Gap Analysis of GloBI: Identifying Research and Data Sharing Opportunities for Species Interactions

Mariana Cains[1], Nuria Altimir[2], Srini Anand[3], William Liao[3], and Sean Shiverick[3]

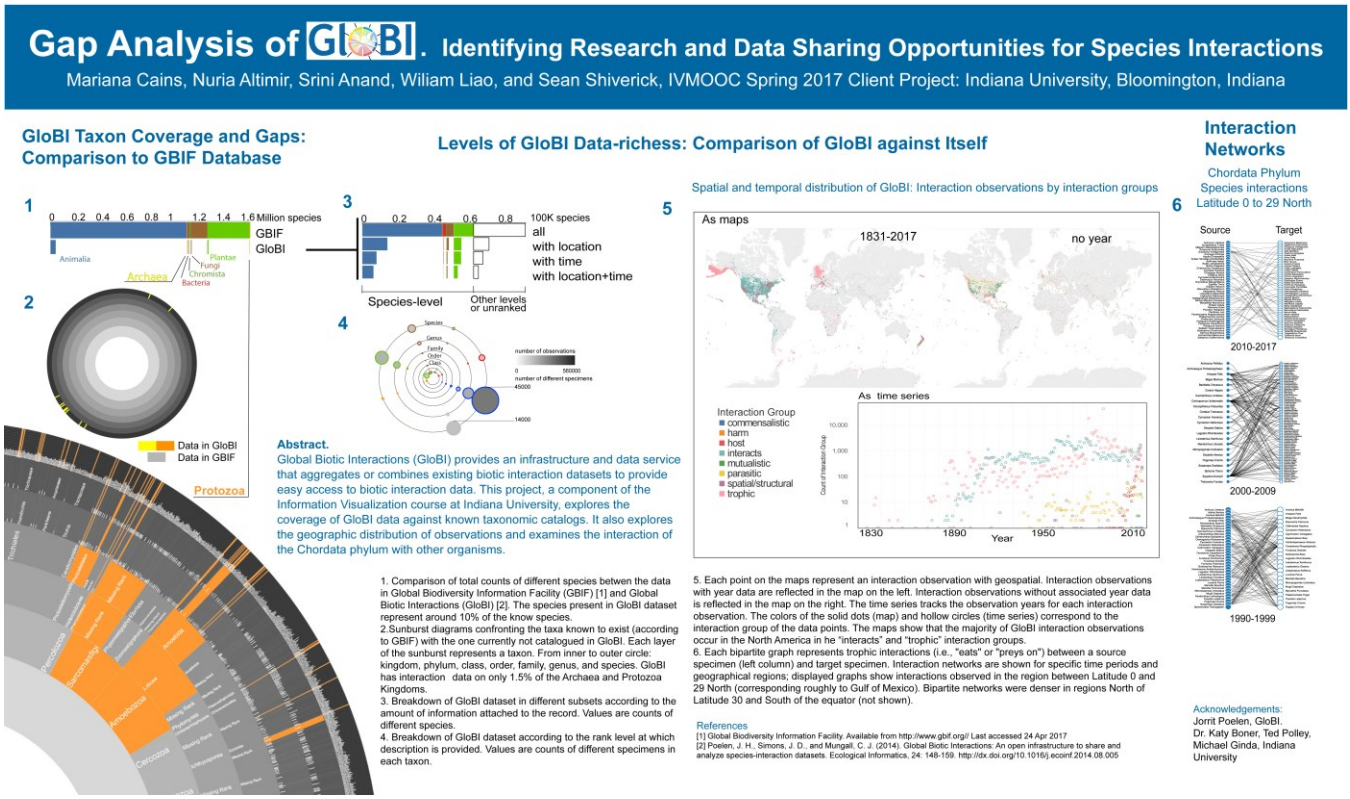

Figure 1. Final Visualization Poster for Information Visualization Massive Open Online Course 2017 GloBI Client Project (High resolution at https://doi.org/10.5281/zenodo.814922)

Abstract— Global Biotic Interactions (GloBI) provides an infrastructure and data service that aggregates and archives known biotic interaction databases to provide easy access to species interaction data. This project explores the coverage of GloBI data against known taxonomic catalogs in order to identify 'gaps' in knowledge of species interactions. We examine the richness of GloBI's datasets using itself as a frame of reference for comparison, and explore interaction networks according to geographic regions over time. The resulting analysis and visualizations intend to provide insights that may help to enhance GloBI as a resource for research and education (Figure 1).

Index Terms— GloBI, GBIF, Taxonomy, Interactions, Visualization, Sunburst, Bubble chart, Bipartite network, Geospatial, Temporal, Neo4j, Graph database, Chordate, Eltonian Shortfall

1. INTRODUCTION

1.1 OVERVIEW

The web of relations among biological organisms in the natural world represents a complex system of interactions. Efforts to catalog the rich nature of species interactions reveals fundamental gaps in biological diversity (biodiversity) knowledge. Global Biotic Interactions (GloBI) is an ambitious project undertaken by Poelen, Simons, and Mungall [1] to consolidate observations and research data of this rich phenomenon. GloBI provides an interactive infrastructure for aggregating databases and archiving species interaction datasets for research and education. An interactive browser allows GloBI users to query its database according to species taxa, scientific name, region, and visualize species interactions in tree graphs and interaction networks. Direct applications of GloBI's data services are found in the Encyclopedia of Life (EOL) [2], and the Gulf of Mexico Species Interaction Database (GoMexSI) [3]. The scope of interaction data in GloBI's databases is vast, ranging from field observations of specimen stomach contents to digitally archived data. According to Poelen and his colleagues, the sampling density between aggregated databases is highly variable across taxonomic rank, region, and time [1].

[1] *Mariana Cains is with Indiana University School of Public and Environmental Affairs. mgcains@indiana.edu*

[2] *Nuria Altimir is with the University of Helsinki.*

[3] *Srini Anand, William Liao, and Sean Shiverick are with Indiana University School of Informatics and Computing.*

The main goal of the present project is to examine the extent of data coverage in GloBI and identify potential gaps in biodiversity data. We constructed several visualizations to show (a) the extent of the hierarchical taxonomy in GloBI compared to larger, more complete taxonomies, (b) the completeness, or data-richness, of GloBI within its own datasets as a frame of reference, and (c) the network structure of species-level interactions in GloBI according to geospatial and temporal information. These three levels of analysis each serve to identify gaps in GloBI's data structure and point to areas of improvement or expansion. The resulting visualizations seek to enable users to understand potential strengths and shortcomings of GloBI's datasets. Overall, our efforts are intended to offer suggestions for helping to make GloBI a more effective tool for research and encourage the development of GloBI as a data resource.

1.2 BACKGROUND and RELATED WORK

1.2.1 Taxonomy of Biological Classifications

Taxonomy is the practice and science of organizing components of a system into a hierarchical structure based on a variety of shared traits between the components. Humans have spent centuries trying to develop a mutually exhaustive taxonomy that could hierarchically organize all of the known and unknown organisms in the world.

Carl Linnaeus, who is considered the "father of taxonomy", published the first edition of *Systema Naturae* in 1735 which offered an organization system for the inhabitants and contents of Earth into three kingdoms: Animalia (animals), Vegetabilia (plants), and Mineralia (minerals). Each kingdom was further organized, with increasing specificity, by class, order, genus, and species [4]. In 1768, the 10th edition Systema Naturae introduced the zoological organization of living organisms where each kingdom was biologically and morphologically organized [5]. With the species discoveries and scientific advances since *Systema Naturae,* the Linnaean taxonomic system has been organized into the following taxonomic levels (i.e. taxa) of biological classification (in increasing specificity): kingdom, phylum, class, order, family, genus, and species. The information and concepts informing how living things should be grouped and categorized has changed over time. The contemporary adaptation of the Linnaeus taxonomy consists of seven kingdoms: Animalia, Archaea, Bacteria, Chromista, Fungi, Plantae, and Protozoa [6].

In this traditional taxonomy the application of a name to a taxon is based on both a type and a rank. There is also a phylogenetic approach to taxonomy that ties names to a clade, which is a group consisting of its ancestor and all its descendants. The sorting of clades are presently under constant development, but this approach already results in the use of nested taxa nomenclature coexisting with the traditional one.

The taxonomic hierarchical organization of the biological world has enabled scientists to better understand the similarities and differences between species in addition to aiding the study of biodiversity.

1.2.2 Biodiversity and Interactions

Incompleteness of Biodiversity Data

Aggregated data infrastructures bring together dispersed sources into a common framework and facilitate the capitalization of the available knowledge. In the case of biodiversity databases, the available knowledge is clearly incomplete. Indeed, available biodiversity data is known to be full of biases and gaps. This situation is due to the complex and extensive nature of biodiversity as well as its recording, making it overall impossible to achieve complete knowledge. The data shortfalls are of many types, with seven being identified and their consequences and remedies analyzed (Table 1) [7].

Table 1. The Seven Main Shortfalls of Biodiversity Knowledge [7]

Linnean shortfall – Most of the species on Earth have not been described and cataloged
Wallacean shortfall – The knowledge on the geographic distribution of most species is incomplete, being most times inadequate at all scales
Prestonian shortfall – Lack of data on species abundances and their dynamics in space and time are often scarce
Darwinian shortfall – Lack of knowledge about the tree of life and evolution of species and their traits
Raunkiæran shortfall – Lack of knowledge on species' traits and their ecological functions
Hutchinsonian shortfall – Lack of knowledge about the responses and tolerances of species to abiotic conditions
Eltonian shortfall – Lack of enough knowledge on species' interactions and their effects on individual survival and fitness

Biotic interactions are practically impossible to be completely described, due to the complexity of the phenomena. Biotic interactions could actually be inferred from proxies rather than direct observations [8]. Still, accurate records are needed to validate such models. Also, the more local the scale of study the more actual observations are useful and are needed. Table 2 details the consequences of lack of biotic interaction data and the strategies proposed to tackle with it.

Identifying gaps and biases in our knowledge, existing databases, and datasets serves several purposes. First, it helps identify areas underrepresented and guide prioritization efforts [9]. It also provides indispensable information to interpret the available data. This information comes under many formats, for example: data quality, quality flags, uncertainty estimates, or so-called maps of ignorance. To furnish biodiversity related datasets with such information is a work in development [10].

1.2.3 Identifying Gaps in GloBI's Datasets

A major challenge for identifying gaps in data coverage is to get a handle on the large volume of data in GloBI's aggregated databases. In order to grasp what is missing, first we must explore what data there are, and identify sources the data were obtained from. As Poelen and his colleagues describe, "spatial, temporal, and taxonomic coverage of the combined datasets shows that the aggregation of the described data sources covers about 50,000 taxa (or 8% of total number of ITIS taxa) in a period from 1897 until the present"[1]. As described above, taxonomic coverage varies considerably between datasets, with the majority of coverage contributed by three principle sources [11] [12] [13], with the largest coverage provided by data mining to extract species interactions from text objects in EOL [11].

Table 2. Eltonian Shortfall of Biotic Interactions: Consequences and Strategies [7]		
Eltonian shortfall : Consequences	Short-term strategies to account for uncertainty	Long-term strategies for filling in the shortfall
• Lack of ability to predict species' responses to global change • Lack of knowledge about assembly rules • Inability to predict processes in non-analog communities • Difficulty of restoration processes • Inability to predict diseases • Inability to characterize community structure	• Concentrate efforts on the best-studied interactions and well-resolved taxa • Produce careful meta-analyses of the best datasets • Prioritize studies on interaction networks at sites which hold basic data from other studies (e.g. permanent forest plots)	• Set clear and widely applicable definitions of interaction types • Develop standards for field procedures to ensure minimum comparability, either longitudinal, across sites or across systems • Allocate resources for large-scale field work, prioritizing interactions that are clearly linked to key ecosystem processes and services (e.g. pollination) • Invest in applying new technologies to interaction surveys (e.g. fingerprinting or molecular profiling of gut contents)

In addition, most of the spatiotemporal interaction data were provided by stomach sampling in International Council for the Exploration of the Sea datasets [14, 15]. Geographically, the spatiotemporal interaction data are concentrated in Europe, North America, the Southern Ocean, New Zealand, with highest densities of observations obtained from the Gulf of Mexico, North Sea, and Weddell Sea [16] [1]. Furthermore, variability in the consistency of measurements obtained from field observations versus archived interaction databases may constrain visualization of species interactions overall.

1.2.4 Past IVMOOC GloBI Projects

Species interactions can be visualized using a variety of approaches. Previous teams of Information Visualization Massive Open Online Course (IVMOOC) students mapped GloBI's data resources using interactive platforms to query taxonomic information, static representations of network associations, or a combination of both. Baron et al. [17] created an interactive Ecosystem Explorer designed for educational use with high school students using the GloBI web API. The GloBI Explorer tool allowed users to explore interactions among species and food webs with information cards that displayed a photograph, scientific name, common name, and interactions for each organism. Relations between organisms were represented as bipartite networks in "drill-down" columns, which helped reduce millions of possible interactions down to those of particular interest by region and taxa.

In 2014 Slyusarev et al. [18] created a GloBI food web map as a static network graph with a geospatial choropleth map insert. This graph represented a food web of predator-prey relationships instantiated at species level. A second network graph with weighted edges was generated to identify community structure. In the network of predator-prey interactions, node size corresponded to number of species, edge width to number of original connections. Color was also used to identify marine and terrestrial ecoregions. Integration of geospatial information with the network graph helped to relate specific interactions to world regions. A third team created a spatial food web map illustrating key predator-prey interactions at different depths and locations in the Gulf of Mexico based on data from GoMexSI [3]. None of the past IVMOOC projects delineated gaps in GloBI's data coverage, which is the goal of the present project.

1.3. OBJECTIVES of Visualizations

1.3.1 Compare GloBI with External Reference GBIF

In order to visualize what GloBI "knows" and "doesn't know", we proposed the development of a comparative visualization of the species represented in the GloBI dataset versus all the species known and identified within an external reference system. Ideally, this visualization would help inform researchers and potential contributors to the GloBI dataset about where their data could supplement the expansion of the dataset.

The Global Biodiversity Information Facility (GBIF) [19], which catalogs taxonomic rank-based hierarchy of all 1,643,948 species known to exist or have existed (accessed March to April 2017), is used as the reference system on which the extent of GloBI species data is visually encoded. Figure 2 shows a simplified mock-up that depicts the difference between GloBI and GBIF taxon coverage.

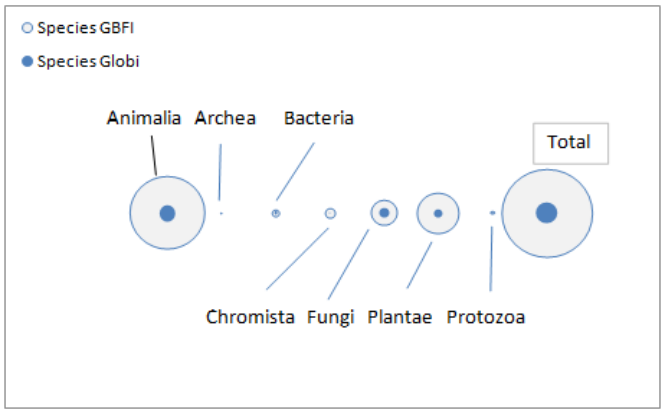

Figure 2. Simplified mockup of taxon coverage of GloBI compared with GBIF data.

1.3.2 Analysis of GloBI Data-richness

GloBI catalogs interaction data at various taxonomic levels (e.g. kingdom, species), and if available, the location and date of the field observation of the interaction. The more specific the data, the more detailed analysis of an interaction can be performed. We analyzed the completeness of the data in terms of presence or absence of records at the species level, availability of information on geolocation, and interaction observation date. For this purpose, interaction observations

containing these three attributes are categorized as “golden data”. As part of this analysis we present an overview of the counts for different data-rich sets and, as well as geospatial maps of interactions observations over time.

1.3.3 Interaction Networks

Another way to identify gaps in knowledge is based on the idea that as our queries into GloBI’s data resources become more focused, some of the richness of species interactions is reduced as data are filtered by species, location, or time. For example, researchers using GloBI as a source of data will extract information relevant to specific research questions, according to their perspective and needs. Such queries may return different cross sections of interaction data, by specific species, in particular regions, over a certain period of time, returning smaller portions of the entire dataset. Given practical considerations, it may prove difficult to visualize interactions at the Kingdom level. Therefore, we explore interaction data for the Chordata phylum, creating various subsets at the species level to represent source-target interactions as bipartite networks.

The concept of a “golden dataset” is used to represent the datasets that include interactions with geospatial location and temporal information. In addition to getting a sense of where the majority of interaction occurrences are observed, we will also gain some understanding of where observation are lacking. In addition to location, time slices of interactions will show when observations were obtained, and how they may have changed over time. Overall, several types of visualizations were produced to address different aspects of our gap analysis in an attempt to understand the extent of data gaps in GloBI,

2. METHODS

2.1 THE DATA

2.1.1 Extraction of Datasets from GloBI

The GloBI dataset was accessed via a neo4j database, an implementation of a graph database [1]. It contains information from different studies during which researchers studied interactions amongst different organisms, called *specimens* in this dataset. There are various types of interactions and at different taxonomic levels.

The major data entities in GloBI are depicted in Figure 3. The GloBI database has 115,358 unique specimens (across various taxa levels) for 36 types of interactions as verified from the Tab Separated Values (TSV) extract. Currently (as of April 2017) there is a total of 2,263,653 specimen observations across all taxa. GloBI data are made available in different formats – graph database (as a Neo4j database), Tab Separated Values (TSV), Resource Description Framework (RDF), and Darwin Core [21]. We selected both the Neo4j database the TSV file for our project. The TSV file is a listing of all interactions, with each row representing one type of interaction between two specimens. The general format of a row of data is:

[sourceSpecimen] [interacts with] [targetSpecimen] [optional observation date of interaction record] [optional location of interaction record] [reference citation of study that recorded interaction]

In addition to the name of the specimen, there are additional characteristics such as the taxon level, the full taxon path, and identification numbers to taxonomic catalogs. This TSV file was imported into a table named *INTERACTIONS*. We added a number of indexes to this table as we examined the data from various perspectives. The columns upon which we added indexes were:

sourceTaxonName, *sourceRank*, *sourceTaxonID*, *targetTaxonName*, *targetRank*, *targetTaxonID*, *interaction*, *locality*, and *eventDateUnixEpoch*.

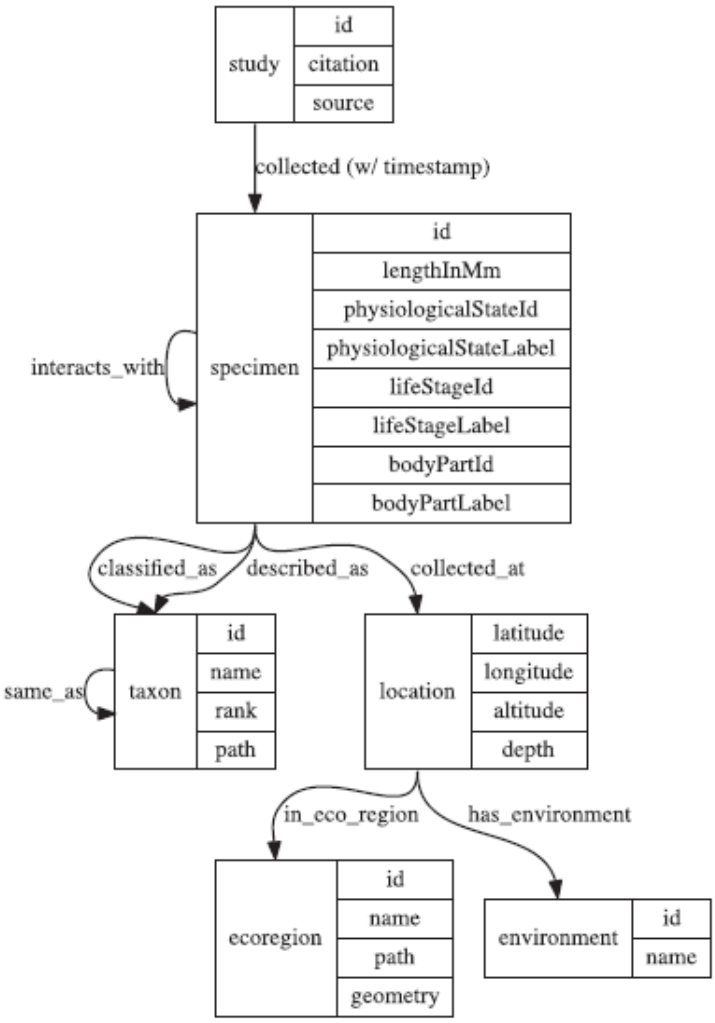

Figure 3. GloBI Database Architecture (from Poelen et al. [1])

The *INTERACTIONS* table was by far the largest table in the SQL Server database, about 4 GB. To aid querying, we added a table, *GloBI_TAXON_PATH* that contained a parsed list of the taxon ranks for each specimen. A third table, *TAXON_MAPPING*, was created to assist the process of normalizing the various taxon ranks to a small but widely used subset. For example, GloBI has 17 variations for the class taxon, such as cl., Cohort, klase, onderklasse etc.

The SQL Server database is available here: https://doi.org/10.5281/zenodo.804103

2.1.2 Extraction from GBIF Reference System

Choice of reference system. The project originally planned on comparing GloBI with the Catalogue of Life (CoL) taxonomy [22]. However, CoL identification number are not consistent across CoL releases and these IDs are not in the GloBI dataset. Matches would have to be found via specimen names and that would have resulted in false negative matches due to various naming schemes. The client suggested that we use GBIF as the

reference system since the GBIF IDs are already available in GloBI. Also, GBIF contains information from CoL and other taxonomies, making GBIF a superset of taxonomic data. In GBIF, a number of datasets have been brought together. The GBIF database has 31,886 datasets from 1,186 institutions, and 1,643,948 species (it does however note that 1,174,586 species (71%) are still under review).

Extraction of data from GLoBI and GBIF. By merging GloBI and GBIF datasets into a common data store, we could compare the GloBI records with GBIF IDs to all of the GBIF records. We chose to use SQL Server as the common data platform. Associated with each GloBI specimen, is its GBIF ID. However, this GBIF ID is embedded within a long delimited field that contains the ID for all taxon levels for this specimen. In addition to GBIF, it contains IDs for CoL, EOL, World Registry of Marine Species, Integrated Taxonomic Information System and others [19]. We pre-processed the specimen data from GloBI to only extract the corresponding taxon level GBIF ID. For instance, if the specimen was at the species taxon, we located the GBIF ID for the corresponding species. If the specimen was at the genus taxon, we chose the ID for the genus taxon. GBIF IDs were extracted and stored in a separate table, *EXTERNAL_IDS*, linked to the TaxonID that is unique to GloBI. The table *EXTERNAL_IDS* also contains a source identifier, allowing it to be used, without change, against other taxonomic catalogs.

GBIF contains taxonomic information and all occurrences known to them. Our initial investigation into GBIF led us to downloads that contained all occurrence data for a species, resulting in large dataset for even the smallest of kingdoms. For example, the kingdom Archaea has 523 known species, but the dataset had 18,888 records. The phylum Chordata with 120,428 species turned out to contain 455,403,193 occurrence records and was about 700 GB in size. Lacking the computing power to extract and process this volume of data, we chose to visualize GloBI's coverage of the Archaea and Protozoa kingdoms However, close to the end of this project, we discovered as a proof of concept. Taxon store within GBIF, available at http://doi.org/10.15468/39omei. This dataset was much smaller than the occurrence set.

2.1.3 Extracting GloBI Data-Richness and Interaction Network Visualizations

Unique species-level, geolocation, and time observations. The overview bar graphs were built with data extracted April 4, 2017 from GloBI using Neo4j. It contained a total of 556,496 unique species-level records with the following attributes: source taxon ID, source taxon rank, source taxon path name, decimal latitude, decimal longitude, and event date Unix Epoch.

Geolocation/temporal interactions. Interaction observations that contained geolocation and or time information were extracted across all seven domains. The geolocation and time interaction dataset was extracted April 17, 2017 from GloBI using Neo4j and contained 1,064,757 records (rows) with 20 attributes (columns): source taxon ID, source taxon rank, source taxon path names, source taxon path ids, source taxon path rank names, interaction type, interaction type id, target taxon ID, target taxon rank, target taxon path names, target taxon path ids, target taxon path rank names, decimal latitude, decimal longitude, locality, event date Unix Epoch, reference citation, and references.

Chordata species-level interactions. The Animalia kingdom was a very large dataset, and therefore it was more practical to extract species level interactions for the Chordata phylum. The Chordata interaction dataset was extracted on March 30, 2017 from GloBI using Neo4j, and contained 940,941 records with the same 20 attributes listed above.

2.2 DATA PREPARATION

2.2.1 Data Cleaning

The data cleaning described below was performed on geolocation/time interactions and Chordata interactions datasets the using the statistical computing software R [23]. Raw and clean datasets are available here: https://doi.org/10.5281/zenodo.814912

Unix Epoch. The calendar date of field observations was recorded in the Unix Epoch time unit, which is the signed number of seconds or milliseconds since January 1, 1970, thus the value must be converted into its respective calendar year in order to be used in our visualizations. Conventionally, 10 digit values are dates recorded in seconds, 13 digit values are recorded in milliseconds, and negative values represent dates prior to 1970. Preliminary temporal analysis showed that there were seven to 14 digit negative and positive values. With the various number of digits the 10 and 13 digit, conversion rules no longer held true. We contacted the client who was unaware of the varying number of digits in the Unix Epoch attribute. Within a day or so, the client was able to confirm that all of the values, regardless of the number of digits were measured in milliseconds.

The POSIXct function in R was used to convert the Unix Epoch milliseconds to calendar years. After importing the interactions datafile (`data`) into R, the numeric attribute of Unix Epoch milliseconds (`eventDateUnixEpoch`) was converted to seconds (`seconds`), and then converted into calendar dates (`dates`) using the as.POSIXct function. The as.POSIXct function requires the input of Unix Epoch seconds (`seconds`) and the origin date (`origin = '1970-01-01'`) corresponding to the Unix Epoch seconds. The year of the field observations (`year`) was extracted from the calendar date by using the format function to return of the year in YYYY format (`%Y`). Below is an example of the R code used (with comments) to determine the year field observations.

```
> data$eventDateUnixEpoch <-
as.numeric(as.character(data$eventDateUnixEpoc
h)) ## convert character string to numeric
> data$seconds <- data$eventDateUnixEpoch/1000
## creates new variable named "seconds" i.e.
converts the UnixEpoch from milliseconds to
seconds
> data$date <- as.POSIXct(data$seconds, origin
= '1970-01-01') ## creates a new variable
named "date" using the as.POSIXct function
> data$year <- format(data$date, format="%Y")
## creates a new variable named "year" by
pulling out the YYYY from the "date" column
```

The resulting `year` column was then used to perform temporal analysis on a yearly basis.

0° latitude, 0° longitude. Several interaction observations were cataloged as occurring at 0° latitude, 0° longitude in the Gulf of Guinea. This seemed erroneous because the reference citations for these interaction observations referenced locations other than the Gulf of Guinea. This issue was brought to the client's attention and they confirmed that the interaction observation of interest did not have an associated location, meaning the zeros were an artifact of the data extraction process. The 0° latitude, 0° longitude attribute value was

replaced with the null value in R, NA. Below is an example of the R code used (with comments) to replace 0° latitude, 0° longitude with NA, NA.

```
> data[which(data$decimalLatitude == 0 & data$decimalLongitude == 0), c("decimalLatitude", "decimalLongitude")] <- NA
## replaces 0,0 location with null value
```

2.2.1 Data Aggregation

One way to facilitate the computation and interpretation of GloBI dataset is via aggregation into meaningful groups. Aggregation does not decrease the number of records, but arranges them in fewer categories and simplifies and unifies analysis. The nature of the information in the database is highly hierarchical, thus it was possible to organize data into higher order groups. The below detailed data aggregation for taxa groups and interaction types was performed on the unique species-level dataset in Excel and on the geolocation/temporal dataset using the statistical computing software R [23] due to record limits in Excel.

Taxa groups. In GloBI, the taxonomic level at which specimens are described vary. That means that specimens are not always described at the taxon level. The Rank Paths (kingdom, phylum, etc) are restricted to the main 7 rank (i.e. taxa), nor the same structure for all the records. To aggregate the data, taxa labels were grouped to the immediate higher rank, the ranks of reference Kingdom, Phylum, Class, Order, Family, Genus and Species, as shown in Table 3. Below is an example of the R code used to aggregate the taxa groups.

```
> data[,"sourceTaxonGroup"] <- NA ## create new column
> data$sourceTaxonGroup <- ifelse(
+     data$sourceTaxonRank == "subfamily" |
+         data$sourceTaxonRank == "infrafamily" |
+         data$sourceTaxonRank ==  "tribe" |
+         data$sourceTaxonRank == "subtribe" |
+         data$sourceTaxonRank ==
+"infratribe",
+     "Family",
+     data$sourceTaxonGroup) ## if the sourceTaxonRank is any of the listed values, replace the value of sourceTaxonGroup with "Family"
```

The above code was repeated for each taxon aggregation detailed in Table 3.

Table 3. Taxonomic levels. Taxa labels existing in the GloBI dataset used by this project and how it was homogenized and aggregated.

Kingdom	Phylum	Class	Order	Family	Genus= Geslacht=gen.	Species=Soort= Sp.	no rank
subkingdom	Division	subclass	suborder	Subfamily	Subgenus= subgen.	Subsp.	Unranked clade
infrakingdom	subdivision	Infraclass	Infraorder	infrafamily	Section (bot)	Subspecies	Unknown
Superdivision	subphyllum	superlegion	parvorder	Tribe	series (bot)	Infraspecies	
infradivision	Infraphylum	legion	Section (zoo)	Subtribe	Species hybrid	Microspecies	
superphyllum	microphyllum	sublegion	series (zoo)	infratribe	Generic hybrid	cultivar	
	Superclass	infralegion	Superfamily		hybrid	Forma= F. =Form	
		supercohort	epifamily			Forma specialis= F.sp.	
		cohort				Nothovariety	
		subcohort				infraspecificname	
		infracohort				Var.	
		gigaorder etc				varietas	
		Superorder				variety	
						VariÍÇteit	
						F.Species	
						Var.	
						varietas	
						variety	
						VariÍÇteit	
						Species pro parte	
						Species sensu lato	
						Species aggregate	
						Species group	
						Species specialis	
						Species Species	

Interaction types. The description of the type of interaction is provided at different levels of details depending on the original aim of the collection study. Presently there are 36 types of interactions in GloBI. GloBI aims to make all terminology machine-readable and therefore attempts to match terms with existing ontologies, also in case of the interaction types. We consulted Ontobee [24] to track the position of each interaction type within the existing ontologies and grouped them as shown in Table 4. Based in ecological principles there would be several hierarchies possible, this one is consistent with the present terminology used in GloBI. No attempts were made at collecting all the existing interaction terminology, just the current terminology in GloBI in order to establish aggregation criteria. Below is an example of the R code used to aggregate the interaction types into interaction groups.

```
> data[,"interactionGroup"] <- NA ## create new column
> data$InteractionGroup <- ifelse(
+     data$interactionTypeName == "pollinatedBy" |
+     data$interactionTypeName == "pollinates",
+     "mutualistic",
+     data$InteractionGroup) ## if the interactionTypeName is any of the listed values, replace the NA value of interactionGroup with "mutualistic"
```

The above code was repeated for each interaction type aggregation detailed in Table 4.

Source Kingdom and Target Kingdom. GloBI provides the source and target taxon paths (e.g. Kingdom | Phylum | Class | Order | Family | Genus | Species) at varying levels of specificity. GloBI also provides the source and target taxon path names (e.g. Animalia | Chordata | Elasmobranchii | Rajiformes | Arhynchobatidae | Bathyraja | Bathyraja parmifera). In order to facilitate the visualization of interaction observations, two additional attributes were created: Source Kingdom and Target Kingdom. Below is an example of the R code used to extract the source and target kingdom for each interaction observation.

Table 4. Interaction level. "Hierarchy "of ecological interaction types, here details only for the biotic-biotic interactions that are currently reflected in GloBI. The list of the 36 types of interactions present in GloBI in the rightmost column. (Terms added for completeness during aggregation: consumer /provider, destroys/is destroyed, uses as habitat). Aggregation groups are differentiated by colors. http://www.ontobee.org/ used for interaction aggregation.

Interacts with					source "takes"	target "gives"
	Genetically					
	Molecularly					
	Biotically	Abiotic-biotic				
		Biotic-biotic				
			Trophically		consumer preys on eats farms acquires nutrients from	provider preyed upon by is eaten by farme by provides nutrients for
			Symbiotically (in the broad sense, ecompassing long-term relationships)		Host of (The term host is usually used for the larger (macro) of the two members of a symbiosis) = guest of	Symbiont of (The term symbiont is used for the smaller (macro) of the two members of a symbiosis) = has host
				Commensually (one benefits and the other is unaffected)	is vector of	has vector
				Mutualistically (benefit for both)	pollinates	pollinated by
				Parasitically* (association disadvantaged or destructive to one of the organisms) *While parasites partake in symbiotic and trophic interactions, we chose to organize them under symbiotic.	parasite of (also known as guest) hyperparasitized by pathogen of hyperparasitoid of endoparasite ectoparasite of kleptoparasite parasitoid of	parasitized by (also known as host) hyperparasite of has pathogen has hyperparasitoid has endoparasite has ectoparasite has kleptoparasite has parasitoid
			Spatially/structurally (related but consequences or degree of dependence not specified)		uses as habitat perching on inhabits lives inside of lives on lives near lives under lives with co occurs with adjacent to visits flowers of lays eggs in guest of lays eggs on visits	is habitat of = creates habitat for perches on by inhavited by lives inside of by lived on by lives near by lives under by lives with co occurs with adjacent to has flowers visited by has eggs laid in by has guest of has eggs laid on visited by
			Direct harm (benefits not specified)		destroys damaged by kills	is detroyed damaged is killed by

```
> data[,"SourceKingdom"] <- NA ## create new
column
> data$SourceKingdom <-
ifelse(grepl("Animalia",
data$sourceTaxonPathRankNames), "Animalia",
data$SourceKingdom) ## if the
sourceTaxonPathRankNames contains the string
"Animalia", replace the NA valued of
SourceKingdom with "Animalia"
> data[,"TargetKingdom"] <- NA ## create new
column
> data$TargetKingdom <-
ifelse(grepl("Animalia",
data$sourceTaxonPathRankNames), "Animalia",
data$TargetKingdom) ## if the
targetTaxonPathRankNames contains the string
"Animalia", replace the NA valued of
TargetKingdom with "Animalia"
```

The above code was repeated for each kingdom. Raw and clean datasets are available here: https://doi.org/10.5281/zenodo.814912

2.2.3 Data Partitioning, Subsets, Construction of Golden Dataset

Golden Dataset for all Kingdoms. A summary of the cleaned and aggregated geospatial/temporal dataset for all kingdoms is detailed below in Table 5. The table breakdowns the percent of interactions per source kingdom that qualifies as golden data (i.e. species-species interaction with geospatial and temporal data). It should be noted that these numbers only reflect the interactions that were captured in the April 17, 2017 data extract that resulted in 1,064,757 interaction records and not GloBI in its entirety (n = 2,263,653 interactions). Computational limitations prevented the Animalia and Plantae kingdom from being queried and extracted in their entirety. Of the interactions extract (n = 1,064,757 interaction) only 7.50% of data extract would be considered golden data. Of the 7.50%, the Animalia extract accounts for the 97% of the golden dataset at 7.29% golden data. The Plantae extract account for 2% of the golden dataset at 0.15% golden data. The Bacteria, Chromista, Fungi, and Protozoa kingdoms collectively account for the remaining 1% of the golden dataset at 0.06% golden data.

Table 5. GLOBI Geospatial & Temporal Interaction Observations for Kingdom Extracts									
Source Kingdom	Animalia[a]	Archaea	Bacteria	Chromista	Fungi	Plantae[a]	Protozoa	NA	Total
Extracted Interactions	804,520	134	99,508	6,996	52,077	93,032	5,157	3,333	1,064,757
Geospatial AND OR Temporal Interactions (all taxa)	567,336	0	32	1,522	1,855	38,104	370	2,042	611,261
Subset 1: Geospatial	530,012	0	32	1,522	1,830	37,981	370	1895	573,642
Subset 2: Temporal	262,136	0	1	34	1,561	2,158	3	170	266,063
Subset 3: Geospatial AND Temporal	224812	0	1	34	1,536	2,035	3	23	228,444
Percent Geospatial AND Temporal	36.78	0	1.6E-4	5.6E-3	0.25	0.33	4.9E-4	3.8E-3	37.37
Species-Species Interactions									
Subset 1: Geospatial	231,309	0	1	432	817	14,752	86	805	248,202
Subset 2: Temporal	91,234	0	1	19	620	1,575	2	0	93,451
Subset 3: Geospatial AND Temporal	77,613	0	1	19	615	1,573	2	0	79,823
Percent Golden Data[a] within Geospatial & Temporal Interaction Extract	12.70	0	0.00016	0.0031	0.10	0.26	0.00033	0	13.06
Percent GloBI Extract Golden Data[b] (n = 1,064,757)	**7.29**	**0**	**9E-5**	**2E-3**	**0.06**	**0.25**	**2E-4**	**0**	**7.50**

[a] Computational limitations prevented the Animalia and Plantae kingdom from being queried and extracted in their entirety.
[b] A golden data interaction observation is species-species interaction with geospatial and temporal data.

Interaction Subsets for Chordata Phylum. To represent species-level interactions as source to target bipartite graphs, the Chordata phylum interaction dataset was partitioned into subsets. The source and target taxon rank was filtered to select records with species-level or species-associated taxon ranks (e.g., “super species”, “sub species”, “infra species”). Target taxon rank was filtered to select records at the species-level, species-associated ranks, and miscellaneous ranks, (e.g., “variety”). In terms of the interaction types, almost 95% were trophic (“eats”, 92%; “preys upon”, 2.7%). More ambiguous interactions “interacts with” (3.7%), “visits flowers of” (1%), and miscellaneous (0.5%) were filtered out.

The resulting subset of species-species trophic interactions consisted of 378,272 records. Additional analysis examined the number of records with complete information for (a) geospatial location: latitude, longitude (b) temporal period: year, or (c) with both geospatial and temporal data: 'golden dataset'. The frequency and proportion of counts for each step in partitioning the Chordata species-level trophic interactions dataset are shown in Table 6.

As described above, a large portion of records (23%) indicated 0 latitude, 0 longitude; whereas the citation references indicated a different locations. For example, many observations were from Fish stomach records in the North Sea [25, 26] or taxonomic information extracted from EOL [11], Records missing location information or with 0, 0 were filtered from the Geospatial subset. Similarly, species-species records with no timestamp data (51.8%) were filtered from the temporal subset.

Following this process of partitioning, an all-inclusive 'golden dataset’ of species-species source-target trophic interactions with geospatial and temporal data for the Chordata Phylum consisted of 127,916 records. Of these, an additional 186 had unidentified target taxon name and were removed, resulting in a final golden dataset of 127,030 records. Examination of latitude locations for the golden dataset revealed that the majority of records (91%) were obtained from above latitude 30 North, between the time period of 1980 to 2017. The second largest portion of records (5%) came from the region South of the equator between 1961 to 2017. The smallest portion (3%) was from the region between the equator and latitude 29 North. between 1998 to 2017 (See Table 6 for subset counts).

2.3 CONSTRUCTION of VISUALIZATIONS

2.3.1 Comparison GloBI with External Reference GBIF: Construction of Sunburst

The sunburst diagram was created to provide a snapshot of how GloBI’s database compares to GBIF. The sunburst diagram was created using java’s Data-Driven Documents (D3), a javascript framework.

Table 6: GloBI Species-Level Trophic Interactions: Data Subsets of Phylum Chordata					
	Records Included		Records Excluded		Total
	Count	Proportion	Count	Proportion	
Chordata Interactions	940941	1.000	0	0.000	940941
Species-Species Interactions	376272	0.400	564669	0.600	940941
Subset 1: Geospatial	289555	0.770	86717	0.230	376272
Subset 2: Temporal	181370	0.482	194902	0.518	376272
Subset 3: Geospatial AND Temporal	127916	0.340	248356	0.660	376272
Golden Data Subsets	127030	0.99	886	0.01	127916
Subset 4: Latitude > 30N	116021	0.91	11009	0.09	127030
Subset 5: Latitude >0 <30N	4182	0.03	122848	0.97	127030
Subset 6: Latitude < 0	6827	0.05	120203	0.95	127030

Data were initially provided in .csv format. In order to make it compatible with D3, the data were converted to a json hierarchy. Sample below:


```
{"children":[{"name":"Archaea","class":"Kingdom","children":[{"name":"Crenarchaeota","class":"Phylum","children":[{"name":"Thermoprotei","class":"Class","children":[{"name":"Acidilobales","class":"ORDER","children":[{"name":"Acidilobus","class":"Genus"
```


Scripts from bl.ocks were referenced to create features of visualization (See Appendix C). Final formatting of the sunbursts was performed with Adobe Illustrator.

2.3.2 Analysis of GloBI Data-richness

Stacked Bars and Bubble Diagram. Stacked bars were used to provide an overview of GloBI data coverage. The comparison with GBIF was done based on total unique species number. Amount of known species by kingdom were read from the GBIF webpage and the amount for GloBI calculated from the sum of unique species in the working database. The bars were also used to compare different sets of GloBI between themselves, namely: description at species level, the previous plus presence of location and/or time stamp. The only care required while constructing these graphics was to avoid duplicated instances. Processing was done via de-duplication and pivot tables in Excel.

A bubble diagram was used to provide an overview of the granularity of the GloBI dataset, based on the extracted data aggregated according to Table 3. Using Excel pivot tables, we obtained the counts of records at each taxon level per phylum and per kingdom. This was done for both the amount of unique specimens as well as the actual total observations, which were used to size and shade of the circles, respectively.

Data were written in .html tree, read by Sci2 and visualized through GUESS [27]. Final formatting of the graphs was performed with Adobe Illustrator.

Geospatial and Temporal Distribution. In order to understand the geospatial and temporal coverage of interaction observation records in the GloBI database, we constructed geospatial/temporal maps and time series graphs by importing the geolocation/temporal interactions dataset into Tableau Desktop [28]. Three sets of interaction observation maps and graphs were constructed: interaction groups, source kingdoms, and target kingdom. Figures 4 and 5 show the construction of the interaction group map and times series (respectively) within the Tableau GUI.

The color of the solid dot (map) and hollow circle (time series) visually encodes the interaction group (e.g. commensalistic, harm, host, interacts, mutualistic, parasitic, spatial/structural, and trophic) of the geolocation (map) or year (time series) for the interactions visualizations and the source/target kingdoms (Animalia, Archaea, Bacteria, Chromista, Fungi, Plantae, and Protozoa).The maps progress from year to year using the “Year” slider. Geospatial interaction observations without an observation time were grouped into one group at the end of the “Year” slider named “No Year”. The maps progress cumulative by selecting the “Show History” option on the “Year” slider. Single year views are achieved by deselecting the “Show History” option. The map was formatted in order to filter on a variety of attributes: year, interaction group, source and target taxon rank (e.g. kingdom, species), and source and target kingdom (e.g. Animalia, Plantae).

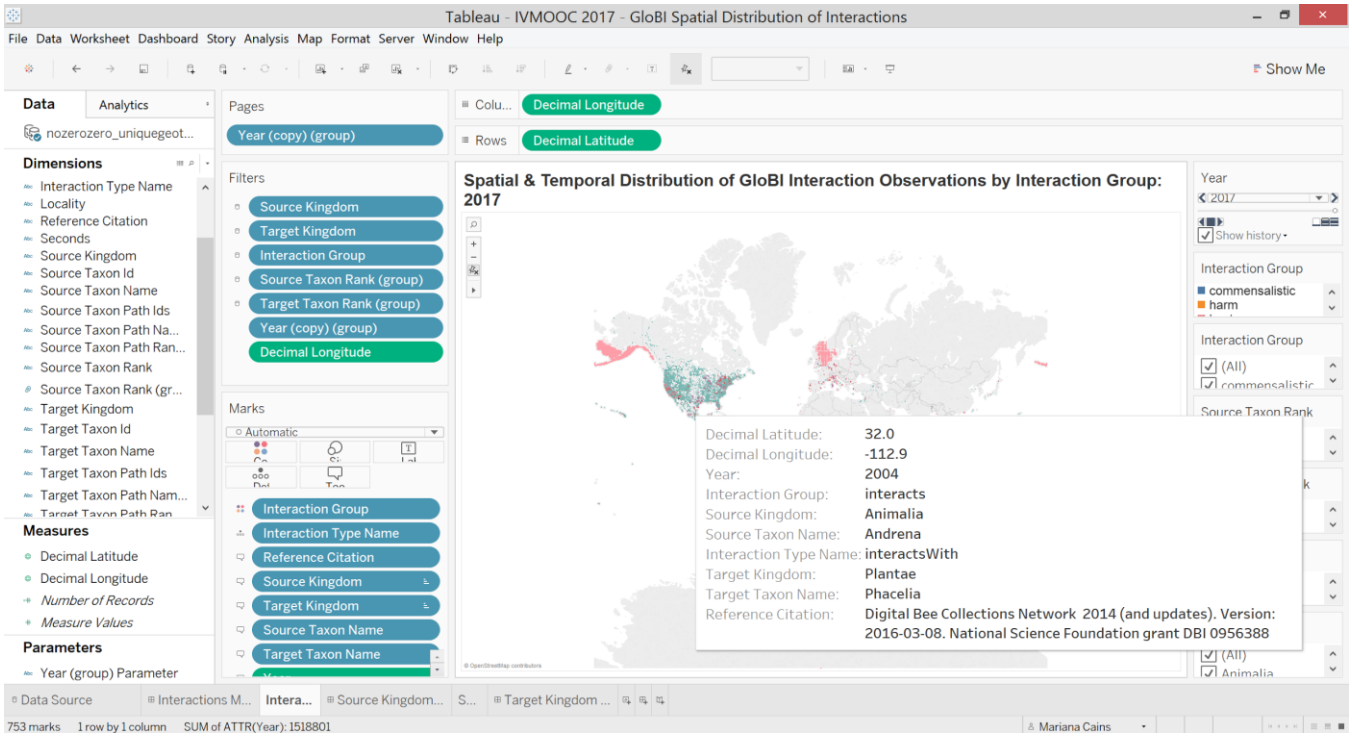

Figure 4. Tableau GUI used to construct geospatial and temporal distribution map of GloBI interaction group observations.

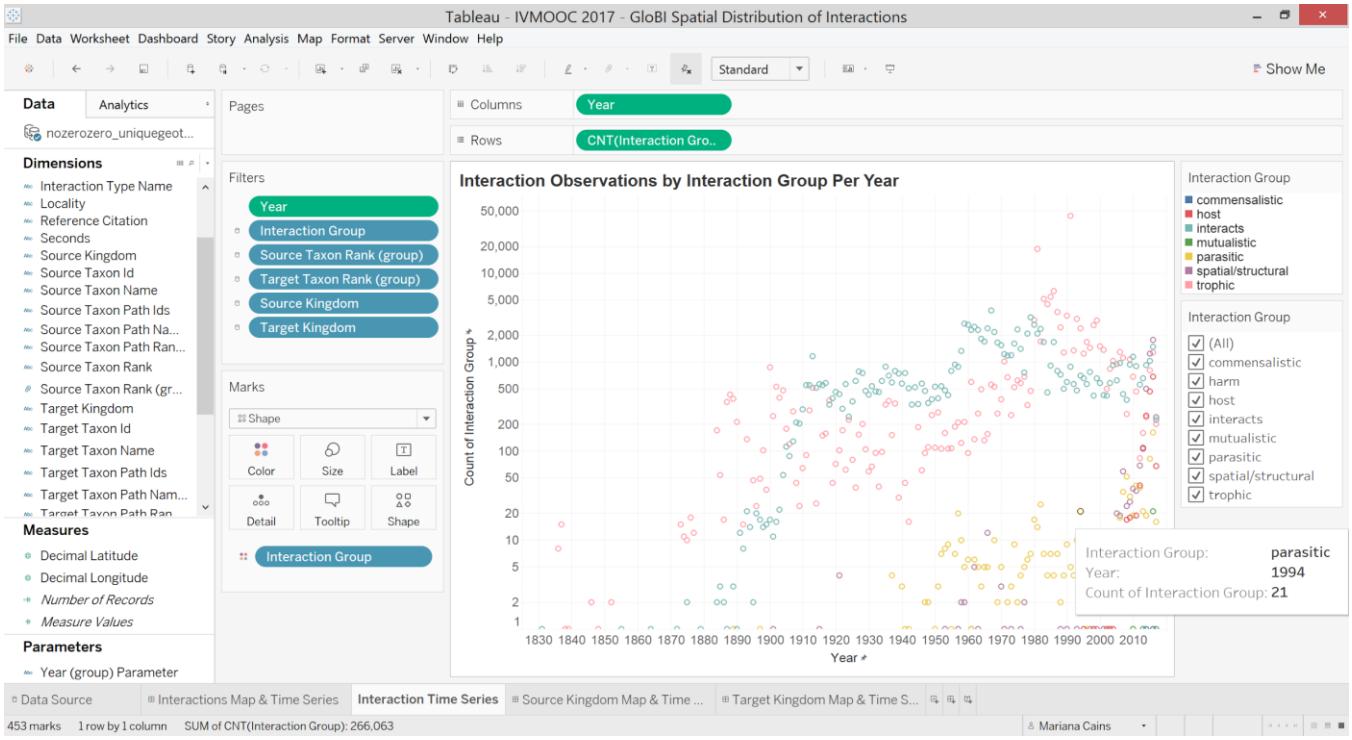

Figure 5. Tableau GUI used to construct time series graph of GloBI interaction group observations.

The “Tooltip” feature of Tableau highlights geolocation and year specific information in order to aid the geospatial and temporal exploration of GloBI. When the mouse hovers over a geolocation on the map the tooltip produces a callout that contains the following information: latitude, longitude, observation year, interaction group, source kingdom, source taxon name, interaction type name, target kingdom, target taxon name, and reference citation. When the mouse hovers over a data point on the graph the tooltip callout contains year, interaction group, and the count of that interaction group for the year. Three Tableau dashboards were constructed, one for each set of visualization (Figure 6). The dashboard allows the map (top of dashboard) and the time series graph (bottom of dashboard) to be simultaneously displayed and synchronously filtered.

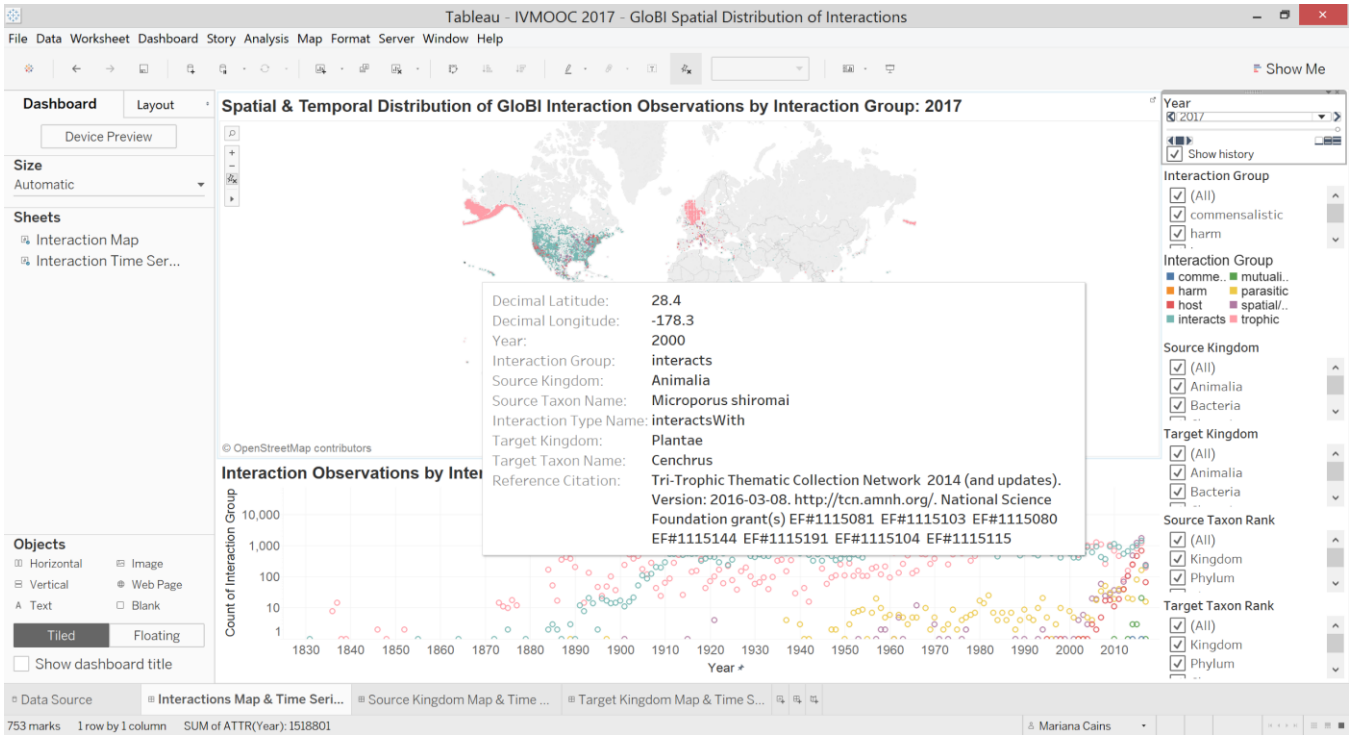

Figure 6. Tableau GUI used to construct dashboard containing geospatial and temporal distribution map and time series graph of GloBI interaction group observations.

2.3.3 Interactions Networks: Time Slicing and Bipartite Graphs by Region.

This section describes network extraction and visualization of bipartite graphs by time periods for major geographical regions in the Chordata phylum subset of species-level trophic interactions. Procedures for network extraction and time slicing are described in the Visual Insights Textbook [20], using Sci2 [27], and demonstrated by Ted Polley in IVMOOC Hands-On Exercises [29]. After preparing the chordata interaction subsets (Table 6), separate files were created for three geographical regions (e.g., North of Latitude 30 N., Latitude 0 to 29 North, South of Latitude 0). Time slices were created for each geographical region for the corresponding years of interaction occurrences (e.g., 1990-1999, 2000-2009, 2010- 2017, for the region North of Latitude 0 to 29 North).

After opening Sci2, the following steps were taken to slice the datasets by year: (1) Load the file: `Chordata_golden_lat>0<30.csv`; (2) Preprocessing > Temporal > Slice a Table by Time: Create three 10-year slices: Date/Time Column: Year; Slice into: Years: 10; From Time: beginning of 1990 / beginning of 2000 / beginning of 2010, To Time: end of 1999 / end of 2009 / end of 2017. After creating the time slices, bipartite networks were extracted for each time period: (3) Data Preparation > Extract Bipartite Network. Set parameters: Source Target. Extracted Network on Column Source. Then, the analysis toolkit was used to return descriptive information about each network: (4) Analysis > Networks > Network Analysis Toolkit (5) Visualization > Networks > Bipartite Network Graph: Subtitle: Bipartite Graph Chordata Latitude >0 <30 1990-99. Left side node type: Source No node weight, no edge weight Left column label: Source Species Right column label: Target Species Select Simplified Layout. Three bipartite graphs were constructed for each time period in the specified geographical region.

3. RESULTS

The following sections present the resulting visualizations for each of the objectives listed in section 1.3.

3.1 COMPARISON of GLOBI with GBIF VISUALIZATIONS

Due to computational capacity limitations, the GloBI vs GBIF comparative sunbursts were only constructed for species within the Archaea and Protozoa kingdoms. The Archaea and Protozoa kingdoms were selected as our “proof of concepts” since our computational power could only analyze these relatively smaller kingdoms.

Archaea Sunburst. In the Archaea sunburst analysis, the vast majority of known Archaea species are missing (i.e. greyed out in Figure 7). GloBI has interaction data for only 10 of the known 337 Archaea species in GBIF. Currently, GloBI only covers 3% of the Protozoa kingdom. The visualization also shows that GloBI does not have the complete taxon path for each of the 10 Archaea species, hence the grey layers from phylum to genus.

Protozoa Sunburst. In the Protozoa sunburst analysis, the majority of known Protozoa species are missing (i.e. greyed out in Figure 8). GloBI has interaction data for only 183 of the known 12,145 Protozoa species. Currently, GloBI only covers 6.5% of the Protozoa kingdom. The visualization also shows that GloBI does not have the complete taxon path for all of the 183 Protozoa species, hence the grey layers from phylum to genus. Only 14 Protozoa species have complete taxon paths and are colored teal from phylum to species (teal sunrays in at the 12 o’clock position in Figure 8).

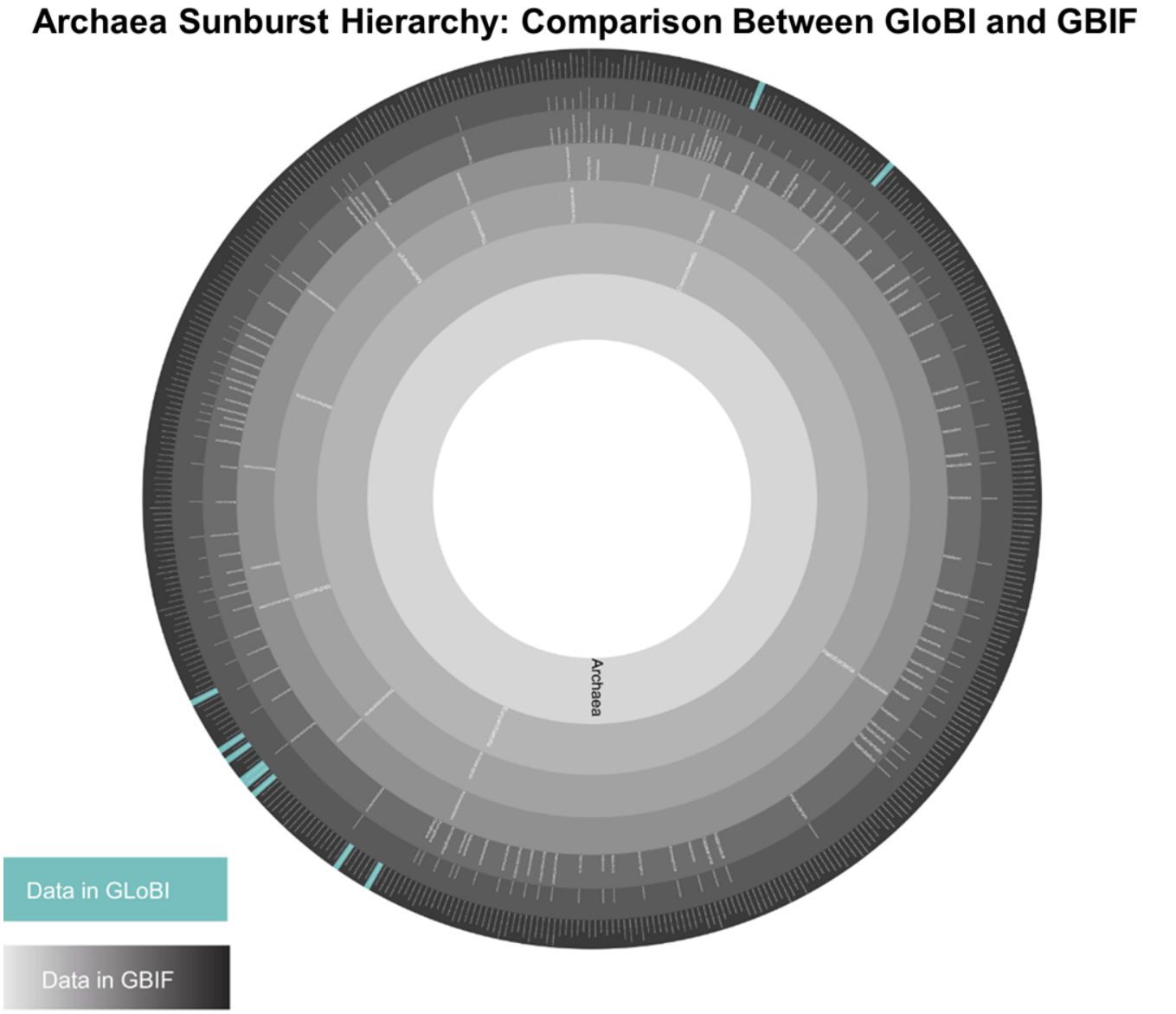

Figure 7. Archaea Sunburst Hierarchy. The grey portions of the sunburst represent taxa known to exist (as cataloged by GBIF) and are currently not cataloged in GloBI. The teal portions of the sunburst represent the taxa currently cataloged in GloBI. Each layer of the sunburst represents a taxon. The taxa are as follows from inner to outer circle: kingdom, phylum, class, order, family, genus, and species. (High resolution at https://doi.org/10.5281/zenodo.814922)

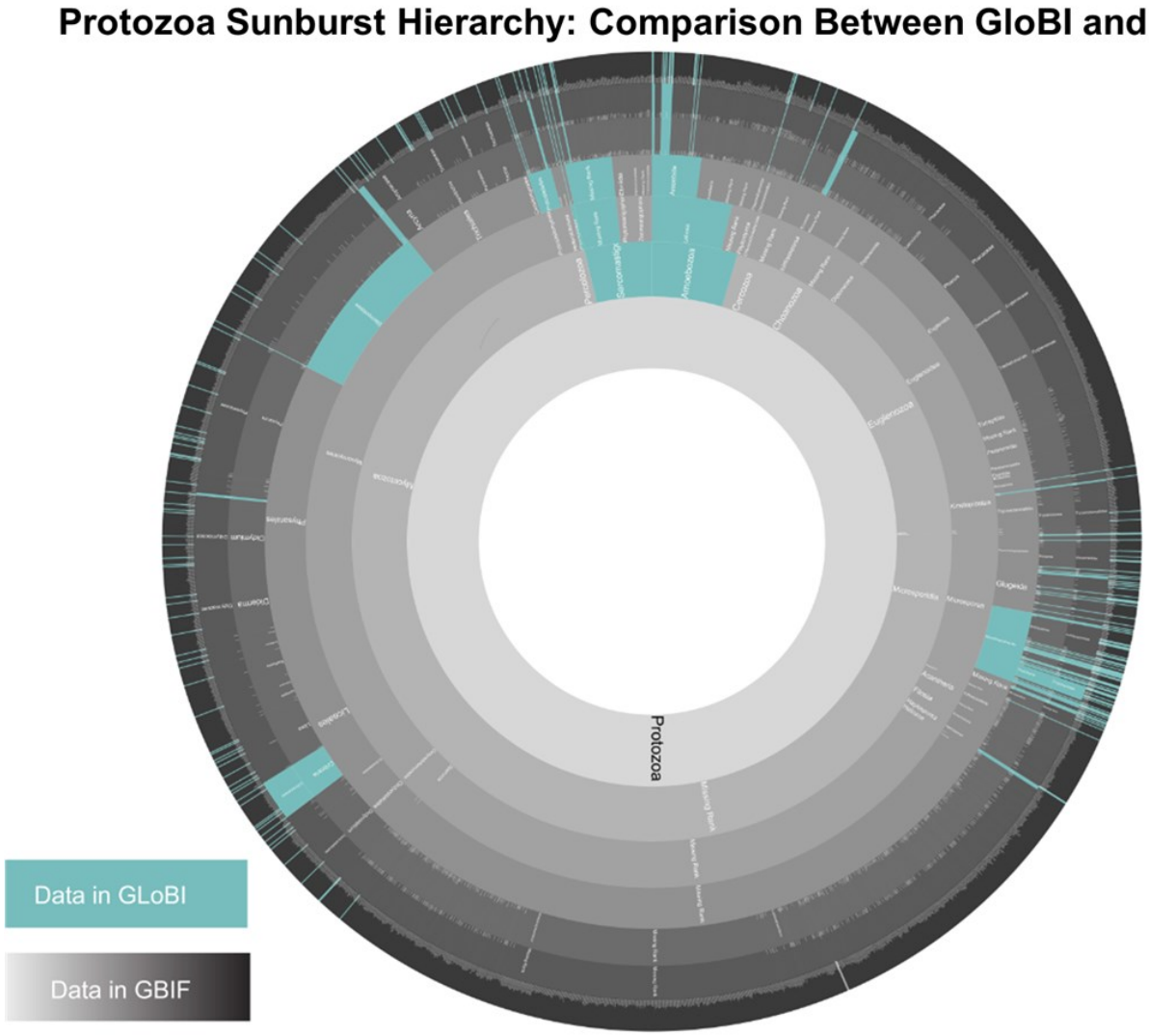

Figure 8. Protozoa Sunburst Hierarchy. The grey portions of the sunburst represent taxa known to exist (as cataloged by GBIF) and are currently not cataloged in GloBI. The teal portions of the sunburst represent the taxa currently cataloged in GloBI. Each layer of the sunburst represents a taxon. The taxa are as follows from inner to outer circle: kingdom, phylum, class, order, family, genus, and species. (High resolution at https://doi.org/10.5281/zenodo.814922)

3.2 GLOBI DATA-RICHNESS VISUALIZATIONS

3.2.1 Stacked bars

An overview of data contents is displayed in Table 7 and Figure 9. The comparison of species between GBIF and GloBI is provided for a sense on the scale of coverage, and it shows that GloBI coverage is one order of magnitude smaller than the amount of known species, that is, only about 10% of species in GBFI are recorded in GloBI. Further, we show the breakdown of GloBI dataset in different subsets and the loss of data records as the data gets more information-rich. The narrowest set would corresponds to golden dataset criteria.

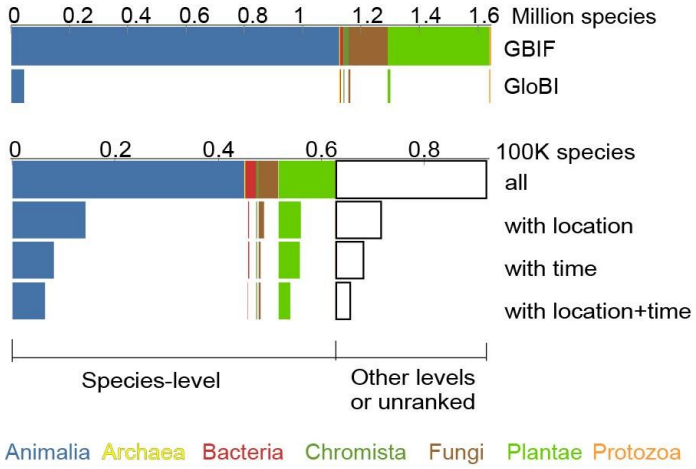

Figure 9. Upper: Comparison of total counts of different species between the database from GBFI (including all known and identified species) and data in GloBI. **Lower:** breakdown of GloBI dataset in different subsets according to the amount of information attached to the record: whether observation is at the species level, and whether it contains or not location and time stamps. Taxa data color-coded by kingdoms. All values are counts of different species (except data at other levels or unranked are counts of non-duplicated specimen types).

Table 7. Unique species in GloBI with geospatial and temporal data compared to known species in GBIF.

Unique Species	Animalia	Archaea	Bacteria	Chromista	Fungi	Plantae	Protozoa	Total
Species in GloBI	45,034	15	2,288	469	3,859	10,962	323	62,950
Geospatial	14,196	0	56	148	1,017	4,296	60	19,773
Temporal	8,016	0	3	44	337	4,172	7	12,573
Geospatial AND Temporal	6,436	0	2	22	335	2,211	5	9,011
Species in GBIF	1,125,250	377	9,982	19,785	132,848	349,812	2,708	1,640,762
% GloBI Golden Data Coverage to Total GBIF Species	0.39	0	1.2E-4	0.0013	0.020	0.13	3E-4	0.55

3.2.2 Hierarchical Bubble Graph

An overview of the data granularity in GloBI is further shown as a circular hierarchy (Figure 10). By granularity we refer to the fact that the information in the records is not homogeneous but there is a different amount of detail depending on the interaction source study. The diagram shows the amount of records that are detailed at different taxa levels, and how the distribution differs between kingdoms. All except Archaea have records described at some lower taxon.

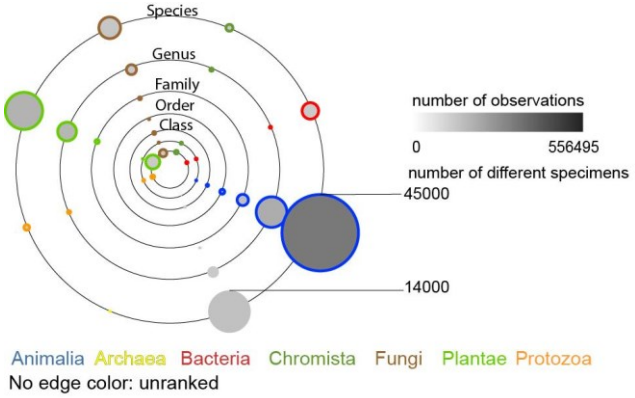

Figure 10. Breakdown of GloBI dataset according to the rank at which description is provided. Values are counts of different specimens in each taxon as represented by circle size. Internal shading of circles denote the total amount of observations. Edge is color-coded for kingdoms.

Also conspicuous is the amount of data that is not properly ranked. These are records that are not containing information on the kingdom in their Rank Paths immediately after download. To complete information on such records, additional data cleaning steps would be needed. Figure 10, therefore, provides a snapshot of the data granularity encountered by database users.

3.2.3 Geospatial and Temporal Distribution

The geospatial and temporal distribution of interaction observations by interaction group from 1813-2017 is shown in Figure 11 below. Figure 11 shows that the majority of GloBI interaction observations occur in the North America. The “interacts” and “trophic” interaction groups represent the majority of the interactions cataloged by GloBI.

The geospatial and temporal distribution of interaction observations by source kingdom from 1813-2017 is shown in Figure 12 below. Figure 12 shows that of GloBI interaction observations occur in the North America. The vast majority of interactions cataloged by GloBI have Animalia as the source kingdom.

The geospatial and temporal distribution of interaction observations by target kingdom from 1813-2017 is shown in Figure 13 below. Figure 13 shows that the majority of GloBI interaction observations occur in the North America. The vast majority of interactions cataloged by GloBI have Plantae kingdom as the target kingdom.

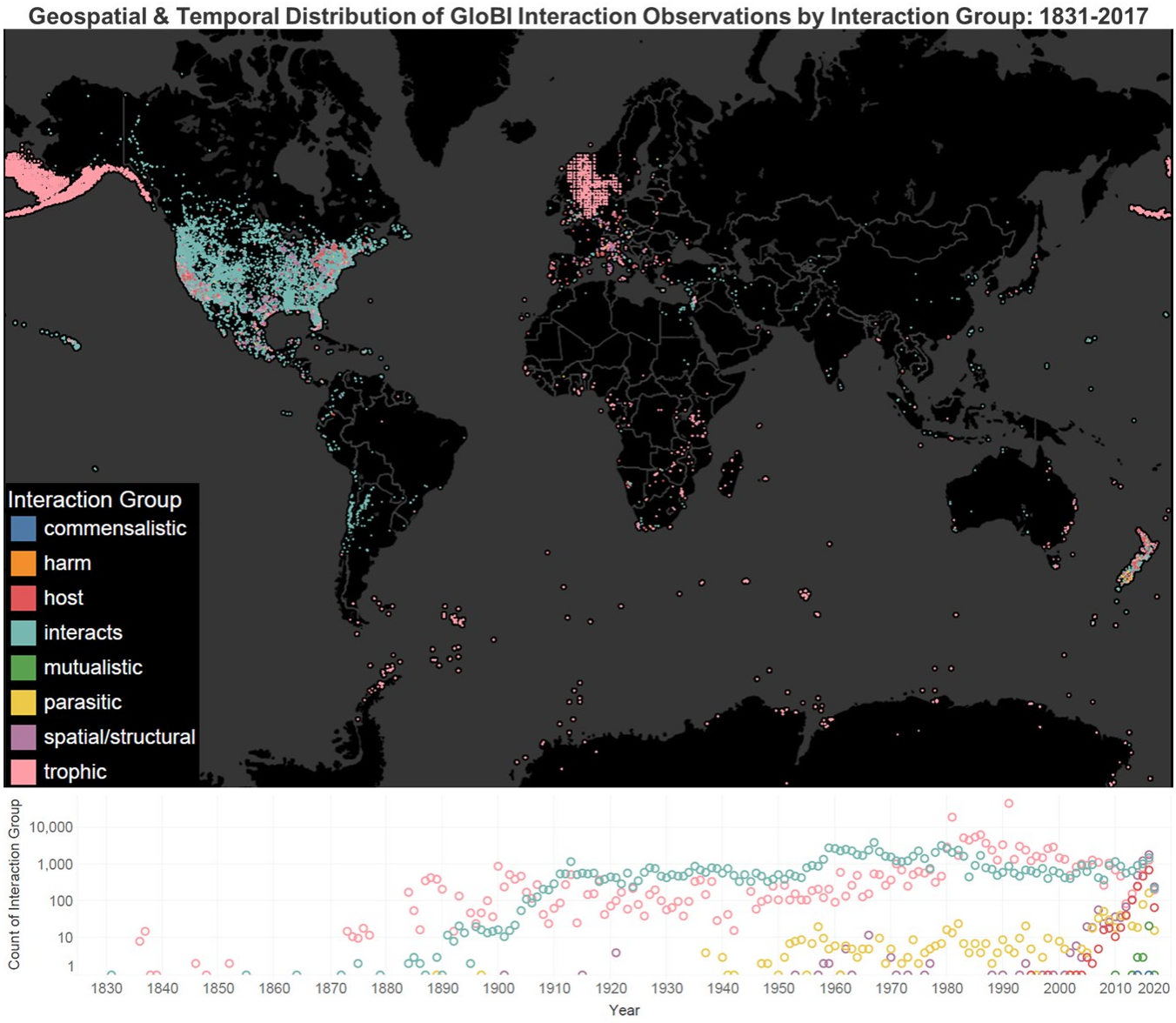

Figure 11. Geospatial and Temporal Distribution of GloBI Interaction Observations by Interaction Group: 1831-2017. Each point on the map represents an interaction observation with geospatial and temporal data. The time series tracks the observation years for each interaction observation. The colors of the solid dots (map) and hollow circles (time series) correspond to the interaction group of the data points.

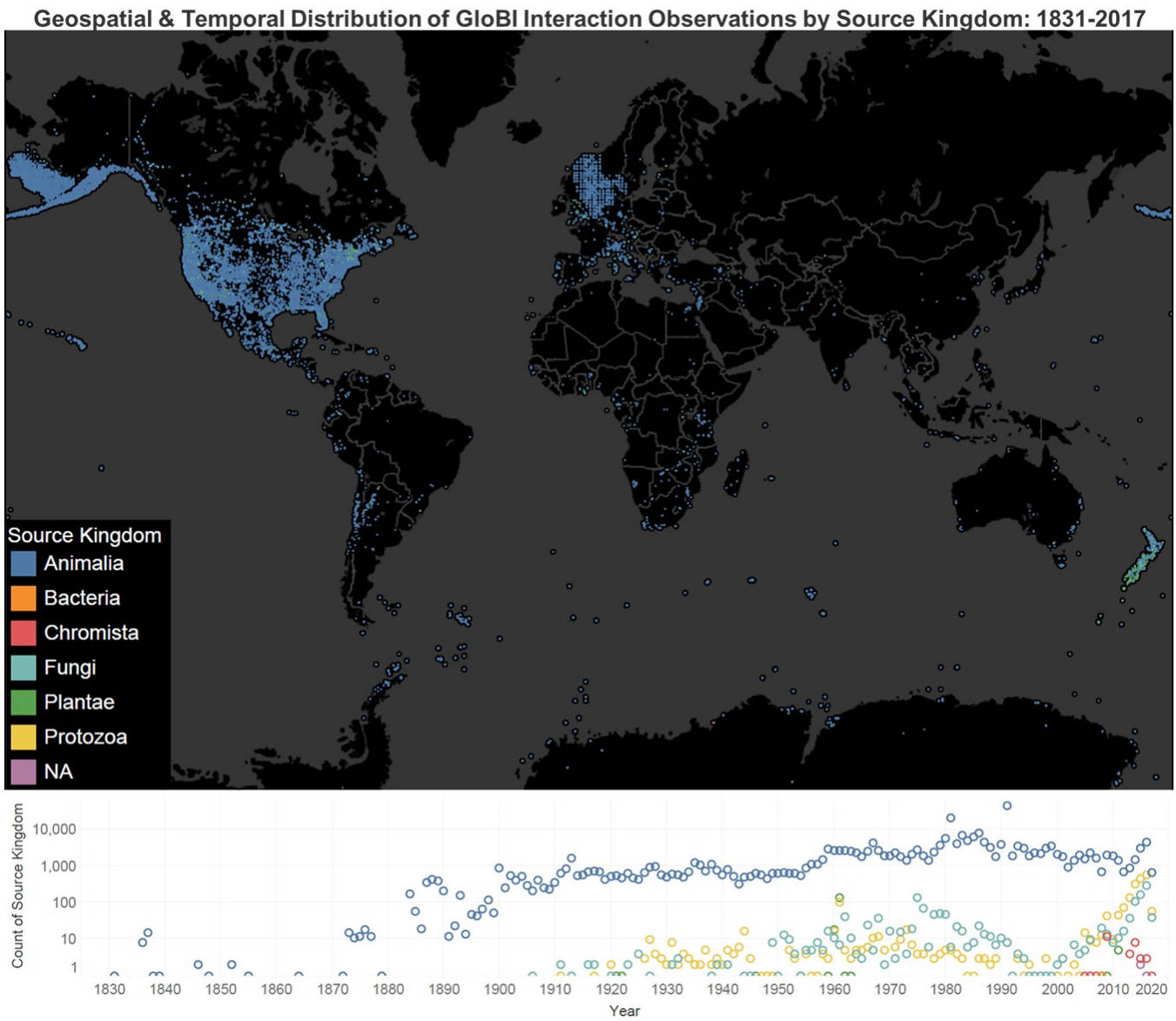

Figure 12. Geospatial and Temporal Distribution of GloBI Interaction Observations by Source Kingdom: 1831-2017. Each point on the map represents an interaction observation with geospatial and temporal data. The time series tracks the observation years for each interaction observation. The colors of the solid dots (map) and hollow circles (time series) correspond to the source kingdom of the data points.

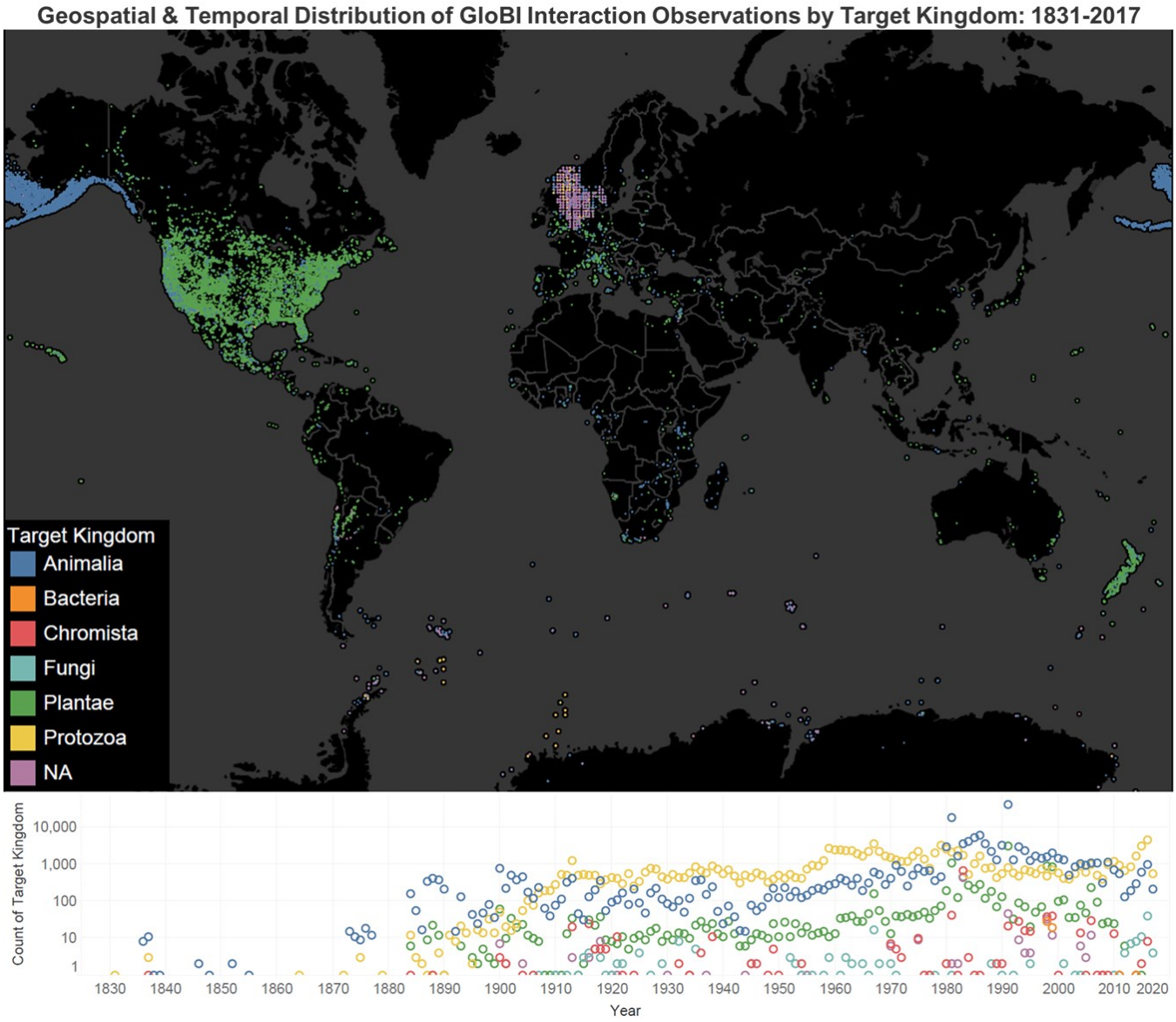

Figure 13. Geospatial and Temporal Distribution of GloBI Interaction Observations by Target Kingdom: 1831-2017. Each point on the map represents an interaction observation with geospatial and temporal data. The time series tracks the observation years for each interaction observation. The colors of the solid dots (map) and hollow circles (time series) correspond to the target kingdom of the data points.

A publically available interactive version of all three geospatial and temporal visualizations can be found at Tableau Public via:

https://public.tableau.com/profile/publish/IVMOOC2017-GloBISpatialDistributionofInteractions/InteractionsMapTimeSeries#!/publish-confirm

The respective data are available here: https://doi.org/10.5281/zenodo.814912

Data preparation details can be found in section 2.2 of this paper.

Figure 14 is a screenshot of the interactive dashboard powered by Tableau Public. Users can filter the interaction group, source kingdom, and target kingdom visualizations for different combinations the following attributes: observation year, interaction group (e.g. mutualistic, trophic), source and target kingdoms (e.g. Animalia, Protozoa), and source and target taxon rank (e.g. kingdom, species).

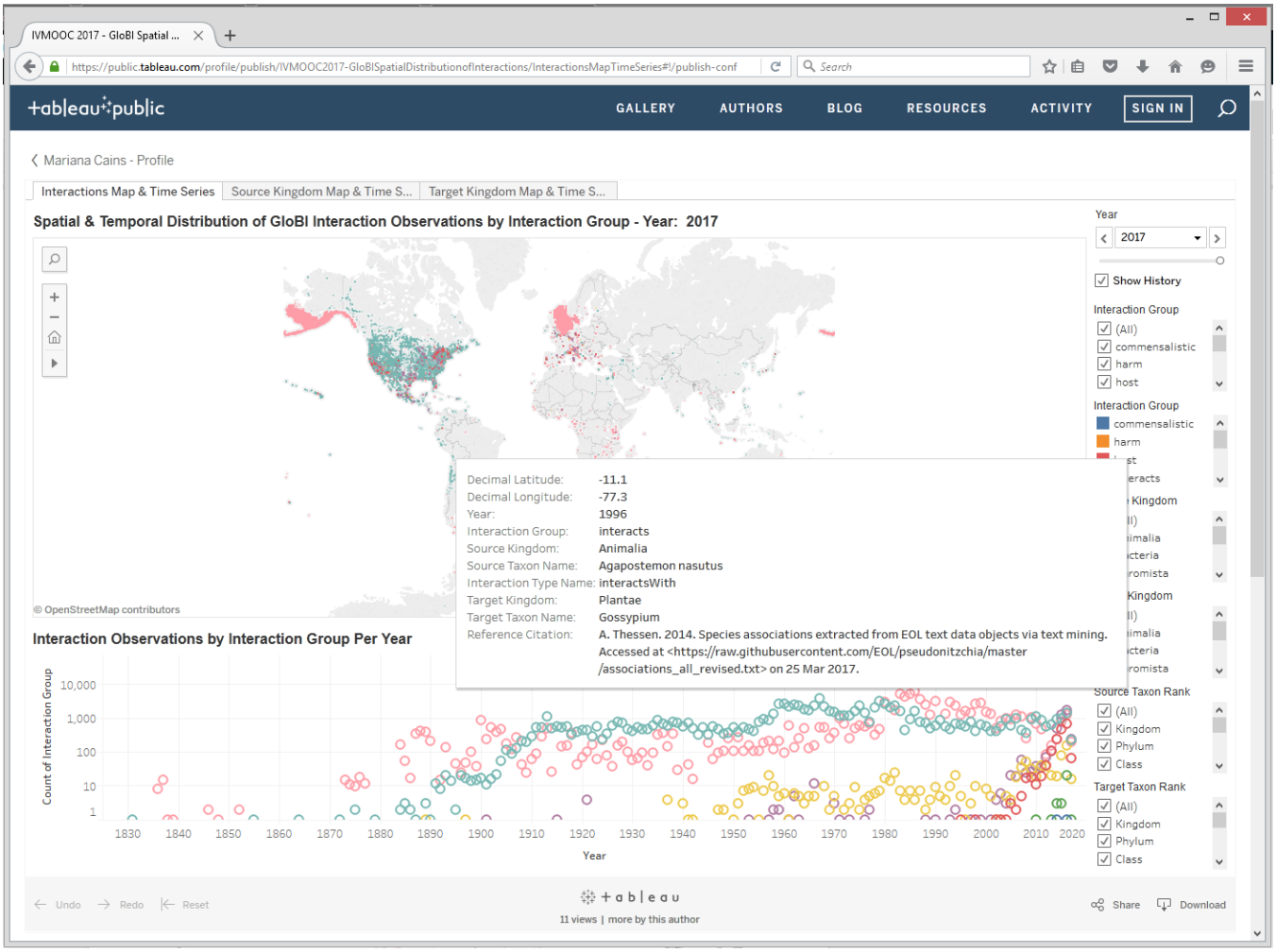

Figure 14. Tableau Public dashboard (with active tooltips) for interaction group, source kingdom, and target kingdom visualizations where users can filter on the following attributes: observation year, interaction group (e.g. mutualistic, trophic), source and target kingdoms (e.g. Animalia, Protozoa), and source and target taxon rank (e.g. kingdom, species).

3.3 INTERACTION NETWORKS VISUALIZATION

Species-level interactions for the Chordata phylum were visualized as bipartite networks. Figure 15 shows three bipartite graphs of source-target trophic Interactions for the region between Latitude 0 to 29 North (approximately corresponding to the Gulf of Mexico), from 1990 to 2017. These graphs show the food web of associations ("eats", "preys on") between specimens identified by scientific names. Table 8 shows the network characteristics for all three regions according to time periods for the species level occurrence records available for the Chordata golden data subset. Given that the number of edges in the bipartite networks for the larger geographical regions ranged from 406 to 2173, the source-target interaction graphs were increasingly dense and it was difficult to identify the source and target specimens by name in the graph.

Therefore, the bipartite network visualizations were included only for the smallest geographical region as shown in Figure 15.

4. DISCUSSION

4.1 OBJECTIVES of VISUALIZATION and INSIGHTS

This project maps out gaps in biodiversity knowledge and visualizes some of the 'unknowns' in GloBI's datasets. The results provide different levels of analysis and visualizations that help to map the nature of the gaps and potential shortcomings of aggregated data from very different sources. We identify some areas of completeness and incompleteness in GloBI. In the following sections, we summarize some of the insights gained in the course of this project and describe some challenges that could be addressed in future work.

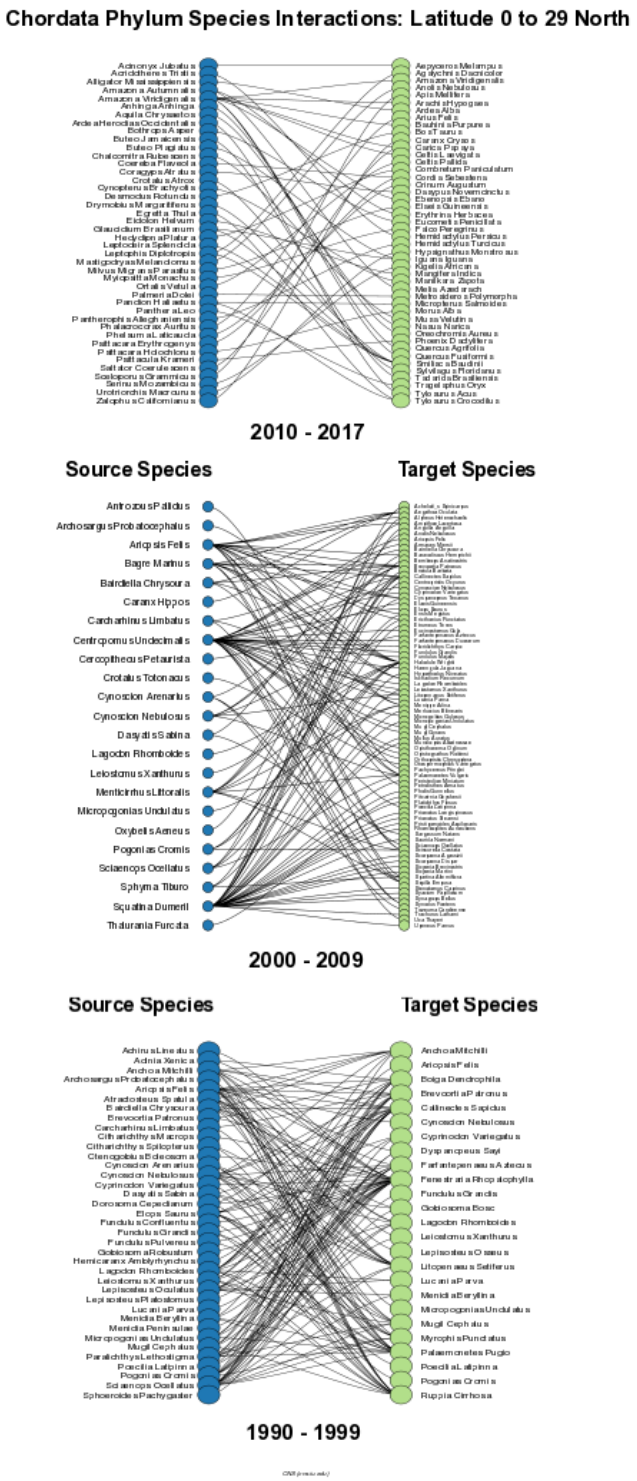

Figure 15. Bipartite Graphs of Species-Species Trophic Interactions for Chordata Phylum for the Region Between Latitude 0 to 29 North, between 1990 to 2017. (High resolution at https://doi.org/10.5281/zenodo.814922)

4.1.1 Compare GloBI with External Reference GBIF

As the sunburst visualizations show, GloBI's data coverage for Protozoa is far from complete, and captures only a small portion of the larger known taxonomy represented in GBIF. Coverage for Archaea is even smaller. Unfortunately, owing to constraints of time and processing capacity, we were not able to map out Kingdoms where GloBI's coverage is more intact. However, it is possible that, with greater processing capacity and additional time, a more complete mapping of GloBI's data coverage could have been achieved.

4.1.2 Analyze GloBI Data-richness

Stacked Bar Graphs. Table 7, which was used to construct the bar graph in Figure 9, shows that only 0.55% of the 1,638,054 unique species in GBIF have geospatial and temporal data associated with interaction observations in GloBI. So only 0.55% of known species have golden data in GloBI. This means if a researcher, or for that matter a curious person, wanted to know when and where an interaction occurred, GloBI can only answer that question for 0.55% of the species on the planet. The Animalia kingdom accounts for 71% of the golden data (0.39% of known species) and the Plantae kingdom accounts for 24% of the golden data (0.13% of the known species).

Geospatial and Temporal Distribution. The geospatial and temporal maps of interaction observations show that the majority of data in GloBI come from studies conducted in North America, specifically the United States (Figure 11, 12, 13). Studies observing interactions in Northern Europe and New Zealand are also well documented. South America, Africa, Asia, Australia, and Antarctica are sparingly peppered with interactions as compared to the United States, Northern Europe, and New Zealand. The maps illustrate the need for contribution from researchers who have and or are conducting biological interaction research in South America, Africa, Asia, Australia, and Antarctica.

The interaction group map (Figure 11) shows that the "interacts" and "trophic" interaction groups represent the majority of observed interactions in the GloBI database. The source kingdom map shows that the vast majority of interactions in GloBI have Animalia as the kingdom source. Alternatively, the target kingdom shows Plantae as the major target kingdom in the continental United States and New Zealand, while Animalia is the major target kingdom in Alaska. The juxtaposition of the source and kingdom maps provide the inside that herbivores dominate the GloBI interaction observations from the continental United States and New Zealand, while carnivores dominate the GloBI interaction observations in Alaska.

Anyone curious about the geospatial and temporal distribution of the GloBI interaction observations can explore the relationships between interaction groups, source and target kingdoms, and source and target taxon rank by visiting https://public.tableau.com/profile/publish/IVMOOC2017-GloBISpatialDistributionofInteractions/InteractionsMapTimeSeries#!/publish-confirm

Table 8. Bipartite Graphs for Geographical Region and Time Period: GloBI Chordata Phylum Trophic Interactions				
	Network Characteristics			
Region 1: Above Latitude 30 N	Nodes	Edges	Avg. Degree	Density
2010 - 2017	443	433	1.96	0.002
2000 - 2009	166	406	4.89	0.015
1990 - 1999	465	1172	5.04	0.005
1980 - 1989	279	680	4.88	0.009
Region 2: Latitude 0 to 29 N				
2010 - 2017	88	53	1.21	0.001
2000 - 2009	102	127	2.49	0.01
1990 - 1999	62	148	4.77	0.04
Region 3: Below Latitude 0				
2010 - 2017	401	727	3.63	0.005
1980 - 1999	445	2173	9.77	0.01
1960 - 1979	97	165	3.40	0.02

4.1.3 Interaction Networks

The bipartite graphs of species interactions for the Chordata Phylum show that network visualizations were interpretable for a only small geographical region, and included a very small portion of the interactions overall. As the size of the region and the numbers of source-target interactions increased, the network visualization became denser and the nature of the connections between species become more difficult to interpret. In partitioning the Chordata phylum to construct a golden dataset with complete geospatial and temporal information, a substantial portion of interactions were filtered from the dataset. This reveals a gap in our knowledge of species interactions, in the Eltonian sense.

Overall, there was a tradeoff in terms of the volume of interaction data available in the GloBI datasets and our ability to represent those interactions in visualizations that are interpretable on a human level. In terms of constructing bipartite networks, finding a balance between grain size and data inclusivity was a major challenge. A more effective way to visualize networks of species interaction, using a more complete section of the data, is perhaps to represent interaction associations as directed networks.

4.4 LIMITATIONS and CHALLENGES:

4.4.1 Conceptual Complexity

Conceptually, understanding the scope of data gaps with the GloBI database is feasible, however the execution for whole database requires both time and computational resources beyond our capabilities.

The ideal visualization would have consisted of a grey-colored tree diagram representing the GBIF reference system. Each level of the tree diagram, with increasing tree depth and organizational specificity, would represent the taxa: kingdom, phylum, class, order, family, genus, with the leaves of the tree representing the species taxon. The tree branch would be highlighted and colored if GloBI contained interaction information for that taxon.

GloBI has a large volume of data, over 2 million recorded interactions between 110,000 taxons. Our plan was to compare these taxons against the known complete set of taxons found in GBIF. GBIF has cataloged over 2 million taxons. The volume of data is very large for conventional analysis and visualization methods. To illustrate the GloBI data gaps within our time and skill constraints, we chose to compare two of the smaller kingdoms, Archaea and Protozoa, against GBIF Archaea and Protozoa kingdoms. The sizes of each kingdom were reasonable enough to perform the analysis and visualization with the time and computational resource readily available to us.

The GloBI vs GBIF comparison for Archaea and Protozoa will serves as a visualization methodology proof of concept. The sunburst methodology described in this paper can be applied to the larger kingdoms provided the availability and accessibility of time and computational resources.

4.4.2 Data Preprocessing

We expected the data analysis would require data preprocessing, but the amount of necessary data preprocessing was unexpected and set back our project schedule. Fortunately we were able to “course correct” for the needed amount of data preprocessing and are confident we have produced our desired end products, albeit modified from what we originally brainstormed. Below we detail some of the challenges faced with data preprocessing.

Inconsistent labeling. Some of the taxon labels were in languages other than English, which further emphasizes the need to have unique identifiers consistent across the reference system and the analyzed dataset.

Inconsistent information structure. In some cases, species data are accompanied with kingdom, phylum, class and order information. In other cases, species data were provided without details on its associated taxa. This inconsistency had to be addressed in preprocessing before the visualizations could be made.

Location. Some interactions provided “0,0” for the latitude and longitude of the geospatial locations while the citation source claims that interaction to have been observed at a different location. For example, “0,0” is recorded as the latitude and longitude for a study that was conducted in the North Sea, which is actually centered around a latitude and longitude of “56, 3”. Our first thought is that the “0,0” was entered to represent missing data or non-reported data. The problem is that “0,0” actually means something in latitude and longitude. Visualizing the geospatial data on a map will help highlight such issues.

Pragmatic Solutions. Performing a user experience (UX) assessment of all processes involved in data entry/submission may reduce inconsistent data entry. Inputting reminders and limiting the type of data that can be entered into submission fields can reduce chance of error and streamline the data analysis and aggregation processes.

4.4.3 Technical Challenges

Capacity Constraint. Both GloBI and GBIF have large datasets. The Neo4j database is about 780 MB in size and holds the information in approximately 9.8 million nodes and associations. The authors of this database have used Lucene indexes to speed up queries. These text indexes are over species name, the taxonomic rank, studies and locations. Large queries that explore interactions amongst specimens, along with the studies from which this data were extracted, can consume 3GB or more of memory. This is especially true when exploring reasonably sized kingdoms such as Fungi or Bacteria. We were unable to query the larger kingdoms such as Plantae or Animalia in their entirety. In addition to the memory requirements, queries against Neo4j would run for a very long time, sometimes over an hour to return interaction information.

We pre-processed the data in order to ease the process of comparison. Figure 16 below shows the process, the sources of information, and the various database tables involved. GloBI data imported into SQL Server was much easier to work with than the Neo4j database. Though it contained over 2 million interactions, with the addition of multiple indexes, the database would return information in just a few minutes. With the added indexes, the database was about 7 GB in size. As a high watermark, it would consume about 2 GB of memory.

When we were ready to compare GloBI with GBIF data, we attempted to import taxon information from GBIF. GBIF occurrence data is very large. The entire dataset is about 770 million records. As mentioned in a previous section, the Chordata phylum alone was 700 GB in size. Our personal computing space did not permit this data to be imported into the SQL Server database and then processed. This analysis only used the GBIF Backbone Taxonomy and not all of the occurrence data. It is recommended that future teams exploring the data in its entirety plan to store their data on a server with approximately 2TB of hard drive space and about 12-16 GB of memory.

4.4.4. Validation and Redesign

Rank-based Taxonomy vs Evolution-based Phylogeny Reference System. To understand the coverage GloBI has across the known taxons, we planned to compare it against a dataset that is considered complete. The original idea was to compare GloBI against the phylogenic Tree of Life. Initially we envisioned a tree (in a tree-of-life manner) with all living components where the ones existing in GloBI dataset would be highlighted. We ultimately decided to use a rank-based taxonomy (e,g. GBIF) to allow increase the potential for cross-reference. If should be noted that GloBI links to the Open Tree of Life, so the presented visualization methods could be performed for the phylogenic coverage of GloBI.

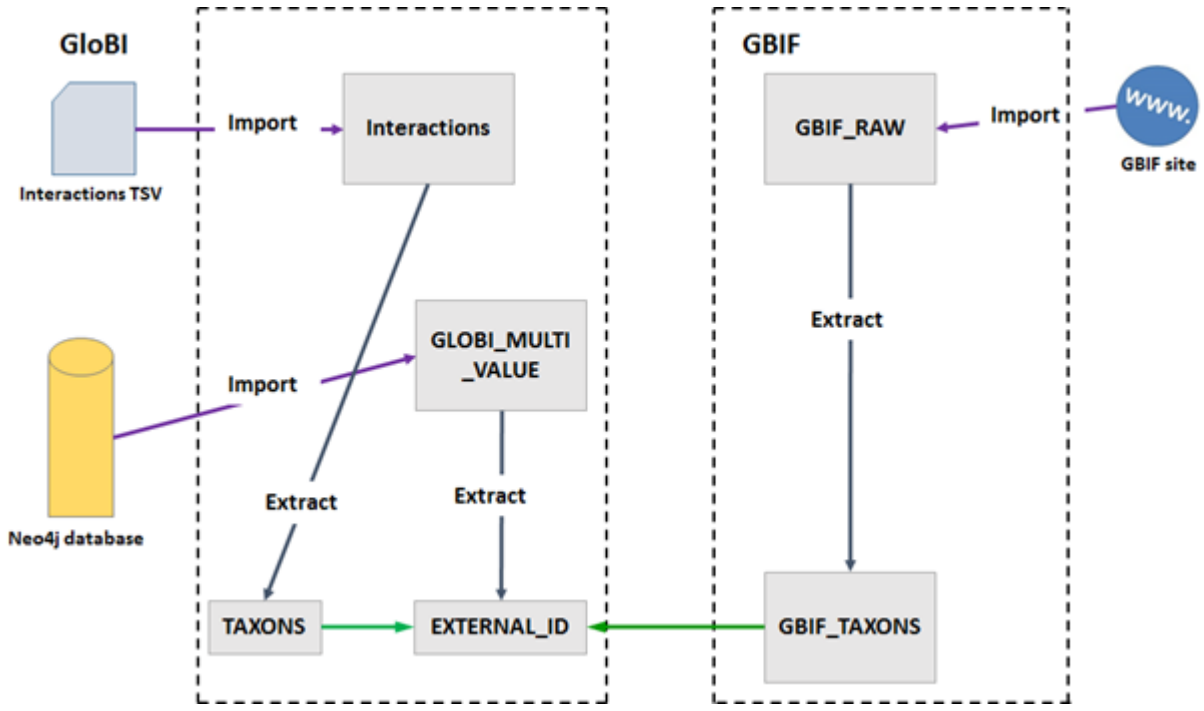

Figure 16. Data Extraction from GloBI and GBIF Databases

QGIS vs Tableau Public. Initially the geospatial and temporal maps were constructed using QGIS, an open-source geographic information system. The number of geospatial and temporal data points was too large to properly represent with the static image QGIS produced. The maps appeared clutter and it was difficult to distinguish the visual encoding. A proper visualization would allow the user to zoom into an area with the ability to distinguish between data points. Due to its interactive nature, Tableau was chosen as to construct the final geospatial and temporal maps. In addition to users being able to zoom in on data points with increased geographic resolution, the Tableau dashboards allow for the filtering of numerous attributes.

The original QGIS maps only visualized three static attributes: interaction group, location, year, and interaction group. The interactive Tableau dashboards visualize the following attributes: location, year, interaction group, interaction type, source and target kingdom, source and target taxon name, and study reference citation. Tableau allows the user to have an immersive experience exploring the geospatial and temporal coverage of GloBI.

Missing Taxon Paths in Sunbursts. In the original data format, data with missing taxon paths are NA values -- as a result the sunbursts initially appeared fragmented and disjoint. Rather than a smooth circle, the outer edge had a stair-step appearance. To address this, the missing taxon paths were replaced with the value 'Missing Rank'. For example, if the data only had kingdom and species info, the path would look like the following on the sunburst: "kingdom→ missing value → missing value → missing value→ missing value → missing value→ species".

5. CONCLUSION

The web of life on earth is a complex system of interactions. In order to understand what we do not know, inevitably, we must begin by describing what we do know. In a sense, the goal of this project was to describe the sparseness of human knowledge in relations to the vastness of life itself. The analysis and visualizations presented here demonstrate the problem of grain size or granularity of data contained in GloBI. This relates to the Eltonian 'gap in knowledge', which speaks to the lack of data or knowledge about species interactions. One realization described above is that, as our queries become more focused and refined, we lose the large scale nature of species interactions. GloBI is an ambitious project that seeks to address this gap by providing an interactive infrastructure for the extraction and collation of archived databases with the hope to motivate researchers to publish their datasets [1].

In conclusion, there are two specific areas where we think GloBI could enhance its data services and broaden its appeal as a resource for research data: First, to persuade researchers to view GloBI as a data source for research, which could save considerable time collecting data in the field. A second challenge is to convince researchers collecting data in the field as to the benefits of data sharing. To a considerable degree, research is a competitive undertaking with limited resources; however, if more scholars and researchers came to recognize the value of sharing data in the spirit of collaboration, perhaps more researchers would be inclined to contribute their data to GloBI and help flesh out some of its gaps.

ACKNOWLEDGMENTS

We would like to acknowledge our client sponsor Jorrit Poelen director of GloBI; Dr. Katy Borner, Professor of Information Science at IU Bloomington and instructor for the Information Visualization Massive Open Online Course (IVMOOC)'; Michael Ginda, IVMOOC co-instructor; and Ted Polley, former IVMOOC co-instructor.

SUPPLEMENTAL MATERIAL

This paper is digitally archived at:
https://doi.org/10.5281/zenodo.814979
Datasets and high resolution visualizations files are available at
https://zenodo.org/communities/ivmooc-2017-globi/

REFERENCES

[1] Poelen, J. H., Simons, J. D., and Mungall, C. J. (2014). Global Biotic Interactions: An open infrastructure to share and analyze species-interaction datasets. *Ecological Informatics, 24*: 148-159. http://dx.doi.org/10.1016/j.ecoinf.2014.08.005

[2] Encyclopedia of Life. Available from http://www.eol.org. Last accessed 15 Apr 2017.

[3] Simons, J.D., Yuan, M., Carollo, C., Vega-Cendejas, M., Shirley, T., Palomares, M.L., Roopnarine, P., Arenas, L.G.A., nez, A.I., Holmes, J., Schoonard, C.M., Hertog, R., Reed, D., Poelen, J., (2013). Building a fisheries trophic interaction database for management and modeling research in the Gulf of Mexico large marine ecosystem. Bull. Mar. Sci. 89 (1), 135–160. http://dx.doi.org/10.5343/bms.2011.1130 (link to the article) https://www.researchgate.net/profile/Peter_Roopnarine/publication/236132068…/links/00b7d52263ba7b629b000000.pdf

[4] Linnaeus, Carl (1735). Systema naturae per regna tria naturae :secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis (in Latin) (1st ed.). Stockholm: Laurentius Salvius.

[5] Linnaeus, Carl (1758). Systema naturae per regna tria naturae :secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis (in Latin) (10th ed.). Stockholm: Laurentius Salvius.

[6] Ruggiero, Michael A.; Gordon, Dennis P.; Orrell, Thomas M.; Bailly, Nicolas; Bourgoin, Thierry; Brusca, Richard C.; Cavalier-Smith, Thomas; Guiry, Michael D.; Kirk, Paul M.; Thuesen, Erik V. (2015). "A higher level classification of all living organisms". PLOS ONE. 10 (4): e0119248. doi:10.1371/journal.pone.0119248. PMC 4418965 Freely accessible. PMID 25923521.

[7] Hortal J, F de Bello, J Al F. Diniz-Filho, T M. Lewinsohn, J M. Lobo and R J. Ladle. 2015 Seven shortfalls that beset large-scale knowledge on biodiversity *Annu. Rev. Ecol. Evol. Syst.* 2015. 46:523–49. 3 doi:10.1146/annurev-ecolsys-112414-054400

[8] Morales-Castilla I, M G. Matias, D Gravel, and MB. Araujo. 2015 "Inferring biotic interactions from proxies" TREE dx.doi.org/10.1016/j.tree.2015.03.014

[9] Meyer C, H Kreft, R Guralnick & W Jetz 2015 2Global priorities for an effective information basis of biodiversity distributions" Nature Comms DOI: 10.1038/ncomms9221

[10] GBIF (2015). Final Report of the Task Group on GBIF Data Fitness for Use in Distribution Modelling Gaps. *Global Biodiversity Information Facility*. http://www.gbif.org/resource/82612 (last accessed March 20, 2017)

[11] Thessen, A. (2014). *pseudonitzchia: biodiversity informatics code relevant to the encyclopedia of life*.

https://github.com/eol/pseudonitzchia (last accessed March 18, 2017).

[12] Sachs, J., Parr, C., Parafiynyk, A., Pan, R., Han, L., Ding, L., Finin, T., Hollander, A., Wang, T., (2006). Using the semantic web to support ecoinformatics. Proceedings of the AAAI

[13] Storey,M. (2014). *Bioinfo: food webs and species interactions in the biodiversity of UK and Ireland*. http://bioinfo.org.uk (last accessed March 18, 2017)

[14] International Council for the Exploration of the Sea (ICES, 1989). Data base report of the stomach sampling project 1981. N. Daan (Ed.). Cooperative research report no. 164 http://www.ices.dk/sites/pub/Publication%20Reports/Cooperative%20Research%20Report%20(CRR)/crr164/CRR164.pdf (last accessed March 18, 2017)

[15] International Council for the Exploration of the Sea (ICES, 1996). Data base report of the stomach sampling project 1981. Cooperative research report no. 219 http://www.ices.dk/sites/pub/Publication%20Reports/Cooperative%20Research%20Report%20(CRR)/crr219/CRR219.pdf (last accessed March 18, 2017)

[16] Raymond, B., Marshall, M., Nevitt, G., Gillies, C.L., van den Hoff, J., Stark, J.S., Losekoot, M., Woehler, E.J., and Constable, A.J. (2011). A Southern ocean dietary database. *Ecology 92 (5),* 1188. http://dx.doi.org/10.1890/10-1907.1 http://esapubs.org/archive/ecol/E092/097/

[17] Baron, D., Caragol, R., Furrer, S., MacMurchy, P. and Stark, A. (2015). GloBI Explorer: Interactive Ecosystem Explorer. IVMOOC Course Project Final Paper. (last accessed March 18, 2017) https://figshare.com/articles/GloBI_Explorer_Interactive_Ecosystem_Explorer/1414253/

[18] Slyusarev, S., Kontopoulos, D.-G., Taysom, W., Guzman, A., and Wadhwa., B. (2014): *Global Biotic Interactions food web map*. http://dx.doi.org/10.6084/m9.figshare.1297762 (last accessed March 19, 2017)

[19] Global Biodiversity Information Facility. Available from http://www.gbif.org// Last accessed 24 Apr 2017

[20] Borner, K., and Polley, D. E. (2014). Visual insights: A practical guide to making sense of data. The MIT Press. Cambridge, MA.

[21] Poelen (2017) Accessing Species Interaction Data. https://github.com/jhpoelen/eol-globi-data/wiki#accessing-species-interaction-data (last accessed April 23, 2017)

[22] Roskov Y., Abucay L., Orrell T., Nicolson D., Bailly N., Kirk P.M., Bourgoin T., DeWalt R.E., Decock W., De Wever A., Nieukerken E. van, Zarucchi J., Penev L., eds. (2017). Species 2000 & ITIS Catalogue of Life, 2017 Annual Checklist. Digital resource at www.catalogueoflife.org/annual-checklist/2017. Species 2000: Naturalis, Leiden, the Netherlands. ISSN 2405-884X.

[23] R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[24] Xiang Z, Mungall C, Ruttenberg A, He Y. Ontobee: A Linked Data Server and Browser for Ontology Terms. Proceedings of the 2nd International Conference on Biomedical Ontologies (ICBO), 2011, USA. URL: http://ceur-ws.org/Vol-833/paper48.pdf.

[25] CEFAS, Center for Environment, Fisheries, and Aquaculture Science. Fish Stomach Records. https://www.cefas.co.uk/cefas-data-hub/fish-stomach-records/ (last accessed April 19, 2017).

[26] Pinnegar, J. K. (2014). DAPSTOM. An Integrated Database and Portal for Fish Stomach Records. https://www.cefas.co.uk/media/41463/dapstom-phase-4-report-2014-dlm.pdf (last accessed April 19, 2017).

[27] Sci2 Team. (2009). Science of Science (Sci2) Tool. Indiana University and SciTech Strategies, https://sci2.cns.iu.edu.

[28] Tableau. Tableau Desktop. http://www.tableausoftware.com, (accessed April 10, 2017)

[29] IVMOOC. (2013, Jan 10) Bipartite Networks: Mapping CTSA Centers: https://www.youtube.com/watch?v=Dp9aJJrC33Q (last accessed April 24, 2017)

APPENDICES

APPENDIX A: Upgrading the Neo4j database

The Neo4j database available from the GloBI website has been configured to work with Neo4j version 1.9.9. However, the latest version of the database is 3.1.2 (as of March 2017) and there are significant differences in the architecture between the two versions.

1 Navigate to the database and click Open
2 Select Options…
3 Click the Edit button for Database Configuration
4 Search for the configuration entry *dbms.allow_format_migration* and set it to true
5 Save and close the configuration file
6 Close the configuration window
7 Start the database

On occasion, the allow_format_migration option may not be sufficient. If the database fails to start, follow these steps

1 Download Neo4j 2.3.10 from the Other Releases page and install it in a different location than version 3.1.3
2 Start Neo4j v 2.3.10
3 Follow steps identified above to open the GloBI database
4 Start the database
5 When the database starts successfully, stop the database. This will migrate the database to a newer architecture
6 Shut down Neo4j v 2.3.10
7 Now open the GloBI database in Neo4j v 3.1.3

These steps are required just one time. Once the migration process is complete, you can open the database with Neo4j v 3.1.3 or higher.

APPENDIX B: Database scripts

Function Wordparser

The function takes a string and a delimiter as input and returns a table of parsed words

```
CREATE FUNCTION [dbo].[Wordparser]
(
  @multiwordstring VARCHAR(MAX),
  @delimiter      CHAR(1) = ','
)
returns @parsedwords TABLE
      (
         line NUMERIC IDENTITY(1, 1),
         word VARCHAR(MAX)
      )
AS
  BEGIN
      DECLARE @remainingstring VARCHAR(MAX)
      DECLARE @numberofwords NUMERIC
      DECLARE @word VARCHAR(MAX)

      SET @remainingstring=@multiwordstring

      SET @numberofwords=(LEN(@remainingstring) - LEN(REPLACE(@remainingstring, @delimiter, '')) + 1)


      WHILE @numberofwords > 1
        BEGIN
            SET @word=LEFT(@remainingstring, CHARINDEX(@delimiter, @remainingstring) - 1)

            IF(LEN(@word)>0)
         BEGIN
            INSERT INTO @parsedwords(word)
            SELECT @word
         END

         SET @remainingstring = RIGHT(@remainingstring, LEN(@remainingString)-LEN(@word)-1)

            SET @numberofwords=(LEN(@remainingstring) - LEN(REPLACE(@remainingstring, @delimiter, '')) +1)

            IF @numberofwords = 1
              BREAK

            ELSE
```

```
            CONTINUE
        END

      IF @numberofwords = 1
        SELECT @word = @remainingstring
      INSERT INTO @parsedwords(word)
      SELECT @word

      RETURN
  END
```

Extract Taxon path from GloBI

```
INSERT INTO GloBI_TAXON_PATH (GloBI_ID, TAXON_RANK, TAXON_NAME, SEQ)
SELECT t.TaxonID AS GloBI_ID, LTRIM(RTRIM(resultPRN.word)) AS TAXON_RANK, LTRIM(RTRIM(resultPN.word)) AS
TAXON_NAME, resultPRN.line AS SEQ
FROM dbo.TAXONS as t
CROSS APPLY dbo.Wordparser(t.TaxonPathName, '|') as resultPN
CROSS APPLY dbo.Wordparser(t.TaxonPathRankName, '|') as resultPRN
WHERE ISNULL(resultPN.word, '') <> ''
AND ISNULL(resultPRN.word, '') <> ''
AND resultPN.line = resultPRN.line;
```

Extract ExternalIDs from GloBI

```
DECLARE @tbl TABLE
(
   GloBI_ID    VARCHAR(256),
   line        BIGINT,
   WORD        VARCHAR(256)
);

DECLARE @tbl2 TABLE
(
   GloBI_ID          VARCHAR(256),
   [EXTERNAL_SYSTEM]    VARCHAR(64),
   WORD              VARCHAR(256),
   External_ID          VARCHAR(64)
);

INSERT INTO @tbl
SELECT stuff.externalId as GloBI_ID, Results.line, RTRIM(LTRIM(Results.word)) as [ExternalID]
FROM [Import].[GLOBI_MULTI_VALUE_FIELDS] as Stuff
CROSS APPLY dbo.WordParser(Stuff.externalIDs, '|') as Results
WHERE ISNULL(Results.word, '') <> ''
ORDER BY 1;

INSERT INTO @tbl2
SELECT t.GloBI_ID, 'GBIF' AS [EXTERNAL_SYSTEM], word, RIGHT(ltrim(rtrim(word)), LEN(ltrim(rtrim(word)))-5)
as External_ID
from
(
   SELECT GloBI_ID, MAX(line) AS MAX_LINE
   FROM @tbl
   WHERE word like '%GBIF:%'
   GROUP BY GloBI_ID
) as ln
   ,@tbl as t
WHERE ln.GloBI_ID = t.GloBI_ID AND ln.MAX_LINE = t.line;

INSERT INTO dbo.EXTERNAL_IDS (GloBI_ID, [EXTERNAL_SYSTEM], External_ID)
SELECT LTRIM(RTRIM(GloBI_ID)), [EXTERNAL_SYSTEM], LTRIM(RTRIM(External_ID))
FROM @tbl2;
```

Extracting interactions from SQL Server

```
SELECT DISTINCT sourceTaxonName, sourceTaxonRank, sourceTaxonPathNames,
sourceTaxonPathRankNames
   , interactionTypeName
   , targetTaxonName, targetTaxonRank, targetTaxonPathNames, targetTaxonPathRankNames
   , decimalLongitude as Longitude, decimalLatitude as Latitude
   , eventDateUnixEpoch
FROM dbo.interactions
WHERE sourceTaxonPathNames LIKE '%Plantae%'
ORDER BY sourceTaxonPathNames, sourceTaxonName, targetTaxonPathNames, targetTaxonName;
```

Extracting interactions from Neo4j

Use this script if you know the specific species, as it uses the index *node:taxons* to speed up the query

```
START sourceTaxon = node:taxons(name="Picea pungens")
MATCH (sourceTaxon)<-[x:CLASSIFIED_AS]-(sourceSpecimen)-
[r:INTERACTS_WITH|:ATE|:ECTOPARASITE_OF|:ECTOPARASITOID_OF|:ENDOPARASITE_OF|:ENDOPARASITOID_OF|:EPIPHITE_O
F|:FARMED_BY|:FARMS|:HAS_ECTOPARASITE|:HAS_ECTOPARASITOID|:HAS_ENDOPARASITE|:HAS_ENDOPARASITOID|:HAS_EPIPH
ITE|:HAS_HOST|:HAS_HYPERPARASITE|:HAS_HYPERPARASITOID|:HAS_KLEPTOPARASITE|:HAS_PARASITE|:HAS_PARASITOID|:H
AS_PATHOGEN|:HAS_VECTOR|:HOST_OF|:HYPERPARASITE_OF|:HYPERPARASITOID_OF|:KLEPTOPARASITE_OF|:PARASITE_OF|:PA
RASITOID_OF|:PATHOGEN_OF|:POLLINATED_BY|:POLLINATES|:PREYS_UPON|:SYMBIONT_OF|:VECTOR_OF]-
>(targetSpecimen)-[:CLASSIFIED_AS]->(targetTaxon)
OPTIONAL MATCH (sourceSpecimen)-[:COLLECTED_AT]->(location)
OPTIONAL MATCH ()-[ti:COLLECTED]-(sourceSpecimen)
WHERE EXISTS(targetTaxon.externalId)
AND EXISTS(sourceTaxon.externalId)
RETURN DISTINCT sourceTaxon.externalId as sourceTaxonId
            , sourceTaxon.name as sourceTaxonName
            , sourceTaxon.rank as sourceTaxonRank
            , sourceTaxon.path as sourceTaxonPathNames
            , sourceTaxon.pathNames as sourceTaxonPathRankNames
            , sourceTaxon.pathIds as sourceTaxonPathIds
                    , type(r) as interactionTypeName
                    , id(r) as interactionTypeId
            , targetTaxon.externalId as targetTaxonId
            , targetTaxon.name as targetTaxonName
            , targetTaxon.rank as targetTaxonRank
            , targetTaxon.path as targetTaxonPathNames
            , targetTaxon.pathNames as targetTaxonPathRankNames
            , targetTaxon.pathIds as targetTaxonPathIds
                    , location.latitude as decimalLatitude, location.longitude as decimalLongitude,
location.locality as locality
                    , ti.dateInUnixEpoch as eventDateUnixEpoch
                    , sourceTaxon.citation as referenceCitation
ORDER BY targetTaxon.name
```

Use this query, if you want all species within a Kingdom or a high level taxonomic rank. It uses the *node:taxonPaths* Lucene index.

```
START sourceTaxon = node:taxonPaths('path:Animalia')
MATCH (sourceTaxon)<-[x:CLASSIFIED_AS]-(sourceSpecimen)-
[r:INTERACTS_WITH|:ATE|:ECTOPARASITE_OF|:ECTOPARASITOID_OF|:ENDOPARASITE_OF|:ENDOPARASITOID_OF|:EPIPHITE_O
F|:FARMED_BY|:FARMS|:HAS_ECTOPARASITE|:HAS_ECTOPARASITOID|:HAS_ENDOPARASITE|:HAS_ENDOPARASITOID|:HAS_EPIPH
ITE|:HAS_HOST|:HAS_HYPERPARASITE|:HAS_HYPERPARASITOID|:HAS_KLEPTOPARASITE|:HAS_PARASITE|:HAS_PARASITOID|:H
AS_PATHOGEN|:HAS_VECTOR|:HOST_OF|:HYPERPARASITE_OF|:HYPERPARASITOID_OF|:KLEPTOPARASITE_OF|:PARASITE_OF|:PA
RASITOID_OF|:PATHOGEN_OF|:POLLINATED_BY|:POLLINATES|:PREYS_UPON|:SYMBIONT_OF|:VECTOR_OF]-
>(targetSpecimen)-[:CLASSIFIED_AS]->(targetTaxon)
OPTIONAL MATCH (sourceSpecimen)-[:COLLECTED_AT]->(location)
OPTIONAL MATCH ()-[ti:COLLECTED]-(sourceSpecimen)
WHERE EXISTS(targetTaxon.externalId)
AND EXISTS(sourceTaxon.externalId)
RETURN DISTINCT sourceTaxon.externalId as sourceTaxonId
            , sourceTaxon.name as sourceTaxonName
            , sourceTaxon.rank as sourceTaxonRank
            , sourceTaxon.path as sourceTaxonPathNames
            , sourceTaxon.pathNames as sourceTaxonPathRankNames
            , sourceTaxon.pathIds as sourceTaxonPathIds
             , type(r) as interactionTypeName
              , id(r) as interactionTypeId
            , targetTaxon.externalId as targetTaxonId
            , targetTaxon.name as targetTaxonName
            , targetTaxon.rank as targetTaxonRank
            , targetTaxon.path as targetTaxonPathNames
            , targetTaxon.pathNames as targetTaxonPathRankNames
            , targetTaxon.pathIds as targetTaxonPathIds
               , location.latitude as decimalLatitude
               , location.longitude as decimalLongitude
               , location.locality as locality
               , ti.dateInUnixEpoch as eventDateUnixEpoch
               , sourceTaxon.citation as referenceCitation
ORDER BY targetTaxon.name
```

Extract Golden Data Master from Neo4j

```
START sourceTaxon = node:taxonPaths('path:Archaea')
MATCH (sourceTaxon)<-[x:CLASSIFIED_AS]-(sourceSpecimen)<-[:COLLECTED]-(study)
OPTIONAL MATCH (sourceSpecimen)-[:COLLECTED_AT]->(location)
OPTIONAL MATCH ()-[ti:COLLECTED]-(sourceSpecimen)
WHERE EXISTS(sourceTaxon.externalId)
RETURN distinct sourceTaxon.externalId
```

```
              , sourceTaxon.rank
              , sourceTaxon.path
              , sourceTaxon.pathNames
              , sourceTaxon.name
              ,location.type as location
              , ti.dateInUnixEpoch
              , study.citation
```

Joining GBIF and GloBI datasets

```
SELECT GBIF.*, GloBI.*
FROM
(
   SELECT DISTINCT species, taxonkey
   FROM [dbo].[GBIF_TAXON]
   WHERE species is NOT NULL
) as GBIF
LEFT JOIN
(
   SELECT t.TaxonID as GloBI_ID
              , t.TaxonName as GloBI_NAME
              , ei.ExternalID as ExternalStringID
              , SUBSTRING(ei.ExternalID, 6, LEN(ei.ExternalID)-5) AS ExternalID
   FROM dbo.[EXTERNAL_IDS] ei
      ,dbo.[TAXONS] t
   WHERE t.TaxonID = ei.GloBI_ID
   AND ei.ExternalID LIKE 'GBIF%'
) AS GloBI ON GBIF.TaxonKey = GloBI.ExternalID
WHERE GloBI.GloBI_NAME IS NOT NULL;
```

APPENDIX C: D3 Referenced Scripts (bl.ocks)

```
<!DOCTYPE html>
<head>
    <script src="https://d3js.org/d3.v4.min.js"></script>
</head>
<body>
    <svg></svg>
</body>

<script>
    // JSON data
    var nodeData = (Full JSON data for Protozoa and Archaea available in project
folder)


   // Variables
    var width = 500;
    var height = 500;
    var radius = Math.min(width, height) / 2;
    var color = d3.scaleOrdinal(d3.schemeCategory20b);

    // Create primary <g> element
    var g = d3.select('svg')
        .attr('width', width)
        .attr('height', height)
        .append('g')
        .attr('transform', 'translate(' + width / 2 + ',' + height / 2 + ')');

    // Data strucure
    var partition = d3.partition()
        .size([2 * Math.PI, radius]);
```

```
    // Find data root
    var root = d3.hierarchy(nodeData)
        .sum(function (d) { return d.size});

    // Size arcs
    partition(root);
    var arc = d3.arc()
        .startAngle(function (d) { return d.x0 })
        .endAngle(function (d) { return d.x1 })
        .innerRadius(function (d) { return d.y0 })
        .outerRadius(function (d) { return d.y1 });

    // Put it all together
    g.selectAll('path')
        .data(root.descendants())
        .enter().append('path')
        .attr("display", function (d) { return d.depth ? null : "none"; })
        .attr("d", arc)
        .style('stroke', '#fff')
        .style("fill", function (d) { return color((d.children ? d :
d.parent).data.name); });
</script>
```