20 GB in 10 minutes: A case for linking major biodiversity databases using an open socio-technical infrastructure and a pragmatic, cross-institutional collaboration

Anne E Thessen[1]*, Jorrit H. Poelen[2], Matthew Collins[3], Jen Hammock[4]

[1]The Ronin Institute for Independent Scholarship, Montclair, NJ 07043, USA
[2]400 Perkins Street, Apt. 104, Oakland, CA 94610, USA
[3]University of Florida, 968 Center Drive, Gainsville, FL, 32611, USA
[4]National Museum of Natural History, 10th and Constitution Ave., Washington, DC
*Corresponding author annethessen@gmail.com

Abstract

17   Biodiversity information is made available through numerous databases that each have their own
18   data models, web services, and data types. Combining data across databases leads to new
19   insights, but is not easy because each database uses its own system of identifiers. In the absence
20   of stable and interoperable identifiers, databases are often linked using taxonomic names. This
21   labor intensive, error prone, and lengthy process relies on accessible versions of nomenclatural
22   authorities and fuzzy-matching algorithms.
23
24   To approach the challenge of linking diverse data, more than technology is needed. New social
25   collaborations like the Global Unified Open Data Architecture (GUODA) that combine skills
26   from diverse groups of computer engineers from iDigBio, server resources from the Advanced
27   Computing and Information Systems (ACIS) Lab, global-scale data presentation from EOL, and
28   independent developers and researchers are what is needed to make concrete progress on finding
29   relationships between biodiversity datasets.
30
31   This paper will discuss a technical solution developed by the GUODA collaboration for faster
32   linking across databases with a use case linking Wikidata and the Global Biodiversity
33   Interactions database (GloBI). The GUODA infrastructure is a 12-node, high performance
34   computing cluster made up of about 192 threads with 12 TB of storage and 288 GB memory.
35   Using GUODA, 20GB of compressed JSON from Wikidata was processed and linked to GloBI
36   in about 10-11 minutes. Instead of comparing name strings or relying on a single identifier,
37   Wikidata and GloBI were linked by comparing graphs of biodiversity identifiers external to each
38   system. This method resulted in adding 119,957 Wikidata links in GloBI, an increase of 13.7%
39   of all outgoing name links in GloBI. Wikidata and GloBI were compared to Open Tree
40   Taxonomy to examine consistency and coverage. The process of parsing Wikidata, Open Tree
41   Taxonomy and GloBI archives and calculating consistency metrics was done in minutes on the
42   GUODA platform. As a model collaboration, GUODA has the potential to revolutionize
43   biodiversity science by bringing diverse technically minded people together with high
44   performance computing resources that are accessible from a laptop or desktop. However,
45   participating in such a collaboration still requires basic programming skills.
46

Introduction

48   Biodiversity databases provide global access to information about species via the Web. These
49   databases contain information as varied as observation records, text descriptions, images, maps,
50   genetic sequences, phylogenetic trees, and trait data (Table 1). All of these data become much
51   more useful if they can be linked. Many biodiversity databases share information with each other
52   (Bingham et al. 2017), but creating the links can be very difficult for several reasons including

53 the size of the databases, the heterogeneous nature of the data, and the heterogeneous nature of
54 the identifiers used by the different resources (Page 2008).
55
56 The more popular methods for linking biodiversity databases include taxonomic names, LSID
57 (Life Sciences Identifier), and DOI (Digital Object Identifier). The Encyclopedia of Life uses
58 taxonomic names to automatically aggregate data from hundreds of providers (Parr et al. 2014).
59 BioNames links data using LSID, DOI, handles, bibliographic citations, and taxonomic names
60 (Page 2013). The iPhylo LinkOut service mapped identifiers used by the NCBI taxonomy
61 database (which provides the taxonomic backbone for GenBank) to Wikipedia pages using
62 taxonomic names, including synonyms (Page 2011). TBMap provides links from TreeBase
63 across several taxonomic databases, such as ITIS and NCBI (Page 2007). This mapping was also
64 achieved using taxonomic names, but in some cases GenBank Accession numbers and museum
65 specimen codes were available for supplement. The use of taxonomic names to aggregate data
66 can lead to errors and requires significant *a priori* knowledge either in the form of curators or an
67 authoritative nomenclature.
68
69 Many databases expose their own internal identifiers, such as the WoRMS Aphia ID, so others
70 can link their data to those resources within their own systems, often by providing a URL.
71 Databases like WoRMS provide web services that allow users to look up an identifier for a taxon
72 in question, one at a time. While this makes linking easier, it is still difficult to scale across all
73 databases. For example, a list of all the taxon identifiers in EOL is 300 MB compressed. No
74 system of identifiers is universal across biodiversity databases and none of them are easy to
75 implement at scale.
76
77 While the data would be much more useful if linked, there is a lack of tools for linking data
78 across databases at scale. Most mappings are done at great expense and then are made available
79 as a separate file or incorporated into the resources themselves. LinkOut, BioNames, GBIF, and
80 EOL take more than a day to link across their entire body of aggregated content. This paper
81 discusses links made between GloBI and Wikidata (WD) in 10 minutes using GUODA, a high
82 performance computing system available for analysis of large biodiversity data sets.

83 Methods

84 Description of Resources

85 GUODA

86 Following an iDigBio hack-a-thon in June 2015, GUODA was created as a pragmatic way to
87 compute over multiple large biodiversity databases in a mutually beneficial collaboration
88 between iDigBio, EOL, Kew Garden, and independent developers. Catalyzed by various
89 presentations at conferences, hardware provided by ACIS, 20+ meetings, and several prototypes
90 (e.g., http://effechecka.org, https://gimmefreshdata.github.io), a general access biodiversity data

91   integration and analysis environment was created. This environment, with the aggregated
92   experience and perspectives of all the collaborators, was used to produce the results of this paper.
93
94   Housed at the ACIS Lab at the University of Florida, the GUODA infrastructure consists of 12
95   IBM HS22 blades each with 8 cores, 24 GB of memory, and 1 TB of storage each. This makes a
96   total of 192 threads, 288 GB of memory and 12 TB of disk space available for processing jobs
97   using Apache Spark (Fig 1; Zaharia et al. 2016). The cluster is managed under Apache Mesos
98   (Hindman et al. 2011) which is a distributed scheduling system for periodic jobs. For long
99   running processes, such as web APIs or databases, the Marathon
100   (https://github.com/mesosphere/marathon) framework is run within Mesos. Marathon facilitates
101   running always-up services with monitoring, automatic deployment of code, re-scaling to
102   multiple nodes, and other management features. Mesos is responsible for accepting requests to
103   start Spark frameworks, processes which do the actual computation and may span multiple
104   servers, and allocation of resources requested by the framework.
105
106   Hadoop HDFS (Shvachko et al. 2010) is installed outside of Mesos directly on all 12 nodes of
107   the cluster and provides redundant parallel shared storage to all nodes as well as the Jupyter
108   notebook (Kluyver et al. 2016) server that provides a programming interface to end users. Each
109   node has 1 TB of local disk storage for a total of about 3.5 TB of usable storage space for data
110   files in Apache Parquet format. Spark is aware of the placement of data on an HDFS cluster and
111   will divide processing among nodes in a way that prefers to read and write data that is local to
112   the node to minimize network traffic.

Wikidata

114   Wikidata (WD) is a free and open knowledge base that provides structured data for WikiMedia
115   projects (www.wikidata.org; Vrandečić & Krötzsch 2014). Similar to Wikipedia, anyone can
116   read or edit the resource. Information, including links to other resources, can be added to
117   Wikidata using bots and batch imports through their Data Import Hub
118   (https://www.wikidata.org/wiki/Wikidata:Data_Import_Hub). Wikidata information about taxa
119   can be conceptualized as a graph linking related taxa to each other and identifiers from other
120   databases to the taxa they represent (Fig 2). Every taxon in Wikidata is issued a Wikidata
121   identifier. While a public Wikidata SPARQL endpoint and associated tools (Voß 2016) exist,
122   these APIs are not suitable for batch processing. For example, when attempting to retrieve all
123   taxa using the public SPARQL endpoint, a query timeout error was reported. In addition, the
124   APIs are expected to return different results over time, so reproducing results is difficult if not
125   impossible. This is why we used a json archive to access Wikidata (Wikidata 2018).

GloBI

127   GloBI is a database of biotic interactions recorded as Organism_1:has_relationship:Organism_2
128   (Poelen et al. 2014). GloBI uses a combination of web APIs, taxon archives, and name
129   correction/parsing methods in an attempt to link names from species interaction datasets to

130   existing sources. Spatial, temporal, and taxonomic coverage in GloBI is sparse and unevenly
131   distributed (see Eltonian shortfall, Hortal et al. 2015), with spatial concentrations in Europe and
132   North America and taxonomically concentrated in Arthropods, Fungi, and Plants. Only 8% of
133   taxa in ITIS are also in GloBI. A detailed technical description of the GloBI data model and
134   services has been published elsewhere (Poelen et al. 2014). GloBI maintains a graph of related
135   taxa and their identifiers from different databases (Poelen et al. 2014). GloBI does not introduce
136   its own taxon ids. Instead, it records how names were mapped from a source name into an
137   external taxonomic database using a taxon graph (see
138   https://globalbioticinteractions.org/references). We used GloBI Taxon Graph v0.4.2 (Poelen
139   2018b). This taxon graph links names and identifiers hierarchically and across resources.

Open Tree Taxonomy

141   To assess taxonomic id coverage, the taxa in Wikidata and GloBI were compared to Open Tree
142   Taxonomy (OTT 3.0; http://files.opentreeoflife.org/ott/ott3.0/ott3.0.tgz; Rees & Cranston 2017).
143   OTT was built using an automated algorithm with informed choices to aggregate and link
144   existing naming authorities into a reasonably comprehensive, artificial, taxonomy. OTT contains
145   4,385,000 external links for 3,594,550 taxa aggregated and linked over 5 authorities (i.e., GBIF,
146   IF, SILVA, WoRMS, NCBI).

147

Linking Wikidata And GloBI

149   Both Wikidata and GloBI have taxon graphs that map to identifiers from external databases (e.g.
150   NCBI, ITIS, GBIF, EOL, Index Fungorum (IF), Fishbase and WoRMS). A Wikidata dump was
151   loaded into GUODA and processed to extract taxon items (about 2.3 million) and their links to
152   NCBI, ITIS, GBIF, EOL, IF, Fishbase and WoRMS. This was the Wikidata taxon graph. This
153   taxon graph was loaded into a lookup table where each row contained an NCBI, ITIS, GBIF,
154   EOL, IF, Fishbase or WoRMS identifier and the corresponding Wikidata identifier. The GloBI
155   taxon graph was already in a similarly formatted lookup table. The taxon graphs in GloBI and
156   Wikidata were mapped to each other with a join of the NCBI, ITIS, GBIF, EOL, IF, Fishbase or
157   WoRMS identifiers of the respective lookup tables (Fig. 3). So, for each external identifier that
158   occurred in both Wikidata and GloBI, the corresponding Wikidata identifier inserted in the
159   GloBI lookup table. For instance, consider Wikidata taxon item Q140
160   (https://www.wikidata.org/wiki/Q140 accessed on 30 March 2018; *Panthera leo*) points to
161   ITIS:183803. With the matching algorithm used, GloBI now considers WD:Q140 to be linked to
162   all taxon entries that are considered the same as, or synonymous to, ITIS:183803.

163

164   This final joined graph was saved into HDFS as a Parquet file and linked entries were appended
165   to GloBI Taxon Graph from v0.3.0 onward (Poelen 2018c). In addition, the GloBI ingestion
166   engine was updated to automatically perform the taxon graph matching for future updates. This
167   linkage enabled lookups of diet items of lions by Wikidata identifier via
168   https://www.globalbioticinteractions.org/?interactionType=eats&sourceTaxon=WD%3AQ140
169   and facilitates future integration of species interaction data with Wikidata.

Taxon Graph Overlap and Consistency

OTT, Wikidata, and GloBI taxon graphs maintain links to GBIF, IF, NCBI and WoRMS identifiers (referred to as external identifiers). The taxon graphs are considered to (partially) overlap if individual taxon ids from different graphs have at least one external identifier in common. In addition, a taxon graph is inconsistent if a taxon id links to multiple external identifiers from the same identifier scheme. Similarly, overlapping taxon ids are said to be inconsistent if they link to multiple external identifiers from the same identifier scheme. Where overlap is a measure for taxon graph similarity, consistency can be seen as a way to measure the relative quality of (overlapping) taxon graphs.

For instance, let's say that OTT:1087695 is linked to NCBI:191633, WoRMS:156905, and GBIF:1449280. In addition, WD:Q7247420 (https://www.wikidata.org/wiki/Q7247420) points to WORMS:156905, GBIF:1449280, and NCBI:191633. This would mean that links of these OTT and WD ids overlap and are consistent, because they do not point to different names in same naming schemes (Fig. 4). However, when considering the GloBI taxon "id" "GLOBI:null@Procladius sp1 M_PL_014", multiple links to external ids were found (e.g., NCBI:1981571, NCBI:1981569, NCBI:1981572, NCBI:1981573, NCBI:1981574, NCBI:1981570). In this case, the GloBI taxon id is inconsistent. The high number of external NCBI identifiers is due to the NCBI taxonomy containing many "provisional" taxa derived from environmental samples.

Data Access

All of the input data sets can be found at:
https://doi.org/10.5281/zenodo.755513 (GloBI Taxon Graph),
http://files.opentreeoflife.org/ott/ott3.0/ott3.0.tgz (Open Tree of Life Taxonomy)
http://doi.org/10.5281/zenodo.1211767 (Wikidata)

A selection of intermediary and result datasets are available online (Poelen 2018d; Poelen 2018a).

All of the scripts used to make the statements in the results can be found here (https://github.com/bio-guoda/guoda-datasets/tree/master/wikidata) with instructions on how to duplicate the analysis.

Results

After 10 minutes of processing, GloBI was linked to Wikidata using pre-existing identifier mappings. The Wikidata dump was 20 GB of compressed JSON with 40-50 million data items. It took about 10 minutes for GUODA to extract taxa (about 2.3 million) and their links in Wikidata and then less than one minute to map the Wikidata taxon graph to the GloBI taxon graph. The 119,957 WikiData links that were added to GloBI increased its outgoing name links by 13.7%

208 (Poelen 2018d). Eighty-seven percent (86.7%) of the external identifiers in Wikidata overlap
209 with the external identifiers in OTT (Fig. 5). Eighty-six percent (86.1%) of the external
210 identifiers in GloBI overlap with the external identifiers in OTT (Fig. 5). Wikidata provided
211 mappings for 65.2% of the external identifiers in GloBI (Fig. 5). Out of the 77,000 external
212 identifiers that occurred only in OTT and GloBI, only 56 were inconsistent
213 (https://github.com/bio-guoda/guoda-
214 [datasets/blob/master/wikidata/inconsistentNameIdsGloBI_OTT.tsv](https://github.com/bio-guoda/guoda-datasets/blob/master/wikidata/inconsistentNameIdsGloBI_OTT.tsv)). These 56 links pointed to
215 seven OTT "taxa". No inconsistent links were found between WD and GloBI. Out of the 38,000
216 links only found in GloBI, 9,000 were inconsistent (https://github.com/bio-guoda/guoda-
217 [datasets/blob/master/wikidata/inconsistentNameIdsGloBIOnly.tsv](https://github.com/bio-guoda/guoda-datasets/blob/master/wikidata/inconsistentNameIdsGloBIOnly.tsv)). The OTT, Wikidata, and
218 GloBI identifier graphs related to this coverage analysis is a 74 MB compressed tab separated
219 values file consisting of about 12 million identifier mapping records (see
220 https://zenodo.org/record/1213477/files/links-globi-wd-ott.tsv.gz). The resulting Wikidata taxon
221 objects were merged into GloBI's Taxon Graph (Poelen 2018d).
222
223 In order for a mapping to be considered consistent, there can only be one identifier per resource
224 included in each local graph. Thus, after removing the inconsistent identifiers, the external id
225 overlap can be interpreted as an estimate of the number of shared taxon names between two
226 databases (Table 2). This cannot be interpreted as total taxa in each resource.

Discussion

228 GUODA is a high performance computing resource for biodiversity science that provides
229 scalable solutions for working with large data sets in a collaborative, online environment. The 10
230 minute processing time for 20 GB of compressed JSON is far faster than any current mapping
231 method used in biodiversity; however, it does benefit from the mapping already completed inside
232 Wikidata. For example, the Wikidata entry for *Panthera leo*
233 (https://www.wikidata.org/wiki/Q140) has 25 links to external databases, not all of them
234 biodiversity-related. This linking may be based on matching name strings. Other efforts using
235 name-string-matching to link biodiversity databases take much longer to map resources together.
236 For instance, EOL takes more than a day to map the content it receives from providers to a
237 unified classification (Rice pers. comm.). Similarly, the taxon matching in BioNames and
238 LinkOut took days to complete (Page pers. comm.). Projects like OTT, Wikidata, and GloBI that
239 keep identifier-based taxonomic graphs make it easier to link databases at scale.
240
241 Despite the notoriously poor nature of taxon names as identifiers, they are still commonly used to
242 link biodiversity data. A much-discussed solution has been the use of universal, unique,
243 persistent, resolvable identifiers across the biodiversity data landscape, but the social barrier to a
244 universal identifier system has, thus far, proven insurmountable (Nimis 2001; Hardisty et al.
245 2013). Rather than rely on name strings or a universal identifier system, this method uses the
246 graph of identifiers to map taxa across two databases. This identifier-based method has the
247 potential to be faster and easier than name-string matching without some of the social difficulties
248 of a single identifier system.

249

250   Most biodiversity databases and nomenclatural authorities expose their data in idiosyncratic
251   ways that are not suitable for batch processing. If data sources published their taxon identifier
252   graph as a lookup table (as described in this paper) integrating across databases would be much
253   easier (Fig. 6). Now, users have to learn a unique format for every data source. These lookup
254   tables have the advantage of being easy to version and integrate.

255

256   In addition to fast linking of biodiversity databases, comparison of identifier graphs may be a
257   scalable way to find inconsistencies, especially when multiple biodiversity databases/identifiers
258   are included. By linking GloBI to OTT and WD, inconsistent names or false positive name
259   matches were detected by considering the (lack of) overlap of GloBI names with OTT and WD
260   external identifier schemes. These inconsistencies might be introduced by a dataset or a name
261   resolution method that produces ambiguous results. In addition, inconsistencies can indicate a
262   disputed / outdated name like "GLOBI:null@Senecio pectinatus" which maps to GBIF:8317096
263   and GBIF:8414746. This would be considered an inconsistent mapping and suggests that
264   Senecio pectinatus is an outdated name. A related method using a variation of the PageRank
265   algorithm (Page et al. 1999; Brin & Page 1998) to identify the most legitimate taxonomic name
266   to apply to a fossilized specimen (Huber & Klump 2009) gives further legitimacy to this concept.
267   Combining the speediness with the promise of scalability, a near-real-time name consistency
268   check can be implemented to detect inconsistencies across various systems in the biodiversity
269   data-ecosystem introduced by integration bugs, taxonomy updates or differences of
270   interpretation.

271

272   GUODA has been available since 2015 and contains data dumps from GBIF, EOL TraitBank,
273   iNaturalist, iDigBio, and BHL which are all accessible via a Jupyter notebook, web services, or
274   Apache Spark shell on the command line. Despite its computing power and successful
275   demonstrations at major conferences, GUODA has not been used to its full potential. The barrier
276   of learning new programming and computing paradigms as well as developing an understanding
277   of large dataset work flows seems to be a barrier to many in the biodiversity community. Despite
278   this, GUODA is being used in several capacities. The Effechecka application generates
279   taxonomic checklists using a web interface that allows a user to draw a polygon on a map and
280   returns a deduplicated list of taxa aggregated from observation data held in GBIF, iNaturalist,
281   etc. The EOL Freshdata project uses it to enable the detection of new occurrence records given
282   geospatial and taxonomic and data source constraints and notifies interested users via email.
283   Several workshops have used it to teach Spark programming skills to students at the University
284   of Florida.

285

286   Future work on the GUODA infrastructure includes training and evaluating neural network
287   models on image data, containerization of the GUODA components to allow the system to be run
288   in additional data centers, and refinement of the end-user interface to integrate programming,
289   source code, and publication to make research more reproducible. GUODA's most impactful
290   contribution has likely been the availability of readily formatted biodiversity data and new data

291 sets will continue to be added to the collaboration platform, enabling domain experts and
292 technical experts to answer new questions in the future.
293
294 The bottlenecks in processing for Hadoop File System and Apache Spark are the number of
295 CPUs, amount of memory, and available storage space allocated to the computer cluster. Both
296 HDFS and Spark are designed to scale horizontally by adding commodity servers (aka nodes) to
297 increase the processing power, working memory, and storage space. Thus, this problem is
298 immediately solvable. Internet bandwidth to transfer the data archives from Open Tree of Life,
299 Wikidata, and GloBI does not scale and is not something that can be addressed solely within our
300 research group. At the moment, it takes longer to download the Wikidata resource than it does to
301 run the linking process discussed in this manuscript. Socio-technical bottlenecks include
302 resource-limitation user education. Increased usage and operational support is expected to
303 positively impact processing performance by encouraging pro-active bug fixing and
304 infrastructure maintenance. In addition, while the technical complexity of operating and using a
305 compute cluster have been dramatically reduced since the introduction of Hadoop in 2006, some
306 re-education may be needed to effectively use these powerful data tools (e.g., jupyter notebooks,
307 HDFS, scala).
308
309 GUODA, and hosted data analytics infrastructure in general, has the potential to drastically
310 improve biodiversity science by making multiple biodiversity databases accessible to scientists
311 for analysis on their laptop or desktop. Users still need to have some programming skills, which
312 have now become an essential skill in biodiversity science.
313

Conclusions

315 Sharing information between biodiversity databases can be difficult because of the amount and
316 heterogeneity of the data and the identifiers. Most mappings are done using taxonomic name
317 strings at great expense. We were able to map Wikidata to GloBI in 10 minutes using identifier
318 graphs and GUODA, a high performance computing infrastructure developed through
319 collaboration between diverse players. The mapping increased GloBI's outgoing name links by
320 13.7%. This method of mapping across databases using identifier graphs is faster than comparing
321 name strings and can help find inconsistencies that point to a disputed or outdated name.
322 GUODA, and systems like it, have the potential to revolutionize biodiversity science by bringing
323 diverse technically minded people together with high performance computing resources that are
324 accessible from a laptop or desktop.

Acknowledgements

326 The authors would like to acknowledge support and resources provided by the ACIS lab. The
327 authors would like to acknowledge José A.B. Fortes for providing infrastructure and creating
328 room for collaboration. The authors would like to thank the two reviewers for their insightful

329 comments that greatly improved the manuscript. The Encyclopedia of Life and iDigBio helped
330 establish an informal yet pragmatic cross-institutional collaboration.

References

332 Bingham HC, Doudin M, Weatherdon LV, Despot-Belmonte K, Wetzel FT, Groom Q, Lewis E,
333   Regan E, Appeltans W, Güntsch A, Mergen P, Agosti D, Penev L, Hoffmann A, Saarenmaa
334   H, Geller G, Kim K, Kim H, Archambeau AS, Häuser C, Schmeller DS, Geijzendorffer I,
335   García Camacho A, Guerra C, Robertson T, Runnel V, Valland, N, Martin CS. 2017. The
336   Biodiversity Informatics Landscape: Elements, Connections and Opportunities. *RIO*
337   3:e14059. DOI: https://doi.org/10.3897/rio.3.e14059.
338 Brin S, Page L. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer*
339   *Networks and ISDN Systems* 30(1-7):107-117.
340 Hardisty A, Roberts D, The Biodiversity Informatics Community. 2013. A decadal view of
341   biodiversity informatics: Challenges and priorities. *BMC Ecology* 13 DOI:
342   https://doi.org/10.1186/1472-6785-13-16.
343 Hindman B, Konwinski A, Zaharia M, Ghodsi A, Josepyh AD, Katz R, Shenker S, Stoica I.
344   2011. Mesos: a platform for fine-grained resource sharing in the data center. In:
345   *Proceedings of the 8th USENIX conference on Networked systems design and*
346   *implementation*. Berkeley: USENIX Association, 295–308.
347 Hortal J, de Bello F, Diniz-Filho JAF, Lewinsohn TM, Lobo JM, Ladle RJ. 2015. Seven
348   Shortfalls that Beset Large-Scale Knowledge of Biodiversity. *Annual Review of Ecology,*
349   *Evolution, and Systematics* 46(1):523–549 DOI: https://doi.org/10.1146/annurev-ecolsys-
350   112414-054400.
351 Huber R, Klump J. 2009. Charting taxonomic knowledge through ontologies and ranking
352   algorithms. *Computers & Geosciences* 35:862-868 DOI: 10.1016/j.cageo.2008.02.016.
353 Kluyver T, Ragan-Kelley B, Pérez F, Granger B, Bussonnier M, Frederic J, Kelley K, Hamrick J,
354   Grout J, Corlay S, Ivanov P, Avila D, Abdalla S, Willing C, Jupyter Development Team.
355   2016. Jupyter Notebooks -- a publishing format for reproducible computational workflows.
356   In: Loizides F, Schmidt B, eds. *Positioning and Power in Academic Publishing: Players,*
357   *Agents and Agendas*. Amsterdam: IOS Press, 87–90.
358 Nimis PL. 2001. A tale from Bioutopia: Could a change of nomenclature bring peace to biology's
359   warring tribes? *Nature* 413(6851):21 DOI: https://doi.org/10.1038/35092637.
360 Page L, Brin S, Motwani R, Winograd T. 1999. The pagerank citation ranking: Bringing order to
361   the web. Technical Report, Stanford Digital Library Technologies Project, Palo Alto, CA.
362 Page RDM. 2008. Biodiversity informatics: the challenge of linking data and the role of shared
363   identifiers. *Briefings in Bioinformatics* 9(5):345–354 DOI:
364   https://doi.org/10.1093/bib/bbn022.
365 Page RDM. 2013. BioNames: linking taxonomy, texts, and trees. *PeerJ* 1:e190 DOI:
366   https://doi.org/10.7717/peerj.190.
367 Page RDM. 2011. Linking NCBI to Wikipedia: a wiki-based approach. *PLoS Currents*
368   3:RRN1228 DOI: 10.1371/currents.RRN1228.
369 Page RDM. 2007. Tbmap: a taxonomic perspective on the phylogenetic database treebase. *BMC*
370   *Bioinformatics* 8(1):158 DOI: https://doi.org/10.1186/1471-2105-8-158.
371 Parr CS, Wilson N, Leary P, Schulz KS, Lans K, Walley L, Hammock JA, Goddard A, Rice J,
372   Studer M, Holmes JTG, Corrigan Jr RJ. 2014. The encyclopedia of life v2: providing global

373   access to knowledge about life on earth. *Biodiversity Data Journal* 2:e1079 DOI:
374   https://doi.org/10.3897/BDJ.2.e1079.
375   Poelen J. 2018a. 20 GB in 10 minutes: Data linking across major biodiversity databases: Data
376   supplements. Version 0.1 Zenodo DOI: http://doi.org/10.5281/zenodo.1213477.
377   Poelen J. 2018b. Global Biotic Interactions: Taxon Graph. Version 0.4.2 Zenodo DOI:
378   http://doi.org/10.5281/zenodo.1210315.
379   Poelen J. 2018c. Global Biotic Interactions: Taxon Graph. Version 0.3.0 Zenodo DOI:
380   http://doi.org/10.5281/zenodo.1210308.
381   Poelen J. 2018d. Global Biotic Interactions: Taxon Graph. Version 0.3.1 Zenodo DOI:
382   http://doi.org/10.5281/zenodo.1213465.
383   Poelen JH, Simons JD, Mungall CJ. 2014. Global biotic interactions: An open infrastructure to
384   share and analyze species-interaction datasets. *Ecological Informatics* 24:148–159 DOI:
385   https://doi.org/10.1016/j.ecoinf.2014.08.005.
386   Rees JA, Cranston K. 2017. Automated assembly of a reference taxonomy for phylogenetic data
387   synthesis. *Biodiversity Data Journal* 5:e12581 DOI: 10.3897/BDJ.5.e12581.
388   Shvachko K, Kuang H, Radia S, Chansler R. 2010. The Hadoop distributed file system. In *MSST*
389   *'10 Proceedings of the 2010 IEEE 26th Symposium on Mass Storage Systems and*
390   *Technologies*. Washington DC: IEEE Computer Society, 1–10.
391   Voß J. 2016. wikidata-taxonomy 0.2.7. Version 0.2.7 Zenodo DOI:
392   http://doi.org/10.5281/zenodo.60708.
393   Vrandečić D, Krötzsch M. 2014. Wikidata: A free collaborative knowledgebase.
394   *Communications of the ACM* 57(10):78–85 DOI: 10.1145/2629489.
395   Wikidata. 2018. Wikidata dump 2017-12-27. Zenodo DOI:
396   http://doi.org/10.5281/zenodo.1211767.
397   Zaharia M, Xin RS, Wendell P, Das T, Armbrust M, Dave A, Meng X, Rosen J, Venkataraman
398   S, Franklin MJ, Ghodsi A, Gonzalez J, Shenker S, Stoica I. 2016. Apache Spark: a unified
399   engine for big data processing. *Communications of the ACM* 59(11):56–65 DOI:
400   10.1145/2934664.
401

Figure 1

GUODA Infrastructure

Data from biodiversity databases is loaded into GUODA as Parquet files (Storage). When a user working in a Jupyter Notebook (Front-end Server) triggers a job interactively or via GitHub and Jenkins, the data are analyzed using Apache Spark (Compute Cluster). This infrastructure allows a user working from a laptop or desktop to compute over multiple biodiversity databases at once. All logos are provided by the organizations they represent and are used with permission.
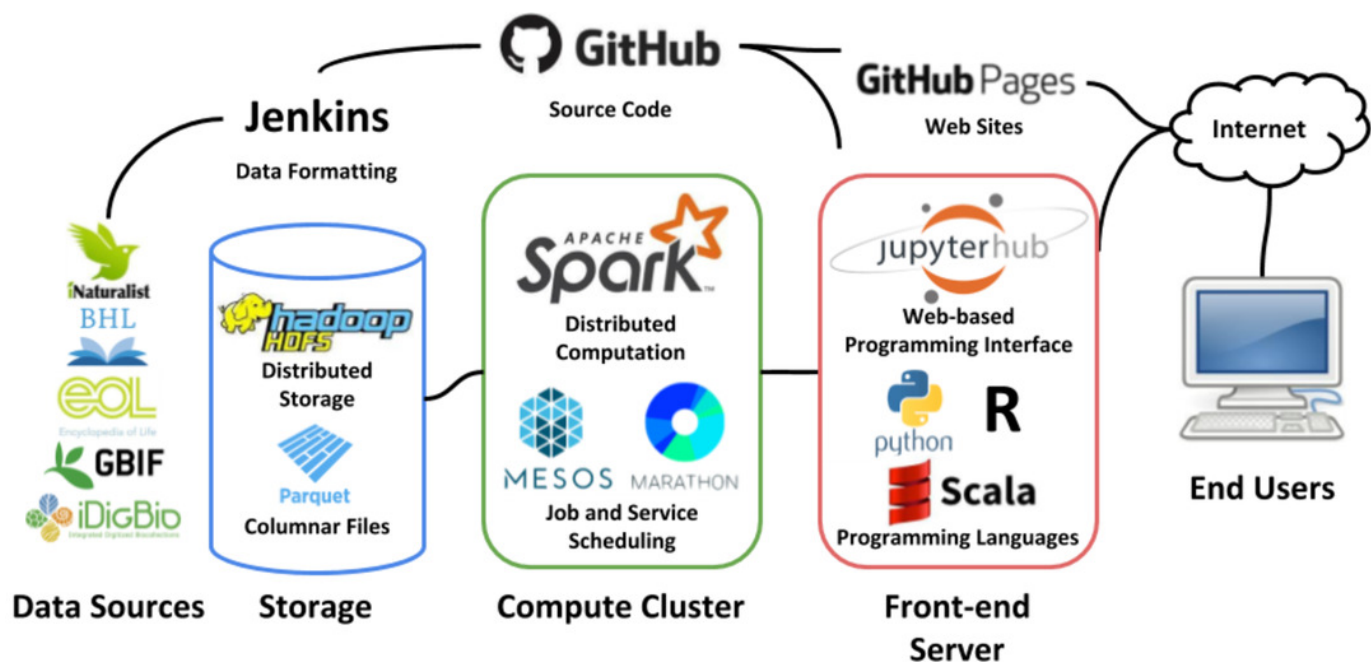
Figure 2

Frequency of Wikidata taxa linked to biodiversity databases

This graph shows the proportion of the approximately 2.3 million Wikidata taxa with 0,1, 2, etc. links to external biodiversity databases (NCBI, ITIS, GBIF, EOL, FishBase, Index Fungorum and iNaturalist). The majority of Wikidata taxa had at least two links. A little more than 15% of Wikidata taxa had no links to external biodiversity databases.
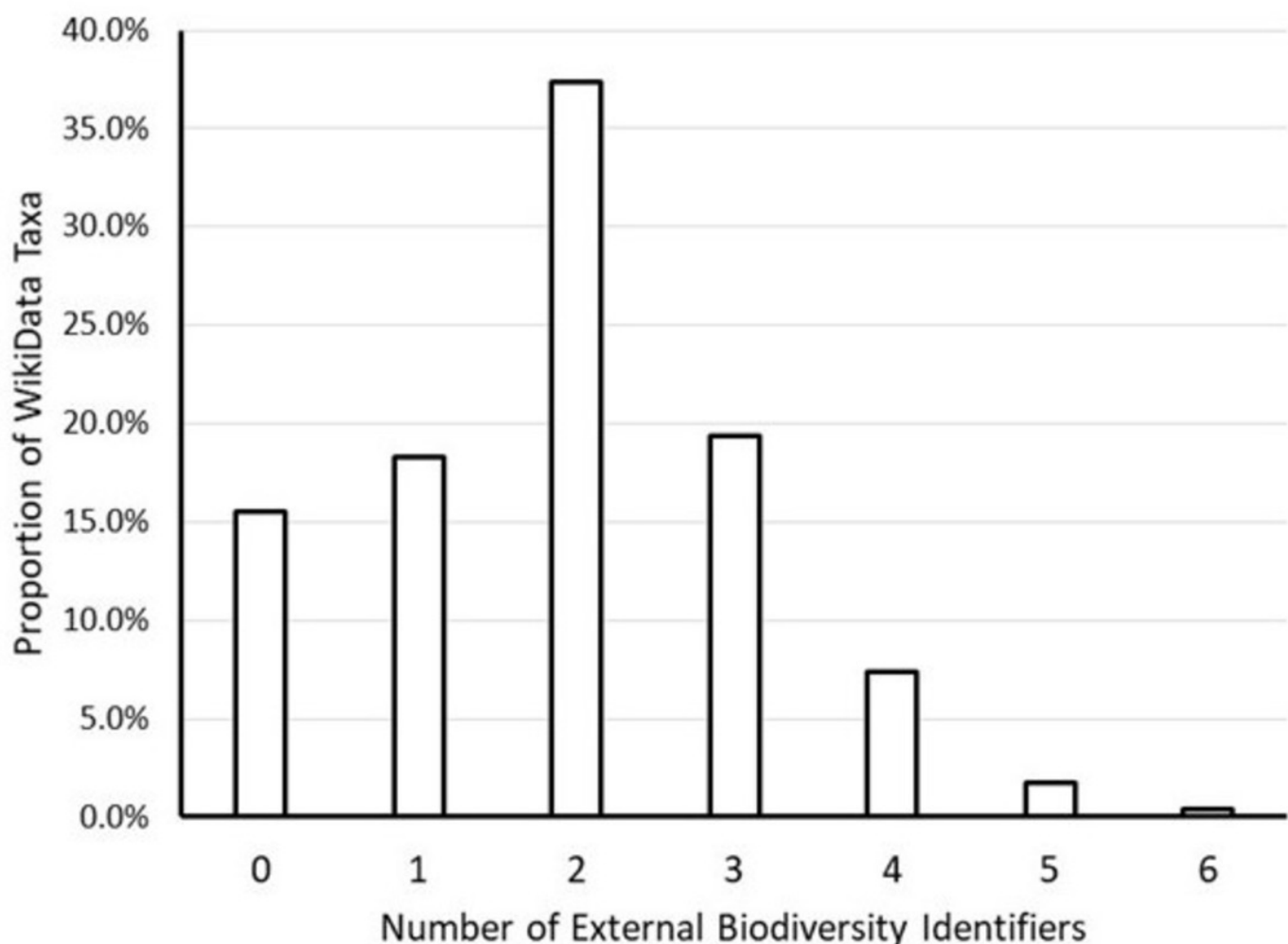
Figure 3

Mapping taxon graphs across resources

Both GloBI and Wikidata contain hierarchical taxon graphs with each taxon having a "star" of external identifiers. The taxa are mapped across these resources by comparing the portion of the graph with the external identifiers between nodes. In this example, the names and identifiers match perfectly, so a relationship between *Panthera leo* in GloBI and *Panthera leo* in Wikidata is inferred.
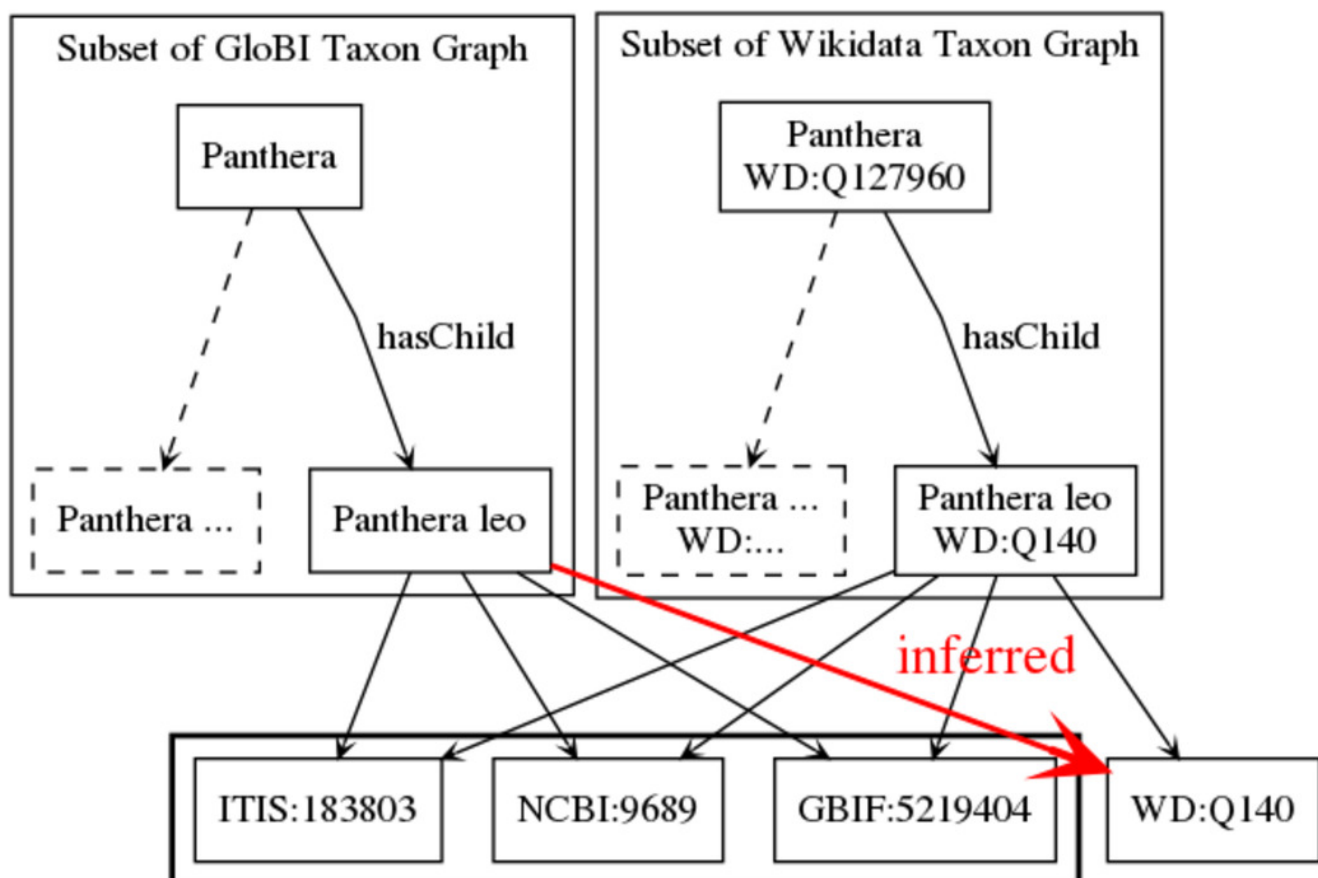
Figure 4

Inconsistent graph matching

When overlapping taxon graphs include multiple name strings, the graph is inconsistent. In this example the *Procladius* genus is present in Wikidata (red), Open Tree (textured fill), and GloBI (blue). The Wikidata, OTT, and GloBI taxon graphs overlap on the NCBI and the GBIF identifiers (purple and textured fill). The WoRMS identifier overlaps the OTT and Wikidata taxon graph (red and textured fill). The *Procladius* graph in GloBI includes NCBI identifiers with a different name string, *Procladius* (*Holotanypus*), which indicates inconsistent usage.
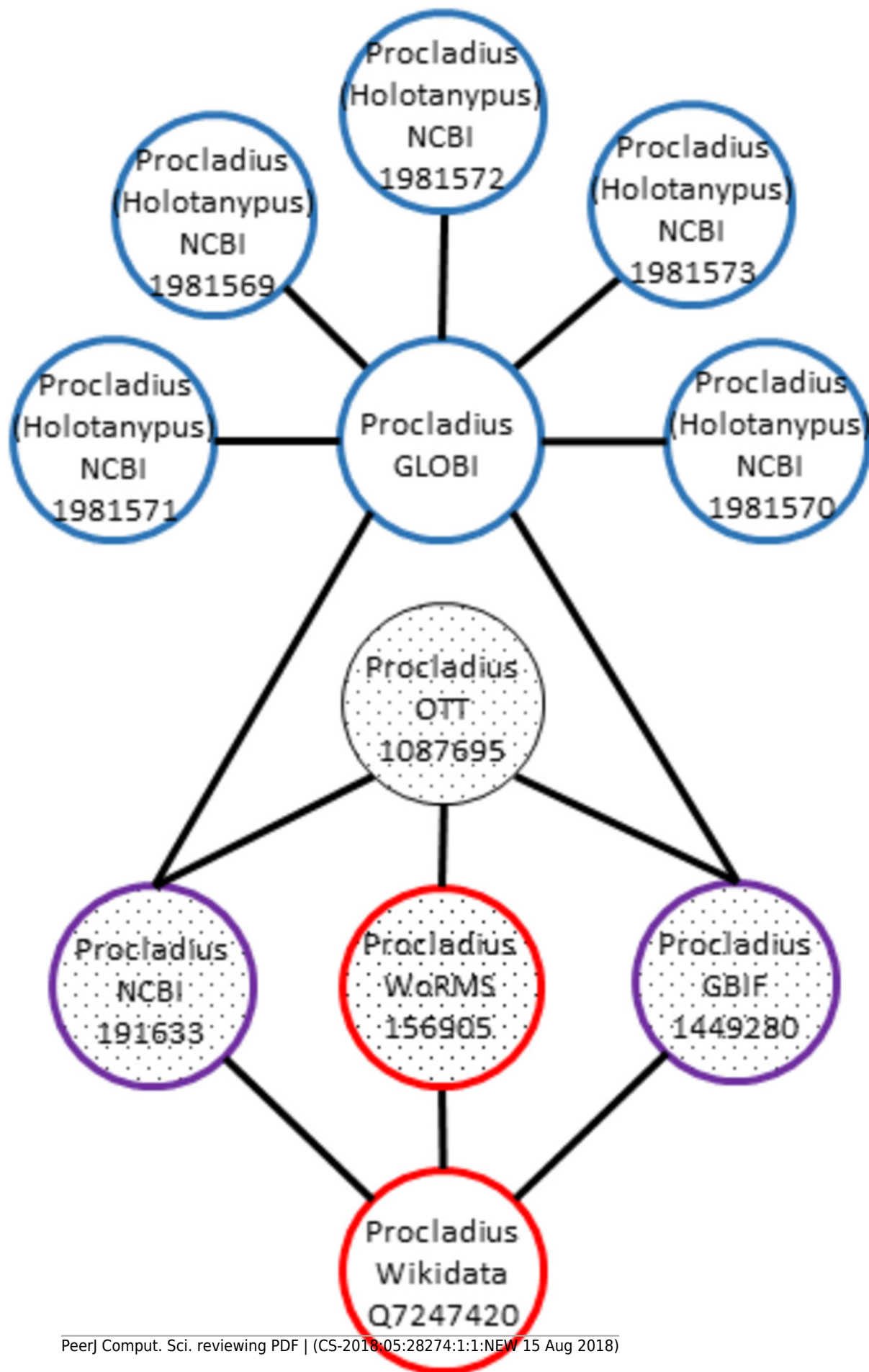
Figure 5

Identifier overlap between Wikidata (WD), OTT, and GloBI

This Venn Diagram shows the number of overlapping external identifiers that can be found in one of three databases. Only 208,000 external ids can be found in all three. These consisted of 22,637 WoRMS links, 71,980 NCBI links, 103,3 GBIF links and 10,04 IF links. Over two million ids are only known to one of the three databases. OTT contains more than half of the external ids in Wikidata and in GloBI, but neither contain half of the external ids in OTT. Mapping Wikidata to GloBI matched 65.2% of the external ids in GloBI.
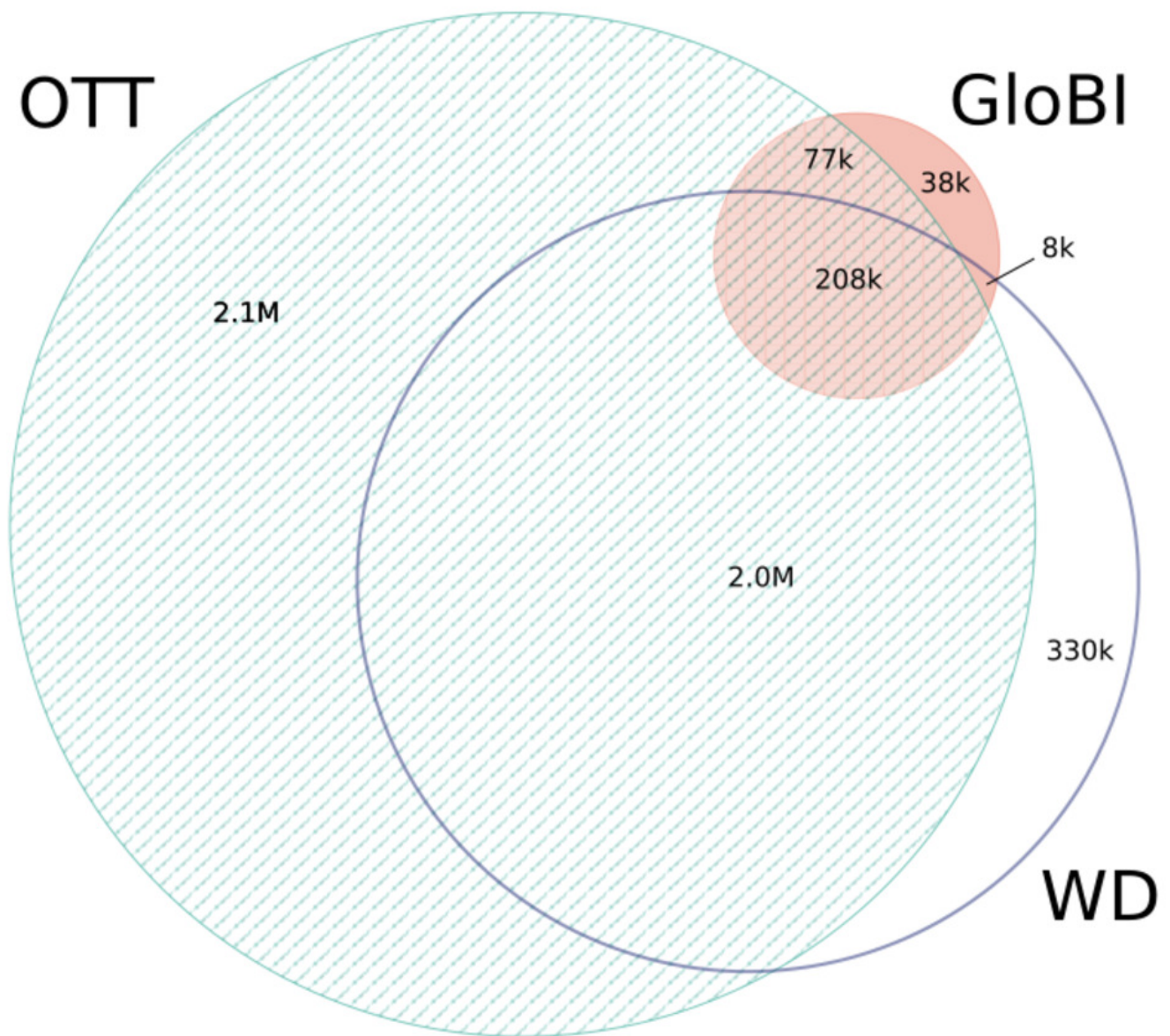
Figure 6

Example look up table

This figure is an excerpt from the GloBI look up table. The providedTaxonId and the providedTaxonName come from the taxon graph external to GloBI. The resolvedTaxonId and the resolvedTaxonName are the names and identifiers that are already mapped within GloBI. Each row represents a mapping from a taxon in an external source (*Pluvialis obscura*) to an identifier from a source already in GloBI, which does not mint its own identifiers.

providedTaxonId	providedTaxonName	resolvedTaxonId	resolvedTaxonName
00003e86-484b-47a0-a4ca-75a5ba6f996e	Pluvialis obscura	EOL:1049098	Charadrius obscurus
00003e86-484b-47a0-a4ca-75a5ba6f996e	Pluvialis obscura	GBIF:2480305	Charadrius obscurus
00003e86-484b-47a0-a4ca-75a5ba6f996e	Pluvialis obscura	INAT_TAXON:4829	Charadrius obscurus
00003e86-484b-47a0-a4ca-75a5ba6f996e	Pluvialis obscura	IRMNG:10195526	Charadrius obscurus
00003e86-484b-47a0-a4ca-75a5ba6f996e	Pluvialis obscura	ITIS:176536	Charadrius obscurus
00003e86-484b-47a0-a4ca-75a5ba6f996e	Pluvialis obscura	NCBI:1424469	Charadrius obscurus
00003e86-484b-47a0-a4ca-75a5ba6f996e	Pluvialis obscura	OTT:3596992	Charadrius obscurus
00003e86-484b-47a0-a4ca-75a5ba6f996e	Pluvialis obscura	WD:Q636452	Charadrius obscurus
00003e86-484b-47a0-a4ca-75a5ba6f996e	Pluvialis obscura	WORMS:366852	Charadrius obscurus
0000b38e-9098-474b-b82d-1bc6e6eb1ace	Lobesia helichrysana	EOL:361371	Lobesia helichrysana

Selected biodiversity databases and their size

Database	Data Quantity (Jan 2018)	Size (compressed)
GBIF	964,547,793 occurrence records	139 GB
Catalogue of Life/ITIS	1.7 million taxa	2.9 GB
GloBI	3,363,528 interactions	206 MB
iDigBio	106,922,498 specimen records	35.5 GB
GenBank	206,293,625 sequences	3 TB
Biodiversity Heritage Library	53,739,062 pages	2.7 GB
WoRMS	243,323 marine species	71 MB
OpenTree	2,722,024 taxa and 6,810 trees	189 MB
EOL TraitBank	Over 11 million records	46 GB uncompressed
EOL	7,705,748 data objects (May 2017)	10 TB uncompressed
Wikidata	42,648,426 data items	20 GB

Absolute and relative link counts from OTT, WD, and GloBI compared to WoRMS, GBIF, Index Fungorum (IF), and NCBI

	WoRMS	GBIF	IF	NCBI	combined
OTT	327929 (100%)*	2451566 (100%)	276262 (100%)	1355207 (100%)	4410964 (100%)
WD	288110 (88%)	1779789 (73%)	76497 (28%)	410092 (30%)	2554488 (58%)
GloBI	68565 (21%)	315173 (13%)	33400 (12%)	704361 (52%)	1121499 (25%)

2 *Overlap between each resource and OTT is set at 100%. The other percentages give a relative
3 estimate of size and scale and should not be interpreted as overlapping ids.