**A peer-reviewed version of this preprint was published in PeerJ on 17 September 2018.**



Thessen AE, Poelen JH, Collins M, Hammock J. 2018. 20 GB in 10 minutes: a case for linking major biodiversity databases using an open socio-technical infrastructure and a pragmatic, cross-institutional collaboration. PeerJ Computer Science 4:e164 https://doi.org/10.7717/peerj-cs.164

# 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** [Corresp., 1] , **Jorrit H Poelen** [2] , **Matthew Collins** [3] , **Jen Hammock** [4]

[1] Ronin Institute for Independent Scholarship, Montclair, New Jersey, USA

[2] Self-employed, Oakland, California, USA

[3] University of Florida, Gainsville, Florida, USA

[4] National Museum of Natural History, Washington DC, USA

Corresponding Author: Anne E Thessen
Email address: annethessen@gmail.com

Biodiversity information is made available through numerous databases that each have their own data models, web services, and data types. Combining data across databases leads to new insights, but is not easy because each database uses its own system of identifiers. In the absence of stable and interoperable identifiers, databases are often linked using taxonomic names. This labor intensive, error prone, and lengthy process relies on accessible versions of nomenclatural authorities and fuzzy-matching algorithms.

To approach the challenge of linking diverse data, more than technology is needed. New social collaborations like the Global Unified Open Data Architecture (GUODA) that combine skills from diverse groups of computer engineers from iDigBio, server resources from the Advanced Computing and Information Systems (ACIS) Lab, global-scale data presentation from EOL, and independent developers and researchers are what is needed to make concrete progress on finding relationships between biodiversity datasets.

This paper will discuss a technical solution developed by the GUODA collaboration for faster linking across databases with a use case linking Wikidata and the Global Biodiversity Interactions database (GloBI). The GUODA infrastructure is a 12-node, high performance computing cluster made up of about 192 threads with 12 TB of storage and 288 GB memory. Using GUODA, 20GB of compressed JSON from Wikidata was processed and linked to GloBI in about 10-11 minutes. Instead of comparing name strings or relying on a single identifier, Wikidata and GloBI were linked by comparing graphs of biodiversity identifiers external to each system. This method resulted in adding 119,957 Wikidata links in GloBI, an increase of 13.7% of all outgoing name links in GloBI. Wikidata and GloBI were compared to Open Tree Taxonomy to examine consistency and coverage. The process of parsing Wikidata, Open Tree Taxonomy and GloBI archives and calculating consistency metrics was done in minutes on the GUODA platform. As a model collaboration, GUODA has the potential to revolutionize biodiversity science by bringing diverse technically minded people together with high performance computing resources that are accessible from a laptop or desktop. However, participating in such a collaboration still requires basic programming skills.

# 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

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26951v1 | CC BY 4.0 Open Access | rec: 22 May 2018, publ: 22 May 2018

# Abstract

Biodiversity information is made available through numerous databases that each have their own data models, web services, and data types. Combining data across databases leads to new insights, but is not easy because each database uses its own system of identifiers. In the absence of stable and interoperable identifiers, databases are often linked using taxonomic names. This labor intensive, error prone, and lengthy process relies on accessible versions of nomenclatural authorities and fuzzy-matching algorithms.

To approach the challenge of linking diverse data, more than technology is needed. New social collaborations like the Global Unified Open Data Architecture (GUODA) that combine skills from diverse groups of computer engineers from iDigBio, server resources from the Advanced Computing and Information Systems (ACIS) Lab, global-scale data presentation from EOL, and independent developers and researchers are what is needed to make concrete progress on finding relationships between biodiversity datasets.

This paper will discuss a technical solution developed by the GUODA collaboration for faster linking across databases with a use case linking Wikidata and the Global Biodiversity Interactions database (GloBI). The GUODA infrastructure is a 12-node, high performance computing cluster made up of about 192 threads with 12 TB of storage and 288 GB memory. Using GUODA, 20GB of compressed JSON from Wikidata was processed and linked to GloBI in about 10-11 minutes. Instead of comparing name strings or relying on a single identifier, Wikidata and GloBI were linked by comparing graphs of biodiversity identifiers external to each system. This method resulted in adding 119,957 Wikidata links in GloBI, an increase of 13.7% of all outgoing name links in GloBI. Wikidata and GloBI were compared to Open Tree Taxonomy to examine consistency and coverage. The process of parsing Wikidata, Open Tree Taxonomy and GloBI archives and calculating consistency metrics was done in minutes on the GUODA platform. As a model collaboration, GUODA has the potential to revolutionize biodiversity science by bringing diverse technically minded people together with high performance computing resources that are accessible from a laptop or desktop. However, participating in such a collaboration still requires basic programming skills.


# Introduction

Biodiversity databases provide global access to information about species via the Web. These databases contain information as varied as observation records, text descriptions, images, maps, genetic sequences, phylogenetic trees, and trait data (Table 1). All of these data become much more useful if they can be linked. Many biodiversity databases share information with each other (Bingham et al. 2017), but creating the links can be very difficult for several reasons including the size of the databases, the heterogeneous nature of the data, and the heterogeneous nature of the identifiers used by the different resources (Page 2008).

The more popular methods for linking biodiversity databases include taxonomic names, lsid, and doi. The Encyclopedia of Life uses taxonomic names to automatically aggregate data from hundreds of providers (Parr et al. 2014). BioNames links data using lsid, doi, handles, bibliographic citations, and taxonomic names (Page 2013). The iPhylo LinkOut service mapped identifiers used by the NCBI taxonomy database (which provides the taxonomic backbone for GenBank) to Wikipedia pages using taxonomic names, including synonyms (Page 2011). TBMap provides links from TreeBase across several taxonomic databases, such as ITIS and NCBI (Page 2007). This mapping was also achieved using taxonomic names, but in some cases GenBank Accession numbers and museum specimen codes were available for supplement. The use of taxonomic names to aggregate data can lead to errors and requires significant *a priori* knowledge either in the form of curators or an authoritative nomenclature.

Many databases expose their own internal identifiers, such as the WoRMS Aphia ID, so others can link their data to those resources within their own systems, often by providing a URL. Databases like WoRMS provide web services that allow users to look up an identifier for a taxon in question, one at a time. While this makes linking easier, it is still difficult to scale across all databases. For example, a list of all the taxon identifiers in EOL is 300 MB compressed. No system of identifiers is universal across biodiversity databases and none of them are easy to implement at scale.

While the data would be much more useful if linked, there is a lack of tools for linking data across databases at scale. Most mappings are done at great expense and then are made available as a separate file or incorporated into the resources themselves. LinkOut, BioNames, GBIF, and EOL take more than a day to link across their entire body of aggregated content. This paper discusses links made between GloBI and Wikidata (WD) in 10 minutes using GUODA, a high performance computing system available for analysis of large biodiversity data sets.

# Methods

## Description of Resources

### GUODA

Following an iDigBio hack-a-thon in June 2015, GUODA was created as a pragmatic way to compute over multiple large biodiversity databases in a mutually beneficial collaboration between iDigBio, EOL, Kew Garden, and independent developers. Catalyzed by various presentations at conferences, hardware provided by ACIS, 20+ meetings, and several prototypes (e.g., http://effechecka.org, https://gimmefreshdata.github.io), a general access biodiversity data integration and analysis environment was created. This environment, with the aggregated experience and perspectives of all the collaborators, was used to produce the results of this paper.

Housed at the ACIS Lab at the University of Florida, the GUODA infrastructure consists of 12 IBM HS22 blades each with 8 cores, 24 GB of memory, and 1 TB of storage each. This makes a total of 192 threads, 288 GB of memory and 12 TB of disk space available for processing jobs using Apache Spark (Fig 1; Zaharia et al. 2016). The cluster is managed under Apache Mesos (Hindman et al. 2011) which is a distributed scheduling system for periodic jobs. For long running processes, such as web APIs or databases, the Marathon (https://github.com/mesosphere/marathon) framework is run within Mesos. Marathon facilitates running always-up services with monitoring, automatic deployment of code, re-scaling to multiple nodes, and other management features. Mesos is responsible for accepting requests to start Spark frameworks, processes which do the actual computation and may span multiple servers, and allocation of resources requested by the framework.

Hadoop HDFS (Shvachko et al. 2010) is installed outside of Mesos directly on all 12 nodes of the cluster and provides redundant parallel shared storage to all nodes as well as the Jupyter notebook (Kluyver et al. 2016) server that provides a programming interface to end users. Each node has 1 TB of local disk storage for a total of about 3.5 TB of usable storage space for data files in Apache Parquet format. Spark is aware of the placement of data on an HDFS cluster and will divide processing among nodes in a way that prefers to read and write data that is local to the node to minimize network traffic.

## Wikidata

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

## GloBI

GloBI is a database of biotic interactions recorded as Organism_1:has_relationship:Organism_2 (Poelen et al. 2014). GloBI uses a combination of web APIs, taxon archives, and name correction/parsing methods in an attempt to link names from species interaction datasets to existing sources. Spatial, temporal, and taxonomic coverage in GloBI is sparse and unevenly distributed (see Eltonian shortfall, Hortal et al. 2015), with spatial concentrations in Europe and North America and taxonomically concentrated in Arthropods, Fungi, and Plants. Only 8% of

taxa in ITIS are also in GloBI. A detailed technical description of the GloBI data model and services has been published elsewhere (Poelen et al. 2014). GloBI maintains a graph of related taxa and their identifiers from different databases (Poelen et al. 2014). GloBI does not introduce its own taxon ids. Instead, it records how names were mapped from a source name into an external taxonomic database using a taxon graph (see https://globalbioticinteractions.org/references). We used GloBI Taxon Graph v0.4.2 (Poelen 2018b).

### Open Tree Taxonomy

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

## Linking Wikidata And GloBI

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

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

## 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% (Poelen 2018d). Eighty-seven percent (86.7%) of the external identifiers in Wikidata overlap with the external identifiers in OTT (Fig 3). Eighty-six percent (86.1%) of the external identifiers

in GloBI overlap with the external identifiers in OTT (Fig. 3). Wikidata provided mappings for 65.2% of the external identifiers in GloBI (Fig. 3). Out of the 77,000 external identifiers that occurred only in OTT and GloBI, only 56 were inconsistent (https://github.com/bio-guoda/guoda-datasets/blob/master/wikidata/inconsistentNameIdsGloBI_OTT.tsv). These 56 links pointed to seven OTT "taxa". No inconsistent links were found between WD and GloBI. Out of the 38,000 links only found in GloBI, 9,000 were inconsistent (https://github.com/bio-guoda/guoda-datasets/blob/master/wikidata/inconsistentNameIdsGloBIOnly.tsv). The OTT, Wikidata, and GloBI identifier graphs related to this coverage analysis is a 74 MB compressed tab separated values file consisting of about 12 million identifier mapping records (see https://zenodo.org/record/1213477/files/links-globi-wd-ott.tsv.gz). The resulting Wikidata taxon objects were merged into GloBI's Taxon Graph (Poelen 2018d).

In order for a mapping to be considered consistent, there can only be one identifier per resource included in each local graph. Thus, after removing the inconsistent identifiers, the external id overlap can be interpreted as an estimate of the number of shared taxon names between two databases (Table 2). This cannot be interpreted as total taxa in each resource.

# Discussion

GUODA is a high performance computing resource for biodiversity science that provides scalable solutions for working with large data sets in a collaborative, online environment. The 10 minute processing time for 20 GB of compressed JSON is far faster than any current mapping method used in biodiversity; however, it does benefit from the mapping already completed inside Wikidata. For example, the Wikidata entry for *Panthera leo* (https://www.wikidata.org/wiki/Q140) has 25 links to external databases, not all of them biodiversity-related. Other efforts using name-string-matching to link biodiversity databases take much longer to map resources together. For instance, EOL takes more than a day to map the content it receives from providers to a unified classification (Rice pers. comm.). Similarly, the taxon matching in BioNames and LinkOut took days to complete (Page pers. comm.). Projects like OTT, Wikidata, and GloBI that keep identifier-based taxonomic graphs make it easier to link databases at scale.

Despite the notoriously poor nature of taxon names as identifiers, they are still commonly used to link biodiversity data. A much-discussed solution has been the use of universal, unique, persistent, resolvable identifiers across the biodiversity data landscape, but the social barrier to a universal identifier system has, thus far, proven insurmountable. Rather than rely on name strings or a universal identifier system, this method uses the graph of identifiers to map taxa across two databases. This identifier-based method has the potential to be faster and easier than name-string matching without some of the social difficulties of a single identifier system.

Most biodiversity databases and nomenclatural authorities expose their data in idiosyncratic ways that are not suitable for batch processing. If data sources published their taxon identifier graph as a lookup table (as described in this paper) integrating across databases would be much

easier. Now, users have to learn a unique format for every data source. These lookup tables have the advantage of being easy to version and integrate.

In addition to fast linking of biodiversity databases, comparison of identifier graphs may be a scalable way to find inconsistencies, especially when multiple biodiversity databases/identifiers are included. By linking GloBI to OTT and WD, inconsistent names or false positive name matches were detected by considering the (lack of) overlap of GloBI names with OTT and WD external identifier schemes. These inconsistencies might be introduced by a dataset or a name resolution method that produces ambiguous results. In addition, inconsistencies can indicate a disputed / outdated name like "GLOBI:null@Senecio pectinatus" which maps to GBIF:8317096 and GBIF:8414746. This would be considered an inconsistent mapping and suggests that Senecio pectinatus is an outdated name. Combining the speediness with the promise of scalability, a near-real-time name consistency check can be implemented to detect inconsistencies across various systems in the biodiversity data-ecosystem introduced by integration bugs, taxonomy updates or differences of interpretation.

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

Future work on the GUODA infrastructure includes traning and evaluating neural network models on image data, containerization of the GUODA components to allow the system to be run in additional data centers, and refinement of the end-user interface to integrate programming, source code, and publication to make research more reproducible. GUODA's most impactful contribution has likely been the availability of readily formatted biodiversity data and new data sets will continue to be added to the collaboration platform, enabling domain experts and technical experts to answer new questions in the future.

GUODA, and hosted data analytics infrastructure in general, has the potential to drastically improve biodiversity science by making multiple biodiversity databases accessible to scientists for analysis on their laptop or desktop. Users still need to have some programming skills, which have now become an essential skill in biodiversity science.

# Conclusions

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

# Acknowledgements

The authors would like to acknowledge support and resources provided by the ACIS lab. The authors would like to acknowledge José A.B. Fortes for providing infrastructure and creating room for collaboration. Funding was provided by David Rubenstein and the Encyclopedia of Life, which also helped establish an informal yet pragmatic cross-institutional collaboration and by iDigBio, NSF award 1547229, which provided for staff, travel, and publication fees for this collaboration.

# References

Bingham, H. et al., 2017. The Biodiversity Informatics Landscape: Elements, Connections and Opportunities. *RIO*, 3, e14059. Available at: https://riojournal.com/article/14059/.

Hindman, B. et al., 2011. Mesos: a platform for fine-grained resource sharing in the data center. In *Proceedings of the 8th USENIX conference on Networked systems design and implementation*. pp. 295–308.

Hortal, J. et al., 2015. Seven Shortfalls that Beset Large-Scale Knowledge of Biodiversity. *Annual Review of Ecology, Evolution, and Systematics*, 46(1), pp. 523–549. Available at: http://www.annualreviews.org/doi/10.1146/annurev-ecolsys-112414-054400.

Kluyver, T. et al., 2016. Jupyter Notebooks -- a publishing format for reproducible computational workflows. In F. Loizides & B. Schmidt, eds. *Positioning and Power in Academic Publishing: Players, Agents and Agendas*. Amsterdam: IOS PRess, pp. 87–90.

Page, R., 2008. Biodiversity informatics: the challenge of linking data and the role of shared identifiers. *Briefings in Bioinformatics*, 9(5), pp. 345–354. Available at: https://academic.oup.com/bib/article/9/5/345/267216.

Page, R., 2013. BioNames: linking taxonomy, texts, and trees. *PeerJ*, 1, p.e190. Available at: https://peerj.com/articles/190/.

Page, R., 2011. Linking NCBI to Wikipedia: a wiki-based approach. *PLoS Currents*, 3, RRN1228.

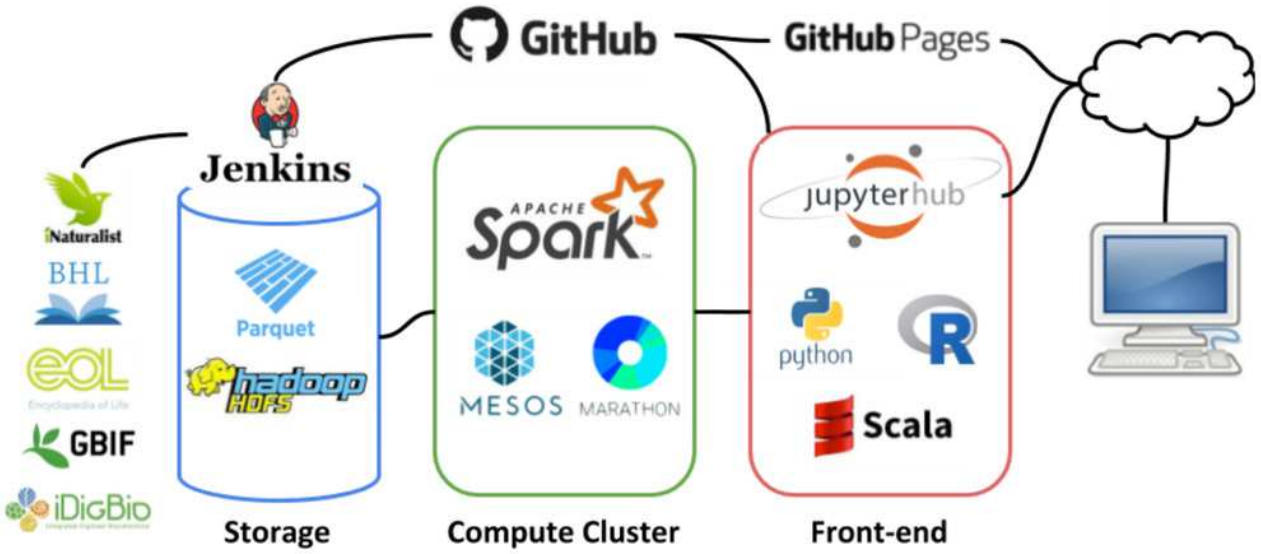
Page, R., 2007. Tbmap: a taxonomic perspective on the phylogenetic database treebase. *BMC*

*Bioinformatics*, 8(1), p. 158.
Parr, C. et al., 2014. The encyclopedia of life v2: providing global access to knowledge about life on earth. *Biodiversity Data Journal*, 2, e1079.
Poelen, J., 2018a. 20 GB in 10 minutes: Data linking across major biodiversity databases: Data supplements. Version 0.1 Zenodo http://doi.org/10.5281/zenodo.1213477
Poelen, J., 2018b. Global Biotic Interactions: Taxon Graph. Version 0.4.2 Zenodo http://doi.org/10.5281/zenodo.1210315
Poelen, J., 2018c. Global Biotic Interactions: Taxon Graph. Version 0.3.0 Zenodo http://doi.org/10.5281/zenodo.1210308
Poelen, J., 2018d. Global Biotic Interactions: Taxon Graph. Version 0.3.1 Zenodo http://doi.org/10.5281/zenodo.1213465
Poelen, J.H., Simons, J.D. & Mungall, C.J., 2014. Global biotic interactions: An open infrastructure to share and analyze species-interaction datasets. *Ecological Informatics*, 24, pp. 148–159.
Rees, J.A. & Cranston, K., 2017. Automated assembly of a reference taxonomy for phylogenetic data synthesis. *Biodiversity Data Journal*, 5, e12581.
Shvachko, K. et al., 2010. The Hadoop distributed file system. In *MSST '10 Proceedings of the 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies*. pp. 1–10.
Voß, J., 2016. wikidata-taxonomy 0.2.7. Version 0.2.7 Zenodo http://doi.org/10.5281/zenodo.60708
Vrandečić, D. & Krötzsch, M., 2014. Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, 57(10), pp. 78–85. Available at: https://cacm.acm.org/magazines/2014/10/178785-wikidata/fulltext.
Wikidata, 2018. Wikidata dump 2017-12-27. Zenodo http://doi.org/10.5281/zenodo.1211767
Zaharia, M. et al., 2016. Apache Spark: a unified engine for big data processing. *Communications of the ACM*, 59(11), pp. 56–65.

# Figure 1

GUODA Infrastructure

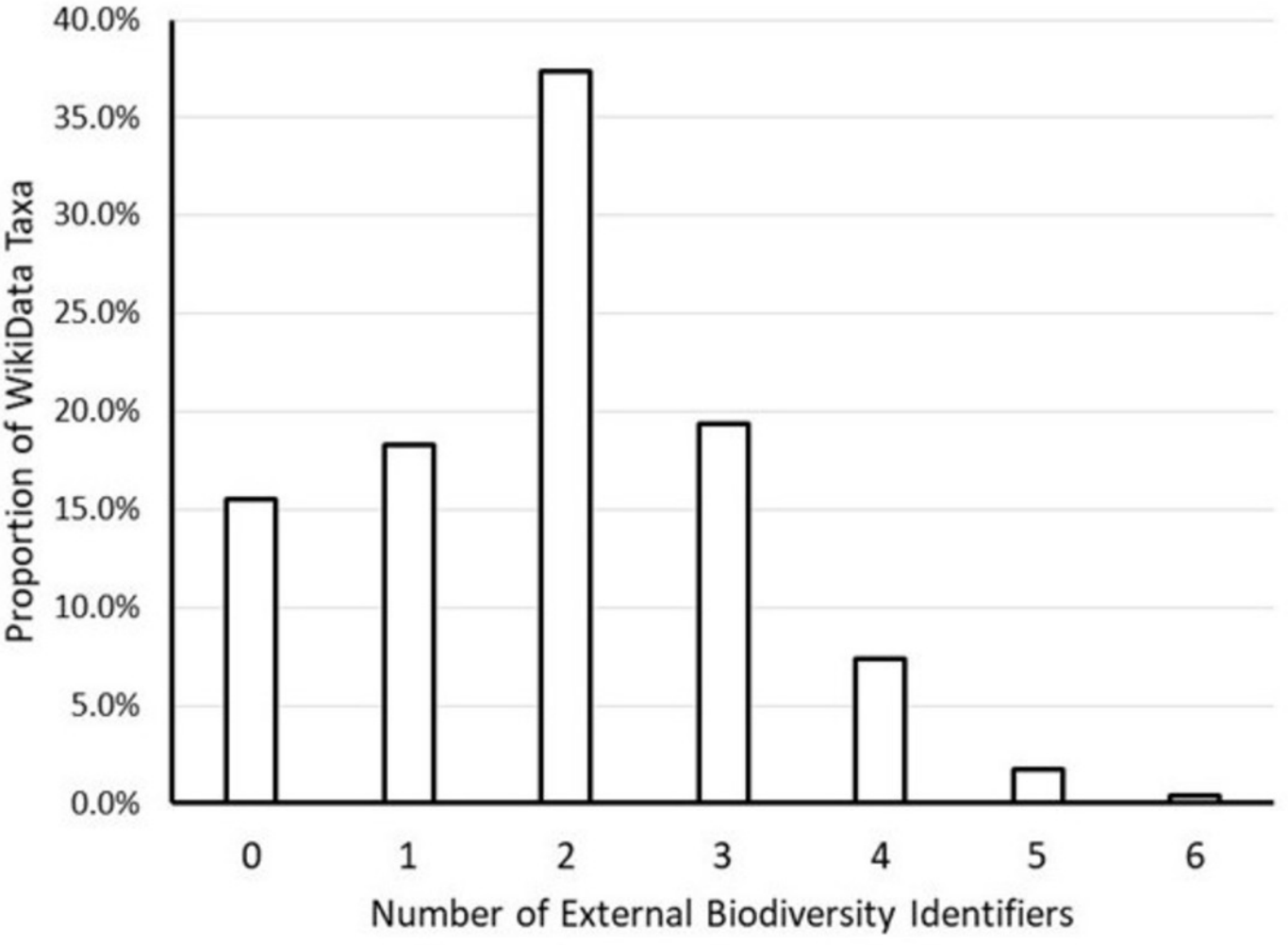
Data from biodiversity databases is loaded into GUODA as Parquet files (Storage). When a user working in a Jupyter Notebook (Front-end Server) triggers a job interactively or via GitHub and Jenkins, the data are analyzed using Apache Spark (Compute Cluster). This infrastructure allows a user working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



# Figure 2

Frequency of Wikidata taxa linked to biodiversity databases

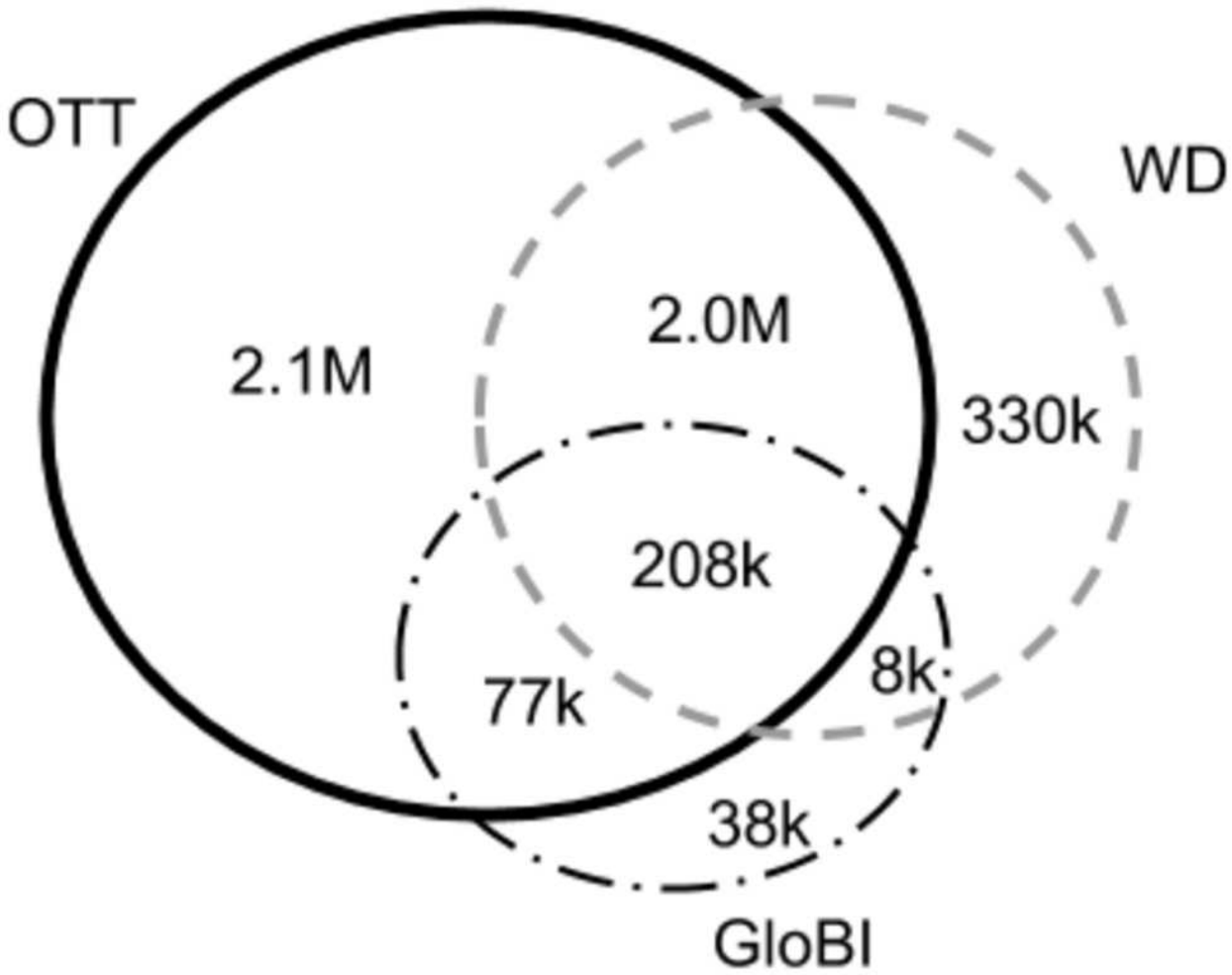
This graph shows the proportion of the approximately 2.3 million Wikidata taxa with 0,1,2, etc. links to external biodiversity databases (NCBI, ITIS, GBIF, EOL, FishBase, Index Fungorum and iNaturalist). The majority of Wikidata taxa had at least two links. A little more than 15% of Wikidata taxa had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



# Figure 3

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 23,000 WoRMS links, 72,000 NCBI links, 103,000 GBIF links and 10,000 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.



## Table 1(on next page)

Selected biodiversity databases and their size

1

| 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 |

2

## Table 2(on next page)

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%) |

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