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

A R T I C L E   I N F O

Article history:
Received 22 July 2014
Received in revised form 16 August 2014
Accepted 24 August 2014
Available online 3 September 2014

Keywords:
Species interactions
Data integration
Taxonomy
Ontology

A B S T R A C T

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 1100 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 (GoMexSI).

© 2014 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-SA license (http://creativecommons.org/licenses/by-nc-sa/3.0/).

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 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 of 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

⁎ Corresponding author.
E-mail address: jhpoelen@xs4all.nl (JH. Poelen).
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 benefits 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 GloBI. The system is implemented in Java (Gosling, 2000) and uses (Neo4j) as a persistent data store and query system. The system 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 (Fig. 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 (Smith et al., 2005). 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 (Abell et al., 2008; Longhurst, 2007; Olson et al., 2001; Spalding et al., 2007), existing ontologies such as EnvO (Buttigieg et al., 2013), Uberon (Mungall et al., 2012), the OBO Relations Ontology (RO) (Smith et al., 2005), and habitat classification vocabularies, such as Coastal and Marine Ecological Classification Standard (CMECS) (F. G. D. Committee, 2012).

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 (Cook, 2012), while other sources have collected data from multiple studies (Raymond et al., 2011; Sachs et al., 2006).

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.

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

Fig. 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 nor 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).
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 (Fig. 2). Taxonomic names are first matched against EOL’s API (Parr et al., 2014). 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, (Spalding et al., 2007)), Freshwater Ecoregions of the World (FEOW, (Abell et al., 2008)), Terrestrial Ecoregions of the World (TEOW, (Abell et al., 2008)), and Longhurst’s Biogeographical Provinces (Longhurst, 2007).

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 reproducible 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 (Wiecezek et al., 2012), Turtle (Beckett and Berners-Lee, 2008), and (Neo4j) 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 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 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 (Gansner and North, 2000) formats for integration into web pages, spreadsheets, R (R Core Team, 2014), or visualization tools like Graphviz (Gansner and North, 2000), Cytoscape (Shannon et al., 2003), or Gephi (Bastian et al., 2009). 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/opensci/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 (Gosling, 2000) 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 predator taxon node: taxonPaths("/path:Elasmobranchii"),
prey taxon node: taxonPaths("/path:Mammalia")
MATCH predator taxon<-[:CLASSIFIED AS] specimen[:ATE:PREYS OR]-:prey,
prey[:CLASSIFIED AS]-:prey taxon
RETURN predator taxon name, prey taxon 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: ecoregion Paths("/path:California")
MATCH taxon<-[:CLASSIFIED AS] specimen[:ATE:PREYS OR]-:prey,
prey[:CLASSIFIED AS]-:prey taxon,
specimen[:COLLECTED_AT]-:collection [:IN_ECOREGION]-:region
RETURN taxon name, prey taxon 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 (Longhurst, 2007), Marine Ecoregions of the World (MEOW) (Spalding et al., 2007), Terrestrial Ecoregions of the World (TEOW) (Olson et al., 2001), and Freshwater Ecoregions of the World (FEOW) (Abell et al., 2008). 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")
WHERE [location:latitude] > 30.0
MATCH taxon<-[:CLASSIFIED AS] specimen[:ATE:PREYS OR]-:prey,
prey[:CLASSIFIED AS]-:prey taxon,
specimen[:COLLECTED_AT]-:collection [:IN_ECOREGION]:region
RETURN taxon name, prey taxon name.
```
indicates the time range in which interaction data was recorded. Data source statistics were retrieved from GloBI on July 16, 2014. CrossRef, as of July 16, 2014.

DOIs out of 1175 references were obtained from the reference text using fi

derive habitat and ecoregions (CMECS, MEOW). No digital object identi-
using decimal degrees and using free-form text locale information

http://inaturalist.org

describe the total and distinct number of interaction observations between a begin taxon (B) and an end taxon (E). Bdistinct and Edistinct are the number of unique taxa that occur within the set of begin and end taxa. Ltotal and Ttotal 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.

<table>
<thead>
<tr>
<th>Data source</th>
<th>Refs</th>
<th>Ototal</th>
<th>Odistinct</th>
<th>Bdistinct</th>
<th>Edistinct</th>
<th>Ltotal</th>
<th>Ttotal</th>
<th>Period</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kelp Forest Food Web (Byrnes et al., 2011)</td>
<td>130</td>
<td>1915</td>
<td>1611</td>
<td>84</td>
<td>210</td>
<td>0</td>
<td>0</td>
<td>n/a</td>
</tr>
<tr>
<td>Species Interactions of Australia (Cassid and Pickering)</td>
<td>1</td>
<td>14,896</td>
<td>13,657</td>
<td>4461</td>
<td>4075</td>
<td>0</td>
<td>0</td>
<td>n/a</td>
</tr>
<tr>
<td>Species Interactions in UK and Ireland (Storey, 2014)</td>
<td>1</td>
<td>50,157</td>
<td>37,612</td>
<td>13,599</td>
<td>5429</td>
<td>0</td>
<td>0</td>
<td>n/a</td>
</tr>
<tr>
<td>Semantic Prototypes in Research Informatics (Sachs et al., 2006)</td>
<td>195</td>
<td>30,101</td>
<td>21,191</td>
<td>3335</td>
<td>2616</td>
<td>27,152</td>
<td>0</td>
<td>n/a</td>
</tr>
<tr>
<td>Southern Ocean Diet Database (Raymond et al., 2011)</td>
<td>324</td>
<td>26,462</td>
<td>10,368</td>
<td>341</td>
<td>898</td>
<td>26,391</td>
<td>25,858</td>
<td>1961-2011</td>
</tr>
<tr>
<td>Cymothoa excisa parasitizes Micropogonias undulatus (Cook, 2012)</td>
<td>1</td>
<td>315</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>335</td>
<td>335</td>
<td>2010-2012</td>
</tr>
<tr>
<td>Food Webs of Three California Estuaries (Hechinger et al., 2011)</td>
<td>1</td>
<td>1,976</td>
<td>7129</td>
<td>234</td>
<td>247</td>
<td>13,966</td>
<td>0</td>
<td>n/a</td>
</tr>
<tr>
<td>Food Webs of Three Coral Reef Systems (Rogosin et al., 1989; Hertog, 2013)</td>
<td>1</td>
<td>96,647</td>
<td>32,257</td>
<td>659</td>
<td>730</td>
<td>19,39</td>
<td>0</td>
<td>n/a</td>
</tr>
<tr>
<td>Predator and Prey Body Sizes in Marine Food Webs (Barnes et al., 2008)</td>
<td>24</td>
<td>308</td>
<td>529</td>
<td>93</td>
<td>167</td>
<td>25,858</td>
<td>0</td>
<td>n/a</td>
</tr>
<tr>
<td>Polytraits: Biological Traits of Polychaetes (Faulwetter et al., 2011)</td>
<td>24</td>
<td>349</td>
<td>529</td>
<td>93</td>
<td>167</td>
<td>34,931</td>
<td>0</td>
<td>n/a</td>
</tr>
<tr>
<td>Gulf of Mexico Species Interactions (Simons et al., 2013)</td>
<td>53</td>
<td>34,902</td>
<td>4810</td>
<td>659</td>
<td>730</td>
<td>19,39</td>
<td>0</td>
<td>n/a</td>
</tr>
<tr>
<td>EOL Text Mining (Thessen, 2014)</td>
<td>1</td>
<td>24,436</td>
<td>9915</td>
<td>2497</td>
<td>3543</td>
<td>0</td>
<td>0</td>
<td>n/a</td>
</tr>
<tr>
<td>Avian Diet Database (Hurlbert, 2014)</td>
<td>68</td>
<td>4,658</td>
<td>1365</td>
<td>63</td>
<td>625</td>
<td>44,280</td>
<td>44,280</td>
<td>1979-2004</td>
</tr>
<tr>
<td>EOL. Taxon Mining (Thessen, 2014)</td>
<td>1</td>
<td>183,872</td>
<td>1611</td>
<td>84</td>
<td>210</td>
<td>0</td>
<td>0</td>
<td>n/a</td>
</tr>
<tr>
<td>Tropical Plant Herbivores (Sachs et al., 2006)</td>
<td>1</td>
<td>793</td>
<td>544</td>
<td>150</td>
<td>64</td>
<td>0</td>
<td>0</td>
<td>n/a</td>
</tr>
<tr>
<td>ICES North Sea Fish Stomach Data (I.C. for the Exploration of the Sea, 1989; I.C. for the Exploration of the Sea, 1996)</td>
<td>1</td>
<td>183,935</td>
<td>1781</td>
<td>824</td>
<td>799</td>
<td>1938</td>
<td>1980</td>
<td>n/a</td>
</tr>
<tr>
<td>Body Sizes of Consumers and Their Resources (Brose et al., 2005)</td>
<td>1</td>
<td>793</td>
<td>544</td>
<td>150</td>
<td>64</td>
<td>0</td>
<td>0</td>
<td>n/a</td>
</tr>
<tr>
<td>Who Eats Whom in the Barents Sea (Planque et al., 2014)</td>
<td>1</td>
<td>793</td>
<td>544</td>
<td>150</td>
<td>64</td>
<td>0</td>
<td>0</td>
<td>n/a</td>
</tr>
<tr>
<td>Total</td>
<td>1173</td>
<td>719,536</td>
<td>302,803</td>
<td>34,993</td>
<td>41,725</td>
<td>437,983</td>
<td>232,117</td>
<td>1897-2014</td>
</tr>
</tbody>
</table>

3.2. Data coverage

Spatial, temporal, and taxonomic coverage of the combined datasets (Fig. 4, Tables 1, 2) 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. 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 (I.C. for the Exploration of the Sea, 1989; I.C. for the Exploration of the Sea, 1996), GoMexSI (Simons et al., 2013), and Southern Ocean (Raymond et al., 2011) provide most of the spatiotemporal interaction data, datasets such as Thessen (2014), SPIRE (Sachs et al., 2006), and BioInfo (Storey, 2014) 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) (Cook, 2012), offers the smallest taxonomic coverage with a spatiotemporal resolution. The largest taxonomic coverage is provided by a study that used a text-mining technique to extract species interactions from text objects in Encyclopedia of Life (EOL) taxon pages (Thessen, 2014). Spatial distributions of species interactions (Fig. 4) are by no means uniform. The maps in Fig. 4a and c 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 (Fig. 4). In regards to the density of data sources, North American and European regions provide most coverage (Fig. 4).

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 (Wieczorek et al., 2012). 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 (Fig. 5a,b). Also, a full list of available interactions (from the GloBI collection) is included in the Data tab of EOL’s taxon page (Fig. 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 (Fig. 6a). This statistical analysis summarizes information about the number of references, interactions, and taxon specific to GoMexSI datasets. Second, query and discovery pages allow the user to request and display specific interaction observations (Fig. 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

<table>
<thead>
<tr>
<th>Any rank</th>
<th>% ITIS</th>
<th>Species</th>
<th>Genus</th>
<th>Family</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arthropoda</td>
<td>17,201</td>
<td>5.5%</td>
<td>13,428</td>
<td>2867</td>
</tr>
<tr>
<td>Fungi</td>
<td>9378</td>
<td>199.5%</td>
<td>7762</td>
<td>725</td>
</tr>
<tr>
<td>Plantae</td>
<td>9382</td>
<td>6.5%</td>
<td>6959</td>
<td>1940</td>
</tr>
<tr>
<td>Chordata</td>
<td>6737</td>
<td>5.9%</td>
<td>5473</td>
<td>740</td>
</tr>
<tr>
<td>Mollusca</td>
<td>1027</td>
<td>5.6%</td>
<td>736</td>
<td>195</td>
</tr>
<tr>
<td>Annelida</td>
<td>485</td>
<td>7.8%</td>
<td>301</td>
<td>128</td>
</tr>
<tr>
<td>Bacteria</td>
<td>109</td>
<td>19.0%</td>
<td>76</td>
<td>28</td>
</tr>
<tr>
<td>Other</td>
<td>5749</td>
<td>13.0%</td>
<td>4379</td>
<td>1047</td>
</tr>
<tr>
<td>Total</td>
<td>50,068</td>
<td>7.8%</td>
<td>39,114</td>
<td>7670</td>
</tr>
</tbody>
</table>

Fig. 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° was used to aggregate spatial statistics shown in the figures. A color-coded log_{10} scale visualizes the values contained in each bin.
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 string (Fig. 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 stand-alone, dynamic, HTML page. GloBIs website showcases three kinds of functionality: First, the “What do ... eat?” page (Fig. 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 (Parr et al., 2014) or WoRMS (Costello et al., 2013), or an associated ontology such as the EnvO (Buttigieg et al., 2013). Second, the reference page (Fig. 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 (Bostock et al., 2011) 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 is 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.

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 in 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.

Fig. 6. Screenshots of GoMexSI pages using GloBI interaction data: (a) homepage, (b) prey statistics, and (c) name suggestions.
On our part, several improvements can be made to show 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) (Gray et al.) 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 (Cohen and Ecologists co-operative web bank (ecoweb), 2010), Webs on the Web (Yoon et al., 2004), Animal Diversity Web (Myers et al., 2014), and Interaction Web Database (Vázquez, 2005) 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 (Michener et al., 2011), LifeWatch (Basset and Los, 2012), (Dryad), PANGAEA (Diepenbroek et al., 2002)) 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 isn’t 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 processing step is required after data publication to allow for automated discovery and aggregation of relevant datasets for a specific purpose. The EcoData Retriever (Morris and White, 2013), 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 (Poisot et al., 2014) implements an R package and web API for accessing and analyzing interaction data. In the future, we will provide an interoperated layer with the rmganel package.

Elaborate data-transformation platforms like Galaxy (Goecks et al., 2010) or Pentaho Kettle (Bouman and Van Dongen, 2009) 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. EOSi and GoMexSi 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 (Poelen, 2014a,2014b). 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 1100 studies, from 19 data sources. With GloBI 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) for graph database), and low or no-cost platforms (e.g., Amazon S3 for storage; GitHub for source-code repository, wiki documentation and issue management; and Travis 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.

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, Göran 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 the data contributors that have opened up 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-Parris; Anne Thessen; Institute for Marine Resources and Ecology 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 Encyclopedia of Life, EOL Rubenstein Fellows Program and the David M. Rubenstein Grant (CRDF EOL-33066-13/F33066), Smithsonian Institution.

References


Bogutskaya, N., Bogutsky, V., Jarrett Byrnes; Colt Cook; Jos Ferrer-Paris; Anne Thessen; Institute for Exploration of the Sea (ICES); UK Species Inventory at the Marine Resources and Ecosystem Studies (IMARES); International Council for Exploration of the Sea (ICES); and many others. This work has been supported by the Encyclopedia of Life, EOL Rubenstein Fellows Program and the David M. Rubenstein Grant (CRDF EOL-33066-13/F33066), Smithsonian Institution.

