutions are all part of a single, shared knowledge space. Motivated by the 2018 GBIF Ebbe Nielsen Challenge I explore the feasibility of constructing a "biodiversity knowledge graph" for the Australian fauna. These steps involved in constructing the graph are described, and examples its application are discussed. A web interface to the knowledge graph (called "Ozymandias") is available at https://ozymandias-demo.herokuapp.com.

Keywords: knowledge graph; biodiversity informatics; linked data; identifiers;

Introduction
"Linnaeus would have been a 'techie'" - (Godfray, 2007)

The recent announcement that the Global Biodiversity Information Facility (GBIF) has reached the milestone of one billion occurrence records reflects the considerable success the biodiversity community has had in mobilising data. Much of this success comes from standardising on a simple column-based data format (Darwin Core) (Wieczorek et al., 2012) and indexing that data using three fields: taxonomic name, geographic location, and date (i.e., "what", "where", and "when"). By flattening the data into a single table, Darwin Core makes data easy to enter and view, but at the cost of potentially obscuring relationships between entities, relationships that may be better represented using a network. In this paper I explore the representation of biodiversity data using a network or "knowledge graph".

A knowledge graph is a network or graph where nodes represent entities or concepts ("things") and the links or edges of the graph represent relationships between those things (Bollacker et al., 2008). Each node is labelled by a unique identifier, and may have one or more attributes or properties. Each edge of the graph is labelled by the name of the relationship it represents. A common representation of a knowledge graph is the linked data triple of subject, predicate, and object, where the subject (e.g., a publication) is connected to an object (e.g., a person) by a predicate (e.g., "author"). Triples are not the only way knowledge graphs can be modelled, but adopting triples means we can use existing technologies such as triple stores and the SPARQL query language (W3C SPARQL Working Group, 2013).

Knowledge graphs are potentially global in scope, hence rely on global identifiers. Most datasets will have their own local identifiers for the entities they contain, such as species, publications, specimens, or collectors. These identifiers are adequate for local use, but local identifiers also serve to keep data isolated in distinct silos. Hence we need to map identifiers for the same thing between the different silos. This can be done by establishing a "broker" service that asserts identify between a set of identifiers, or by mapping local identifiers to a single global identifier. The case for mapping to a single global identifier ("strings to things") is attractive in terms of scalability (mapping each local identifier to a single global identifier is easier than managing cross mappings between multiple identifiers), and is even more attractive if there are useful services built around that global identifier. For example, Digital Object Identifiers (DOIs) are becoming the standard for identifying academic publications. Given a DOI we can retrieve metadata about the work from CrossRef ("CrossRef"), we can get measures of attention from services such as Altmetric ("Altmetric"), and we can discover the identities of the work's authors from ORCID ("ORCID"). Furthermore, by agreeing on a centralised identifier we effectively decentralise the building of the knowledge graph: given a DOI, anybody that links local information to that DOI is potentially contributing to the construction of the global knowledge graph.

Mapping strings to things give us a way to refer to the nodes in the knowledge graph, but we also need a consistent way to label the edges of the graph. There has been an explosion in vocabularies and ontologies for describing both attributes of entities and their interrelationships. While arguments can be made that domain-specific ontologies enable us to represent knowledge with greater fidelity, the existence of multiple vocabularies comes with the cognitive overhead of having to decide which term from what vocabulary to use. In contrast to, say, (Senderov et al., 2018) who use several ontologies to model taxonomic publications, the approach I have adopted here is to try and minimise the number of vocabularies employed, and to avoid domain-specific vocabularies where ever possible. For this reason the default vocabulary used is schema.org ("Schema.org"), being developed by a consortium of search engine vendors including Google, Microsoft, and Yahoo. In addition to simplifying development, adopting a widely used vocabulary increases the potential utility of the knowledge graph. One motivation for the development of schema.org is to encourage the inclusion of structured data in web pages, helping search engines interpret the contents of those pages. By adopting schema.org in knowledge graphs we can make it easier for developers of biodiversity web sites to incorporate structured data from those knowledge graphs directly into their web pages.

There are several different categories of applications that can be built on top of a knowledge graph, for example data reconciliation, data augmentation, and meta-analyses. Reconciliation involves either matching strings to things, or matching entities from different data sources. An example of reconciliation is matching author names to identifiers. Augmentation involves combining data for the same entities from different sources that individually may be incomplete, but together yield more extensive coverage of those entities. An example is supplementing existing imagery of species with figures published in the taxonomic literature. Meta analyses make use of the data aggregated in the knowledge graph to explore larger patterns. There have been numerous studies of patterns of taxonomic activity (Joppa, Roberts & Pimm, 2011; Costello, Wilson & Houlding, 2013; Bebber et al., 2013; Grieneisen et al., 2014; Sangster & Luksenburg, 2014; Tancoigne & Ollivier, 2017), typically these studies assembled a custom database, and often this data is not made more widely available, or the data is not actively updated. Having a biodiversity knowledge graph would enable users to ask similar questions but for different taxonomic groups, or different time periods.

In response to the GBIF 2018 Ebbe Nielsen Challenge I constructed a knowledge graph for the Australian fauna, based on data in the Atlas of Living Australia (ALA) ("Atlas of Living Australia") and the Australian Faunal Directory (AFD) ("Australian Faunal Directory"). This regional-scale dataset was chosen to be sufficiently large to be interesting, but without being too distracted by issues of scalability. The knowledge graph combines information on taxa and their names, taxonomic publications, the authors of those publications together with their interrelationships, such as publication, citation, and authorship. Constructing the knowledge graph required extensive data cleaning and cross linking. These steps are described below, and examples of the application of the knowledge graph are discussed.

Materials and Methods

Knowledge graph

The general structure of the knowledge graph is based on (Page, 2013, 2016a). The core entities are taxa, taxonomic names, publications, journals, and people. Figure 1 summarises the relationships between those entities.

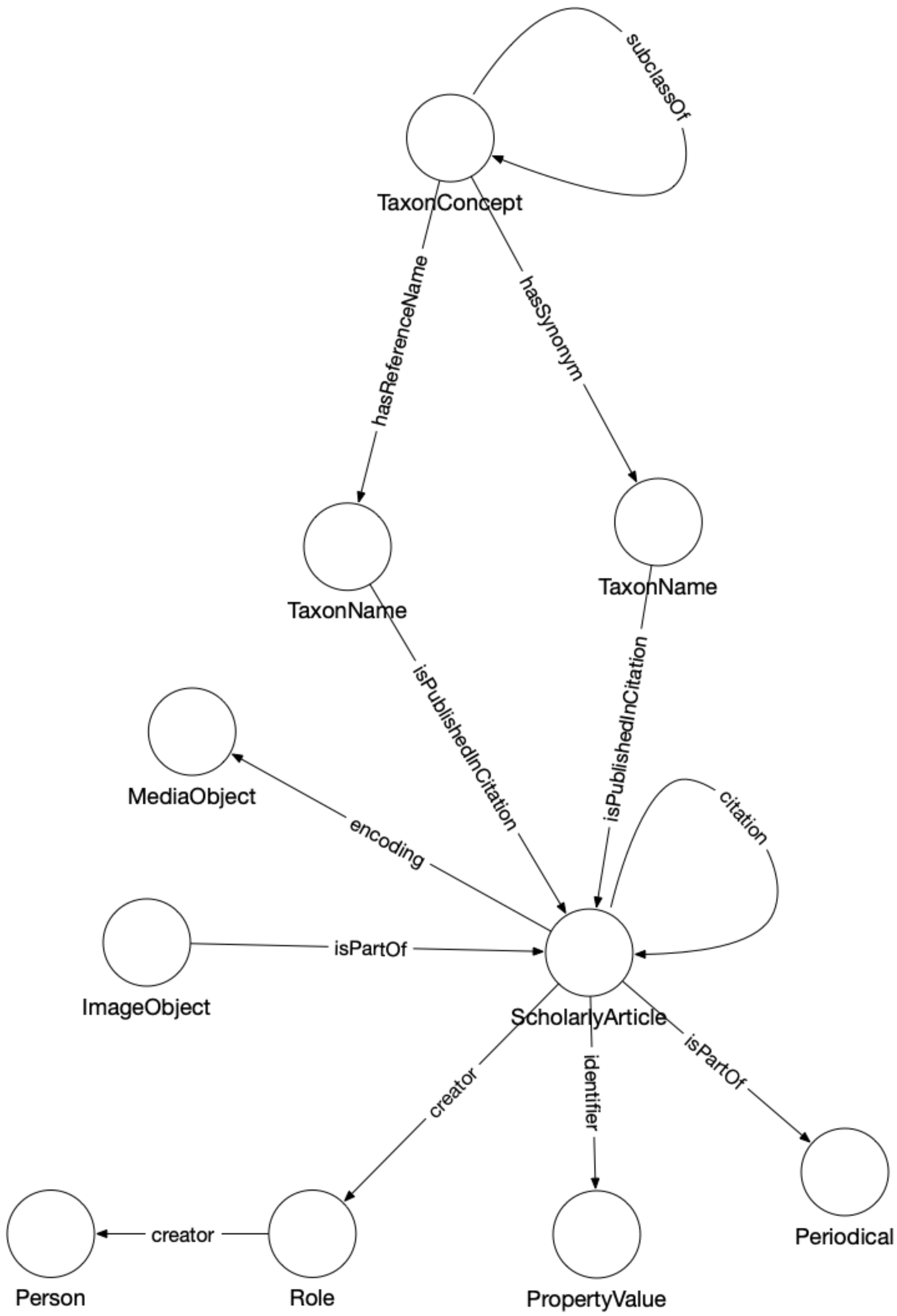

Figure 1. The knowledge graph model used in Ozymandias. Nodes in the graph are represented by circles and are labelled with the *rdf:type* of that node. Nodes are connected by edges in the graph which are labelled by a term from a vocabulary, typically schema.org.

Taxa and taxonomic names were modelled using the TDWG LSID vocabulary based on (Kennedy et al., 2006). Taxa are nodes in a tree representing the taxonomic classification and are instances of the type *tc:TaxonConcept*. The taxonomic classification is represented by *rdfs:subClassOf* relationship between parent and child taxa (a child is a *rdfs:subClassOf* its parent).

Taxonomic names (type *tn:TaxonName*) are connected to the corresponding taxa using relations from the TAXREF vocabulary (Michel et al., 2017) and are typically either accepted names or synonyms. This vocabulary was adopted to because it enables a more direct way of expressing the relationship between taxa and taxonomic names than is possible using the TDWG LSID vocabulary.

Taxonomic names are published in publications, which were represented using terms from the schema.org vocabulary. In cases where the full text of an article is available as a PDF file I make use of the *schema:encoding* property to link the publication to a *schema:MediaObject* representing the PDF. Articles are linked to the journals they were published in by the *schema:isPartOf* property.

Representing ordered lists in RDF is not straightforward, which presents a challenge for expressing relationships such as authorship. Not only is the order of authorship an important feature when formatting a published work for display, it is also useful information when trying to reconcile author names (see below). The approach adopted here is to use the *schema:Role* type (Vicki Tardif Holland & Jason Johnson, 2014) . Rather than directly connect a publication to an author using, say, the *schema:creator* property, instead the creator of a work is a Role, which in turn has the author as a creator property. The position of author in the list of authors is stored using the *schema:roleName* property (e.g., “1”, “2”, etc.) (Figure 2).

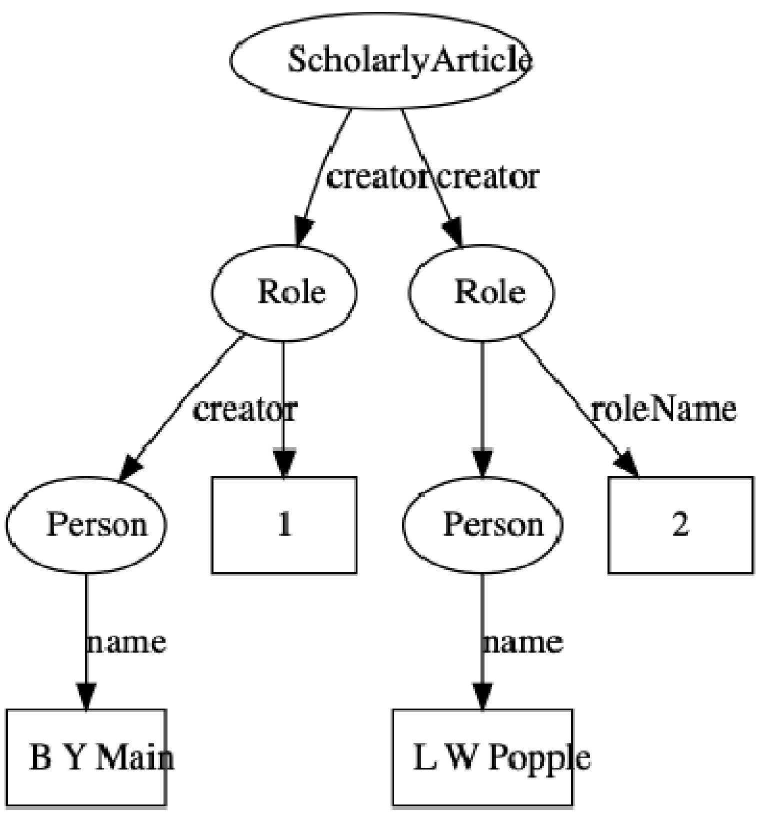

Figure 2. An example of modelling order of authorship using *schema:Role*. Each author is linked to the article they authored via a *schema:Role* node, which specifies the order of authorship for each author. In this example, "B Y Main" is the first author, "L W Popple" is the second author.

Identifiers

Identifiers are both central to any attempt to link data together, and at the same time can be one of the major obstacles to creating links. Ideally identifiers should be globally unique, persistent, and each entity would have only a single identifier. In reality, entities may have many identifiers, typically minted by different databases, and identifiers may change, or at least have multiple representations. For example, DOIs may contain upper and lowercase letters, but are actually case insensitive. Some databases may choose to store DOIs in lower case form, others in upper case, or any combination in between. Identifiers typically require dereferencing and the mechanism for this may evolve over time, often for reasons outside the control of the organisation that minted the identifier. DOIs are dereferenced ("resolved") using the web proxy https://doi.org. This proxy recently switched from the HTTP to the HTTPS protocol, immediately rendering out of date any DOIs stored URLs starting with the prefix "http://". To minimise the impact of these kinds of changes, Ozymandias stores identifiers both as URLs (where appropriate) but also as property-value pairs (*schema:PropertyValue*) where the *schema:value* property stores the identifier string stripped of any dereferencing prefix. For example, a DOI https://doi.org/10.5134/176044 would be stored as a *schema:PropertyValue* with *schema:propertyID* "doi" and *schema:value* "10.5134/176044" (Figure 3).

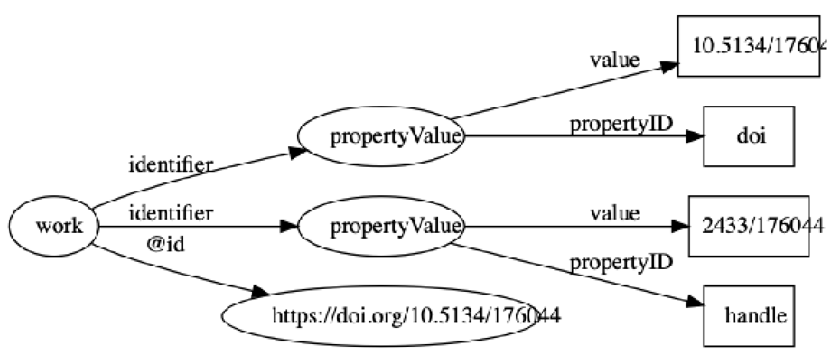

Figure 3. Storing identifiers using *schema:PropertyValue*. The work has two identifiers, a DOI https://doi.org/10.5134/176044 and a Handle https://hdl.handle.net/2433/176044. These are stored as *schema:PropertyValue* pairs.

Citations

One paper citing another can be represented by a direct link between two identifiers, for example a link between the DOIs of the citing and the cited work. CrossRef provides lists of literature cited for many of the works in its database, and many of these cited works themselves have DOIs Hence if we have a DOI for a work we can immediately populate the triple store with citation links. This works well if both works have a DOI, but many taxonomically relevant works do not have these identifiers. Even for those works that do have DOIs, these may not have been available at the time the citing work was deposited by a publisher, for example, if the cited work has only recently been assigned a DOI.

To expand the citation links beyond just those works with DOIs I also generated an identifier for each work modelled on the Serial Item and Contribution Identifier (SICI). This identifier comprised the International Standard Serial Number (ISSN) of the journal, together with the volume, and the starting page. This triple uniquely identifies most articles, and is easy to generate. SICIs were generated for works harvested from the Australian Faunal Directory, and from the lists of literature cited obtained from CrossRef, and were stored as *schema:PropertyValue* pairs in the same way as DOIs and other identifiers. By matching SICIs it was possible to expand citation links beyond those where both works had DOIs.

Populating the knowledge graph

Perhaps the biggest challenge in constructing a knowledge graph is to map names or descriptions of entities to one or more globally unique identifiers. In some cases the sources data will already have identifiers. Taxa in the ALA each have a unique identifier (a LSID), as do taxa and publications in the AFD (which use UUIDs). The ALA and AFD share the same taxon identifiers, which makes linking the two databases straightforward. However, these identifiers are local in the sense that they are primary keys for local databases that have been converted into URLs. The knowledge graph can only grow if we use external identifiers that are shared by other databases, or at least map local identifiers onto those external identifiers. For publications this is straightforward in the sense that a publication in a database of Australian animals can be unambiguously mapped onto the publication in, say, a database for Japanese animals. However, a taxon as defined in the Australian Faunal Directory may not correspond exactly to a taxon with the same name in another.

Reconciling works

For the works in AFD I searched for DOIs using the API provided by CrossRef. If a reference was found the associated DOI was assigned to that reference. CrossRef is not the only registration agency for DOIs, there are several others that are used by digital libraries and publishers, such as DataCite, the multilingual European Registration Agency (mEDRA), and Airiti (華藝數位). Most of these agencies lack the discovery services provided by CrossRef, so for these DOIs I harvested the article metadata using a combination of web services and screen scraping, created a local MySQL database to store the metadata, and used that database to match references to DOIs. This database was also used to match articles to other classes of identifiers, such as Handles and URLs.

Australian natural history institutions are significant publishers of biodiversity literature, and much of this has been scanned by the Biodiversity Heritage Library in Australia. As a consequence many of the articles in the knowledge graph were available in my BioStor project (Page, 2011). Identifiers for these articles were found by matching using the BioStor OpenURL service.

Reconciling authors

Multiple approaches were used to match author names to external identifiers. Metadata for DOIs from CrossRef would sometimes include ORCID ids for authors. The ORCID record for each ORCID id was retrieved using the ORCID API and converted to a set of RDF triples linking the identifiers for a work (e.g., DOI) to a person's ORCID. These triples modelled the order of authorship using *schema:Role* as described above. Similarly, I parsed Wikispecies pages and extracted bibliographic records for works identified by a DOI, and constructed triples linking the work to its authors where those authors had their own Wikispecies page. Hence to match authors in the knowledge graph to authors in ORCID or Wikispecies, we can ask whether the same pairing of work and author name appears in both databases. For example, we can retrieve the second author of a work in the knowledge graph and in ORCID by querying by DOI for the work and restricting the value of *schema:roleName* to "2" (Figure 4). As a final check we can compare the author names and accept only those names whose similarity exceeds a threshold. In this way we can automate the matching authors across databases.

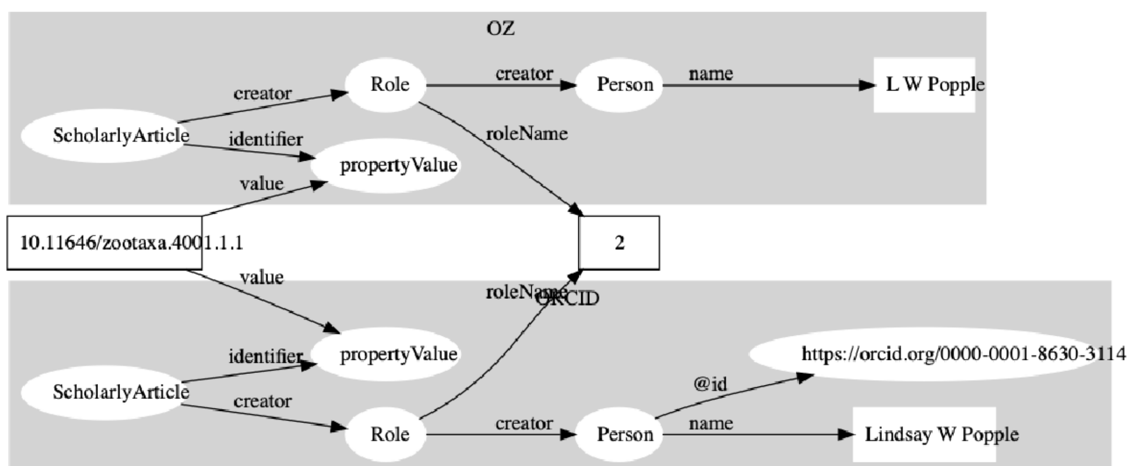

Figure 4. Matching author records in two different databases. In this example the article with DOI 10.11646/zootaxa.4001.1.1 occurs in both Ozymandias (OZ) and ORCID. Using a SPARQL query we retrieve the name of the second author in the two databases: "L W Popple" in Ozymandias and "Lindsay W Popple" in ORCID. Given the similarity in the names, we conclude that the two authors are the same, and we can assign the ORCID for Lindsay W Popple (https://orcid.org/0000-0001-8630-3114) to "L W Popple" in Ozymandias.

Data sources

I used several different strategies to convert data into the triples required for the knowledge graph. If the source data was in the form of CSV files (e.g., the Australian Faunal Directory) it was imported into a MySQL database, and PHP scripts were written to further clean the data and map it to any external identifiers. Once the data was cleaned and linked, a PHP script was used to export the data in N-triples format.

Several sources of data (Atlas of Living Australia, CrossRef, ORCID, Wikispecies, and Biodiversity Literature Repository) were accessed via their APIs. For ALA a list of all animal taxa was obtained from the ALA web site, then the JSON record for each taxon was harvested. For CrossRef, data was harvested for just those DOIs found by the bibliographic string to DOI mapping process described above. These DOIs were also submitted to a custom script that queried the ORCID database to discover whether any authors had works with those DOIs in their ORCID profile. If this was the case, the corresponding ORCID profile was downloaded. Each DOI was also used as a query term for searching Wikispecies using its API with the "list" parameter set to "exturlusage" to find wiki pages that mentioned that DOI. Pages found were retrieved in XML format using the API, any references on that page parsed and converted into JSON. All JSON documents obtained from these sources were stored in CouchDB databases and custom CouchDB views were written in Javascript to convert the JSON documents into N-triples.

By default Ozymandias treats individual publications as a single, monolithic entity. However, some publishers such as PLoS and Pensoft provide DOIs for component parts of an article, such as individual figures. (Egloff et al., 2017) have argued that even if a taxonomic article itself is copyrighted, the individual figures are not eligible for copyright, and hence extract and assign DOIs to large numbers of figures extracted from journals such as *Zootaxa*. These figures, together with ones sourced from open access journals are available through the Biodiversity Literature Repository ("Biodiversity Literature Repository") (BLR). The BLR is hosted by Zenodo (https://zenodo.org) and each publication and figure has a unique identifier (typically a DOI), and metadata for each publication and figure is available as JSON-LD. This means data from the BLR can be directly incorporated into a triple store. However for this project I wanted just a subset relevant to publications on the Australian fauna, and so I created a CouchDB version of the BLR and write scripts to match publications from the AFD to the corresponding record in the BLR. Metadata for each matching publication and its associated figures were then retrieved directly from Zenodo.

Knowledge graph

The knowledge graph was implemented as a triple store using Blazegraph 2.1.4 running on a Windows 10 server, with a nginx web server acting as a reverse proxy. N-triples for different categories of data (e.g., taxa, publications, etc.) were partitioned using named graphs and uploaded to the triple store. This made it easier to manage sets of data, for example the bibliographic data could be deleted and reloaded by simply deleting all triples in the corresponding named graph, rather than having to delete the entire knowledge graph. It also facilitated some queries, such as author matching across multiple data sources where distinguishing between data source was an essential part of the query.

Search

Being able to simply search for relevant documents by typing in one or more terms is a feature users expect from almost any web site. To implement search, basic information on taxa and publications was encoded into a simple JSON document (one per entity) and these JSON documents were indexed using an instance of Elasticsearch 6.3.1 hosted on Google's Compute Engine.

Web interface

Designing a semantic web browser to display a richly interconnected data set is a challenging task (Quan & Karger, 2004). For Ozymandias the goal was to have a simple interface which encouraged the user to explore connections between taxa, publications, and people. Apart from the home page, there are two main page types in the web interface for Ozymandias. The first is the search interface which is a simple list of the entities that best match the search terms. Clicking on any member of that list leads to the second page type, which is a display of the entity itself. This display comprises three columns. The left column displays core facts about the entity. These are typically triples that have the entity as their subject, or are one edge away in the knowledge graph (such as thumbnail images), and so can be retrieved from the knowledge graph using either SPARQL DESCRIBE or CONSTRUCT queries. The middle column displays connections between the main entity on the page and related entities in the knowledge graph (such as authors of a paper, taxonomic names mentioned in a work, etc.), and is populated by SPARQL queries. The rightmost column is used to display the result of searching external sources for information relevant to the entity displayed on the page. Hence, unlike columns one and two, these queries are not SPARQL queries to the local knowledge graph.

Results

Ozymandias can be viewed at https://ozymandias-demo.herokuapp.com. Source code is available on GitHub https://github.com/rdmpage/ozymandias-demo. Below I describe the web interface to Ozymandias, and outline some of the exploratory analyses that can be undertaken using the underlying knowledge graph. Where the results are based on SPARQL queries, those queries are listed in the Supplementary material.

Web interface

A screenshot of the web interface is shown in Figure 5. This shows the three-column layout used to display an entity, its relationships within the knowledge graph, and any known external relationships.

Figure 5. Web interface to Ozymandias knowledge graph displaying information for an article. The left column displays a summary of the article, and a PDF viewer (only available if content is freely accessible). The middle column displays related content from the knowledge graph, such as taxa mentioned in the article. The right column shows the result of searches in external web sites for related information, in this case is displays the identifier for Wikidata item that corresponds to this article. To view this page live go to https://ozymandias-demo.herokuapp.com/?uri=https://biodiversity.org.au/afd/publication/3e0c1402-de05-4227-9df3-803e68300623.

The first example is a publication, in this case (Nakabo, 1982). The first column summarises basic data about the publication, and if the full text is available it is displayed using either a PDF viewer, or a simple image viewer in the case of scanned images. The second column lists taxa associated with the publication. For publications with identifiers such as DOIs the third column displays whether a record with that DOI exists in external sources such as Wikidata and ORCID.

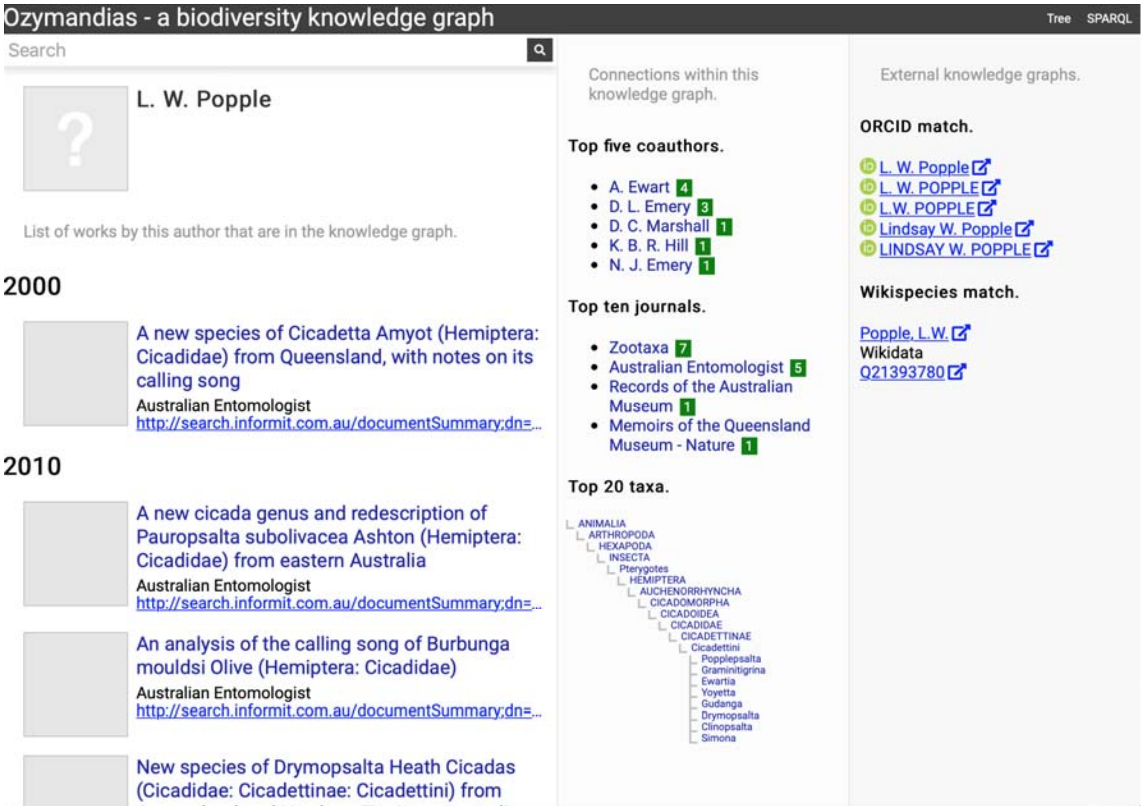

Figure 6. Information about an author displayed in Ozymandias. The left column lists the author's publications, the middle column uses the knowledge graph to identify coauthors, venues for publication, and the taxonomic expertise of the author, the right column displays information from external sources. To view live go to https://ozymandias-demo.herokuapp.com/?uri=https://biodiversity.org.au/afd/publication/%23creator/l-w-popple.

The second example (Figure 6) displays information for an author, including a list of publications, coauthors, journals the author publishes in, and a summary of their taxonomic expertise. This later diagram is computed by using a SPARQL query to find the top 20 taxa the author has published on. For each taxon the query uses a property path expression to retrieve the list of higher taxa each taxon belongs to, and a Javascript script assembles those lists into a tree. The third panel displays the results of matching the author to author identifiers using external web services, in this case discovering the author's ORCID id and entry in Wikidata.

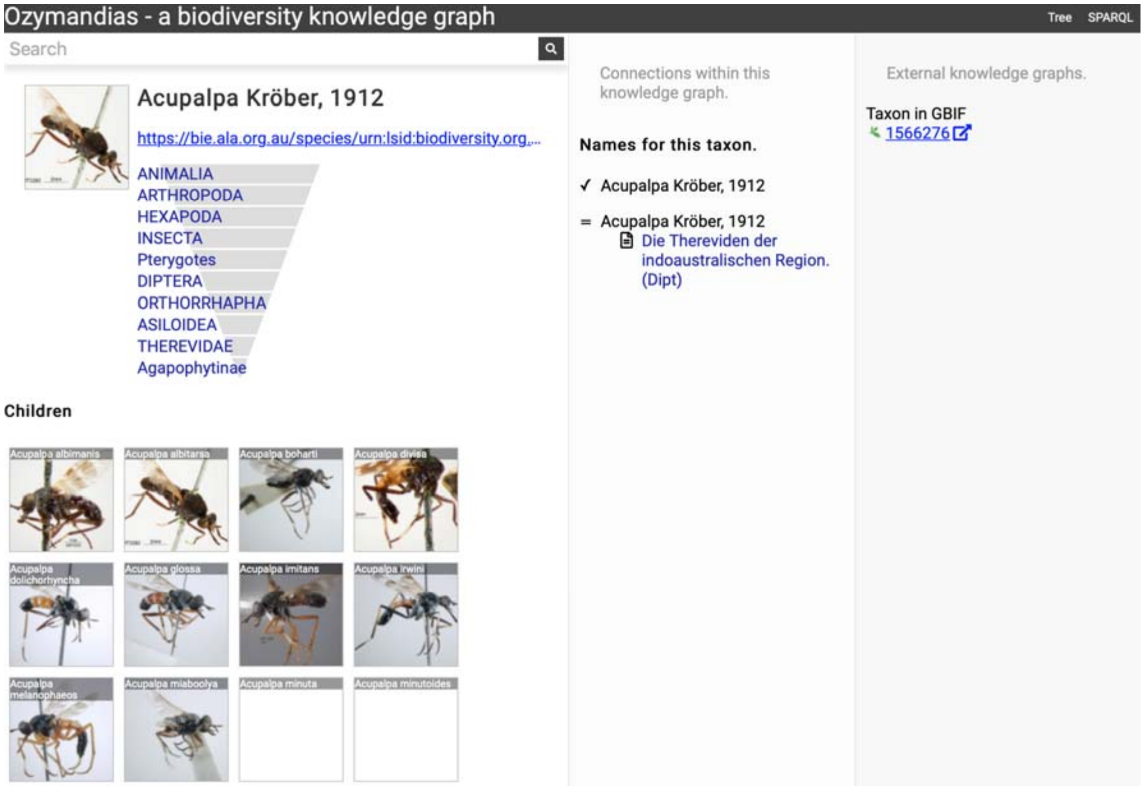

Figure 7. Information about the genus *Acupalpa* Kröber, 1912 displayed in Ozymandias. The display includes the species in the genus, details about the publication of the name *Acupalpa* , and a link to the taxon in GBIF. Live version at https://ozymandias-demo.herokuapp.com/?uri=https://bie.ala.org.au/species/urn:lsid:biodiversity.org.au:afd.taxon:111fc7e9-0265-453e-8e60-1761e42efc9a.

Figure 7 shows the view of a taxon, in this case genus *Acupalpa* Kröber, 1912. We see the member species of the genus, the taxonomic hierarchy of the genus (generated using a SPARQL property path query) and, where available, a thumbnail image from the ALA. The second column lists the taxonomic names associated with the genus, together with the publications that made those names available. The third column shows the results of mapping the taxon to one or more external taxonomic databases, in this case GBIF.

Wherever possible, Ozymandias uses thumbnail images from ALA to illustrate taxa. However, many taxa lack images. Figure 8 shows an example where the ALA has no image for a taxon (*Trigonopterus cooktownensis*). Because the taxon, its name, the publication, and the figures in that publication extracted by the Biodiversity Literature Repository are all part of the knowledge graph, we can traverse the graph and discover that an image for that species was published in (Riedel & Tänzler, 2016) .

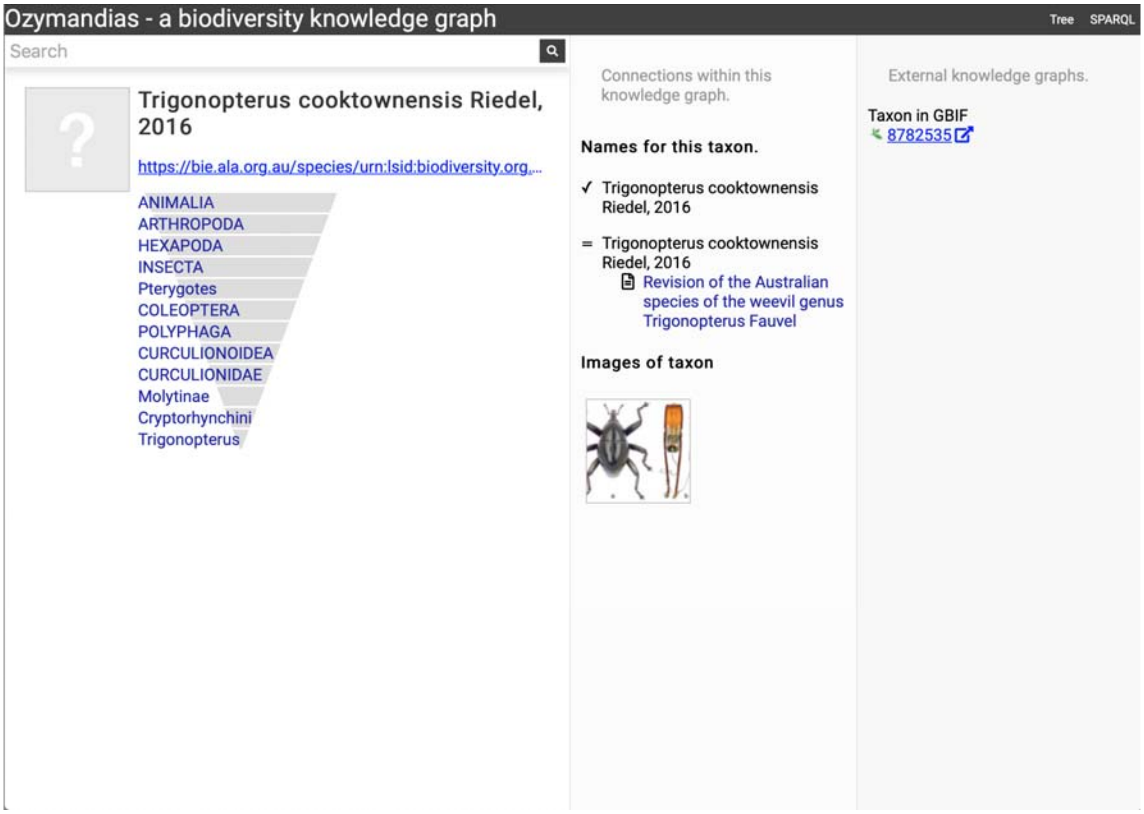

Figure 8. Augmenting data using knowledge graph. The Atlas of Living Australia did not have an image for *Trigonopterus cooktownensis* at the time it was harvested by Ozymandias, hence the "?" displayed in the square in the left column. However, the original description of that species did include images which are available in the Biodiversity Literature Repository, and hence are displayed by Ozymandias in the middle column. Live example https://ozymandias-demo.herokuapp.com/?uri=https://bie.ala.org.au/species/urn:lsid:biodiversity.org.au:afd.taxon:14feec1f-9d2a-496b-9b98-ec691289b5ce.

Strings to things

Most of the work on data cleaning and linking was devoted to matching string representations of publications to the corresponding digital identifiers. The result of this matching provides us with an estimate of how many publications have been digitised and hence are potentially available online.

Figure 9 shows the distribution of publications over time, together with the numbers that have been matched to digital identifiers. The pattern of publication shows three prominent dips. The first two correspond to the two world wars in the twentieth century, the third dip occurs from the mid-1990's to the present day. Given that the AFD is retrospectively collecting publication data, it is not clear to what extent this decline in recent publications represents an actual decline in activity versus a under sampling the most recent literature.

Many publications lack a digital identifier, suggesting that a considerable amount of the relevant literature has not been digitised. However, this may be overstated as the matching was done by a single individual working to a deadline (in this case the Ebbe Nielsen Challenge submission date). As more effort is expended on matching records the gap between the number of publications and the number of publications online is likely to decrease.

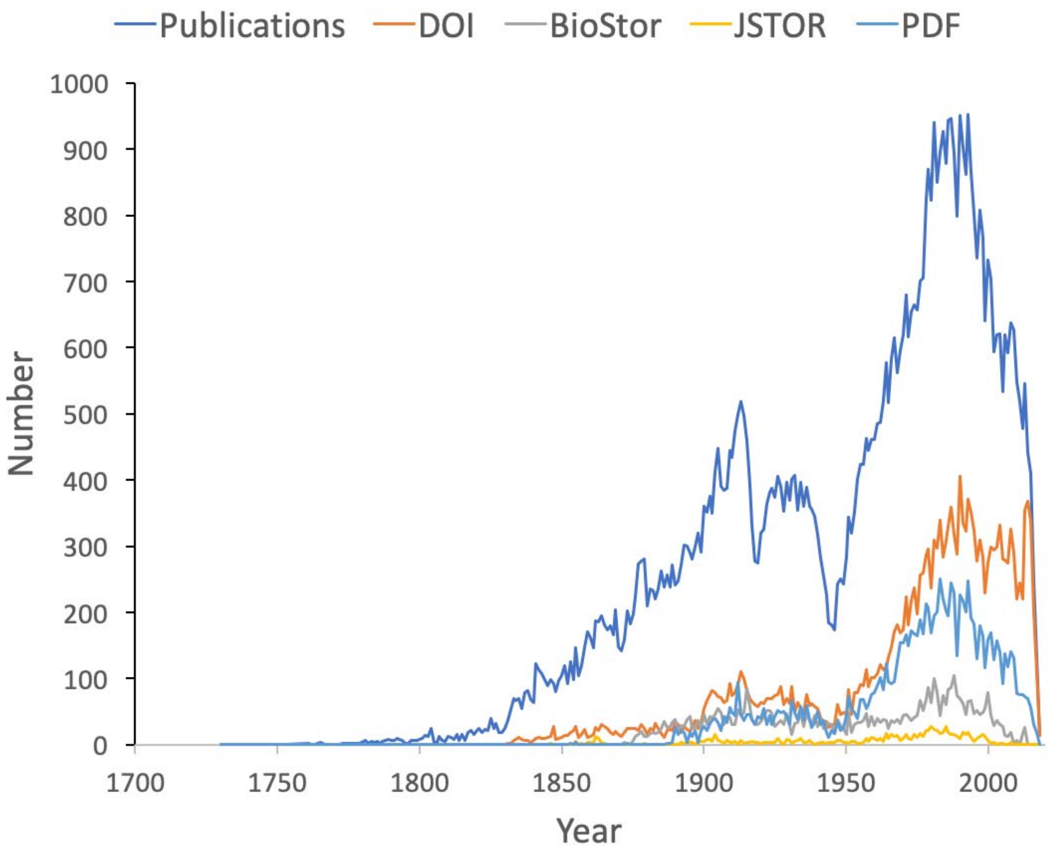

Figure 9. Plot of publications over time. As well as the total number of publications for each year, the chart shows the numbers of publications that have a digital identifier (DOI, BioStor, or JSTOR) or have a PDF available online.

Linking authors to identifiers

We can measure the uptake of ORCID ids for researchers working on the Australian fauna by using DOIs to match works in the knowledge graph to works in the ORCID database. ORCID was launched in 2012, for period from 2012 to the present day the knowledge graph contains 2302 distinct author names. Matching DOIs for the works those authors published to the ORCID database shows that 346 (15%) of authors publishing in that time period have ORCID ids. This number is likely to be an underestimate as not all works in ORCID have DOIs (and ORCID records sometimes omit DOIs for works that have them), but it suggests limited adoption of ORCIDs amongst taxonomists and other biodiversity researchers.

Changes in taxonomic publications over time

To explore the publication history of taxonomic research on Australian animals for each decade from 1820 to 2020 I found the ten journals that had the most articles in the knowledge graph. The numbers of articles in each journal were plotted for each decade (Figure 10). Over time different journals have been dominant venues for publishing taxonomic work. In the 18th century British or other European journals dominated, such as *Proceedings of the Zoological Society* and *Annals and Magazine of Natural History*, although the local journal *Proceedings of the Linnean Society of New South Wales* (establish 1875) was a major outlet for taxonomic work. In the mid to late 20th century Australian journals, typically published by museums or by the Commonwealth Scientific and Industrial Research Organisation (CSIRO) were the primary venues for taxonomic papers on the Australian fauna. However, the last decade has seen the spectacular rise of the “megajournal” *Zootaxa*, published in New Zealand but with global coverage. Hence, taxonomic publication in Australia has gone from an early colonial period where much of it was published overseas, to a national period where many papers were published in local journals, culminating in the present situation where international journals such as *Zootaxa* and, to a lesser extent *Zookeys*, dominate.

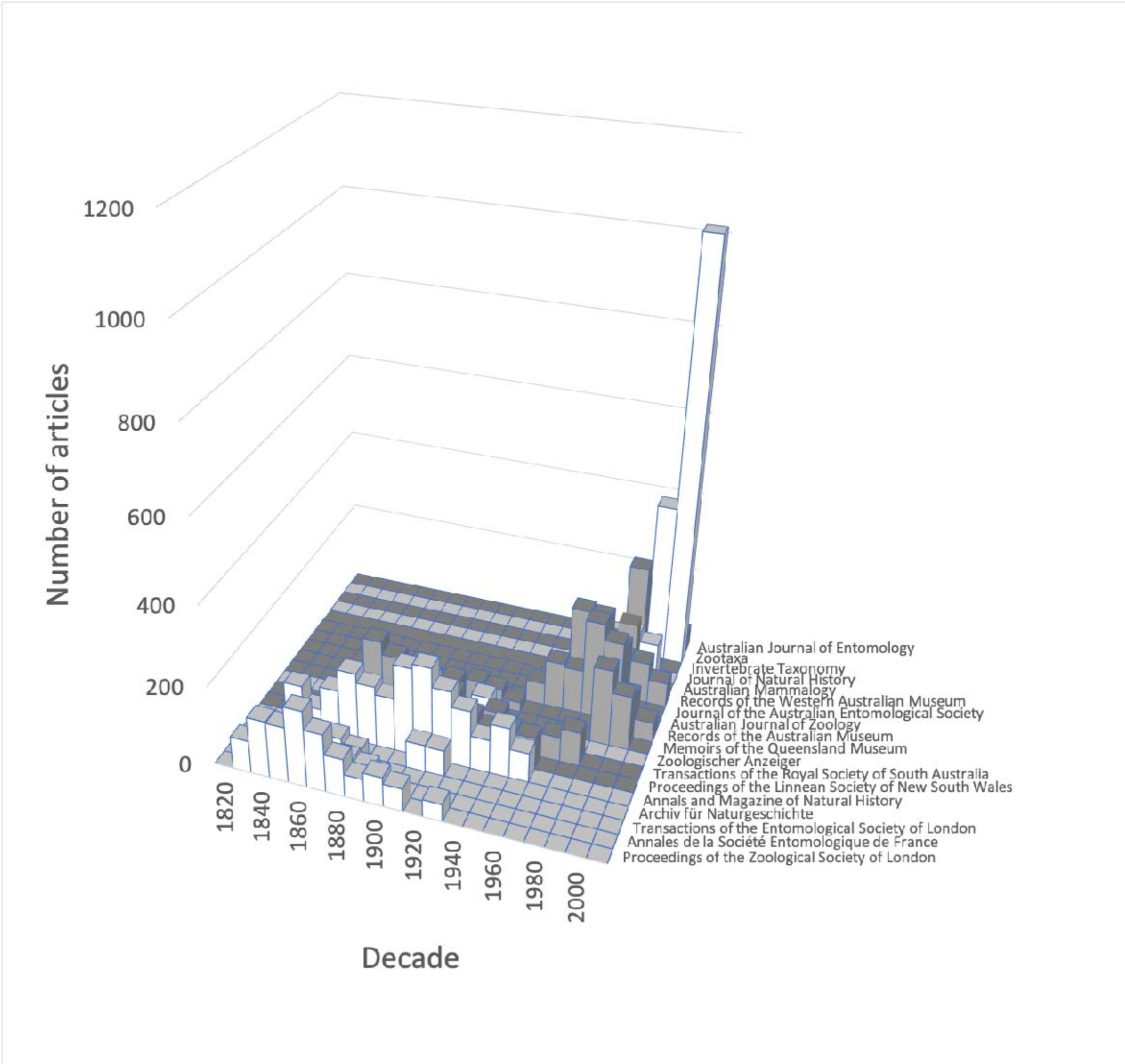

Figure 10. Patterns of publication of taxonomic work on Australian animals 1820-2020. Chart shows the numbers of publications in the top ten journals for each decade. The 19th and early 20th centuries are dominated by European journals, by the mid 20th century most taxonomy was published in Australian journals, more recently international journals such as *Zootaxa* are increasingly important.

Citations and taxonomy as long data

Taxonomy is a "long data" discipline (Page, 2016b). In some scientific fields published papers have a short citation half-life and hence are relatively ephemeral, quickly losing their relevance as the "research front" moves on (de Solla Price, 1965). The rise of academic search engines such as Google Scholar may increase the discoverability of the older literature (and hence increasing its likelihood of being cited, (Verstak et al., 2014)), but for many fields the older literature fades from importance. In contrast, the taxonomic literature is essentially ageless - any published work is potentially relevant. Part of this relevance reflects the importance of priority in biological nomenclature, given competing names for the same taxon in general the oldest name wins. Another factor is the sheer number of species and the relative paucity of published knowledge on many of those species. May (1988) estimated that for publications in the period 1978 to 1987 for insects there were on average 0.02 papers per species per year, for beetles it was 0.01 papers. Hence a researcher may have to search back through a hundred years of literature in order to find mention of a specific beetle species.

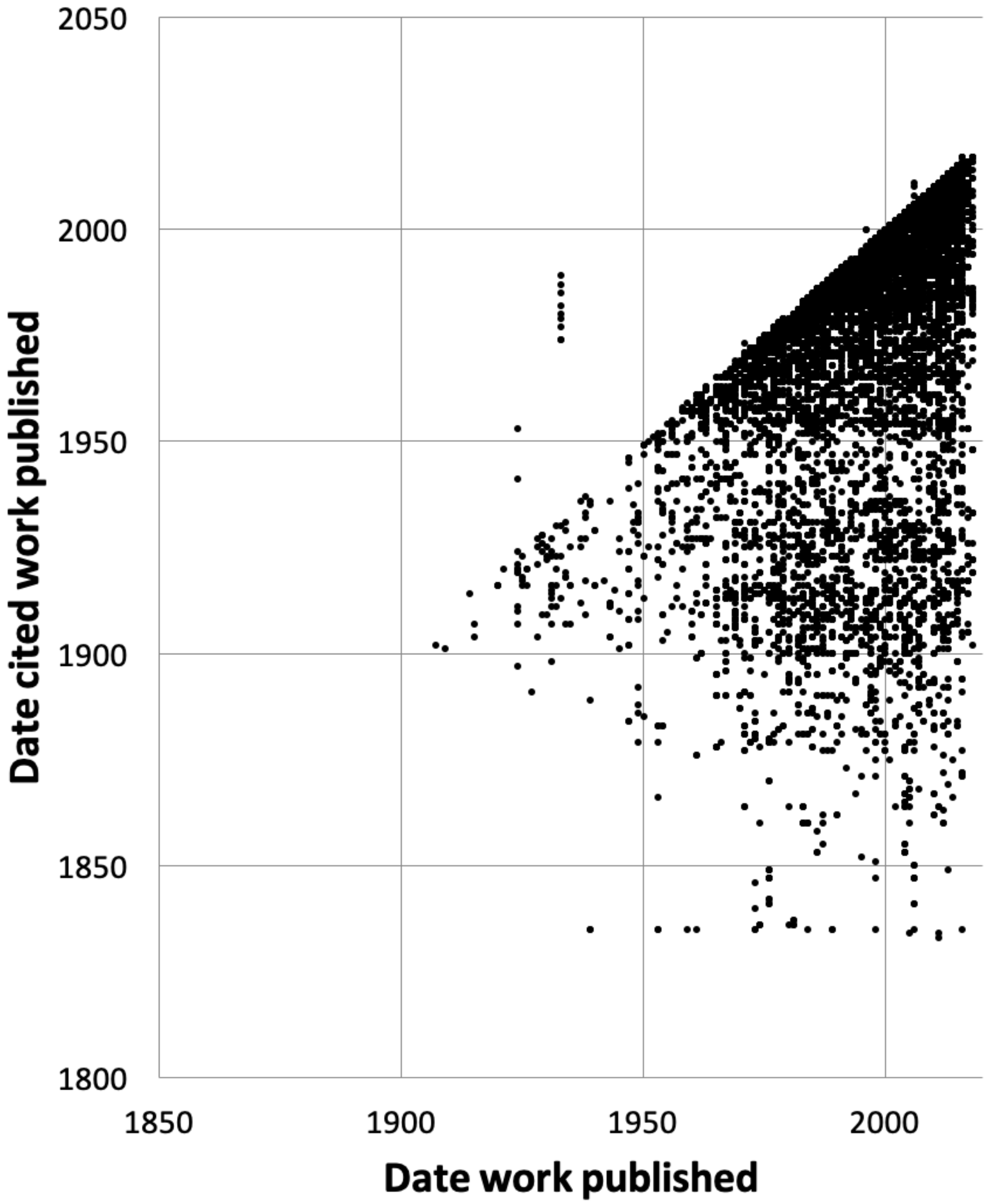

Figure 11. Dates of publication of works cited against the date of publication of the cited work. Each point represents the (*x*, *y*) pair (publication date, cited publication date). All cited works must, by definition, be published in the same year or earlier, and hence the points fall

on or below the diagonal. The few points that are above the diagonal represent errors in CrossRef's metadata.

To explore the citation graph for publications on the Australian fauna I queried each citation relationship for the dates of publication of the citing and the cited works. The relationship between these two dates (Figure 11) highlights the enduring value of the older taxonomic literature. If taxonomic work cited only recent publications then the points in Figure 11would fall on or close to the diagonal. However, even papers published recently (top right of the chart) cite older literature (represented by the vertical columns of dots below each year), and hence much of the area below the diagonal is occupied.

History of species discovery in different taxonomic groups

The knowledge graph enables exploration of the taxonomic history of any taxon of interest. (Pullen, Jennings & Oberprieler, 2014) recently reviewed the history of weevil taxonomy in Australia. Ozymandias has some 3958 accepted weevil species. For each accepted taxon in the ALA classification I used a SPARQL query to retrieve the date the species was originally described, and the dates where then grouped by year. The plot of cumulative numbers of accepted species over time (Figure 12) closely matches that reported by Pullen et al.

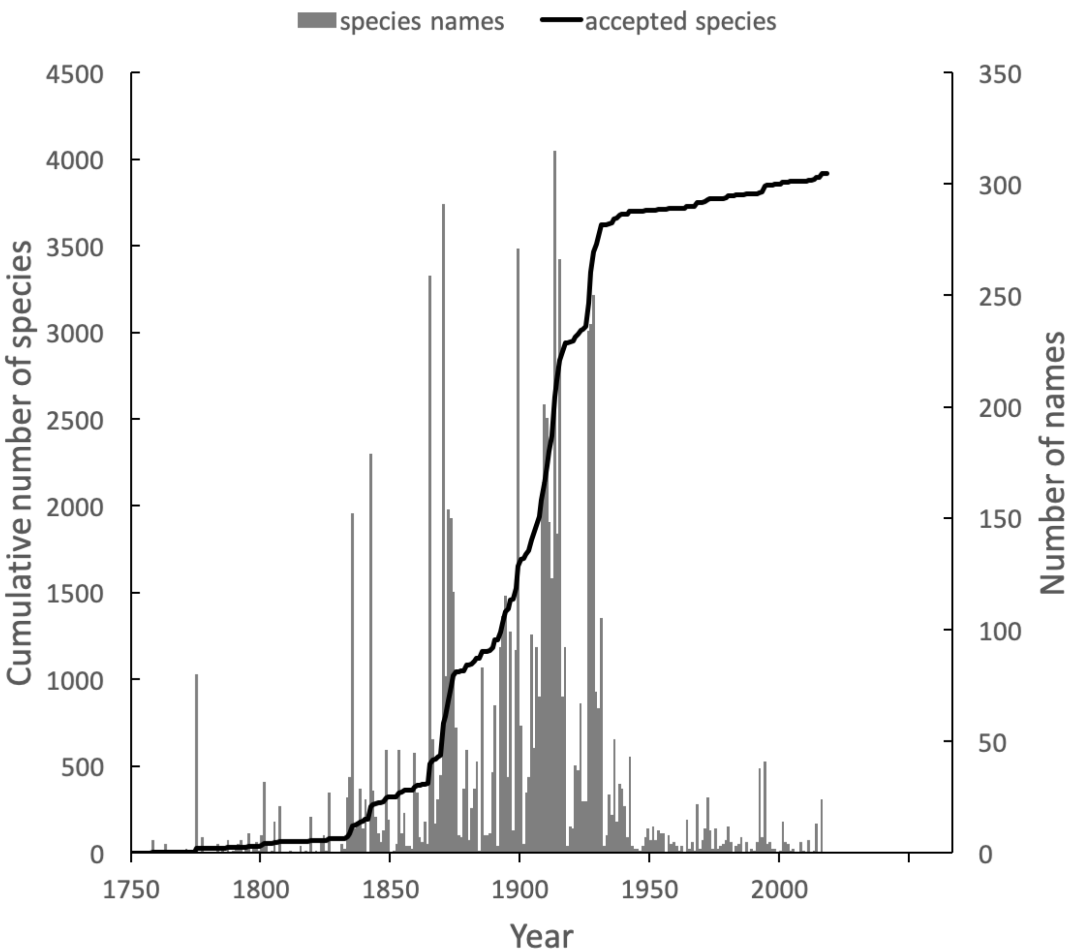

Figure 12. Plot of the history of species discovery for Australian weevils. The solid line is the cumulative number of weevil species that are currently accepted. The vertical bars record the number of new weevil species names published each year. Note the relatively modest increase in names and taxa since the 1930's.

The same chart also shows the number of weevil species names published each year, including synonyms. This chart shows that the bulk of weevil discovery took place in the mid-19th to mid-20th centuries. The sharp drop in species discovery since the 1930's may indicate that the bulk of the Australian weevil fauna has been described, but this seems unlikely given that weevils are typically small and cryptic, and many species leaf-litter and other habitats may remain undiscovered (Stork et al., 2008; Riedel & Tänzler, 2016)

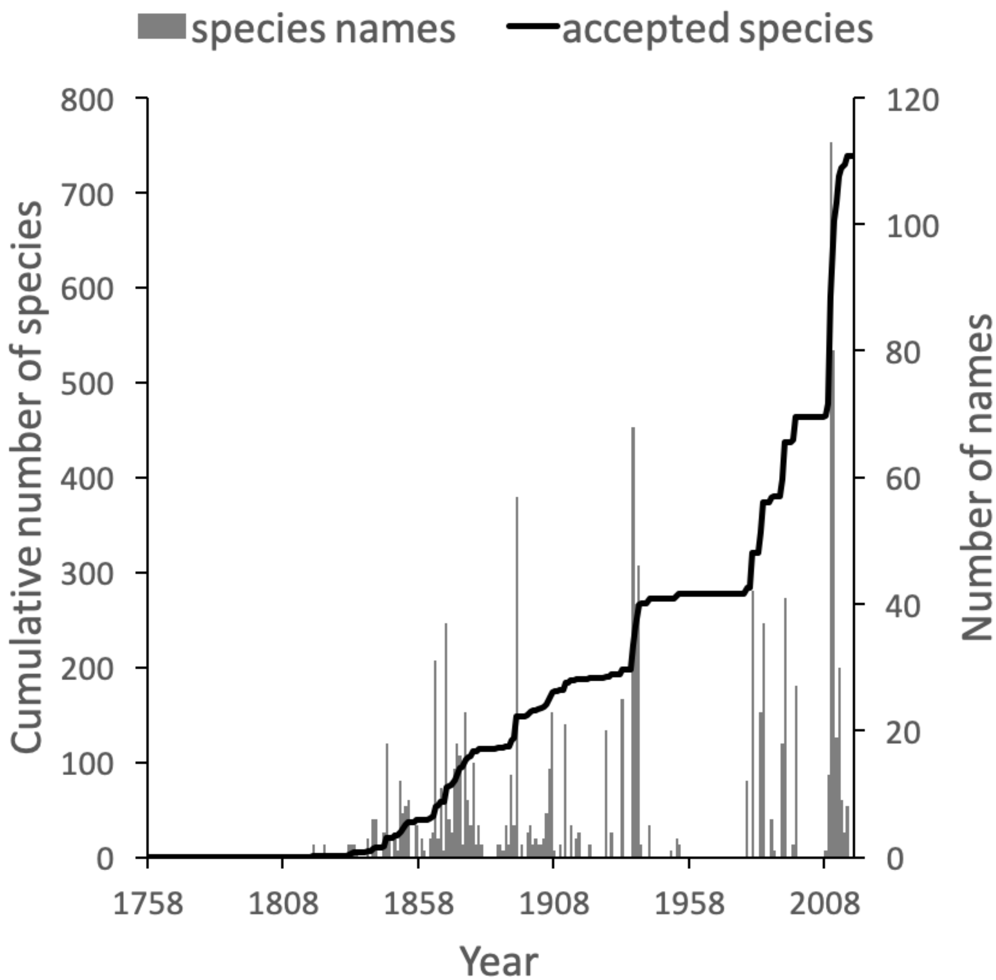

Figure 13. Plot of the history of species discovery for Australian snails in the family Camaenidae. The solid line is the cumulative number of camaenid species that are currently accepted. The vertical bars record the number of new camaenid species names published each year. In contrast to the weevils (Fig. weevils) new Camaenidae species are continuing to be discovered.

These same queries can be used on other taxonomic groups, enabling us to compare the state of knowledge for different taxa. For example, the land snail family Camaenidae (Figure 13) shows a similar pattern of discovery in the mid-19th to mid-20th centuries to that seen in weevils. However, in contrast to weevils these snails have been the subject of ongoing study with over 200 new species being described in the last decade (Köhler, 2010, 2011) a rate of discovery that shows no sign of declining.

Discussion

Building a knowledge graph requires mapping textual representations of entities to identifiers that are shared across data sources ("strings to things"). Creating this mapping is tedious and time consuming to construct, and in a time limited project such as a challenge entry like Ozymandias the mapping is likely to be incomplete before the deadline for the project. Despite its necessarily incomplete state I think the project illustrates some of the ways a network approach can enrich our knowledge of a topic. The web interface exposes many more connections between taxa, publications and people than are evident in the Atlas of Living Australia and Australian Faunal Directory that were used as source databases.

The underlying knowledge graph can be used to support queries exploring the history of taxonomic publishing and discovery. Some of these queries could be used to help prioritise future work. For example, the pattern of citations (Figure 11) confirms that the older taxonomic literature is still relevant today, reinforcing the case for digitising the legacy taxonomic literature. We could further explore the citation data to prioritise which journals should be scanned first: for example, by focusing on those journals that have been cited the most. Given that the bulk of taxonomic publications in the 20th century appeared in Australian journals, initiatives such as the Biodiversity Heritage Library in Australia would seem well placed to make the case that this work should be scanned and made openly available. Citation counts can also be used more directly. For example, the International Institute for Species Exploration annually issues a manually curated list of the "top 10" species discovered the previous year. Such a list could be automatically generated from a knowledge graph using, for example, the number of citations (or other measures of attention) that each work publishing a new species has received.

Some analyses of the knowledge graph are more focussed on the state of the knowledge graph itself. For example, querying for author identifiers such as ORCIDs reveals a limited uptake of that identifier. This has implications for proposals to use ORCID as the basis for tracking the broader activities of taxonomists, such as specimen collection and identification (Shorthouse). Perhaps the development of such tools may help raise awareness of the possible benefits of authors registering with ORCID.

Expanding the knowledge graph

The knowledge graph in Ozymandias features only a subset of the entities depicted in earlier work sketching the "biodiversity knowledge graph" (Page, 2013, 2016a). There are several entities that are obvious candidates to be added to Ozymandias, such as specimens and nucleotide sequences. However, the number of specimens that could potentially be added has implications for the scalability of the knowledge graph. Bearing this in mind, we could add a subset of specimens, such as type specimens, or those which have been sequenced. Fontaine et al. reported that the average lag time between the discovery of a specimen representing a new species and the description of that species is 21 years. The generality of this observation could be evaluated using a knowledge graph that contains both the taxonomic literature and type specimens with dates of collection.

The Biodiversity Literature Repository highlights the potential of treating scientific articles not as monolithic entities but rather as assembles of component parts, including figures. We can drill down further and start to annotate individual parts including fragments of text. The idea of annotating and interlinking fragments of text has a long history, pioneered by people such as Ted Nelson (Douglas R. Dechow & Daniele C. Struppa, 2015), and tools such as Hypotheses.is ("Hypothes.is") now make this possible. We could view the "micro citations" used by taxonomists to specify the page location of a taxonomic name as a form of annotation, hence a logical next step is to map these micro citations onto publications in the knowledge graph so that we can locate these micro citations in the context of the taxonomic literature that they refer to.

The future of knowledge graphs

To the extent that Ozymandias is judged to be a success it suggests that knowledge graphs have potential to improve the way we aggregate and interface with biodiversity data. However, it is worth noting that the biodiversity informatics community has been aware of knowledge graphs and semantic web technologies for a decade or more, and several taxonomic databases have been serving data in RDF since the mid-2000's. Yet it is hard to point to successful applications of these approaches to the study of biodiversity, and there has been limited uptake of linked data beyond a few databases.

There is a considerable cost involved in cross linking datasets, and to date the rewards for this effort are, perhaps, unclear. At the same time, there is growing concern within biology in general (McDade et al., 2011) and in taxonomy in particular, that existing measures of the output of researchers, such as citations, are poor metrics of activity (cite citation papers). There are also concerns that existing data aggregators do not pay enough attention to tracking the provenance and authorship of information (Franz & Sterner, 2018). Researchers may do much more than write papers, they may clean, prepare, and publish datasets, collect specimens, curate collections, identify specimens, etc. Keeping track of these activities is greatly facilitated by the use of stable identifiers for people and the objects they work with (e.g., specimens, collections, datasets), and a knowledge graph would be an ideal data structure to quantify the work done, and trace the provenance of data and associated annotations. Projects such as Scholia (Nielsen, Mietchen & Willighagen, 2017) already demonstrate the potential of Wikidata to explore the output of scholars. Hence, it may be that the best way to bootstrap the adoption of biodiversity knowledge graphs is to focus on the implications for being able to give appropriate credit to researchers for all the activities that they undertake.

There is considerable enthusiasm for the potential of identifiers to help evaluate research (Haak, Meadows & Brown, 2018) and yield insights into the behaviour of researchers (Bohannon, 2017). However, the ease with which measures of research activity (such as citation-based measures) switch from being tools for insight into targets to be met suggests we should consider the possibility that metrics developed to create incentives to build knowledge graphs may ultimately harm the researchers being measured.

Beyond internal drivers, such as documenting the provenance of taxonomic information, and quantifying the contributions of researchers, there are also external drivers for considering knowledge graphs. Wikidata (Vrandečić & Krötzsch, 2014) is an open, global knowledge graph with an enthusiastic community of editors, and many of the entities taxonomists care about are already included in the graph, such as taxa, people, and publications. This means that we can use Wikidata to help define the scope of a knowledge graph. Anyone constructing a knowledge graph rapidly runs into the problem of scope, in other words, when do you stop adding entities? Once we move beyond specialist knowledge in a given field (such as specimens, rules of nomenclature, sequences and phylogenies) and include more generic entities that other communities may also be interested in (such as publications, natural history collections, people) we reach the point at which we can stop constructing our graph and defer to Wikidata. Hence a key part of the future development of biodiversity knowledge graphs will be to determine the extent to which Wikidata and its community can be responsible for managing biodiversity-related data.

Additional Information and Declarations

Competing Interests

The author declares he has no competing interests.

Author Contributions

Roderic D.M. Page conceived and designed the experiments, performed the experiments, analysed the data, contributed reagents/materials/analysis tools, wrote the paper.

Data Availability

The Ozymandias web site is https://ozymandias-demo.herokuapp.com. This site includes a SPARQL interface to query the knowledge graph directly. Source code for the interface is available from GitHub https://github.com/rdmpage/ozymandias-demo. Source code for the scripts used to harvest and clean the data used to populate the knowledge graph is available from https://github.com/rdmpage/oz-afd-export, https://github.com/rdmpage/oz-ala-harvest, https://github.com/rdmpage/oz-csl, and https://github.com/rdmpage/oz-wikispecies.

Funding

The work described here was an entry in the Global Biodiversity Information Facility 2018 Ebbe Nielsen Challenge. GBIF had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

Constructing the knowledge graph described here would have been impossible without the wealth of freely available and open source software used in the project. Furthermore, it should be obvious that none of this would have been possible without the centuries of taxonomic research by generations of researchers, and the recent efforts to make that research digitally accessible via projects such as Atlas of Living Australia and the Australian Faunal Directory. I'm also indebted to GBIF for running the 2018 Ebbe Nielsen Challenge which gave me a hard deadline to work towards. I also thank Steve Baskauf and Joel Sachs for feedback on the project, and for invitations to present Ozymandias to their colleagues.

References

Altmetric. *Available at https://www.altmetric.com/* (accessed November 28, 2018).

Atlas of Living Australia. *Available at https://www.ala.org.au/* (accessed November 27, 2018).

Australian Faunal Directory. *Available at https://biodiversity.org.au/afd/home* (accessed November 27, 2018).

Bebber DP, Wood JRI, Barker C, Scotland RW. 2013. Author inflation masks global capacity for species discovery in flowering plants. *New Phytologist* 201:700–706. DOI: 10.1111/nph.12522.

Biodiversity Literature Repository. *Available at http://plazi.org/resources/bibliography-of-life-bol/biodiversity-literature-repository-blr/* (accessed November 28, 2018).

Bohannon J. 2017. Vast set of public CVs reveals the world's most migratory scientists. *Science*. DOI: 10.1126/science.aal1189.

Bollacker K, Evans C, Paritosh P, Sturge T, Taylor J. 2008. Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge. In: *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*. SIGMOD '08. New York, NY, USA: ACM, 1247–1250. DOI: 10.1145/1376616.1376746.

Costello MJ, Wilson S, Houlding B. 2013. More Taxonomists Describing Significantly Fewer Species per Unit Effort May Indicate That Most Species Have Been Discovered. *Systematic Biology* 62:616–624. DOI: 10.1093/sysbio/syt024.

CrossRef. *Available at https://www.crossref.org/* (accessed November 27, 2018).

Douglas R. Dechow, Daniele C. Struppa (eds.). 2015. Intertwingled: The work and influence of Ted Nelson. *History of Computing*. DOI: 10.1007/978-3-319-16925-5.

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.

Franz NM, Sterner BW. 2018. To increase trust, change the social design behind aggregated biodiversity data. *Database* 2018. DOI: 10.1093/database/bax100.

Godfray HCJ. 2007. Linnaeus in the information age. *Nature* 446:259–260. DOI: 10.1038/446259a.

Grieneisen ML, Zhan Y, Potter D, Zhang M. 2014. Biodiversity, Taxonomic Infrastructure, International Collaboration, and New Species Discovery. *BioScience* 64:322–332. DOI: 10.1093/biosci/biu035.

Haak LL, Meadows A, Brown J. 2018. Using ORCID, DOI, and Other Open Identifiers in Research Evaluation. *Frontiers in Research Metrics and Analytics* 3. DOI: 10.3389/frma.2018.00028.

Hypothes.is. *Available at https://web.hypothes.is/* (accessed November 28, 2018).

Joppa LN, Roberts DL, Pimm SL. 2011. The population ecology and social behaviour of taxonomists. *Trends in Ecology & Evolution* 26:551–553. DOI: 10.1016/j.tree.2011.07.010.

Kennedy J, Hyam R, Kukla R, Paterson T. 2006. Standard Data Model Representation for Taxonomic Information. *OMICS: A Journal of Integrative Biology* 10:220–230. DOI: 10.1089/omi.2006.10.220.

Köhler F. 2010. Uncovering local endemism in the Kimberley, Western Australia: description of new species of the genus Amplirhagada Iredale, 1933 (Pulmonata: Camaenidae). *Records of the Australian Museum* 62:217–284. DOI: 10.3853/j.0067-1975.62.2010.1554.

Köhler F. 2011. Descriptions of new species of the diverse and endemic land snail Amplirhagada Iredale, 1933 from rainforest patches across the Kimberley, Western Australia (Pulmonata: Camaenidae). *Records of the Australian Museum* 63:167–202. DOI: 10.3853/j.0067-1975.63.2011.1581.

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

McDade LA, Maddison DR, Guralnick R, Piwowar HA, Jameson ML, Helgen KM, Herendeen PS, Hill A, Vis ML. 2011. Biology Needs a Modern Assessment System for Professional Productivity. *BioScience* 61:619–625. DOI: 10.1525/bio.2011.61.8.8.

Michel F, Gargominy O, Tercerie S, Faron Zucker C. 2017. A Model to Represent Nomenclatural and Taxonomic Information as Linked Data. Application to the French Taxonomic Register, TAXREF. In: *S4Biodiv 2017 - 2nd International Workshop on Semantics for Biodiversity co-located with ISWC 2017*. Vienna, Austria, 1–12.

Nakabo T. 1982. REVISION OF GENERA OF THE DRAGONETS (PISCES□: CALLIONYMIDAE). *Publications of the Seto Marine Biological Laboratory* 27:77–131. DOI: 10.5134/176044.

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. 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 RDM. 2016a. Towards a biodiversity knowledge graph. *Research Ideas and Outcomes* 2:e8767. DOI: 10.3897/rio.2.e8767.

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.

Pullen KR, Jennings D, Oberprieler RG. 2014. Annotated catalogue of Australian weevils (Coleoptera: Curculionoidea). *Zootaxa* 3896:1. DOI: 10.11646/zootaxa.3896.1.1.

Quan DA, Karger R. 2004. How to make a semantic web browser. In: ACM Press,. DOI: 10.1145/988672.988707.

Riedel A, Tänzler R. 2016. Revision of the Australian species of the weevil genus Trigonopterus Fauvel. *ZooKeys* 556:97–162. DOI: 10.3897/zookeys.556.6126.

Sangster G, Luksenburg JA. 2014. Declining Rates of Species Described per Taxonomist: Slowdown of Progress or a Side-effect of Improved Quality in Taxonomy? *Systematic Biology* 64:144–151. DOI: 10.1093/sysbio/syu069.

Schema.org. *Available at https://schema.org/* (accessed December 3, 2018).

Senderov V, Simov K, Franz N, Stoev P, Catapano T, Agosti D, Sautter G, Morris RA, Penev L. 2018. OpenBiodiv-O: ontology of the OpenBiodiv knowledge management system. *Journal of Biomedical Semantics* 9. DOI: 10.1186/s13326-017-0174-5.

Shorthouse DP.Bloodhound. *Available at https://bloodhound.shorthouse.net* (accessed November 28, 2018).

de Solla Price DJ. 1965. Networks of Scientific Papers. *Science* 149:510–515. DOI: 10.1126/science.149.3683.510.

Stork NE, Grimbacher PS, Storey R, Oberprieler RG, Reid C, Slipinski SA. 2008. What determines whether a species of insect is described? Evidence from a study of tropical forest beetles. *Insect Conservation and Diversity* 1:114–119. DOI: 10.1111/j.1752-4598.2008.00016.x.

Tancoigne E, Ollivier G. 2017. Evaluating the progress and needs of taxonomy since the Convention on Biological Diversity: going beyond the rate of species description. *Australian Systematic Botany* 30:326. DOI: 10.1071/sb16017.

Verstak A, Acharya A, Suzuki H, Henderson S, Iakhiaev M, Lin CCY, Shetty N. 2014. On the Shoulders of Giants: The Growing Impact of Older Articles. *arXiv:1411.0275 [cs]*.

Vicki Tardif Holland, Jason Johnson. 2014.Introducing "Role." *Available at http://blog.schema.org/2014/06/introducing-role.html* (accessed November 29, 2018).

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

W3C SPARQL Working Group. 2013.SPARQL 1.1 Overview. *Available at https://www.w3.org/TR/sparql11-overview/* (accessed November 27, 2018).

Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, Robertson T, Vieglais D. 2012. Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. *PLoS ONE* 7:e29715. DOI: 10.1371/journal.pone.0029715.

Supplementary Information

Publications and identifiers

Get count of number of published works for each year, and number of works with identifiers.

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT ?work_date (COUNT(?w) as ?c) (COUNT(?doi) as ?c_doi)
(COUNT(?biostor) as ?c_biostor) (COUNT(?jstor) as ?c_jstor)
(COUNT(?pdf) as ?c_pdf)
WHERE
{
  ?w <http://schema.org/datePublished> ?work_date .

  # just articles
  ?w <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://schema.org/ScholarlyArticle> .

  # DOI?
  OPTIONAL {
  ?w <http://schema.org/identifier> ?doi .
  ?doi <http://schema.org/propertyID> "doi" .
  }

  # BioStor?
  OPTIONAL {
  ?w <http://schema.org/identifier> ?biostor .
  ?biostor <http://schema.org/propertyID> "biostor" .
  }

  # JSTOR?
  OPTIONAL {
  ?w <http://schema.org/identifier> ?jstor .
  ?jstor <http://schema.org/propertyID> "jstor" .
  }


  # PDF?
  OPTIONAL {
  ?w <http://schema.org/encoding> ?pdf .
```

```
  ?pdf <http://schema.org/fileFormat> "application/pdf" .
  }


  FILTER regex(?work_date, "^[0-9]{4}$")

  #FILTER (xsd:integer(?work_date) > 1980)
}
GROUP BY ?work_date
ORDER BY ?work_date
```

Data in publications.tsv

Journal ranks

Query to retrieve top 10 journals for a given decade (in this case 1910)

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX tc: <http://rs.tdwg.org/ontology/voc/TaxonConcept#>
SELECT   ?journal ?issn  (COUNT(?journal) AS ?count) WHERE
{
 ?work  <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://schema.org/ScholarlyArticle> .
 ?work <http://schema.org/isPartOf> ?container .
  ?container <http://schema.org/name> ?journal .
 ?work <http://schema.org/datePublished> ?year .

  OPTIONAL {
  ?container <http://schema.org/issn> ?issn .
  }
  FILTER ((xsd:integer(?year) >= 1910) && (xsd:integer(?year)
< " . ($year + 9) . "))
}
GROUP BY ?journal ?issn
ORDER BY DESC(?count)
LIMIT 10
```

Repeat this query for all decades, aggregate results, then filter for journals with > 200 articles.

Data in journals.tsv

Citation patterns

Find all pairs of citing articles and get dates they were published.

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT ?cited_identifier_type (xsd:integer(?w_year) as ?from)
(xsd:integer(?work_year) as ?to)
WHERE
{

?w <http://schema.org/identifier> ?identifier .
  ?w <http://schema.org/name> ?w_name .
?w <http://schema.org/datePublished> ?w_year .
# Identifier (e.g., DOI) for work we are displaying
?identifier <http://schema.org/value> ?identifier_value .

?citing_identifier <http://schema.org/value> ?identifier_value
.
?citing <http://schema.org/identifier> ?citing_identifier .

# What does this work cite (typically from CrossRef data)
?citing <http://schema.org/citation> ?cited .

# Translate the citing work\'s DOI (or other identifier) into
AFD identifier
# Get identifier (typically a DOI) for citing work
?cited <http://schema.org/identifier> ?cited_identifier .
?cited_identifier <http://schema.org/value>
?cited_identifier_value .
?cited_identifier <http://schema.org/propertyID>
?cited_identifier_type .


# Get work(s) with this identifer (may have > 1 if we have
CrossRef record in our triple store
?work_identifier <http://schema.org/value>
?cited_identifier_value .
?work <http://schema.org/identifier> ?work_identifier .
?work <http://schema.org/name> ?name .
?work <http://schema.org/datePublished> ?work_year .

# Just include citing records that are also in ALA
```

```
FILTER regex(str(?work),\'biodiversity.org.au\') .
FILTER regex(str(?w),\'biodiversity.org.au\') .

FILTER regex(?w_year, "^[0-9]{4}$")
FILTER regex(?work_year, "^[0-9]{4}$")
}
```

Data in cites.tsv

Weevils

Number of accepted taxon names per year.

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT ?year (COUNT(?taxonName) AS ?count)
WHERE
{
VALUES ?root_name {"CURCULIONOIDEA"}
?root <http://schema.org/name> ?root_name .
?child rdfs:subClassOf+ ?root .
?child rdfs:subClassOf ?parent .
?child <http://schema.org/name> ?child_name .
?parent <http://schema.org/name> ?parent_name .

  ?child
<http://taxref.mnhn.fr/lod/property/hasReferenceName> ?taxon-
Name .

  ?taxonName
<http://rs.tdwg.org/ontology/voc/TaxonName#rankString> "spe-
cies" .
  ?taxonName <http://rs.tdwg.org/ontology/voc/TaxonName#year>
?year .

}
GROUP BY ?year
ORDER BY ?year
```

Sum these to generate cumulative total.

Number of weevil names published each year:

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT ?year (COUNT(DISTINCT ?name) AS ?c)
WHERE
{
VALUES ?root_name {"CURCULIONOIDEA"}
?root <http://schema.org/name> ?root_name .
?child rdfs:subClassOf+ ?root .
?child rdfs:subClassOf ?parent .
?child <http://schema.org/name> ?child_name .
?parent <http://schema.org/name> ?parent_name .

  ?child
<http://taxref.mnhn.fr/lod/property/hasReferenceName>|<http://
taxref.mnhn.fr/lod/property/hasSynonym> ?taxonName .
  ?taxonName
<http://rs.tdwg.org/ontology/voc/TaxonName#rankString> "spe-
cies" .
  ?taxonName <http://schema.org/name> ?name .
  ?taxonName <http://rs.tdwg.org/ontology/voc/TaxonName#year>
?year .

}

GROUP BY ?year
ORDER BY ?year
```

Combined data in weevils.tsv

Snails

Number of accepted taxon names per year

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT ?year (COUNT(?taxonName) AS ?count)
WHERE
{
VALUES ?root_name {"CAMAENIDAE"}
?root <http://schema.org/name> ?root_name .
?child rdfs:subClassOf+ ?root .
?child rdfs:subClassOf ?parent .
?child <http://schema.org/name> ?child_name .
?parent <http://schema.org/name> ?parent_name .
```

```
  ?child
<http://taxref.mnhn.fr/lod/property/hasReferenceName> ?taxon-
Name .

  ?taxonName
<http://rs.tdwg.org/ontology/voc/TaxonName#rankString> "spe-
cies" .
  ?taxonName <http://rs.tdwg.org/ontology/voc/TaxonName#year>
?year .

}
GROUP BY ?year
ORDER BY ?year
```

Sum these to generate cumulative total.

Number of snail names published each year:

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT ?year (COUNT(DISTINCT ?name) AS ?c)
WHERE
{
VALUES ?root_name {"CAMAENIDAE"}
?root <http://schema.org/name> ?root_name .
?child rdfs:subClassOf+ ?root .
?child rdfs:subClassOf ?parent .
?child <http://schema.org/name> ?child_name .
?parent <http://schema.org/name> ?parent_name .

  ?child
<http://taxref.mnhn.fr/lod/property/hasReferenceName>|<http://
taxref.mnhn.fr/lod/property/hasSynonym> ?taxonName .
  ?taxonName
<http://rs.tdwg.org/ontology/voc/TaxonName#rankString> "spe-
cies" .
  ?taxonName <http://schema.org/name> ?name .
  ?taxonName <http://rs.tdwg.org/ontology/voc/TaxonName#year>
?year .

}

GROUP BY ?year
ORDER BY ?year
```

Combined data in snails.tsv

Authors and ORCIDs

How many authors of works with DOIs post 2011?

```
SELECT (COUNT(DISTINCT ?creator) as ?c)
WHERE
{
  GRAPH <https://biodiversity.org.au/afd/publication> {

?work <http://schema.org/identifier> ?identifier .
?identifier <http://schema.org/propertyID> "doi" .
?identifier <http://schema.org/value> ?doi .

?work <http://schema.org/datePublished> ?datePublished .

?work <http://schema.org/creator> ?role  .
?role <http://schema.org/roleName> ?roleName .
?role <http://schema.org/creator> ?creator .
?creator <http://schema.org/name> ?name .
}

FILTER (xsd:integer(?datePublished) > 2011)
}
```

How many authors of works with DOIs post 2011 had an ORCID?

```
SELECT DISTINCT ?orcid_creator
WHERE
{
  GRAPH <https://biodiversity.org.au/afd/publication> {

?work <http://schema.org/identifier> ?identifier .
?identifier <http://schema.org/propertyID> "doi" .
?identifier <http://schema.org/value> ?doi .

?work <http://schema.org/datePublished> ?datePublished .
```

```
?work <http://schema.org/creator> ?role  .
?role <http://schema.org/roleName> ?roleName .
?role <http://schema.org/creator> ?creator .
?creator <http://schema.org/name> ?name .
}

GRAPH <https://orcid.org>
  {
    ?orcid_identifier <http://schema.org/value> ?doi .
    ?orcid_work <http://schema.org/identifier> ?or-
cid_identifier .

     ?orcid_work <http://schema.org/creator> ?orcid_role  .
    ?orcid_role <http://schema.org/roleName> ?orcid_roleName
.

    ?orcid_role <http://schema.org/creator> ?orcid_creator  .

    ?orcid_creator <http://schema.org/name> ?orcid_name .
  }

  FILTER(?roleName = ?orcid_roleName)
FILTER (xsd:integer(?datePublished) > 2011)
}
```