Serving computational ecology from a digital library

Cynthia Sims Parr
Human-Computer Interaction Lab
University of Maryland
College Park, Maryland 20742
1-301-405-7445
csparr@umd.edu

Roger Espinosa
IT Central Services
University of Michigan
Ann Arbor, MI 48105
1-734-763-4677
roger@umich.edu

Philip Myers
Museum of Zoology
Dept of EEOB, Univ. of Michigan
Ann Arbor, MI 48109
1-734-647-2206
pmyers@umich.edu

ABSTRACT
We describe a case study using a digital library resource to assist ecological research that involves computational approaches. Our purpose is to detail the approach and demonstrate the power of combining encyclopedic content presentation with harvestable data. While acknowledging the advantages and generality of this approach, we also consider the challenges faced before digital libraries can adequately support research in this way.

Categories and Subject Descriptors
H.5.4 [Hypertext/Hypermedia]: Information interfaces and presentation (e.g., HCI) -- *User issues*

General Terms
Design, Standardization

Keywords
Biodiversity, ecology, digital library, encyclopedia

1. INTRODUCTION
While efforts to design, implement, and populate digital libraries for education and literature access are well underway (e.g. NSDL and DLESE), effective use of them for scientific research is not yet common practice. Notable leaders are Unidata, e.g. THREDDS [1], and FishBase (http:///www.fishbase.org). Research taking a synthesizing approach typically involves manual coding of data from tables or text found via literature searches, or downloading and integrating data from multiple, specialized data archives. How can digital libraries improve the process of compiling data for these studies?

We describe a computational approach to ecological interaction analysis, and present results of integrating data from a digital encyclopedia and data archives to support this analysis. Deepening the contribution of digital libraries to such research will require thoughtful structuring and exposure of data to facilitate discovery, export, and integration.

2. BACKGROUND
When one organism regularly eats, parasitizes, or benefits another organism in its community, they have an *ecological interaction*. A food web is a well-known example of a collection of ecological interactions. We chose this domain for testing ideas on the use of digital libraries in scientific research for several reasons. This field has a number of well-established datasets, a history of synthetic studies, and recent theories that are amenable to computational approaches (reviewed in [2]). Also, food webs provide an example familiar to non-biologists.

Our primary digital library resource in this case study is the Animal Diversity Web (ADW) (http://www.animaldiversity.org). Initially designed for education by zoologists at the University of Michigan, this online multimedia collection includes descriptions suitable for general audiences about the physical and reproductive characteristics, behavior, conservation status, and ecological interactions of animals. Coverage is intended to be geographically and taxonomically comprehensive; rich information is currently available for several thousand species. The content management system [3] allows both experts and non-experts (e.g. undergraduates) to build the digital encyclopedia in a highly structured and highly readable manner.

3. APPROACH AND IMPLEMENTATION
Our science goals are to 1) investigate general ecological interaction rules in known food webs, and 2) predict interactions in less-well-known food webs. We begin by combining large numbers of known food webs in a relational database, as described in more detail below. These data on "who eats whom" can come partly from data archives of the results of particular studies, but can also come from aggregated summaries in digital encyclopedias such as ADW. Interactors are identified where possible to the scientific name at the most appropriate taxonomic level. This allows data from different sources to be combined, using scientific names. Additional data tables with traits or attributes such as size, habitat preferences, reproductive characteristics, and nutritive requirements allow the construction of "trait-space" for each organism. Visualization tools, under development, will allow biologists to explore the data for patterns or to select subsets for analysis. Algorithms, to be discussed elsewhere, involve predictive modeling using trait-spaces and inferences across related organisms. Once parameterized by well-studied systems, these algorithms will generate testable hypotheses about unstudied systems.

Our approach requires large quantities of data to be brought together into a single analysis which should expand as new results are added to digital libraries. It does not rely on particular algorithms, but is essentially a blueprint for the workflow of data gathering, analysis, and predictions.

We obtained delimited ASCII or spreadsheets directly from researchers (Webs on the Web, EcoWEB) or from a public data archive (Interaction Web Database, http://www.nceas.ucsb.edu/interactionweb/). Common name searches using ADW, TaxonTree [4], FishBase, ITIS (http://www.itis.usda.gov), and other online sources aided identifications of interactors to scientific name. These sources include both animal and non-animal interactors.

We also obtained delimited ASCII for the entire structured contents of the Animal Diversity Web. This included lists of animal predators and their prey (predator-prey links) in addition to quantitative data such as lifespan or size, as well as natural history keywords applying to each scientific name. These attribute data use a controlled vocabulary associated with an OWL ontology [3]. ADW's controlled vocabulary structured the coding of non-standardized portions (such as location and habitat of food web site) from the other datasets.

4. RESULTS

ADW contributed over 30,000 attribute records (Table 1), representing the distillation of about 10,000 references, compiled by about 1400 authors. A comparison with specialized archive data shows the relative contribution of a digital encyclopedia to predator-prey interaction data (Table 2).

Table 1. Large amounts of structured data can be downloaded from ADW. The 6 most populated, relevant categories are shown.

Attribute category	# records
Reproduction keywords	9858
Habitat keywords	5799
Physical characteristics keywords	4174
Behavior keywords	4170
Food habits (e.g. trophic levels)	3000
Size	1819

Table 2. ADW supplements data from 3 food web data archives.

Source	#webs	#interactors	#links
ADW	n/a	1012	2869
Webs on the Web	17	1537	6328
Interaction Web DB	26	2177	9882
EcoWEB	213	4064	6363

Specialized data archives are compiled by scientists directly from peer-reviewed scientific literature, and thus are likely to be higher quality than ADW, and less sparse. Formats were not difficult to standardize manually. However, in most old and some new food webs interactors were not already identified to scientific name, which posed a significant challenge.

The schema for the ecological interaction analysis database is available at http://www.cs.umd.edu/hcil/biodiversity. ADW data can be obtained in various machine-readable formats using http://scoobydoo.us.itd.umich.edu:8099/dogsled/tools/ui9/inquiry.

5. DISCUSSION

Our approach generalizes to most comparative studies using compiled data. An already aggregated resource such as ADW has disadvantages. One must trust the coding that others have done, which may be subject to hidden biases (though ADW's authoring model should randomize errors). Coding schemes for such an all-purpose resource may not be as effective as a taxon-specific dataset (e.g. focusing only on birds), or one with coding tailored to answer a specific research question.

At the same time, there are many advantages to using a digital encyclopedia. Data are easier to explore before downloading. Compiling data is less time consuming because data are pre-aggregated according to a single standard. Fewer mappings of schema are required in order to integrate the data with other sources. Coding can be checked against accompanying text and references in the encyclopedic source. As digital library collections grow, analyses can be rerun with more data or with additional attributes. Importantly, digital encyclopedia data also serve education and outreach purposes.

Digital encyclopedias can never replace high-quality, specialized archives, but ADW can serve as a model for encyclopedic resources. Currently one cannot easily find nor retrieve the data we used via the National Science Digital Library, though ADW metadata is available there. We recommend that digital collections in general expose data to harvesting and discovery by indexing controlled vocabulary terms, not just the general metadata. Semantic web approaches to data discovery and integration, such as those pursued by the SPIRE project (http://spire.umbc.edu) are also promising for ecological research.

6. ACKNOWLEDGMENTS

Our thanks to J. Cohen (EcoWEB), J. Dunne and N. Martinez (Webs on the Web), for providing food web data, to B. Fagan for discussions on "trait-spaces," to B. Lee for help with the database schema, and to T. Jones and T. Dewey for helpful comments on the manuscript. Funding from NSF IDM/ITR 0219492 (PI Bederson) and IERI REC-0089283 (PI's Songer and Myers).

7. REFERENCES

[1] Domenico, B., Caron, J., Davis, E., Kambic, R., and Nativi, S. Thematic real-time environmental distributed data services (THREDDS): Incorporating interactive analysis tools into NSDL. *Journal of Digital Information, 2,* 4 (2002), No. 114.

[2] Pimm, S. *Food webs.* University of Chicago Press, Chicago, IL, 2002.

[3] Parr, C.S., Espinosa, R., Dewey, T., Hammond, G., Myers, P. Building a biodiversity content management system for science, outreach, and education. Submitted to *Data Science Journal*. <http://animaldiversity.org/site/about/technology/>.

[4] Parr, C.S., Lee, B., Campbell, D., and Bederson, B. Tree visualizations for taxonomies and phylogenies. *Bioinformatics, 20,* 17 (2004), 2997-3004.