Harnessing museum resources for the Census of Marine Life: the FISHNET project

David Vieglais, E.O. Wiley, C. Richard Robins and A. Townsend Peterson
The University of Kansas • Lawrence, Kansas USA

FISHNET[1] is a distributed information system that seeks to provide data towards answering two fundamental questions regarding marine biodiversity: how many vouchered records exist for each species of fishes, and what are their distributions in time and space? Recent advances in information technology make possible efficient access to one of the major resources for such knowledge, and the only resource in which data are backed up by actual specimens of the organisms: natural history museums.

Scientific collections housed at world natural history museums constitute the most complete and authoritative record of global biodiversity (PCAST, 1998; Krishtalka and Humphrey, 1998, in press). The value of scientific collections lies in the fact that records are directly tied to actual voucher specimens. Changes in taxonomic names and discovery of new species can always be related to actual specimens which are available to all qualified investigators. Efficient access to such data can furnish investigators with baseline data regarding where and when collections have been made. Clearly, this information provides an important tool in planning future collecting efforts. Combining these data with environmental data and sophisticated artificial-intelligence algorithms (see Stockwell and Noble, 1991; Stockwell and Peters, 1993) makes it possible to predict species' distributions (Peterson et al., 1999), study the possible impact of the introduction of exotic species, and to model effects of environmental changes such as global climate change. The purpose of FISHNET is seamless integration of diverse data sets available to all scientists.

FISHNET is a consortium of natural history museum fish collections that have agreed to share specimen data openly among all users. As a distributed information system, the data reside on computer servers at each participating institution who thus maintain ownership and stewardship of the data. Because data are served on request to remote users, FISHNET is not an entity but a cooperative community.

The concept of sharing fish specimen data is an old idea that has grown out of the earlier initiatives such as FISHGOPHER[2], MUSE[3] and NEODAT[4] projects. The fish community is almost unique among communities in the extent to which museum holdings have been captured in electronic form and a willingness to share data over the Internet. This particular initiative had its origins when Vieglais began exploring the application of the ANSI/NISO Z39.50 information transfer protocol to overcome key impediments to a true distributed information system. Vieglais, Peterson, and colleagues were developing the North American Biodiversity Information Network (NABIN), a network that has concentrated on distributed information for the North American flora and fauna, in a system called The Species Analyst. FISHNET databases are fully integrated with NABIN databases through a common implementation of ANSI/NISO Z39.50 standards and common use of The Species Analyst.

Direct access to museum data requires their entry in electronic databases. The ichthyological community has been active for over twenty years in electronic capture of specimen data through such initiatives as the MUSE and NEODAT projects. These early steps have positioned the community to play a leadership role in making museum data accessible. However, their full use has been hindered by lack of efficient mechanisms for search and retrieval of information from geographically scattered databases with idiosyncratic database structures. Even though many FISHNET partners make

> *FISHNET is a consortium of natural history museum fish collections that have agreed to share specimen data openly among all users.*

---

[1] http://habanero.nhm.ukans.edu/fishnet/
[2] http://www.ncodat.org/dbs/fgtop.htm
[3] http://www.neodat.org/dbs/MUSEtop2.htm
[4] http://www.neodat.org/

Session Results

#	Query	Targets	# Hits	Summary	BSW	Download
0	@attr 1=1 @attr 2=6 !Cyclothone pallida%! @attr 1=1 @attr 2=6 !!	FloridaFish(32) UMMZFish(2) TulaneFish(0) MCZFish(475)	509	Summary	New	HTML CSV SHP XML Excel

Summary Information

Summarizes the results from this query by listing the unique scientific names, the number of records (actually the number of records with valid year entries) and the earliest and latest years of collection for each name. A direct link to the ITIS and GenBank (nucleotide or protein) databases is provided. Clicking on those links will open a new window.

Scientific Name (ITIS Link)	Num. Records	Earliest Year	Latest Year	GENBANK Link	Zoo Record
Cyclothone pallida	509	1891	1995	Nucleotide Protein	ZR

Distribution Map

This distribution map provides an indication of the global distribution of collection sites for the records identified by your query.
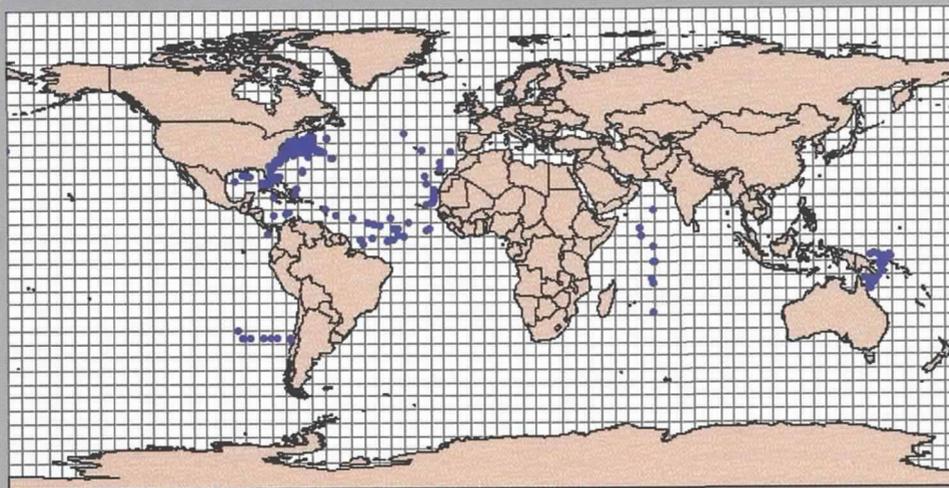

Figure 1. Results of a FISHNET/Species Analyst query of four museum collections for the mesopelagic fish Cyclothone pallida. Four collections have a total of 504 lots of specimens. The Summary page provides a HTML spot map of geo-referenced lots as well as links to the Integrated Taxonomic Information System (IT IS), Genbank, and the Zoological Record. All data (including lots not geo-referenced) can be downloaded to the client computer in a variety of formats: HTML, XML, tab-delimited (CSV), Excel spread sheets, and shape files for GIS applications (SHP).

their data available for searches, single databases rarely contain enough information for detailed analysis of a particular taxon or region, requiring investigators to search each database and compile information piece-by-piece. Some projects, such as NEODAT II, have fulfilled this promise on local scales, but widespread implementation for numerous collection databases is only now being addressed by the community.

As a distributed database system, FISHNET has several advantages. It does not require partners to have data in the same database structure or to use the same database management system. The distributed structure of the network allows partner institutions to maintain control of their data locally, and make available only those data they wish to be queried. Not only are the data distributed, but also the work, with most of the computing burden borne by local data servers, allowing for almost infinite expansibility without slowing down the system. Further, the data are always current since they are available as soon as entered or as soon as the database replica is updated. Finally, as a component of the NABIN partnership, composed of databases for other taxa (mammals, birds, insects, plants, etc.), making cross-taxon queries and analyses is possible. For example, we can query for fish species and for other taxa simultaneously, making possible detection of correlations among organisms, such as

predator and prey, effects of abundance on communities, and other cross-taxon effects.

As a distributed database network, FISHNET has some characteristics of which users should be aware (and which are detailed on the Species Analyst website). The information provided is only as good as the information associated with the specimens and its manifestation in the database. Different collections and different parts of a single collection may vary in quality of taxonomic identifications associated with specimens. Some collections are extensively geo-referenced, others are not, and some geographic references can be erroneous. FISHNET data are no replacement for thorough systematic and taxonomic study of a group, nor are they a replacement for looking at specimens.

How does it work?

A detailed understanding of the structure of distributed database technology is neither necessary nor warranted, but a superficial description may help clarify how the network operates. FISHNET is based on the same technology as that commonly used for bibliographic searches; ANSI/NISO Z39.50 (NISO, 1995; Lynch, 1997),[5] which defines a standard way for two

[5] http://www.niso.org/z3950.html

computers to communicate for the purpose of information retrieval. A computer operating as a "client" submits a request to another computer acting as a "server." Software on the server performs a search on the database, generates a set of records that meets the criteria of the search request, and returns the results to the client for processing.

Distributed data networks that use ANSI/NISO Z39.50 technology are able to share information because they share a common core of information fields that defines the kinds of data that are being searched. FISHNET and NABIN use the Darwin Core[6] as the criterion for searching databases. The Darwin Core is a list of fields common to all museum specimen records: genus name, species epitaph, date of collection, collectors, fields specifying the locality, latitude and longitude, and higher taxonomic names such as family, order, etc. The Darwin Core is dynamic, so a number of fields specific to marine biodiversity will be added, such as ocean basin, and depth. Not all museum records carry all information, nor is it necessary that they do so in order to be queried.

A search is initiated from a client computer using a software program called a "Z-Client application." The application currently used by FISHNET is The Species Analyst Web Interface, developed by Vieglais at the University of Kansas.[7] Included in The Species Analyst Web Interface are options for searching distributed databases for particular items of information. The Species Analyst processes the request and broadcasts it to all linked databases requested by the user. When a linked server (termed a "Z-server") receives the request, a program installed on the server searches its associated database(s), compiles and sorts the information, and returns it to The Species Analyst application. The Species Analyst can then process the information into a format requested by the user or in a format specified by the client software application.

The Z-server software provides a layer of abstraction based on the definitions in the Darwin Core that provides a common interface to the database(s) being served, and a common result set structure for the information being returned. Imagine a single Z-server serving two databases, one for fishes and another for corals. Now, imagine that although these databases contain similar kinds of information, their data structure is different (i.e, they are idiosyncratic relative to each other), or perhaps they are entered into different database management programs. In this case, we would have two pieces of Z39.50 software, each configured to read its associated database, and each configured to put the data into a common format as defined by the Darwin Core. Both packages of information would then be returned to the client computer.

There are two ways of accessing FISHNET. The method that is most accessible at this time is through the Species Analyst Web Interface. At The Species Analyst website, the application processes search requests and holds the results in a temporary cache where the information is available for further processing. For example, the user can preview a summary of the data by viewing an HTML spot map and explore

---

[6] http://habanero.nhm.ukans.edu/Z.X/documents/DarwinCore
[7] Information on The Species Analyst is available at http://habanero.nhm.ukans.edu
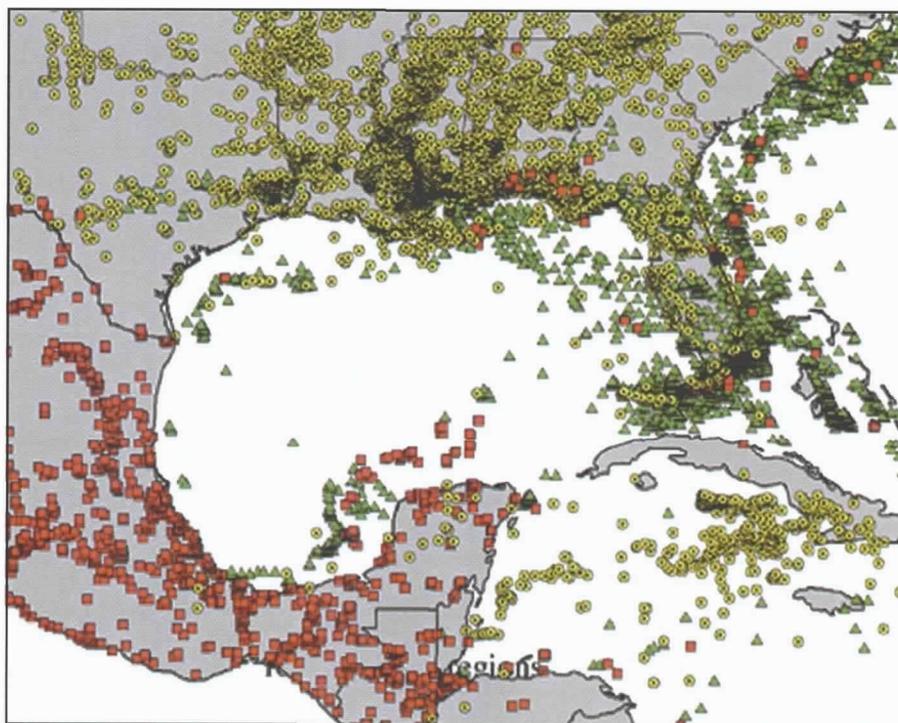

Figure 2. Results of a FISHNET/Species Analyst query in ESRI ArcVeiw of three museum collections for collecting events centered around the Gulf of Mexico. Collections queried are Tulane University (yellow circles), Florida Museum of Natural History (green triangles), and University of Michigan Museum of Zoology (red squares). Each collecting event includes one to many different species.

links to the Integrated Taxonomic Information System (ITIS), GenBank, and the Zoological Record. In addition, the user has a number of options for downloading data; as HTML files, Excel® files, text delimited files, and GIS shapefiles for use in applications such as Excel® and ESRI ArcView® where the data can be further processed and analyzed by the user.

The second method of accessing FISHNET is by attaching The Species Analyst directly as an extension to an application program such as Excel® and ESRI ArcView®. The user interface is similar to that available at the Species Analyst web site, but the data returned by the linked Z-servers are returned directly into the particular application in a format that can be used for further analysis such as a spreadsheet in Excel, or a data table in ArcView®. Since this method eliminates an extra step in the retrieval process, it is much faster than going through the website.

Uses of FISHNET

Within the next year, FISHNET will consist of a consortium of 21 museums and their associated databases, providing access to more than 3 million specimen lots and more than 30 million fish specimens. We hope that, as community consensus builds, we will add even more museums to create a virtual "world museum" that can be accessed by the entire scientific community as well as the interested public. We assume that FISHNET will be a boon to ichthyologists, evolutionary biologists, and marine biogeographers as well as fisheries scientists and environmental managers who can use FISHNET resources to evaluate sampling effort, plan nature pre-serves, assess threats for invasive species, and study the effects of global climatic changes on the marine biota.

> *Within the next year, FISHNET will consist of a consortium of 21 museums and their associated databases, providing access to more than 3 million specimen lots and more than 30 million fish specimens.*

REFERENCES:

Krishtalka, L. and P.S. Humphrey, 1998: Fiddling while the Planet Burns: The Challenge for U.S. Natural History Museum. *Museum News*, 77 (2): 29-35.

Krishtalka, L. and P.S. Humphrey. In press. Can Natural History Museums Capture the Future? *BioScience*, URL:http://www.dlib.org/dlib/march96/briefings/03indexdata.html.

Lynch, C., 1997: The Z39.50 Information Retrieval Standard. *D-LIB Magazine*, April 1997. URL: http://www.dlib.org/dlib/april97/04lynch.html.

NISO, 1995: Information Retrieval Protocol (Z39.50): *Application Service Definition and Protocol Specification.* NISO Press, Bethesda, MD. Available in electronic form at the Z39.50 Maintenance Agency (http://lcweb.loc.gov/z3950/agency.)

PCAST, 1998: *Teaming with Life: Investing in science to understand and use America's living capital.* President's Committee of Advisors On Science and Technology: Panel on Biodiversity and Ecosystems. 86 pp.

Peterson A.T., J. Soberón and V. Sánchez-Cordero, 1999: Conservatism of Ecological Niches in Evolutionary Time. *Science*, 285: 1265-1267.

Stockwell, D.R.B. and I.R. Noble, 1991: Induction of sets of rules from animal distribution data: a robust and informative method of data analysis. *Mathematics and Computers in Simulation*, 32:249-254.

Stockwell D.R.B. and D.G. Peters, 1993: Artificial intelligence predicts biodiversity. *ERINYES*, 18. 📀

OCEANOGRAPHY

SOCIETY · 1988