Ecological Informatics 24 (2014) 148–159

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

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

ARTICLE INFO

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

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 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).

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 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 (J.H. Poelen).

http://dx.doi.org/10.1016/j.ecoinf.2014.08.005

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 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 (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.

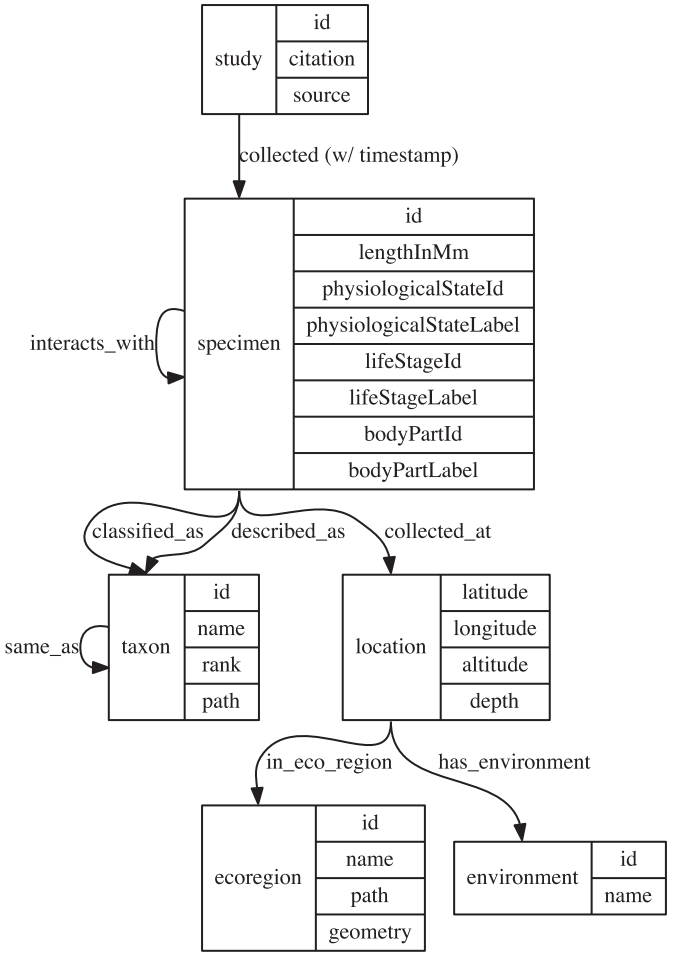

Fig. 1. Interaction data is modeled in terms of study, specimen, taxon, and location concepts. The location has an additional relation to ecoregions and environments to facilitate spatial searches. Most IDs are uniform resource identifiers (URIs) to external ontologies and/or vocabularies. If neither 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).

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