Unleashing EOL's Species Interaction Datasets—Integration, Visualization, and Analysis

by Jorrit Poelen

Our Rubenstein Fellowship 2013 Wish: What is the most effective way to extract existing knowledge about species interactions from biological literature and the biological community in order to make that information available in a useful fashion on the EOL platform? [1]

Overview

Understanding how trophic interactions are connected to climate change and loss of biodiversity may help us to sustain our ecosystems and food supply [2]. However, studying trophic interactions traditionally requires experimental data that is difficult and costly to obtain [3]. This proposal formulates a method to aggregate existing trophic datasets such as Semantic Prototypes in Research Ecoinformatics (SPIRE) [4], BioInfo [5], and Gulf of Mexico Species Interaction Database (GoMexSI) [6] into a single, normalized dataset. Also, three methods are proposed to give easy access to this rich aggregation of trophic data for scientific and educational purposes using the Encyclopedia of Life (EOL) [7].

Introduction

EOL has integrated information from over 200 content providers, creating a vast online collection of species information that contains taxonomies, specimen images, habitat information, and references to scientific papers. EOL makes most of this data accessible through human-readable web pages and machine-readable APIs. The human-readable web pages permit the user to explore and learn about species on a case-by-case basis, whereas the machine-readable APIs allow developers and researchers to build computer applications that enrich non-EOL content with EOL data, such as images or textual description of a species.

In the recent past, EOL acquired two content providers, SPIRE and BioInfo, to enhance existing EOL entries with information about species' feeding habits. This trophic data is currently provided in the form of human-readable text only. Our goal is to build a system that makes this species interaction data available as machine-readable information by extending EOLs application programming interfaces (APIs). Further, we propose to integrate the information with existing ontologies or data models to make it semantically interoperable. This approach will more readily support the integration of non-EOL content into EOL species interaction data.

In addition to providing an open and easy-to-access species interaction dataset, we propose to provide two applications of the resulting dataset: First, a web-based food-web exploration tool that will enable scientists and educators to visualize and navigate species' trophic relationships by way of an interactive network analysis. The food-web tool will originate from and connect back to existing EOL species pages, making it possible to view interactions on a macro level as well as zoom in on specific species.

To complement the food-web browser, a qualitative research tool, we propose building a quantitative tool. The quantitative tool will calculate metrics such as trophic level, average taxonomic predator-prey distance, and shortest predator-prey path. By calculating similar or

identical metrics for different food-web segments, researchers and educators will get a sense of the connectedness of specific species and of the difference in properties of food webs across spatial and temporal dimensions.

All tools, datasets, and computer source code described in this proposal will be available to the public. Moreover, we will encourage contributions and comments in the form of progress updates through a blog; frequent source code commits to an open source repository (e.g., GitHub); and publication of updated versions of datasets, APIs, and tools.

The methods and technologies in this proposal are such that they can easily integrate with EOL's existing PHP [8], Ruby on Rails [9], MySQL [10] and Solr [11] software. That said, we do expect to collaborate with EOL's staff on continuously improving methods and technologies for the proposed features.

Summary of Objectives

Objective 1. Design, implement, and publish an extensible framework to normalize interaction datasets into a machine-readable format.

Objective 2. Normalize and publish an interaction dataset from SPIRE, BioInfo, and GoMexSI. Datasets will be published in Resource Description Framework (RDF) [12] and comma-separated value (CSV) format, in addition to being available with a public web service API.

Objective 3. Design, implement, and deploy a web-based visualization tool to explore species interactions, using the normalized datasets and the public web service API. The API will be made to easily integrate with EOL species pages and EOL's Ecosystem Explorer.

Objective 4. Design, implement, and deploy a tool to calculate specific properties of a food web (e.g., trophic connectance [13] and food resource diversity [14]). This tool will also be designed to easily integrate with EOL pages.

Objective 5. Provide public access to source code, datasets, and APIs.

Methods

The process of unleashing species interaction data can be broken down into the following activities: acquisition, normalization, integration, and consumption. A workflow diagram of these activities can be found in figure 1.

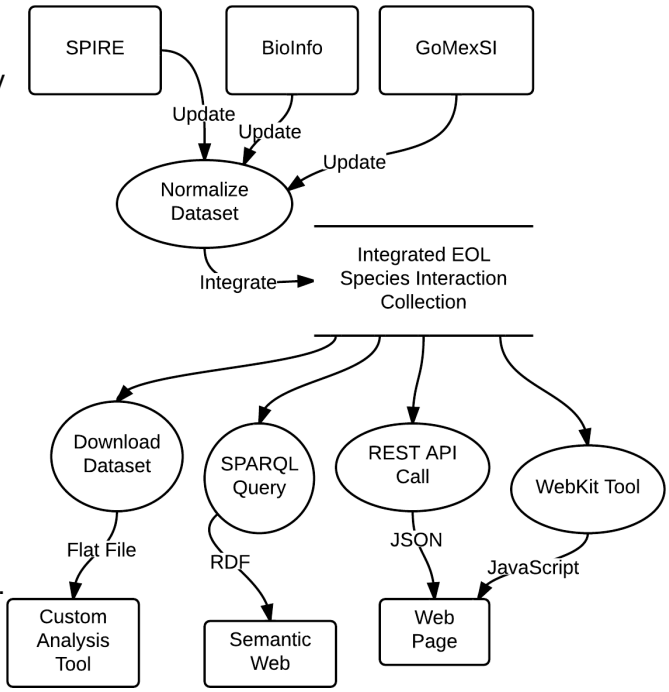

Figure 1 - Shows the normalized and integrated source datasets (SPIRE, BioInfo, and GoMexSI) being accessed by way of four methods: flat file, Representational State Transfer (REST) API, SPARQL Protocol and RDF Query Language (SPARQL), and a web toolkit.

Acquisition

EOL has previously imported species interaction datasets such as SPIRE and BioInfo into its catalog. Currently, the species interaction data is stored as unstructured text, which makes the data hard to parse algorithmically. Until EOL provides access to structured species interaction data, the original datasets will be acquired in digital form from trusted sources. The acquired datasets will then be stored in GitHub, an open source code repository. This storage strategy achieves three goals: First, it makes it possible to track changes made to the datasets. Second, it lets users access earlier versions of datasets even when newer versions are available. Third, it makes datasets easily accessible to others using the GitHub web interface.

Normalization

After species interaction datasets are acquired, they will be normalized. This means that the content of the datasets will be mapped to the species interaction data model, which will be derived primarily from existing data models [15]. A preliminary example of such a model can be seen in figure 2. In addition to what is shown in figure 2, our working model includes references to existing EOL content, such as species pages and images of species for visual identification.

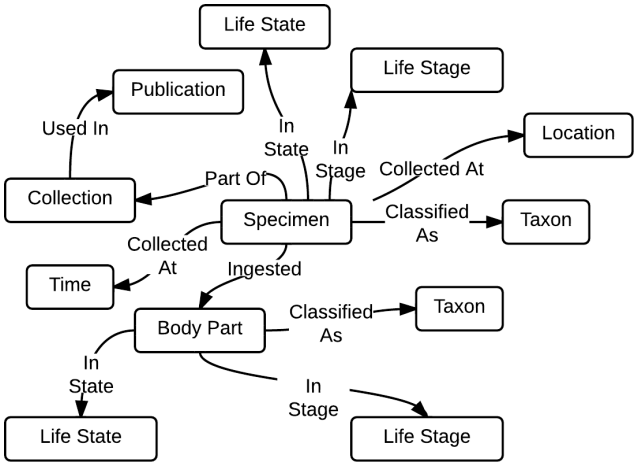

Figure 2 - A preliminary species interaction data model includes, but is not limited to, the relationships among predator, prey, location of collection, and time of collection.

We will use custom-built software programs to automate the normalization of the datasets into our species interaction data model. As part of this automated process, programmatic tests will be executed to verify that the normalization is working as expected.

The automated normalization process helps us to efficiently maintain the normalized dataset. With this process, we can quickly respond to emerging EOL wishes such as acquiring new datasets, extending our data models, and improving our normalization algorithms.

Integration

Once our datasets are normalized, they can be integrated into a single homogeneous dataset. The normalized and integrated dataset will allow us to build analysis tools, create data exports, or provide API access in a way that is decoupled from the original dataset. This means that whenever a dataset is added, updated, or removed, the consumer interfaces don't need to be modified. Since we expect constant updates of the datasets, this decoupling is essential for easy development.

Consumption

The normalized, integrated datasets provide the necessary basis for building species interaction data access, visualization, and analysis tools. The interaction data will be consumable in three ways. First, the full dataset will be downloadable in the form of a flat file. This export method will allow scientists to import and transform the dataset with specialized analysis tools. For instance, a flat file can be read into the R Project for Statistical Computing [16], SPSS [17], or another statistical analysis tool.

Secondly, parts of the interaction dataset will be accessible through both a Representational State Transfer (REST) API [18] and a SPARQL Protocol and RDF Query Language (SPARQL) endpoint [19]. This allows tech-savvy researchers or web developers to programmatically access relevant subsets of interaction data. For example, a web developer can use this functionality to enrich an educational website with species interaction data. The SPARQL endpoint makes the interaction data available as part of the semantic web [20]. This query interface is similar to the REST API endpoint, with the added benefit that the data is semantically interoperable. In other words, the data will not only be accessible by standard web transport mechanisms such as HTTP [21], but it will also be structured in a standardized form. This additional layer of standardization is expected to lower the barrier to linking into the species interaction dataset.

Lastly, two embeddable web tools will be made available. The first web tool will provide a way to explore a graphic representation of a food web, starting from a specific predator or prey species. The food web tool will also make it possible for the user to filter searches by location, time, or specific taxonomic rank. An example of this kind of visualization can be found in figure 3. In the online version, each node in the graph can be clicked on to reveal information about the species and specific sample collection data. This work could potentially enhance EOL's Ecosystem Explorer, or it could be integrated separately.

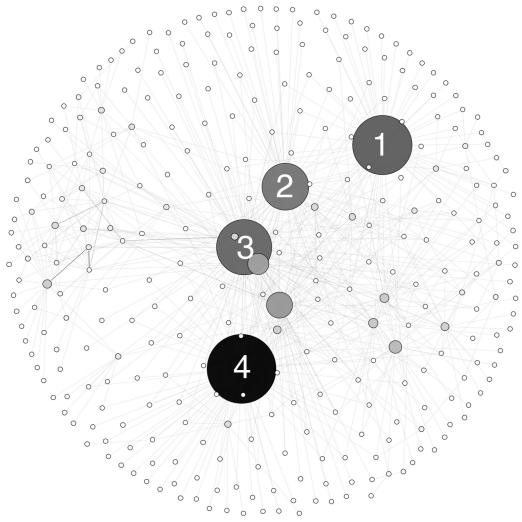

Figure 3 - The graph shown here was generated from two combined and normalized trophic datasets (about 15,000 nodes). The size and darkness of the predator species circle is proportional to the out-degree of its prey species within the dataset. (1=Coelorinchus caribbaeus, 2=Halieutichthys aculeatus, 3=Ariopsis felis, 4=Syacium papillosum.) The graph was created using Gephi 0.8alpha [22] and the Fruchterman-Reingold graph layout algorithm [23].

The second embeddable web tool will provide an information nugget related to a specific predator or prey species. This nugget will consist of one or more configurable food-web statistics such as trophic level, centrality, or taxonomic distance for a specific predator or prey species within a specific food web. Both exploration and statistical tools will be provided as a JavaScript library that leverages the web APIs mentioned above. This toolkit is intended to lower the threshold for web developers to integrate the functionality into their web pages. The JavaScript libraries will be available as an open-source project to encourage reuse and to provide a reference implementation for using the REST APIs and SPARQL endpoint.

Contributor	Institution	Study Title	Predator Species	Prey Taxa	Total Interactions	Period
David A. Blewett	Fish and Wildlife Research Institute, Florida Fish and Wildlife Conservation Commission	Feeding Habits of Common Snook, Centropomus undecimalis, in Charlotte Harbor, Florida	1	51	1130	Mar 2000-Feb 2002
Jenny L. Wrast	Department of Life Sciences Texas A&M University-Corpus Christi	Spatiotemporal And Habitat-Mediated Food Web Dynamics in Lavaca Bay, Texas	24	67	996	July 2006 - April 2007
Ivy E. Baremore	University of Florida, Department of Fisheries and Aquatic Sciences	Prey Selection By The Atlantic Angel Shark Squatina Dumeril In The Northeastern Gulf Of Mexico	1	45	1029	2005

Figure 4 - An example of a simple numerical analysis of a normalized and integrated dataset extracted from a working proof of concept based on a preliminary GoMexSI dataset [24].

Technologies

Normalization and integration algorithms and REST API will be implemented in Java. An instance of the normalized and integrated data will be hosted in a graph database, Neo4j [25], to shorten response times of graph algorithms, such as a calculation of the shortest path between two nodes. A triple store, Virtuoso [26], will host another instance of normalized and integrated data to provide a SPARQL endpoint. All software used and created will be available to the general public at no cost. Our source code, datasets, and deployment infrastructure will be hosted on GitHub, and will be deployable on any platform that is supported by Java 1.6, including, but not limited to, consumer laptops and cloud platforms such as Amazon EC2 [27]. We expect that an initial deployment on Amazon EC2 will gradually transition to the EOL infrastructure.

Proof of Concept

Most methods and technologies described in this proposal have been implemented as a proof of concept to normalize and visualize GoMexSI datasets [24]. This proof of concept uses a Ruby on Rails [24] web server in combination with a Neo4j graph database to present normalized trophic data in web pages. The heterogeneous GoMexSI trophic datasets were normalized using a custom Java application that integrates with National Center for Biotechnology Information Taxonomy [28], Integrated Taxonomic Information System [29], EOL, and WoRMS [30] APIs. The proof of concept did achieve our basic goal of providing easy access to a normalized trophic dataset, demonstrating that proposed methods and technologies are viable.

References

1. **Rubenstein 2013 Wish #6**, [http://eol.org/data_objects/21913492].
2. Savage VM, Webb CT, Norberg J (2007) **A general multi-trait-based framework for studying the effects of biodiversity on ecosystem functioning.** Journal of theoretical biology, 247(2), 213.
3. McCann K (2007) **Protecting biostructure.** Nature, 446(7131), 29-29.
4. Parr CS, Parafiynyk A, Sachs J, Ding L, Dornbush S, Finin, T, Wang D, Hollander, A (2006) **Integrating ecoinformatics resources on the semantic web.** Proceedings of the 15th international conference on World Wide Web, pp. 1073-1074. ACM.
5. **BioInfo**, [http://bioinfo.org.uk].
6. Simons JD, Yuan M, Carollo C, Vega-Cendejas M, Shirley T, Palomares MLD, Roopnarine P, Abarca Arenas L, Ibañez A, Holmes J, Mazza C, Hertog R, Reed D, Poelen JH (2012). **Building a fisheries trophic interaction database for management and modeling research in the Gulf of Mexico large marine ecosystem.** Bulletin of Marine Science, Under revision.
7. **Encyclopedia of Life**, [http://eol.org].
8. **PHP: Hypertext Preprocessor**, [http://www.php.net].
9. **Ruby on Rails**, [http://rubyonrails.org].
10. **MySQL**, [http://www.mysql.com].
11. **Apache Solr**, [http://lucene.apache.org/solr/].
12. **Resource Definition Framework (RDF),** [http://www.w3.org/RDF/].
13. Warren PH (1989) **Spatial and temporal variation in the structure of a freshwater food web.** Oikos 55:299–311.
14. Simpson EH (1949) **Measurement of diversity.** Nature 163:688.
15. Mungall CJ, Torniai C, Gkoutos GV, Lewis SE, Haendel MA (2012). **Uberon, an integrative multi-species anatomy ontology.** Genome Biology, 13(1), R5.
16. The R Project for Statistical Computing, [http://www.r-project.org].
17. SPSS, [http://www-01.ibm.com/software/analytics/spss/products/statistics/].
18. Richardson L, Ruby S (2007) **RESTful Web Services.** O'Reilly Media, Inc.
19. Prud'hommeaux E, Seaborne A (2008) **SPARQL Query Language for RDF.** [http://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/].
20. Bizer C, Heath T, Berners-Lee T (2009) **Linked data-the story so far.** International Journal on Semantic Web and Information Systems (IJSWIS) 5, no. 3: 1-22.
21. Berners-Lee T, Fielding R, Frystyk H (1996) **Hypertext Transfer Protocol -- HTTP/1.0**, [http://www.ietf.org/rfc/rfc1945.txt].
22. **Gephi**, an open source graph visualization and manipulation software, [http://gephi.org]
23. Fruchterman T, Reingold E (1991) **Graph drawing by force-directed placement. Softw.** – Pract. Exp., 21(11):1129–1164.
24. Poelen JH, Reiz R, Simons JD (2012) **Preliminary GoMexSI Dataset Webbased Visualization**, [http://trophic-graph.herokuapp.com].
25. **Neo4j**, a graph database [http://www.neo4j.org].
26. **Vistuoso**, a RDF triple store, [http://virtuoso.openlinksw.com].
27. **Amazon Elastic Cloud**, [http://aws.amazon.com/ec2/].
28. **National Center for Biotechnology Information Taxonomy**, [http://www.ncbi.nlm.nih.gov/taxonomy].
29. **Integrated Taxonomic Information System**, [http://www.itis.gov].
30. **World Register of Marine Species**, [http://www.marinespecies.org].