Global Biotic Interactions: An open infrastructure to share and analyze species-interaction datasets

Jorrit H. Poelen[a,*], James D. Simons[b], Chris J. Mungall[c]

[a] 400 Perkins Street, Apt. 104 Oakland, CA 94610, USA
[b] Center for Coastal Studies Natural Resource Center, Ste. 3200 6300 Ocean Drive, Unit 5866 Corpus Christi, TX 78412-5866, USA
[c] Genomics Division Lawrence Berkeley National Laboratory 1 Cyclotron Road, MS 64-121 Berkeley, CA 94720, USA

Abstract

An intricate network of interactions between organisms and their environment form the ecosystems that sustain life on earth. With a detailed understanding of these interactions, ecologists and biologists can make better informed predictions about the ways different environmental factors will impact ecosystems. Despite the abundance of research data on biotic and abiotic interactions, no comprehensive and easily accessible data collection is available that spans taxonomic, geospatial, and temporal domains. Biotic-interaction datasets are effectively siloed, inhibiting cross-dataset comparisons. In order to pool resources and bring to light individual datasets, specialized research tools are needed to aggregate, normalize, and integrate existing datasets with standard taxonomies, ontologies, vocabularies, and structured data repositories. Global Biotic Interactions (GloBI) provides such tools by way of an open, community-driven infrastructure designed to lower the barrier for researchers to perform ecological systems analysis and modeling. GloBI provides a tool that (a) ingests, normalizes, and aggregates datasets, (b) integrates interoperable data with accepted ontologies (e.g., OBO Relations Ontology, Uberon, and Environment Ontology), vocabularies (e.g., Coastal and Marine Ecological Classification Standard), and taxonomies (e.g., Integrated Taxonomic Information System and National Center for Biotechnology Information Taxonomy Database), (c) makes data accessible through an application programming interface (API) and various data archives (Darwin Core, Turtle, and Neo4j), and (d) houses a data collection of about 700,000 species interactions across about 50,000 taxa, covering over 1,100 references from 19 data sources. GloBI has taken an open-source and open-data approach in order to make integrated species-interaction data maximally accessible and to encourage users to provide feedback, contribute data, and improve data access methods. The GloBI collection of datasets is currently used in the Encyclopedia of Life (EOL) and Gulf of Mexico Species Interactions

[*] Corresponding author
Email address: jhpoelen@xs4all.nl (Jorrit H. Poelen)

(GoMexSI).

Keywords: species interactions data integration taxonomy ontology

1. Introduction

Though relationships between organisms and their environment have been studied for hundreds of years, answering a question as simple as “What do sharks eat near California?” still requires quite some research, even for an experienced marine biologist. If we enter this query into a mainstream search engine, we get back lists of web pages with general information about white sharks (*Carcharodon carcharias*) and leopard sharks (*Triakis semifasciata*) and articles about how to avoid sharks while surfing and why sharks attack humans. The search result closest to providing an answer is a Yahoo! Answers page that addresses the question “What do great white sharks eat?” in free-form text without references to data sources. This results page shows that the search engine lacks the ability to answer a question that requires the knowledge of the interactions between species in a specific environment. What we expect in the search results is one or more reference to a web resource that contains a comprehensive list of shark diets off the coast California. By using the system and methods described in this paper, such web resources can be developed.

We believe that the reasons for the absence of a comprehensive, machine-readable, spatiotemporal species-interaction data collection are (a) the lack of integrated information systems specifically built for capturing and sharing structured species-interaction data, and (b) insufficient incentives for scientists to make their datasets available. In this paper, we discuss a method and system addressing both these obstacles to an open repository of species-interaction data. We describe Global Biotic Interactions (GloBI), an extensible, open-source infrastructure that was tailored for importing, searching, and exporting species-interaction data. The GloBI infrastructure implements an automated workflow in which existing datasets are transformed, integrated, and aggregated into a normalized data collection. GloBI also incentivizes data sharing by providing a framework for increasing the visibility of a contributing researcher; each entry is attributed to a scientist, research institution, or other source. The inclusion of attributions in GloBI has the multiple benefit of encouraging connections among researchers, assigning credit, and creating accountability. Also, an argument can be made that data collection efforts are facilitated by repurposing existing datasets. With access to a large species-interaction data collection, a researcher might decide that no extra data collection is necessary to test a hypothesis. Alternatively, with a clearer assessment of gaps in existing data collections, researchers might decide to target taxa or geographical locations that have not yet been studied.

2. Methods

2.1. GloBI framework

We created an integrated system for the acquisition, normalization, management, and querying of biotic-interaction data called Global Biotic Interactions (GloBI). The system is implemented in Java [1] and uses Neo4j [2] as a persistent data store and query system. The systems architecture consists of (a) a data model capable of representing diverse types of interaction data, (b) an ingestion framework for the acquisition and normalization of data, and a collection of parsers for different data formats, (c) a term matcher to assign vocabulary identifiers to free-form text descriptions, and (d) an application programming interface (API) and web interface.

2.2. Data model

For the basis of the GloBI framework, we designed a data model (Figure 1) to capture species interactions and their associated spatiotemporal information. In our model, an interaction observation is figured as a specimen (or occurrence) that interacts with another specimen, using interaction terms from the OBO Relations Ontology [3]. Each specimen can be related to (or classified as) a specific biotic or abiotic term like a taxon of appropriate rank (e.g., *Homo sapiens*, Elasmobranchii), functional group (e.g., algae, plankton), or environment (e.g., rocks, sediment). In addition, when the information is available, the location at which the interaction was observed is described by its latitude, longitude, altitude and depth properties. To make grouping of locations more meaningful, we made an association between a location and its ecoregion (e.g. Northern Gulf of Mexico), habitat, or environment when possible. Terms used to describe ecoregion, habitat, and environment are taken from published ecoregion classifications [4, 5, 6, 7], existing ontologies such as EnvO [8], Uberon [9], the OBO Relations Ontology (RO) [3], and habitat classification vocabularies, such as Coastal and Marine Ecological Classification Standard (CMECS) [10].

To enable granular citation of interaction data, each specimen is associated with a study, and each study is related to a source or contributor. The study represents a reference to the origins of the data, and the source is a reference to the entity that shared the data in electronic form. Some sources share data related to a single study [11], while others sources have collected data from multiple studies [12, 13].

2.3. Data acquisition

Individual interaction datasets were acquired through web resources (e.g., data journals, web APIs) or received by email after directly contacting data managers or authors. Our only data requirement was that it should be in digital form. Data contributors were encouraged to submit their interaction data in the original file format to preserve as much information as possible. When necessary, we implemented parsers to map these datasets to the GloBI data model.

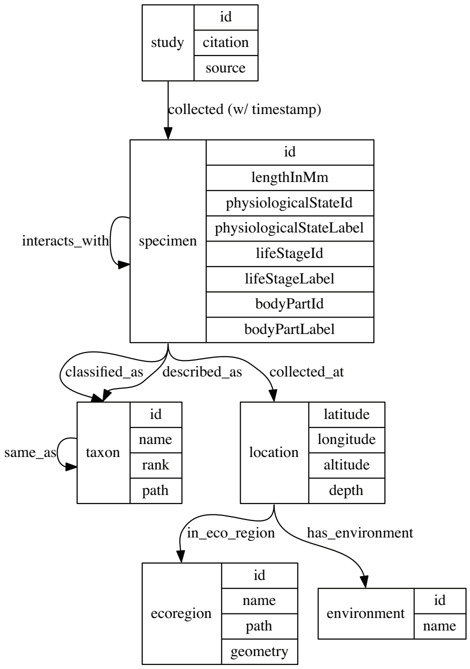


Figure 1: Interaction data is modeled in terms of study, specimen, taxon, and location concepts. The location has an additional relation to ecoregions and environments to facilitate spatial searches. Most IDs are uniform resource identifiers (URIs) to external ontologies and/or vocabularies. If neither ontologies or vocabularies are available, a custom GloBI term is used until a suitable (external) ID is found. Note that only a single interaction type is displayed in the figure, where many interaction types exist (e.g., predator-prey, host-parasite).

2.4. Software and data management

We take advantage of free tools provided by GitHub to share, document, and discuss datasets and associated data processing software (see `https://github.com/jhpoelen/eol-globi-data`). We established a GloBI GitHub wiki to describe data processing and access methods, and created a Git repository to archive original interaction data in case the data has not yet been archived or made available elsewhere. We use GitHubs issue tracker to keep track of promising interaction datasets, discuss new features, or report issues with existing datasets.

2.5. Term matching

In an effort to detect spelling errors and ambiguous or invalid names, all terms used in the interaction data are checked against existing taxonomies, ontologies, and/or vocabularies. Terms that do not match are published in web-accessible tabular comma-separated values (CSV) files. Domain experts use these files to review troublesome names and request corrections or explanations from authors. If an author is unable to correct the name in the source data, GloBI curators can correct a name without changing the original data by adding the original name, the corrected name, and the reason for the correction to a taxon correction list. Because the original data is preserved, the corrections can be undone or updated as needed.

Terms that are mapped include common and scientific names of organisms, life stages, body parts, environments (or habitats), and citations. Term mapping is performed while ingesting the interaction datasets and is part of the normalization process (Figure 2). Taxonomic names are first matched against EOL's API [14]. Then, an additional name service, GlobalNames (`http://resolver.globalnames.org`), is used to cross-reference the name against WoRMS, ITIS, NCBI, and Global Biodiversity Information Facility (GBIF) taxonomies. For life stages and body parts terms, the Uberon metazoan anatomy ontology is used. The Environmental Ontology (EnvO) is used to map environment or habitat terms, while citations are resolved to their digital object identifiers (DOIs) using CrossRef (`http://crossref.org`) wherever possible. Whenever free-form text locales are used in datasets (e.g., "locality:Kansas,US"), they are mapped onto GeoNames (`http://geonames.org`) terms, and location information is extracted in the form of latitude and longitude pairs. These pairs are associated with relevant terms from Marine Ecoregions of the World (MEOW, [4]), Freshwater Ecoregions of the World (FEOW, [6]), Terrestrial Ecoregions of the World (FEOW, [6]), and Longhurst's Biogeographical Provinces [7].

In addition to normalizing data-source text with exact term matching, we also implemented a fuzzy or partial term-matching method based on the Levenshtein Distance algorithm (as provided by Apache Lucene, `http://lucene.apache.org`). This fuzzy search was implemented so users can interactively discover normalized terms in the data collection even when their text strings contain typos. The matching algorithm matches not only exact taxonomic name matches (e.g., *Homo sapiens*), but also slightly invalid or incomplete names (e.g., Homo zapiens and Homo sap, respectively). The algorithm also incorporates higher taxonomic ranks and common names to give the user many ways to find a desired taxon.

2.6. Continuous integration

Integrating interaction data into a single, consistent aggregated dataset involves many steps: data is parsed, mapped, checked, aggregated, and exported. Custom software was developed (see `https://github.com/jhpoelen/eol-globi-data`) to automate all these steps as well as regular quality checks and tests. An automated test suite is executed each time a change is made to the software (see `http://travis-ci.org/jhpoelen/eol-globi-data`), and another automated process rebuilds the aggregated datasets from the original data sources on a daily basis. This continuous integration helps to keep the interaction data up-to-date. The services that the rebuild process relies on (e.g., GlobalNames, EOL API, CrossRef) are used daily to ensure that technical integration and availability issues are caught within a matter of days. As the volume of interaction data grows, it is expected that this process will be optimized as needed to ensure scalability. One of such optimizations is to split up the process into intermediate reproduceable steps to allow for distributed or parallel data processing.

2.7. Data access

The output of the automated dataset normalization and aggregation process can be accessed through a hosted web API. The full datasets can also be published in three file formats: Darwin Core [15], Turtle [16], and Neo4j [2] database archives. The API enables users to build interactive web applications without having to install custom software. The archives allow for bulk data processing by way of custom or existing software, without the limitations of web APIs.

For Darwin Core, we provide different archives: one with all interaction observations and another with distinct interaction observations aggregated by study. The row types included in the export are Occurrence, Taxon, Reference, MeasurementOrFact, and Association. Occurrence tables include information specific to the observed occurrences of taxa. Each occurrence is related to a specific taxon and reference. The Taxon table includes information about the taxonomic classification of observed organisms. The Reference table contains bibliographic references to source studies that recorded the associations between classified occurrences. The MeasurementOrFact table allows users to annotate a recorded occurrence with additional information that isn't shown in the Occurrence table. The Association table captures how occurrences interact with each other. MeasurementOrFact and Association tables are custom extensions created by EOL.

2.8. Competency query results

To integrate GloBI data into web applications, statistical environments, and other interactive computer applications, two methods of data extraction were

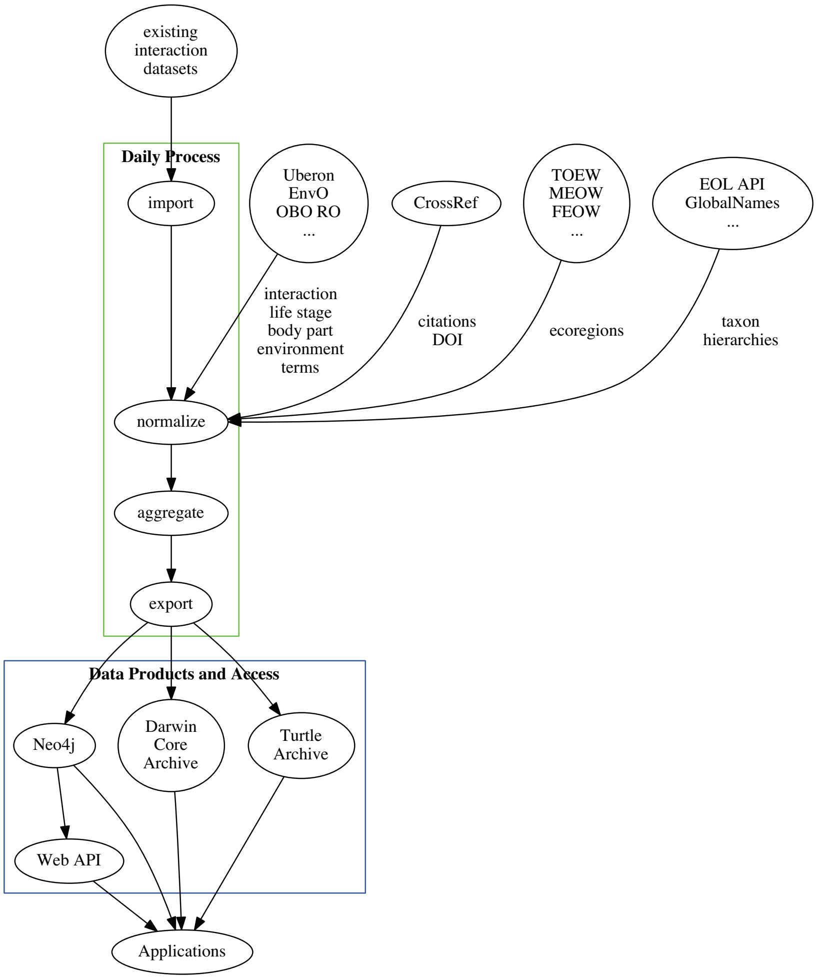


Figure 2: GloBI's data aggregation process: First, existing datasets are imported; then data elements are normalized, using existing taxonomies, ontologies, and vocabularies. Next, the data is added to the data collection. This aggregated data collection forms the basis for data-access methods, such as file exports and web APIs provided by GloBI's web API and Neo4j's Cypher API.

put in place: the GloBI API, which gives web developers a way to integrate interaction data with commonly used web-programming languages such as PHP and JavaScript, and the Cypher query interface, a web-accessible interface that requires knowledge of Cyphers graph query language. The Cypher method was added for specialized uses of GloBI data that are not covered by the GloBI API.

2.8.1. GloBI API

GloBIs web API data extraction method was built with Java programming language to implement a Spring MVC-based service that runs in Jetty, an open-source servlet container. This service receives requests from web clients, executes queries on the graph database, and returns the results in the requested format. For an example, visit the following URL to see GloBIs answer to the question, "What mammals do sharks eat?":`http://api.globalbioticinteractions.org/taxon/Elasmobranchii/preysOn/Mammalia`.

Query results can be requested in JavaScript Object Notation (JSON), CSV, or DOT [17] formats for integration into web pages, spreadsheets, R [18], or visualization tools like Graphviz [17], Cytoscape [19], or Gephi [20]. GloBIs web API also features a bounding box that can be manipulated to limit search results to a specific geographical area, and offers various convenience methods, such as taxon name suggestions, a list of supported interaction types, and dataset statistics (see `https://github.com/jhpoelen/eol-globi-data/wiki/api`).

To make it easier to use this API in web pages or the R statistical environment, a JavaScript library (`https://www.npmjs.org/package/globi-data`) and an R package (`https://github.com/ropensci/rglobi`) are available.

2.8.2. Cypher queries

Neo4j's Cypher query language can also be used to answer questions and create summaries of GloBI species-interaction data. The queries can be executed using Neo4j technologies such as a Java [1] API, a web service, and the web-based data browser. These tools are openly accessible by way of a server hosting an instance of the Neo4j graph database.

The Cypher query language is specially designed for traversing complex directed-graph data in an intuitive way. The Cypher query in Listing 1 queries GloBI to answer the question: "What mammals do sharks eat?" The query defines two starting points: the predator and prey taxon. The predator taxon selects any taxa classified as sharks, skates, and rays (Elasmobranchii) from the taxonPaths index. The prey taxon includes any taxa part of the mammal (Mammalia) family that are present in the taxonPaths index. The match clause specifies how these two taxa should be related. In our case, the predator taxon should be related to a predator specimen that ate or preyed on a prey specimen. This prey specimen should then be classified as the prey taxon. The return clause specifies that only the names of the respective taxon nodes should be returned.

Listing 1: Cypher query to find mammalian (Mammalia) prey of sharks, rays, and skates (Elasmobranchii).

```
START predatortaxon=node:taxonPaths('path:Elasmobranchii'),
    preytaxon=node:taxonPaths('path:Mammalia')
MATCH predatortaxon<-[:CLASSIFIED_AS]-specimen-[:ATE|PREYS_ON]->prey,
    prey-[:CLASSIFIED_AS]->preytaxon
RETURN predatortaxon.name, preytaxon.name
```

So, the start clause selects the nodes that are included as the starting point in the graph traversal. The match clause describes the pattern in which the graph should be searched. The return clause specifies what properties of the matching nodes should be returned.

In Listing 2 an additional pattern is added to answer the question: “What do sharks eat in California?” In our case, California is interpreted as an ecoregion that includes the term “California.”

Listing 2: Cypher query to find prey of Elasmobranchii in ecoregions that include term “California.”

```
START taxon=node:taxonPaths('path:Elasmobranchii'),
    region=node:ecoregionPaths('path:California')
MATCH taxon<-[:CLASSIFIED_AS]-specimen-[:ATE|PREYS_ON]->prey,
    prey-[:CLASSIFIED_AS]->preytaxon,
    specimen-[:COLLECTED_AT]->location-[:IN_ECOREGION]->region
RETURN taxon.name, preytaxon.name
```

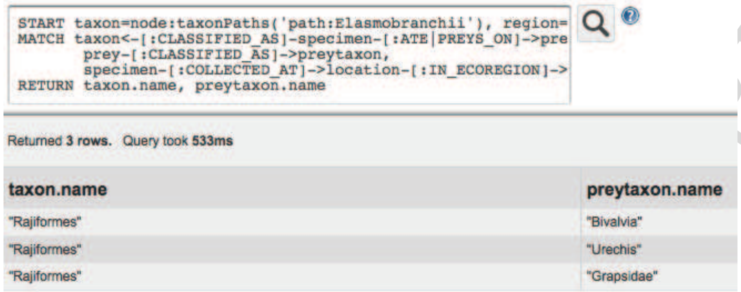
The search indexes used in this example (e.g., taxonPaths, ecoregionPaths) are populated during the GloBI data import process. The ecoregion search index uses the latitude and longitude information provided by the location node in combination with shapefiles from Longhurst’s Biogeographical Provinces [7], Marine Ecoregions of the World (MEOW) [4], Terrestrial Ecoregions of the World (TEOW) [5], and Freshwater Ecoregions of the World (FEOW) [6]. Each location is associated, or indexed, with the ecoregion that contains it, making it possible to constrain searches to ecologically relevant geographical areas.

An alternate approach to selecting a geographical area is to use a WHERE clause with latitude and longitude constraints (see Listing 3). For example, the following listing answers the question, “What do sharks eat above latitude 30?”

Listing 3: Cypher query to find prey of Elasmobranchii found above latitude 30.

```
START taxon=node:taxonPaths('path:Elasmobranchii')
MATCH taxon<-[:CLASSIFIED_AS]-specimen-[:ATE|PREYS_ON]->prey,
    prey-[:CLASSIFIED_AS]->preytaxon, specimen-[:COLLECTED_AT]->location
WHERE has(location.latitude) AND location.latitude > 30.0
RETURN taxon.name, preytaxon.name
```

The three queries above use a fraction of the functionality that the Cypher query language provides, but already answer complex ecological questions. The



```
START taxon=node:taxonPaths('path:Elasmobranchii'), region=
MATCH taxon<-[:CLASSIFIED_AS]-specimen-[:ATE|PREYS_ON]->pre
       prey-[:CLASSIFIED_AS]->preytaxon,
       specimen-[:COLLECTED_AT]->location-[:IN_ECOREGION]->
RETURN taxon.name, preytaxon.name
```

Returned **3 rows.** Query took **533ms**

taxon.name	preytaxon.name
"Rajiformes"	"Bivalvia"
"Rajiformes"	"Urechis"
"Rajiformes"	"Grapsidae"

Figure 3: Screenshot of Cypher query result using Listing 2 using Neo4j's Data Browser tool with GloBI data (see `http://tinyurl.com/globiBrowser`).

queries can be executed against GloBI data by copying the query text into the web tool available at `http://api.globalbioticinteractions.org:7474/webadmin/#/data/search/`. A screenshot of a query result of Listing 2 using this discovery tool is shown in Figure 3.

3. Results

3.1. Data sources

Table 1: Breakdown of GloBI data by source. O_{total} and $O_{distinct}$ describe the total and distinct number of interaction observations between a begin taxon (B) and an end taxon (E). $B_{distinct}$ and $E_{distinct}$ are the number of unique taxa that occur within the set of begin and end taxa. L_{total} and T_{total} are the total number of location and time data elements across all observations O. Period indicates the time range in which interaction data was recorded. Data source statistics were retrieved from GloBI on July 16, 2014.

data source	refs	O_{total}	$O_{distinct}$	$B_{distinct}$	$E_{distinct}$	L_{total}	T_{total}	period
Kelp Forest Food Web [21]	130	1,915	1,611	84	210	0	0	n/a
Species Interactions of Australia [22]	1	14,896	13,657	4,461	4,075	0	0	n/a
Species Interactions in UK and Ireland [23]	1	50,157	37,612	13,599	5,429	0	0	n/a
Semantic Prototypes in Research Informatics [12]	195	30,101	21,191	3,335	2,616	27,152	0	n/a
Southern Ocean Diet Database [13]	324	26,462	10,368	341	898	26,391	25,858	1961-2011
Cymothoa excisa parasitizes *Micropogonias undulatus* [11]	1	335	1	1	1	335	335	2010-2012
`http://inaturalist.org`	1	2,015	1,718	824	799	1,980	2,010	1989-2014
Food Webs of Three California Estuaries [24]	1	13,966	7,129	234	247	13,966	0	n/a
Food Webs of Three Coral Reef Systems [25]	1	96,647	33,257	659	730	1,939	0	n/a
Predator and Prey Body Sizes in Marine Food Webs [26]	24	34,931	529	93	167	34,931	0	n/a
Polytraits: Biological Traits of Polychaetes [27]	146	793	544	150	64	0	0	n/a
Gulf of Mexico Species Interactions [28]	53	34,902	4,810	180	1,063	9,691	8,058	1998-2007
Papilionoidea of the World [29]	1	24,436	9,915	2,497	3,543	0	0	n/a
Avian Diet Database [30]	68	1,658	1,365	63	625	0	510	1897-2004
EOL Text Mining [31]	1	183,872	161,871	21,516	30,957	0	0	n/a
Tropical Plant Herbivore Networks [32]	1	74	52	19	18	19	0	n/a
ICES North Sea Fish Stomach Data [33, 34]	1	183,935	2,628	35	825	44,280	44,280	1979-1991
Body Sizes of Consumers and Their Resources [35]	20	16,865	11,233	951	985	16,181	0	n/a
Who Eats Whom in the Barents Sea [36]	215	1,578	1,371	159	194	1,578	0	n/a
Total	1,173	719,536	302,803	34,993	41,725	437,983	232,117	1897-2014

Table 2: Breakdown of taxa with known interactions provided by datasets included in GloBI. The numbers represent the count of distinct taxa given the higher-order taxon (e.g., Chordata) and taxonomic rank (e.g., any rank, species, genus). The percentage represents the ratio of GloBI taxa count (accessed on July 16, 2014) and ITIS taxa count (accessed on March 3, 2014) for given higher-order taxonomic groups.

	Any rank	% ITIS	Species	Genus	Family
Arthropoda	17,201	5.5%	13,428	2,867	417
Fungi	9,378	199.5%	7,762	725	12
Plantae	9,382	6.5%	6,959	1,940	160
Chordata	6,737	5.9%	5,473	740	249
Mollusca	1,027	5.6%	736	195	65
Annelida	485	7.8%	301	128	47
Bacteria	109	19.0%	76	28	2
Other	5,749	13.0%	4,379	1,047	63
Total	50,068	7.8%	39,114	7,670	1,015

Shared datasets were aggregated from various sources (Table 1). None of the data sources shared a common data format, and custom importers had to be developed to fit the data into our interaction model (Figure 1). While most datasets contained scientific taxon names, only a single source, `http://inaturalist.org`, contained explicit references to a taxonomy reference. References to locations varied between using decimal degrees and using free-form text locale information (e.g., Country: USA). Only GoMexSI used a controlled vocabulary to describe habitat and ecoregions (CMECS, MEOW). No digital object identifiers (DOIs) for references within the data sources were present, but 704 DOIs out of 1,175 references were obtained from the reference text using CrossRef, as of July 16, 2014.

3.2. Data coverage

Spatial, temporal, and taxonomic coverage of the combined datasets (Figure 4, Tables 1, 2) show 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 present. This coverage is larger than any other single open aggregated species-interaction data collection that is available today simply because it aggregates many of the currently available, large, open datasets (Table 1). However, the sampling density across space, time, and taxonomic ranks is highly variable: where datasets provided by ICES [33, 34], GoMexSI [28], and Southern Ocean [13] provide most of the spatiotemporal interaction data, datasets such as Thessen 2014 [31], SPIRE [12], and BioInfo [23] contribute most of the total taxonomic coverage.

Taxonomic coverage (Table 2) varies by multiple orders of magnitude between datasets. A study of a single parasite, *Cymothoa excisa*, and its host, the Atlantic croaker (*Micropogonias undulatus*) [11], offers the smallest taxonomic coverage with a spatiotemporal resolution. The largest taxonomic coverage is

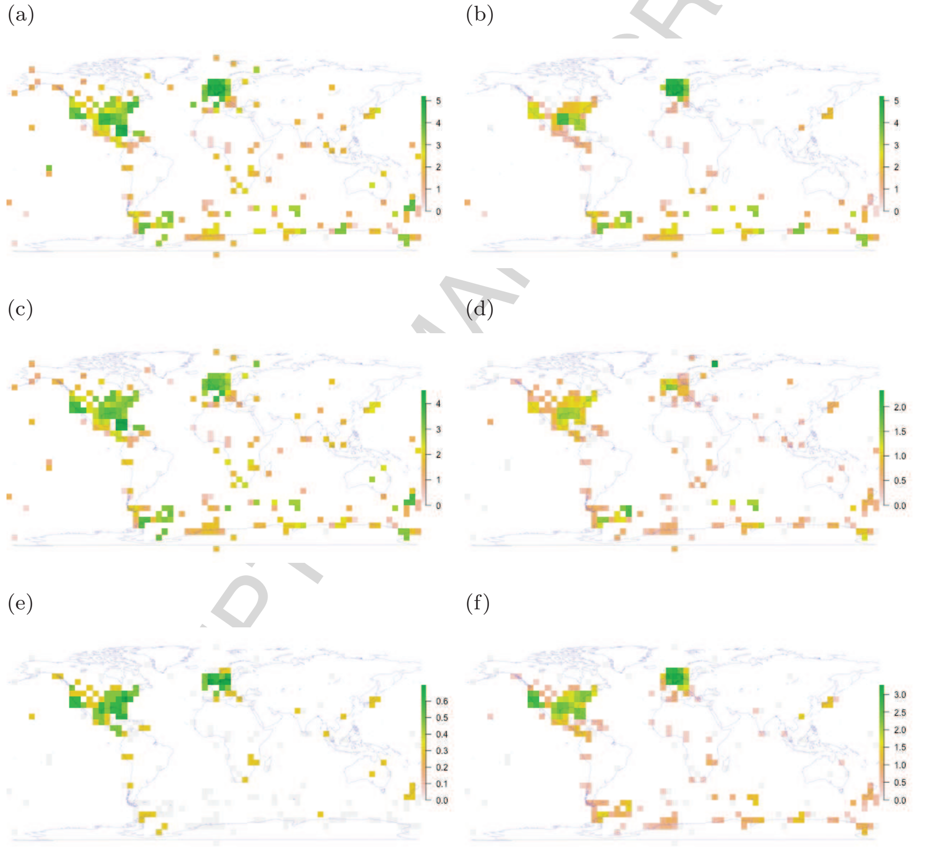


Figure 4: Spatial distributions of the GloBI data collection (as of July 16, 2014). Map (a) shows the global distribution of the number of recorded interactions. Map (b) includes the number of spatially explicit interactions that include time stamps. Map (c) shows the sum of all distinct interactions. A distinct interaction is defined as presence of at least one interaction between two specific taxa. Map (d) contains the number of interaction studies referenced in a specific area. Map (e) indicates the spatial distribution of the number of unique data sources (or contributors) for recorded interactions, and map (f) provides a distribution of how many individual locations were observed within the sample areas. A spatial bin size of 5 by 5 degrees was used to aggregate spatial statistics shown in the figures. A color-coded $\log_{10}$ scale visualizes the values contained in each bin.

provided by a study that used a text-mining technique to extract species interactions from text objects in Encyclopedia of Life (EOL) taxon pages [31]. Spatial distributions of species interactions (Figure 4) are by no means uniform. The maps in Figures 4a and 4c show a high concentration of distinct interactions with spatiotemporal interaction information in Europe, North America, the Southern Ocean, and New Zealand. The highest density of distinct interaction studies can be found in the Gulf of Mexico, the North Sea, and Weddell Sea (Figure 4d). In regards to the density of data sources, North American and European regions provide most coverage (Figure 4e).

3.3. Usage of the GloBI framework and data collection

Besides GloBI web pages, two websites use GloBI data to embed structured species-interaction information: the Encyclopedia of Life (EOL) and Gulf of Mexico Species Interactions (GoMexSI). EOL is a website that provides global access to knowledge about life on earth. Specifically, EOL hosts curated web pages containing information about organisms and the classification of organisms. GoMexSI offers open web access to trophic-interaction datasets observed in the Gulf of Mexico. The GoMexSI website features tools that allow users to query, navigate, and download data on the diets of marine life in the Gulf of Mexico.

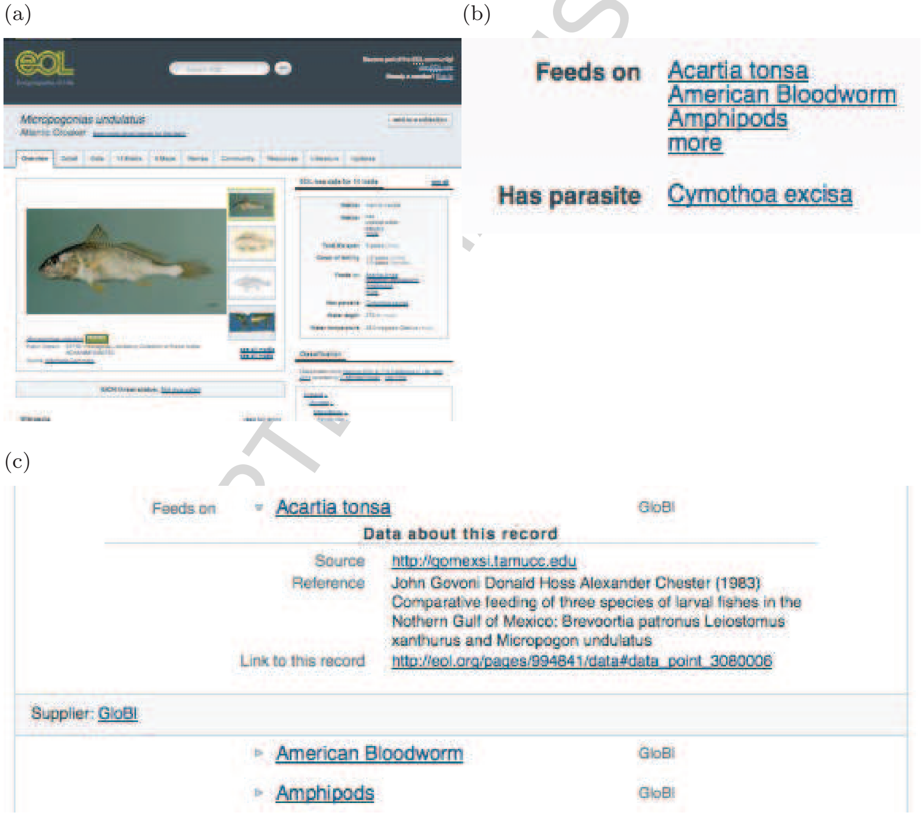
3.3.1. Usage within EOL

EOL integrates GloBI interaction data by periodically importing a publicly available GloBI Darwin Core archive [15]. Of the two Darwin Core archives available, EOL uses the archive that aggregates all interactions by study reference. This way, EOL can cite the study and its data source without having to import every single observed interaction. The interaction data is then integrated into the trait summary section on the relevant taxon overview page (Figures 5a, 5b). Also, a full list of available interactions (from the GloBI collection) is included in the Data tab of EOLs taxon page (Figure 5c).

3.3.2. Usage within GoMexSI

GoMexSI is using the web API provided by GloBI to support various features. First, statistics are retrieved from GloBI and displayed on the home page of GoMexSI (Figure 6a). This statistical analysis summarizes information about the number of references, interactions, and taxa specific to GoMexSI datasets. Second, query and discovery pages allow the user to request and display specific interaction observations (Figure 6b). The data from the queries can also be downloaded into a tabular file for further use. The raw data is provided in a CSV file format with a much richer array of parameters for the interacting species. The explorer mode provides a complete array of all the predators and prey of a selected organism, a favorite feature of some educators who have registered on the site.

To help the user enter a scientific name for available taxa, GoMexSI uses GloBIs fuzzy name-search algorithm, generating close matches to the entered text



Feeds on Acartia tonsa
American Bloodworm
Amphipods
more

Has parasite Cymothoa excisa

Feeds on ▾ Acartia tonsa GloBI

Data about this record

Source http://gomexsi.tamucc.edu
Reference John Govoni Donald Hoss Alexander Chester (1983) Comparative feeding of three species of larval fishes in the Nothern Gulf of Mexico: Brevoortia patronus Leiostomus xanthurus and Micropogon undulatus
Link to this record http://eol.org/pages/994841/data#data_point_3080006

Supplier: GloBI

▸ American Bloodworm GloBI

▸ Amphipods GloBI

Figure 5: Screenshots of EOL pages on the Atlantic croaker (*Micropagonias undulatus*) at `http://eol.org/pages/994841` with GloBI interaction data: (a) overview page, (b) detail of overview page, and (c) detail of the Data tab.

string (Figure 6c). Other enhancements to GloBI are currently being planned to enrich and simplify the data-searching experience on GoMexSI. Planned enhancements include tools that will allow the data to be parsed by locations, life history/size class, or seasonal/temporal modes.

3.3.3. GloBI web pages

An informational, proof-of-concept website, `http://globalbioticinteractions.org`, demonstrates some examples of how to embed interaction data into a standalone, dynamic, HTML page. GloBIs website showcases three kinds of functionality: First, the "What do ... eat?" page (Figure 7a) allows the user to search for all prey of a specific predator, using the name suggestion tool and interaction query service. The resulting prey-items list includes a scientific name, common name, and image for each prey. Each prey item also links to an external website, such as EOL [14] or WoRMS [37], or an associated ontology such as the EnvO [8]. Second, the reference page (Figure 7b) gives an overview of GloBIs aggregated interaction data by study, displaying the citation, data source, number of interactions, and distinct number of source and target taxa for each study. Finally, an interaction browser makes it possible for the user to select a region of interest on a map, and displays a visual representation of the species interactions in the selection in bundle and circular layouts. The visualizations for the interaction browser were created using Google Maps APIs and d3js.org [38] in combination with the GloBI API. The interaction browser is under development and can be found at `http://globalbioticinteractions.org/browse`.

3.4. RDF export

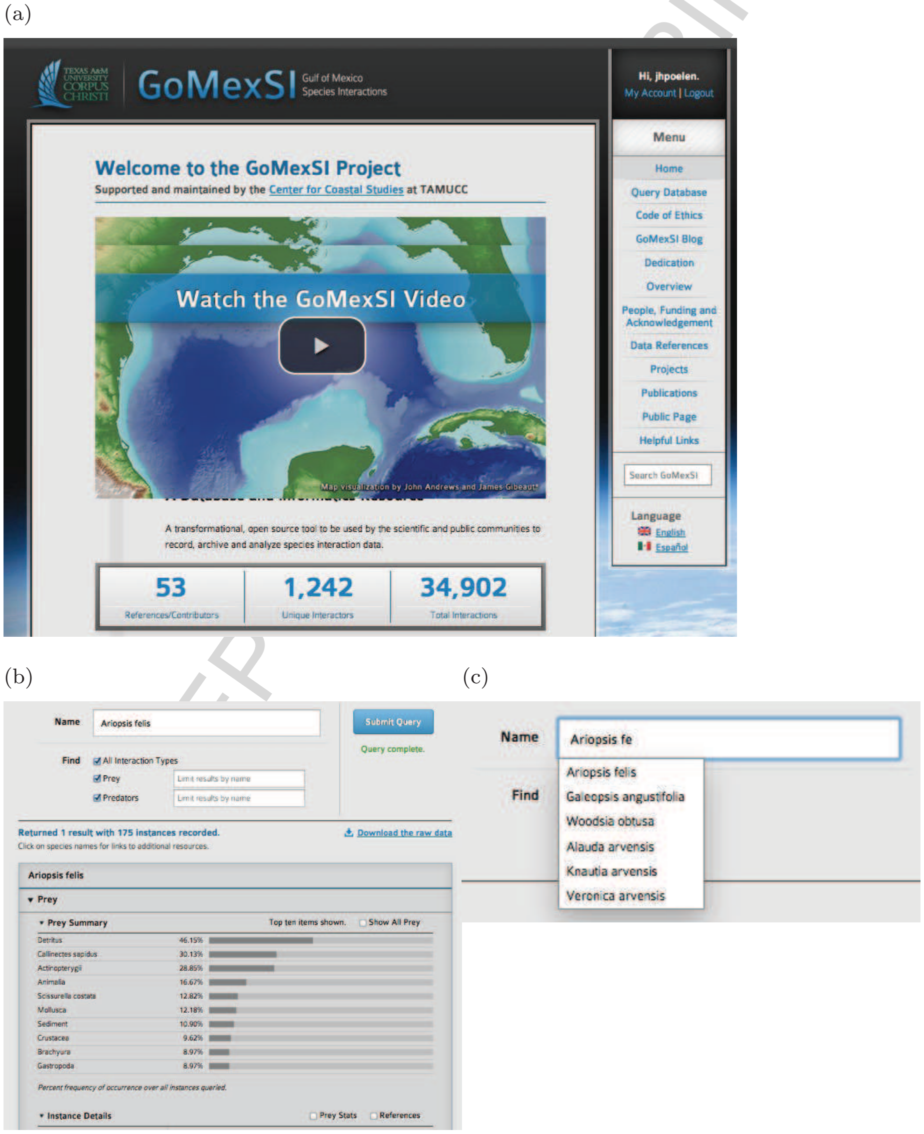
GloBI's data collection is also available as an RDF triple dump, which can be queried via its SPARQL endpoint. The export includes all interactions in GloBI, but, at this time, does not include the full metadata available for each interaction.

Each interaction is modeled as an instance of the Gene Ontology (GO) class "interspecies interaction between organisms" (GO_0044419). The interaction is connected to the two organisms participating in the interaction, each of which is a type of the class "organism" (CARO_0010004), taken from the Common Anatomy Reference Ontology. The two organisms are connected by way of an interaction relation taken from the OBO Relations Ontology (RO), for example, "parasite of" (RO_0002444).

Each of the two organisms are connected by a "member of" relation (RO_0002350) to the taxon object, which is connected to one or more taxonomy references via an OWL sameAs predicate.

If the interaction has environmental context, this is recorded using an "occurs in" relationship (BFO_0000066) to an instance of a class from EnvO.

The triples from GloBI can be combined with these ontologies and logical reasoners to perform powerful knowledge-enhanced queries. For example, querying for all interactions that occur in a "terrestrial biome" (ENVO_00000446) will return interactions that occur in any instance of this EnvO class or its subclasses.



(a)

Welcome to the GoMexSI Project

Supported and maintained by the Center for Coastal Studies at TAMUCC

A transformational, open source tool to be used by the scientific and public communities to record, archive and analyze species interaction data.

53	1,242	34,902
References/Contributors	Unique Interactors	Total Interactions

(b)

Name: Ariopsis felis

Find: ☑ All Interaction Types
☑ Prey — Limit results by name
☑ Predators — Limit results by name

Query complete.

Returned 1 result with 175 instances recorded.

Click on species names for links to additional resources.

Ariopsis felis

▾ Prey

▾ Prey Summary — Top ten items shown. □ Show All Prey

Detritus	46.15%
Callinectes sapidus	30.13%
Actinopterygii	28.85%
Animalia	16.67%
Scissurella costata	12.82%
Mollusca	12.18%
Sediment	10.90%
Crustacea	9.62%
Brachyura	8.97%
Gastropoda	8.97%

Percent frequency of occurrence over all instances queried.

▾ Instance Details — □ Prey Stats □ References

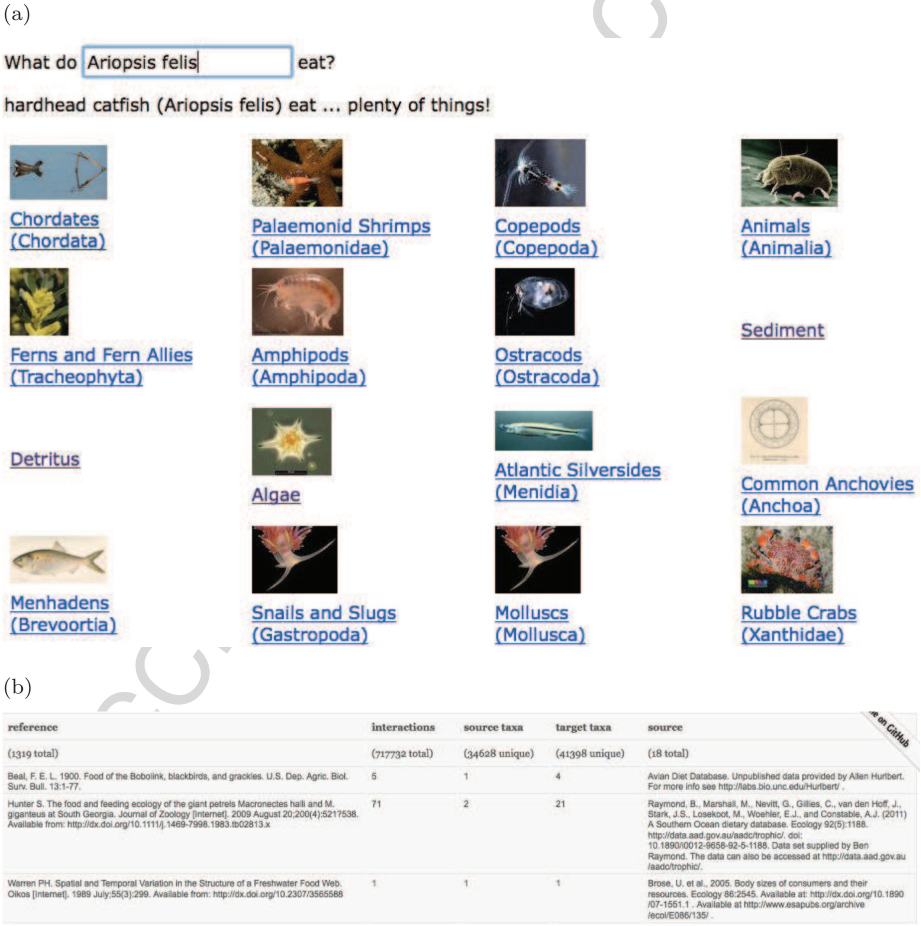
(c)

Name: Ariopsis fe

Find

Ariopsis felis
Galeopsis angustifolia
Woodsia obtusa
Alauda arvensis
Knautia arvensis
Veronica arvensis

Figure 6: Screenshots of GoMexSI pages using GloBI interaction data: (a) home page, (b) prey statistics, and (c) name suggestions.



(a)

What do Ariopsis felis eat?

hardhead catfish (Ariopsis felis) eat ... plenty of things!

(b)

reference	interactions	source taxa	target taxa	source
(1319 total)	(717732 total)	(34628 unique)	(41398 unique)	(18 total)
Beal, F. E. L. 1900. Food of the Bobolink, blackbirds, and grackles. U.S. Dep. Agric. Biol. Surv. Bull. 13:1-77.	5	1	4	Avian Diet Database. Unpublished data provided by Allen Hurlbert. For more info see http://labs.bio.unc.edu/Hurlbert/ .
Hunter S. The food and feeding ecology of the giant petrels Macronectes halli and M. giganteus at South Georgia. Journal of Zoology [Internet]. 2009 August 20;200(4):521?538. Available from: http://dx.doi.org/10.1111/j.1469-7998.1983.tb02813.x	71	2	21	Raymond, B., Marshall, M., Nevitt, G., Gillies, C., 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://data.aad.gov.au/aadc/trophic/. doi: 10.1890/i0012-9658-92-5-1188. Data set supplied by Ben Raymond. The data can also be accessed at http://data.aad.gov.au/aadc/trophic/.
Warren PH. Spatial and Temporal Variation in the Structure of a Freshwater Food Web. Oikos [Internet]. 1989 July;55(3):299. Available from: http://dx.doi.org/10.2307/3565588	1	1	1	Brose, U. et al., 2005. Body sizes of consumers and their resources. Ecology 86:2545. Available at: http://dx.doi.org/10.1890/07-1551.1 . Available at http://www.esapubs.org/archive/ecol/E086/135/ .

Figure 7: Screenshots of GloBI web pages at `http://globalbioticinteractions.org`: (a) What do *Ariopsis felis* eat?, and (b) reference table.

4. Discussion

In the process of building the GloBI infrastructure and integrating datasets, we encountered a number of nontrivial challenges that fall in roughly three categories: (a) data sharing, (b) process automation, and (c) term mapping.

4.1. Data sharing

Despite the advances in technology and a shift in various disciplines to open-data science (e.g., genomics, particle physics, astronomy), scientists and institutions are often unable or unwilling to make available the species-interaction datasets on which their publications are based. For instance, large US government fisheries-survey datasets remain closed due to legal restrictions of the Magnuson-Stevens Fishery Conservation and Management Act; as amended, Public Law 109-479; 16 U.S.C. 1853; implemented at 50 CFR 679.50. Some research bodies, including governmental institutions, claim to have invested too many funds into compiling datasets to give them away for free. Other frequently heard reasons for not sharing data are that, even though research results have been published, the data is not ready to be released because other publications based on the data are pending.

For those who are uncomfortable sharing their datasets, we have instantiated Dark GloBI, a version of GloBI that contains datasets accessible only with the explicit permission of the data contributors. In this way, researchers can still take advantage of GloBIs different functionalities while using restricted datasets.

That said, we hope that GloBIs open-source tools for attributing, accessing, and discussing datasets will both increase the visibility of data contributors and amplify the usefulness of their data. By loosening proprietary restrictions on species-interaction data, researchers will have more opportunities to collaborate and ampler resources for testing hypotheses. These benefits will hopefully provide enough incentive not only to share data, but also to reuse datasets from colleagues around the world.

On our part, several improvements can be made to how data is shared and managed within GloBI. Currently, some Java programming is necessary to add an interaction dataset to GloBI. While our results suggest that this approach is sufficient at current scale, introducing a more user-friendly way to add datasets might lower the threshold for sharing. However, if datasets use standard exchange formats such as Darwin Core, this problem is mitigated.

Another possible improvement would be to build a tool that allows for manual entry of interaction data. This data-entry tool would permit institutions or individuals to transcribe and share interaction records from nondigital sources. We believe that the growing community of data contributors and users will promote the adoption of existing data-exchange formats and guide the development of increasingly effective data-aggregation and access methods.

Attribution is an important part of GloBI - inclusion of attributions can potentially encourage connections among researchers. While contributor visibility is currently increased by an online reference list (see `http://globalbioticinteractions.`

org/references.html), no major technical barriers exist to use more sophisticated methods like assigning digital object identifiers (DOIs) to contributed datasets that individual contributors can list on their profiles at http://orcid.org. We are working with the Semantic Web in Health Case and Life Sciences Interest Group (HCLSIG)[39] to provide enhanced metadata descriptions of each dataset using standard web vocabularies in order to make the data more discoverable.

4.2. Process automation

GloBI's data-transformation algorithms are automated to reduce human error. This automation helps to establish a reproducible data-processing pipeline to reduce variability in outcomes with the same input data. The main source of process variation and delay are introduced by web APIs. While the process is sufficiently reliable, it raises a question about the long-term challenges of reproducing data transformations relying on web APIs that might not be available in the future.

GloBI itself proposes an answer to this question by offering an API and full data archives. These archives contain a versioned copy of all interaction data and are stored in a Maven repository (http://maven.apache.org). They can be embedded in automated processing workflows without the need for internet access, because copies can be cached locally.

4.3. Term mapping

Terms like taxon names, life stages, and locations are mapped to existing ontologies as much as possible. Currently, identical mapping procedures are used across all datasets; term mapping is largely a manual operation that involves inspecting mismatching terms and entering an appropriate mapping using a CSV file. While this method holds up nicely at this time, we anticipate two challenges: First, as more datasets come in, the mapping files will grow to a point where a tool will have to be introduced to curate data mappings. Second, we expect that more sophisticated, dataset-specific name mappings will be needed to avoid mapping conflicts.

4.4. The Prior Art of Data Sharing

The realization that sharing and reusing interaction data facilitates ecological research is far from new. Efforts such as Ecologists' Cooperative Web Bank [40], Webs on the Web [41], Animal Diversity Web [42], and Interaction Web Database [43] have aggregated and published biotic-interaction datasets using various methods to make it easier to access existing interaction datasets using various custom information systems. Unfortunately, few of these initiatives are still active, and their aggregated datasets are at risk of becoming inaccessible due to loss of web-hosting capacity.

The scientific community is getting better at ensuring continued access to scientific datasets beyond a two-to-three year research grant cycle: the advent of open-access data publication clauses in research grants and journal data submission requirements have stimulated the creation of scientific data repositories (DataONE [44], LifeWatch [45], Dryad [46], PANGAEA [47]) and online data journals (ESA's *Ecological Archives*, *Biodiversity Data Journal*, Nature's *Scientific Data*). These new publication methods allow scientific communities to reuse and cite source datasets for purposes of reproducing results, or gaining new insights without having to collect new data. However, these data publication platforms are general purpose and provide data that isnt necessarily machine-readable. For instance, the ability to search inside these sources' published datasets is not provided. In other words, while a wealth of data is now accessible, it cannot be used without a significant effort on the part of the data consumer. Aside from downloading and attempting to import nonstandard file formats, typical manual operations to reuse a published species-interaction dataset include correcting taxonomic names, converting common names to scientific names, and mapping unstructured text that describes terms such as life stage or locale into controlled vocabularies. So, while many ecological datasets are available, putting the data to work is a time-consuming task.

To turn openly accessible data into usable information, a data process step is required after data publication to allow for automated discovery and aggregation of relevant datasets for a specific purpose. The EcoData Retriever [48], an ecological data discovery and retrieval tool, provides a way to search preindexed, tabular, ecological datasets and install them in a relational database. However, this leaves the challenge of aggregating separate datasets into a single dataset. A recent effort [49] implements an R package and web API for accessing and analyzing interaction data. In the future, we will provide an interoperation layer with the rmangal package.

Elaborate data-transformation platforms like Galaxy [50] or Pentaho Kettle [51] automate the process of data discovery and aggregation, but even these sophisticated platforms require significant amount of programming to configure data workflows.

4.5. Assessment of GloBI's benefits

While the success of GloBI is difficult to express quantitatively at this early stage, some qualitative observations can be made. Namely, the quality of interaction datasets was improved when they were introduced into GloBI. In many instances, the GloBI data-aggregation process or GloBI users brought data inconsistencies or errors to light. For example, a user of GloBI data reported a number of suspicious interactions by creating a GitHub issue (`https://github.com/jhpoelen/eol-globi-data/issues/67`). This issue was then used to discuss the steps to be taken to fix the errors in the data. In this case, the issue revealed a bug in the data import and an invalid record in a source database. The bug was fixed and the author of the source data confirmed a data error and corrected the invalid record at the source. Effectively, aggregating species-interaction data into GloBI initiates a crowd-sourced review process for that data.

Another sign of GloBI's success is its use by several scientific institutions and researchers. EOLs and GoMexSIs websites have been using GloBI data access methods since January 2014 and September 2013, respectively, and GloBI provided data access tools that were used to develop novel interaction visualization method [52, 53]. We expect that, over time, the GloBI citation index will give a more quantitative measure of success within the scientific community, and that API usage and data-download statistics will provide a more general measure for success within citizen science and educational communities.

5. Conclusion

By making a large collection of machine-readable species-interaction data available, we can help educators and scientists to better understand how organisms interact with their surroundings. This understanding could inform strategies on how to conserve the natural resources that we rely on for survival.

Through an open, iterative collaboration process between the Encyclopedia of Life community, data contributors, scientists, and software engineers, we built GloBI. GloBI offers methods to discover, analyze, and cite existing species-interaction datasets that currently cover about 700,000 interactions, across 50,000 taxa, over 1,100 studies, from 19 data sources. With GloBIs tools and data collection, we can now ask complex questions about species interactions, and obtain answers with detailed taxonomic, habitat/environment, geographic, and temporal information. The GloBI infrastructure is a modular, scalable infrastructure that relies heavily on openly accessible APIs (e.g., GlobalNames, EOL's API, CrossRef), free technologies (e.g., Apache Maven for software life-cycle and dependency management, and Neo4j [2] for graph database), and low or no-cost platforms (e.g., Amazon S3[1] for storage; GitHub for source-code repository, wiki documentation and issue management; and Travis[2] for continuous integration testing).

As the cross-disciplinary GloBI community continues to share, discuss, improve, and use species-interaction data, we expect that others will step forward and open up their biotic-interaction observations to offer an increasingly valuable ecological research resource to all.

6. Acknowledgments

We thank Jen Hammock, Brian Hayden, Peter Roopnarine, Jeff Holmes, and Cyndy Parr for sharing their ideas, participating in discussions, and supporting our efforts. Also, thanks to Robert Reiz, Gran Bodenschatz, Stephan Gosnell, and Sergey Slyusarev for using and providing feedback on GloBI data as part of developing software and visualizations. Thanks to Pepper Luboff for editing this article. Last but not least, we thank all data contributors that have opened up

[1] http://aws.amazon.com/s3

[2] http://travis-ci.org

their datasets and answered many of our questions: Malcolm Storey; Joel Sachs; Ken-ichi Ueda; Allen Hurlbert; Ben Raymond; Carolyn Barnes; Jarrett Byrnes; Colt Cook; Jos Ferrer-Paris; Anne Thessen; Institute for Marine Resources and Ecosystem Studies (IMARES); International Council for Exploration of the Sea (ICES); UK Species Inventory at the Angela Marmont Centre for UK Biodiversity, the Natural History Museum, London; and many others. This work has been supported by the EOL Rubenstein Fellowship Program 2013.

References

[1] J. Gosling, The Java language specification, Addison-Wesley Professional, 2000.

[2] Neo4j graph database, `http://neo4j.org`, last Accessed on July 17, 2014.

[3] B. Smith, W. Ceusters, B. Klagges, J. Köhler, A. Kumar, J. Lomax, C. Mungall, F. Neuhaus, A. L. Rector, C. Rosse, Relations in biomedical ontologies, Genome Biol 6 (5) (2005) R46. doi:10.1186/gb-2005-6-5-r46.

[4] M. D. Spalding, H. E. Fox, G. R. Allen, N. Davidson, Z. A. F. na, M. Finlayson, B. S. Halphern, M. A. Jorge, A. Lombana, S. A. Lourie, K. D. Martin, E. McManus, J. Molnar, C. A. Rechhia, J. Robertson, Marine ecoregions of the world: A bioregionalization of coastal and shelf areas, BioScience 57 (7) (2007) 573. doi:10.1641/b570707.

[5] D. M. Olson, E. Dinerstein, E. D. Wikramanayake, N. D. Burgess, G. V. Powell, E. C. Underwood, J. A. D'amico, I. Itoua, H. E. Strand, J. C. Morrison, et al., Terrestrial ecoregions of the world: A new map of life on earth a new global map of terrestrial ecoregions provides an innovative tool for conserving biodiversity, BioScience 51 (11) (2001) 933–938. doi:10.1641/0006-3568(2001)051[0933:teotwa]2.0.co;2.

[6] R. Abell, M. L. Thieme, C. Revenga, M. Bryer, M. Kottelat, N. Bogutskaya, B. Coad, N. Mandrak, S. C. Balderas, W. Bussing, M. L. J. Stiassny, P. Skelton, G. R. Allen, P. Unmack, A. Naseka, R. Ng, N. Sindorf, J. Robertson, E. Armijo, J. V. Higgins, T. J. Heibel, E. Wikramanayake, D. Olson, H. L. López, R. E. Reis, J. G. Lundberg, M. H. S. Pérez, P. Petry, Freshwater ecoregions of the world: A new map of biogeographic units for freshwater biodiversity conservation, BioScience 58 (5) (2008) 403. doi:10.1641/b580507.

[7] A. R. Longhurst, Biogeographic partition of the ocean, in: Ecological Geography of the Sea, Elsevier, 2007, pp. 19–34. doi:10.1016/b978-012455521-1/50003-6.

[8] P. Buttigieg, N. Morrison, B. Smith, C. J. Mungall, S. E. Lewis, The environment ontology: contextualising biological and biomedical entities, J Biomed Sem 4 (1) (2013) 43. doi:10.1186/2041-1480-4-43.

[9] C. J. Mungall, C. Torniai, G. V. Gkoutos, S. E. Lewis, M. A. Haendel, Uberon, an integrative multi-species anatomy ontology, Genome Biol 13 (1) (2012) R5. doi:10.1186/gb-2012-13-1-r5.

[10] F. G. D. Committee, Coastal and marine ecological classification standard, http://www.csc.noaa.gov/digitalcoast/publications/cmecs, last Accessed on July 17, 2014 (2012).

[11] C. W. Cook, The early life history and reproductive biology of cymothoa excisa, a marine isopod parasitizing atlantic croaker,(micropogonias undulatus), along the texas coast, Master's thesis, University of Texas at Austin, last Accessed on July 17, 2014 (2012).

[12] J. Sachs, C. Parr, A. Parafiynyk, R. Pan, L. Han, L. Ding, T. Finin, A. Hollander, T. Wang, Using the semantic web to support ecoinformatics, in: Proceedings of the AAAI Fall Symposium on the Semantic Web for Collaborative Knowledge Acquisition, American Associate for Artificial Intelligence, 2006, pp. 56–61.

[13] B. Raymond, M. Marshall, G. Nevitt, C. L. Gillies, J. van den Hoff, J. S. Stark, M. Losekoot, E. J. Woehler, A. J. Constable, A southern ocean dietary database, Ecology 92 (5) (2011) 1188–1188. doi:10.1890/i0012-9658-92-5-1188.

[14] C. S. Parr, N. Wilson, P. Leary, K. Schulz, K. Lans, L. Walley, J. Hammock, A. Goddard, J. Rice, M. Studer, J. Holmes, J. Robert Corrigan, The encyclopedia of life v2: Providing global access to knowledge about life on earth, Biodiversity Data Journal 2 (2014) e1079. doi:10.3897/bdj.2.e1079.

[15] J. Wieczorek, D. Bloom, R. Guralnick, S. Blum, M. Döring, R. Giovanni, T. Robertson, D. Vieglais, Darwin core: An evolving community-developed biodiversity data standard, PLoS ONE 7 (1) (2012) e29715. doi:10.1371/journal.pone.0029715.

[16] D. Beckett, T. Berners-Lee, Turtle-terse rdf triple language, http://www.w3.org/TeamSubmission/turtle/, last Accessed on July 17, 2014 (2008).

[17] E. R. Gansner, S. C. North, An open graph visualization system and its applications to software engineering, Software - Practice and Experience 30 (11) (2000) 1203–1233. doi:10.1002/1097-024X(200009)30:11<1203::AID-SPE338>3.0.CO;2-N.

[18] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria (2014).

[19] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, Cytoscape: A software environment for integrated models of biomolecular interaction networks, Genome Research 13 (11) (2003) 2498–2504. doi:10.1101/gr.1239303.

[20] M. Bastian, S. Heymann, M. Jacomy, Gephi: an open source software for exploring and manipulating networks., in: ICWSM, 2009, pp. 361–362.

[21] J. E. Byrnes, D. C. Reed, B. J. Cardinale, K. C. Cavanaugh, S. J. Holbrook, R. J. Schmitt, Climate-driven increases in storm frequency simplify kelp forest food webs, Global Change Biology 17 (8) (2011) 2513–2524. doi:10.1111/j.1365-2486.2011.02409.x.

[22] G. Cassis, J. Pickering, Species interactions of australia database, http://www.discoverlife.org/siad/, last Accessed on July 17, 2014.

[23] M. Storey, Bioinfo: food webs and species interactions in the biodiversity of uk and ireland., http://bioinfo.org.uk, last accessed on July 17, 2014 (2014).

[24] R. F. Hechinger, K. D. Lafferty, J. P. McLaughlin, B. L. Fredensborg, T. C. Huspeni, J. Lorda, P. K. Sandhu, J. C. Shaw, M. E. Torchin, K. L. Whitney, A. M. Kuris, Food webs including parasites, biomass, body sizes, and life stages for three california/baja california estuaries, Ecology 92 (3) (2011) 791–791. doi:10.1890/10-1383.1.

[25] P. D. Roopnarine, R. Hertog, Detailed food web networks of three greater antillean coral reef systems: The cayman islands, cuba, and jamaica, Dataset Papers in Ecology 2013 (2013) 1–9. doi:10.7167/2013/857470.

[26] C. Barnes, D. M. Bethea, R. D. Brodeur, J. Spitz, V. Ridoux, C. Pusineri, B. C. Chase, M. E. Hunsicker, F. Juanes, A. Kellermann, J. Lancaster, F. Ménard, F.-X. Bard, P. Munk, J. K. Pinnegar, F. S. Scharf, R. A. Rountree, K. I. Stergiou, C. Sassa, A. Sabates, S. Jennings, Predator and prey body sizes in marine foodwebs, Ecology 89 (3) (2008) 881–881. doi:10.1890/07-1551.1.

[27] S. Faulwetter, V. Markantonatou, C. Pavloudi, N. Papageorgiou, K. Keklikoglou, E. Chatzinikolaou, E. Pafilis, E. Pafilis, G. Chatzigeorgiou, K. Vasileiadou, T. Dailianis, L. Fanini, P. Koulouri, C. Arvanitidis, Polytraits : A database on biological traits of marine polychaetes, BDJ 2 (2014) e1024. doi:10.3897/bdj.2.e1024.

[28] J. D. Simons, M. Yuan, C. Carollo, M. Vega-Cendejas, T. Shirley, M. L. Palomares, P. Roopnarine, L. G. A. Arenas, A. I. nez, J. Holmes, C. M. Schoonard, R. Hertog, D. Reed, J. Poelen, Building a fisheries trophic interaction database for management and modeling research in the gulf of mexico large marine ecosystem, Bulletin of Marine Science 89 (1) (2013) 135–160. doi:10.5343/bms.2011.1130.

[29] J. R. Ferrer-Paris, A. Sánchez-Mercado, A. L. Viloria, J. Donaldson, Congruence and diversity of butterfly-host plant associations at higher taxonomic levels, PLoS ONE 8 (5) (2013) e63570. doi:10.1371/journal.pone.0063570.

[30] A. Hurlbert, Avian diet database, unpublished dataset (2014).

[31] A. Thessen, pseudonitzchia: Biodiversity informatics code relevant to the encyclopedia of life, `https://github.com/eol/pseudonitzchia` (2014).

[32] C. García-Robledo, D. L. Erickson, C. L. Staines, T. L. Erwin, W. J. Kress, Tropical plant-herbivore networks: Reconstructing species interactions using dna barcodes, PLoS ONE 8 (1) (2013) e52967. `doi:10.1371/journal.pone.0052967`.

[33] I. C. for the Exploration of the Sea, Data base report of the stomach sampling project 1981, cooperative research report no. 164, `http://www.ices.dk/sites/pub/Publication%20Reports/Cooperative%20Research%20Report%20(CRR)/crr164/CRR164.pdf`, last accessed on July 17, 2014 (April 1989).

[34] I. C. for the Exploration of the Sea, 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 on July 17, 2014 (October 1996).

[35] U. Brose, L. Cushing, E. L. Berlow, T. Jonsson, C. Banasek-Richter, L.-F. Bersier, J. L. Blanchard, T. Brey, S. R. Carpenter, M.-F. C. Blandenier, J. E. Cohen, H. A. Dawah, T. Dell, F. Edwards, S. Harper-Smith, U. Jacob, R. A. Knapp, M. E. Ledger, J. Memmott, K. Mintenbeck, J. K. Pinnegar, B. C. Rall, T. Rayner, L. Ruess, W. Ulrich, P. Warren, R. J. Williams, G. Woodward, P. Yodzis, N. D. Martinez, Body sizes of consumers and their resources, Ecology 86 (9) (2005) 2545–2545. `doi:10.1890/05-0379`.

[36] B. Planque, R. Primicerio, K. Michalsen, M. Aschan, G. Certain, P. Dalpadado, H. Gjosaeater, C. Hansen, E. Johannesen, L. L. Jorgensen, I. Kolsum, S. Kortsch, L.-M. Leclerc, L. Omli, M. Skern-Mauritzen, M. Wiedmann, Who eats whom in the barents sea: a food web topology from plankton to whales, Ecology 95 (5) (2014) 1430–1430. `doi:10.1890/13-1062.1`.

[37] M. J. Costello, P. Bouchet, G. Boxshall, K. Fauchald, D. Gordon, B. W. Hoeksema, G. C. B. Poore, R. W. M. van Soest, S. Stöhr, T. C. Walter, B. Vanhoorne, W. Decock, W. Appeltans, Global coordination and standardisation in marine biodiversity through the world register of marine species (worms) and related databases, PLoS ONE 8 (1) (2013) e51629. `doi:10.1371/journal.pone.0051629`.

[38] M. Bostock, V. Ogievetsky, J. Heer, D3 data-driven documents, IEEE Transactions on Visualization and Computer Graphics 17 (12) (2011) 2301–2309. `doi:10.1109/tvcg.2011.185`.

[39] A. J. Gray, M. Dumontier, M. S. Marshall, J. Baran, Dataset descriptions: Hcls community profile, `http://tinyurl.com/HCLSDatasetDescription`, last Accessed on August 15, 2014.

[40] J. E. Cohen, Ecologists co-operative web bank (ecoweb), version 1.1 machine-readable data base of food webs. new york: The rockefeller university., `http://hdl.handle.net/10209/306` (2010).

[41] I. Yoon, R. Williams, E. Levine, S. Yoon, J. Dunne, N. Martinez, Webs on the web (wow): 3d visualization of ecological networks on the www for collaborative research and education, in: Visualization and Data Analysis 2004, Vol. 5295, 2004, pp. 124–132. `doi:10.1117/12.526956`.

[42] P. Myers, R. Espinosa, C. S. Parr, T. Jones, G. S. Hammond, T. A. Dewey, The animal diversity web, `http://animaldiversity.org`, last Accessed on July 17, 2014 (2014).

[43] D. P. Vázquez, Degree distribution in plant-animal mutualistic networks: forbidden links or random interactions?, Oikos 108 (2) (2005) 421–426. `doi:10.1111/j.0030-1299.2005.13619.x`.

[44] W. Michener, D. Vieglais, T. Vision, J. Kunze, P. Cruse, G. Janée, Dataone: Data observation network for earth preserving data and enabling innovation in the biological and environmental sciences, D-Lib Magazine 17 (1/2). `doi:10.1045/january2011-michener`.

[45] A. Basset, W. Los, Biodiversity escience: Lifewatch, the european infrastructure on biodiversity and ecosystem research, Plant Biosystems - An International Journal Dealing with all Aspects of Plant Biology 146 (4) (2012) 780–782. `doi:10.1080/11263504.2012.740091`.

[46] Dryad: a nonprofit repository for data underlying the international scientific and medical literature, `http://www.datadryad.org/`, last Accessed on August 13, 2014.

[47] M. Diepenbroek, H. Grobe, M. Reinke, U. Schindler, R. Schlitzer, R. Sieger, G. Wefer, Pangaea|an information system for environmental sciences, Computers & Geosciences 28 (10) (2002) 1201–1210. `doi:10.1016/s0098-3004(02)00039-0`.

[48] B. D. Morris, E. P. White, The ecodata retriever: Improving access to existing ecological data, PLoS ONE 8 (6) (2013) e65848. `doi:10.1371/journal.pone.0065848`.

[49] T. E. Poisot, B. Baiser, J. A. Dunne, S. Kéfi, F. Massol, N. Mouquet, T. N. Romanuk, D. B. Stouffer, S. A. Wood, D. Gravel, mangal - making complex ecological network analysis simpler, bioRxiv`doi:10.1101/002634`.

[50] J. Goecks, A. Nekrutenko, J. Taylor, T. G. Team, Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences, Genome Biol 11 (8) (2010) R86. `doi:10.1186/gb-2010-11-8-r86`.

[51] R. Bouman, J. Van Dongen, Pentaho Solutions: Business Intelligence and Data Warehousing with Pentaho and MySQL, Wiley Publishing, 2009.

[52] J. H. Poelen, Exploring antarctic interactions using globis interaction browser, http://globalbioticinteractions.wordpress.com/2014/03/21/exploring-antarctic-interactions-using-globis-interaction-browser/, last accessed on July 17, 2014 (2014).

[53] J. H. Poelen, A food-web map of the world, http://globalbioticinteractions.wordpress.com/2014/06/06/a-food-web-map-of-the-world/, last accessed on July 17, 2014 (2014).

Highlights

Integrates existing species-interaction datasets
Provides access to a large spatiotemporal data collection of biotic interactions
Cross-references existing ontologies, vocabularies, and taxonomies
Used by the Encyclopedia of Life and Gulf of Mexico Species Interactions projects