Toward species interaction networks – Managing, visualizing and synthesizing Gulf of Mexico geo-spatial trophic data

James Simons[1], May Yuan[2], Cristina Carollo[3], Cristina Mazza[4], Sara Gonzalez-Perez[2], Lesley Williams[2], Derek Morris[2], Dave Reed[4], Maru Vega Cendejas[5]
[1]Center for Coastal Studies, Texas A&M University-Corpus Christi
[2]Center for Spatial Analysis, Oklahoma University
[3]Harte Research Institute, Texas A&M University-Corpus Christi
[4]Fish and Wildlife Research Institute, Florida Fish and Wildlife conservation Commission
[5]El Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional
James.simons@tamucc.edu, myuan@ou.edu, Cristina.Carollo@tamucc.edu, cristina.mazza@MyFWC.com, gops@ou.edu, Lesley.N.Williams-1@ou.edu, soonernation1@hotmail.com, Dave.Reed@MyFWC.com, maruvega@mda.cinvestav.mx

***Abstract*--The last 20 years has witnessed the collection, documentation, and storage of massive quantities of biodiversity data. However, interactive networks between species that define and characterize the world's ecosystems have largely been ignored. Continued development of ecoinformatics is critical to species interaction research. Many interactions exist among species including predator-prey, competition, host-parasite, symbiosis and others. Knowledge of these interactions within an ecosystem context allows us to predict the consequences of changes in biodiversity, e.g. trophic cascades. We report on progress toward development of a model database of one type of interaction – predator/prey. Our model ecosystem is the Gulf of Mexico. Trophic data for the Gulf will be extracted from published and unpublished sources and contributed databases. Metadata "lite" has been collected for ~720 trophic references, ~650 have been geocoded and habitat information has been digitized for ~420 references. We anticipate using this network of species interactions in the context of a dynamics geographic information system (GIS) to link biodiversity data collections together in time and space. In addition, creation of an ecosystem based trophic database will have applications toward further development of food web theory, ecosystem-based fisheries models, and directed network research.**

Keywords—trophic, food web, Gulf of Mexico, metadata, CMECS, database, ecoinformatics, Gulf GAME

I. Introduction

Collection, documentation and storage of massive quantities of biodiversity data, including archiving of museum specimens and biodiversity data has evolved and amplified over the past 20 years, yet species interaction networks have largely been ignored [1]. To date, species interaction networks have been studied with small databases in the context of a very low taxonomic, spatial and temporal resolution [2, 3]. However, as larger databases are generated, ecoinformatics research will be critical to advancing the understanding of interactions among species and between species and the environment.

Museum specimen databases (e.g. FishNet2 [4], Ornithological Information System (ORNIS) [5] and others), global biodiversity databases (e.g. FishBase [6], SealifeBase [7], Encyclopedia of Life (EOL) [8] and others), and projects facilitating the use of biodiversity data (e.g Knowledge Network for Biocomplexity (KNB) [9], Global Biodiversity Information Facility (GBIF) [10], Census of Marine Life (CoML) [11] and others) provide a limited amount of species interaction data. The Interaction Web Database [12] and Webs on the Web [13] are species interaction databases for select ecosystems with only presence/absence data for that does not include interaction strength, habitat, environmental, spatial or temporal data. NOAA's Food Web Dynamics Program (FWDP) at Woods Hole, MA [14] and Resource Ecology and Ecosystem Modeling (REEM) in Seattle, WA [15], who's missions include collection, analysis and modeling of trophic interaction data, each have large collections of food habits data on slightly more than 100, mostly commercial, fish species.

Ecoinformatics emphasizes conceptual and practical tools for the understanding, generation, processing and dissemination of ecological data and information [16]. High performance computing, biologically inspired computation, object oriented data, and the internet frame informatics for ecological modeling to integrate climate, environmental, community, phenotypic and genomic data [17, 18]. Ecoinformatics explicitly recognizes the heterogeneous nature of ecological data and seeks to develop tools that consider simultaneously the high resolution and heterogeneity of the data and create added value to large volumes of data at multiple biological levels and spatial scales. Informatics research has resulted in the development of BioGeomancer [19], Lifemapper [20],

This work is licensed under a Creative Commons Attribution 3.0 Unported License (see http://creativecommons.org/licenses/by/3.0).

Aquamaps [21], Webs on the Web (WoW) [13], Interaction Web Database [12] and Ocean Biodiversity Informatics (OBI) [22].

Advances in ecoinformatics depend fundamentally upon database architectures that can represent entities involved in a system and the system structure across multiple taxonomic, spatial, and temporal resolutions. Key challenges posed by trophic dynamics data provide excellent ecological cases for database architecture development. The proposed research will build a trophic database for the Gulf of Mexico (GoM) to support theoretical advances in trophic dynamics. Despite the fact that many data are collected at a high level of spatio-temporal resolution (i.e., individual or size class level in each specific habitat) food web studies are not detailed, and most theory has been developed at species level (or higher) in homogeneous environments [2, 3]. This has inhibited the development of unified datasets and tools to aid development and testing of flexible, first principle, individual-based models able to explore consequences of individual variability and spatio-temporal heterogeneity of raw data which will advance the understanding of ecosystems.

II. Development of the Proposed Database

A. *Database Architecture and Development*

A spatio-temporal database architecture for ecological interactions will be designed to account for the heterogeneity of trophic data. The complexity and diversity of the data creates challenges in building an ecoinformatics database. Because our approaches to data representation and organization will center on complex system processes and ecological interactions, as well as account for data heterogeneity, the database architectures developed will be transferable to other ecological domains.

We will adopt Hierarchy Theory to develop database architectures that address common ecological issues, such as grain and scale, identification of entities, levels of dynamics, and disturbances [23]. Hierarchy by definition imposes ordinations, as from smaller to larger, or from simpler to more complex. These concepts from Hierarchy Theory are central to many complex systems, including ecological systems and weather systems [24]. Database architectures built upon these concepts will provide rich grounds for data mining and knowledge discovery of higher level concepts [25].

Database architecture includes two components: (1) representation of reality; and (2) organization of data. The first component concerns what concepts or objects need to be represented in the database and how to most effectively represent these concepts or objects in database models. Because our proposed research aims to integrate spatial and temporal information for ecological interactions, we need to represent spatial and temporal characteristics of the identified concepts or objects. The second component addresses how different sets of data, such as species, habitat, sea surface temperature, management zones, *etc*, should be organized in the ecoinformatics to support modeling efforts that relate multiple variables to derive new understanding or forecasting. Both components of data representation and data organization need to account for complexity and diversity of ecological systems and the nature of potential data sources.

B. *Data Sources, Acquisition, and Quality Assurance*

This project will encompass the marine and estuarine waters of the GoM, along the United States, Mexico and Cuba. Species that inhabit the Gulf region and its waters for at least part of their life cycle will be included, *eg*. taxonomic groups listed in Table 1. Habitats covered include estuaries and continental shelf as well as the pelagic, mesopelagic, continental slope, and abyssal realms.

TABLE 1, TAXONOMIC GROUPS TO BE INCLUDED IN THE GOM TROPHIC DATABASE AND THE CURRENT STATUS OF IN-HAND AND PERCEIVED REFERENCES ADDRESSING FOOD HABITS.

Taxonomic Group	**Number of References in Hand**	**Estimated References Available**	**Total Species Currently Cited with Diet Data**
Marine Mammals	3	25	1
Sea Turtles	9	10-15	3
Fishes	721	740	~650
Sea and Shore Birds	4	100-200	4
Crustaceans	19	25-50	58
Mollusks	3	25	45
Polychaetes	~25	100-200	99
Ctenophores	5	10	2
Cnidarians	5	10	6

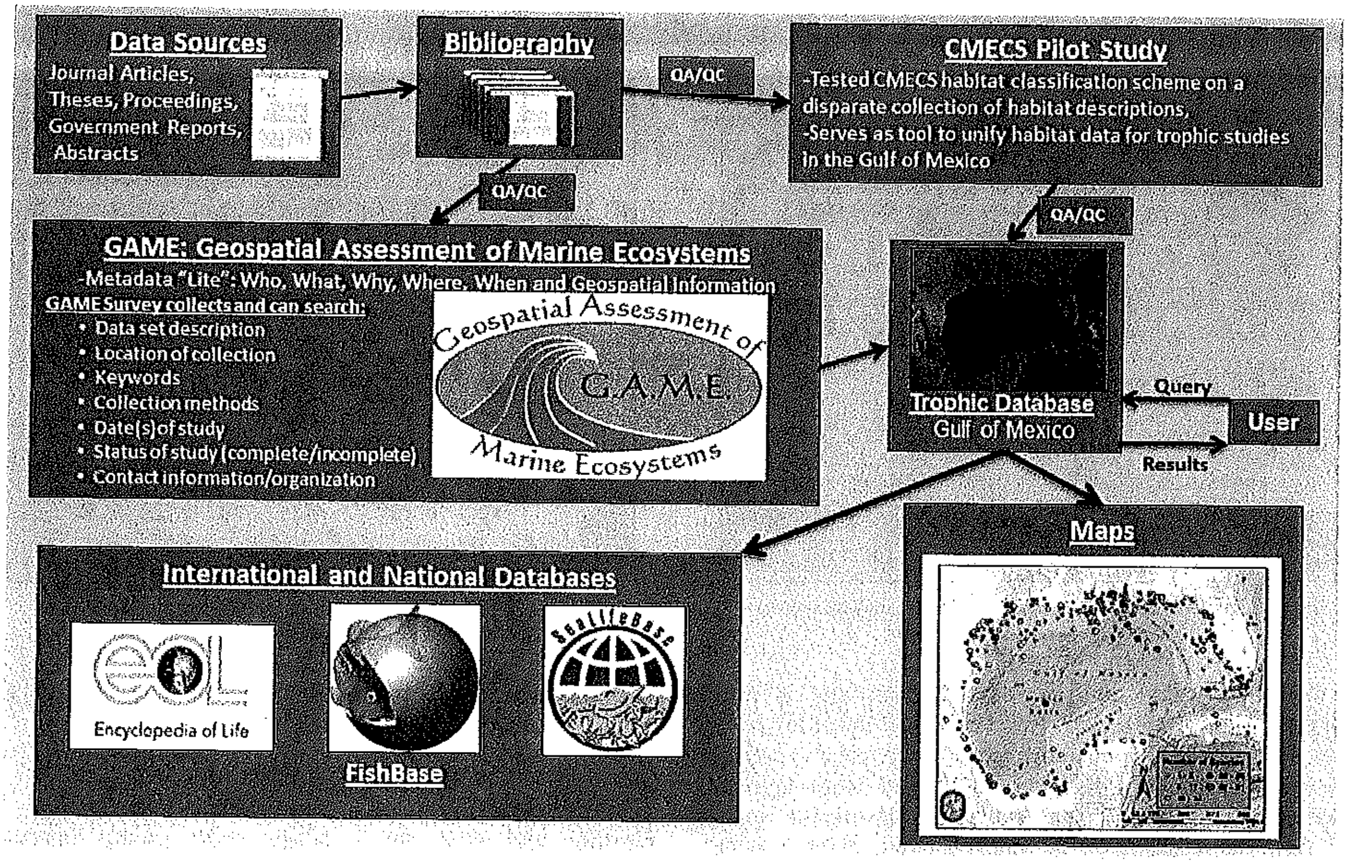
Data Sources
Journal Articles, Theses, Proceedings, Government Reports, Abstracts
Bibliography
QA/QC
CMECS Pilot Study
-Tested CMECS habitat classification scheme on a disparate collection of habitat descriptions,
-Serves as tool to unify habitat data for trophic studies in the Gulf of Mexico
QA/QC
QA/QC
GAME: Geospatial Assessment of Marine Ecosystems
-Metadata "Lite": Who, What, Why, Where, When and Geospatial Information
GAME Survey collects and can search:
• Data set description
• Location of collection
• Keywords
• Collection methods
• Date(s) of study
• Status of study (complete/incomplete)
• Contact information/organization
Geospatial Assessment of
G.A.M.E.
Marine Ecosystems
Trophic Database
Gulf of Mexico
Query
User
Results
Maps
International and National Databases
eol
Encyclopedia of Life
FishBase
SeaLifeBase

Figure 1. Schematic of the GoM trophic database workflows, links and outputs.

The following categories of data will be extracted from each source, when provided: Geopolitical location, Geospatial areas, Habitat, Geographic location, Time, Physico-chemical data, Collection method, Taxonomy, Specimen data, Food description, Stable isotopes and Source. Draft metadata fields as well as data and function requirements analysis will be developed. The database schema will follow the Ecological Metadata Language (EML) [26], an Extensible Markup Language (XML)-based metadata specification, and OBIS schema to ensure we structure marine data properly (Fig. 1). As part of this process, we will contribute metadata standards for trophically related data.

Data will be extracted from peer reviewed articles, government reports, dissertations/theses, abstracts, conference proceedings, electronic databases and unpublished data. Our data entry system will have error checking routines built into a data entry interface. Data available in electronic document will be extracted with wrappers. When feasible, tabular numeric hard copy data will be scanned with optical character recognition (OCR) software and converted to an electronic format for manipulation and extraction. Graphical data will be scanned into digital format. Data quality will, to some extent, be maintained through users reporting errors, similar to The Paleobiology Database [27] and other community-based cyber-infrastructure. Spatial context of the data will be preserved, through maps, names, coordinates or descriptions of sampling locations. Spatial data will be documented with the Federal Geographic Data Committee (FGDC) Biological Profile [28] and metadata made available with the FGDC Clearinghouse mechanism. Metadata will provide the user adequate information to make an assessment of the quality to ensure informed use of the data.

C. *Informatics Tools*

To access, process and create value-added analyses, informatics tools will be developed or links provided to websites with existing tools. We will create an interactive, spatial analyst tool for accessing, analyzing, visualizing, and production of distribution maps of predator and prey and other spatially based graphic displays of diet data. Users will select and access physico-chemical, habitat, geo-political, or other variables relevant to the study of predator-prey relationships. Temporal data will be used to evaluate the effects of environmental and climate change on trophic dynamics and evolutionary processes. In addition, these data will be useful for: assessing bioaccumulation and trophic transfer of historic and newly emerging contaminants [29], joining large biodiversity datasets together for better trophic ecosystem models [30] and drawing various inferences on the ecological functioning and fisheries impacts [31].

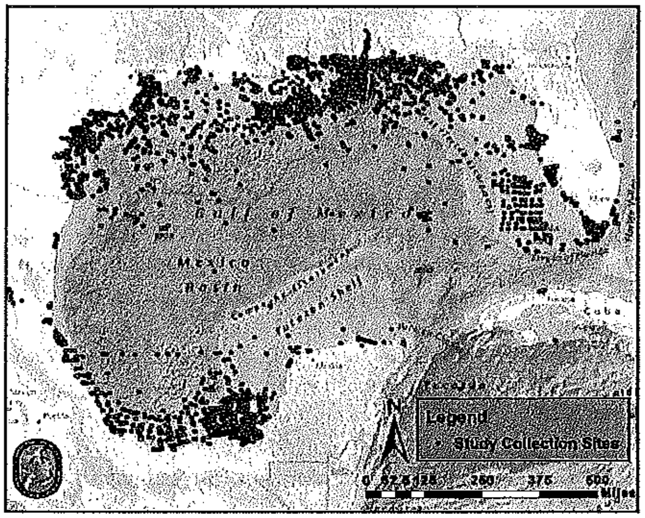

Figure 2. Map showing the location of individual sampling sites for ~520 food habits studies in the GoM.

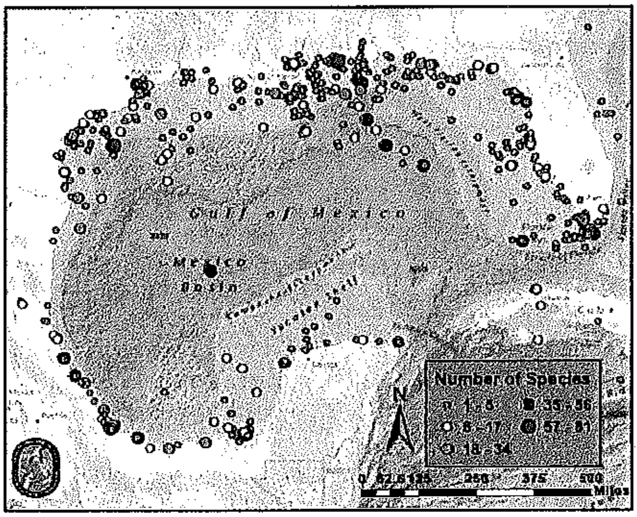

Figure 3. Map showing the centroid location of ~520 food habit studies and the number of fish species examined for food habits.

A metaweb, using the raw data without any a priori aggregation, will be flexible, (i.e., individual based to species level, homogeneous to heterogeneous space, etc.) to explore consequences of individual variability, and spatio-temporal heterogeneity of the raw data, and level of taxonomic, spatial and temporal aggregation for understanding of ecosystems. Fuzzy kriging techniques [32] will incorporate both crisp (certain) and fuzzy data to estimate categorical regions (such as abundance or average) of species distributions or trophic relations. Self-organizing maps (SOM) [33] will be developed to measure similarity of trophic structures in different habitats. A SOM will show clusters of habitats based on their trophic characteristics. Other informatics tools include qualitative reasoning models for trophic interactions among populations [34], genetic algorithms to predict food habits of fishes in unstudied habitats [35] and adaptive agents to simulate food webs [36].

D. *Web Applications*

Data will be publically available through a multi-lingual website with relational database and geographic information system (GIS) entry portals. Data will be available on the website in two formats: 1) table format; and, 2) EML formats for the purpose of information exchange with other databases. To exchange data with other databases, server software such as Distributed Generic Information Retrieval (DiGIR) or Taxonomic Database working Group (TDWG) Access Protocol for Information Retrieval (TAPIR) will be adapted to send/retrieve data on the Internet. Links will be provided to relevant database and informatics websites [5, 6, 7, 8, 9, 10, 11, 12, 13, 18, 20, 22 and others].

E. *Challenges*

Creating and using the proposed GoM trophic database presents several challenges. The various studies were conducted under a wide variety of objectives and methods, requiring units and methods be standardized to the extent possible. Data are reported in a wide array of graphic and tabular formats, which will need to be converted to a single database format. The spatial and taxonomic distribution of the species is clumped (Fig 2), requiring the use rarefaction and interpolation where feasible. While these issues present challenges in analyzing these data, they also identify opportunities for further research.

III. Results and Discussion

A. *Geo-Coding References*

We began by capturing the spatial information for the studies (i.e., station locations, and locations and names of systems where the studies were conducted) and display the results in a GIS (Fig 1). Study collection points, polygons, and centroid points (derived from the study polygons) have been created for ~650 of the ~720 references at the University of Oklahoma and the Florida Fish and Wildlife Research Institute. Attributes of these points, polygons, and centroids include the study's author, study location, number of species studied (Fig 2 and 3), and associated metadata.

B. Coding the Coastal and Marine Ecological Classification Standard (CMECS)

A pilot study was conducted at the University of Oklahoma to unify codification of habitat data in the numerous trophic references using CMECS [37]. Approximately 60% of the references in hand at the time the project was undertaken were coded. This entailed extracting all relevant habitat information reported in the document and adapting those descriptions to the CMECS terminology. The CMECS system first classifies a habitat into one of two systems, and then up to five components (Water Column, Benthic Biotic, Surface Geology, Sub-benthic, GeoForm) can be used to provide detailed information.

C. Metadata and Gulf Geospatial Assessment of Marine Ecosystems (GAME)

Metadata records were created for 690+ peer-reviewed papers organized in a Food Habits of Fishes Bibliography for estuarine and marine environments of the Gulf of Mexico [38]. Metadata were generated using the Gulf GAME survey tool that allows records to be entered through a user friendly interface. These records were incorporated into the Gulf GAME catalog and are available online for search and retrieval [39]. The catalog stores metadata "lite" (i.e. only primary elements are captured) and the records are FGDC compliant. The importance of this work lies in that it allows archival for long-term persistence of information that previously had no attendant metadata. Also, it makes the information discoverable since the majority of the Bibliography studies are not available online.

IV. Conclusions

The trophic informatics system for the GoM will be designed to be compatible and extensible to extant database projects and programs. Coordination and collaborative promises have already been achieved with FishBase [6], SeaLifeBase [7] and EOL [8]. This will allow for data and format sharing so that the maximum accessibility and usefulness of the trophic data are achieved, and that value is added to the existing databases through pre-planned links. The vision is that the structure, methods, and tools will be extensible to other large marine ecosystems. The extensibility and transportability of the model is important to support the development of similar databases globally. Toward this end, this database will prove invaluable in furthering research on directed networks, ecosystem fisheries models and food web theory.

Acknowledgment

This work was supported in part by NatureServe and NOAA for a CMECS pilot study and the U.S. Environmental Protection Agency Gulf of Mexico Program provided funds to create metadata "lite" records and to capture spatial information from a collection references on food habits of GoM fishes provided by Dr. J. Simons. Comments from three anonymous reviewers greatly improved the manuscript.

References Cited

[1] K. McCann, "Protecting biostructure," Nature, vol. 446, p. 9, 2007.

[2] R. M. May, "Network structure and the biology of populations," TREE, vol. 21, number 7, pp. 394-399, 2006.

[3] T.C. Ings, J.M. Montoya, J. Bascompte, N. Blüthgen, L. Brown, C.F. Dormann, F. Edwards, D. Figueroa, U. Jacob, J.I. Jones, R.B. Laurisden, M.E. Ledger, H.M. Lewis, J. Olesen, E.J.F. van Veen, P.H. Warren, and G. Woodward, "Ecological networks -- beyond food webs," J. Anim. Ecol., vol, 78, pp. 253-269, 2009.

[4] FishNet2, 2008, http://fishnet2.net/, Accessed 8/2011.

[5] Ornithological Information System, 2008, http://olla.berkeley.edu/ornisnet/, Accessed 8/2011.

[6] Froese, R., Pauly, D. (Eds), 2009, FishBase, World Wide Web electronic publication, www.fishbase.org, version (06/2009).

[7] Palomares, M.L.D and Pauly, D. (Eds.), 2009. SeaLifeBase. World Wide Web electronic publication. http://www.sealifebase.org/, Version (3/2009).

[8] Encyclopedia of Life, 2008, http://www.eol.org/, Accessed8/2011.

[9] Knowledge Network for Biocomplexity, 2009, http://knb.ecoinformatics.org/index.jsp, Accessed 8/2011.

[10] Global Biodiversity Information Facility, 2008, http://www.gbif.org/, Accessed 8/2011.

[11] Census of Marine Life, 2008, http://www.coml.org/, Accessed 8/2011.

[12] Interaction Web Database, 2008, http://www.nceas.ucsb.edu/interactionweb/, Accessed 8/2011.

[13] Webs on the Web, 2008, http://foodwebs.org/, Accessed 7/2011.

[14] FWDP, 2011, http://www.nefsc.noaa.gov/pbio/fwdp/FWDP.htm, Accessed 8/2011.

[15] REEM, 2011, http://access.afsc.noaa.gov/reem/ecoweb/Index.cfm, Accessed 8/2011.

[16] F.A. Bisby, "The quiet revolution: biodiversity informatics and the internet," Science, vol. 289, pp. 2309-2312, 2000.

[17] W.K. Michener, J.W. Brunt, and K.L.Vanderbilt, "Ecological informatics: a long-term ecological research perspective," in Proceedings Information Systems Development II, N.J. Callaos, J. Porter, and N. Rishe, Eds, 6th World Multiconference on Systemics, Cybernetics and Informatics, 2002.

[18] M.B. Jones, M.P. Schildhauer, O.J. Reichman, and S. Bowers, "The new bioinformatics: Integrating ecological data from the gene to the biosphere," Ann. Rev. Ecol. Evol. Syst., vol. 37, pp. 519-544, 2006.

[19] BioGeoMancer, 2002, BioGeoMancer: Automated Georeferencing for Natural History Collections, http//www.biogeomancer.org, Accessed 8/2011.

[20] LifeMapper, 2008, http://www.lifemapper.org/, Accessed 8/2011.

[21] Aquamaps, 2008, http://www.fishbase.ca/tools/aquamaps/search.php, Accessed 8/2011.

[22] Ocean Biodiversity Informatics, 2008, http://www.vliz.be/events/obi/index.php, Accessed 8/2011.

[23] V. Ahl, and T.F.H Allen, Hierarchy Theory: A Vision, Vocabulary, and Epistemology. New York: Columbia University Press, 1996.

[24] M. Yuan, "Representing geographic information to enhance GIS support for complex spatiotemporal queries," Trans. in GIS, vol. 3, no. 2, pp. 137-160, 1999.

[25] M. Yuan, "Knowledge discovery of geographic dynamics in spatiotemporal data," in Geographic Data Mining and Knowledge Discovery (2nd ed), H. Miller and J. Han, Eds. CRC/Taylor and Francis. 2008.

[26] Ecological Metadata Language, 2008, http://knb.ecoinformatics.org/software/eml/, Accessed 8/2011.

[27] The Paleobiology Database, 2008, http://paleodb.org/, Accessed 7/2011.

[28] Federal Geographic Data Committee (FGDC), 2008, http://www.fgdc.gov/, Accessed 8/2011.

[29] P.A. Sandifer, A.F. Holland, T.K. Rowles, and G.I. Scott, "The oceans and human health," Environ. Health Perspect, vol. 112, no. 8, pp. 454-455, 2004.

[30] R.A. Myers, J.K. Baum, and T.D. Shepherd, "Cascading effects of the loss of apex predatory sharks from a coastal ocean," Science, vol. 315, pp. 1846-1850, 2007.

[31] L. Vidal, and D. Pauly, "Integration of subsystems models as a tool toward describing feeding interactions and fisheries impacts in a large marine ecosystem, the Gulf of Mexico," Ocean Coastal Management, vol. 47, pp. 709-725, 2004.

[32] A. Salaski, "Ecological applications of fuzzy logic," in Ecological Informatics: Scope, Techniques and Applications, 2nd ed. F. Recknagel Ed. New York, NY: Springer, 2006, pp. 3-14.

[33] J.L Giraudel, and S. Lek, "Ecological applications of non-supervised artificial neural networks," in Ecological Informatics: Scope, Techniques and Applications, 2nd ed. F. Recknagel Ed. New York, NY: Springer, 2006, pp 49-67.

[34] P. Salles, B. Bredeweg, S. Araujo, and W. Neto, "Qualitative models of interactions between populations," AI Communications, vol 16, no. 4, pp. 291-308, 2003.

[35] D.J. D'Angelo, L.M. Howard, J.L. Meyer, S.V. Gregory, and Z.L.R. Ashkenas, "Ecological uses for genetic algorithms: predicting fish distributions in complex habitats," Can. J. Fish. Aquatic Sci. vol. 52, pp. 1893-1908, 1995.

[36] F. Recknagel, "Ecological applications of adaptive agents," in Ecological Informatics: Scope, Techniques and Applications, 2nd ed. F. Recknagel Ed). New York, NY: Springer, 2006, pp 109-124.

[37] M. Yuan, L. Williams, S. Gonzalez-Perez, D. Morris, and J. Simons, Data acquisition and meta-analysis of habitat information from Gulf of Mexico trophic studies using CMECS, Final Report submitted to NatureServe. NOAA contract # EA133C-05-CQ-1051 and Fugro EarthData Project # E09-0039-00. 22p. 2010.

[38] J.D. Simons, R.M. Darnell and M.E. Vega-Cendejas, Bibliography of studies of Food Habits of Estuarine and Marine Fishes in the Gulf of Mexico, unpublished manuscript.

[39] FWRI, 2011, http://myfwc.com/research/gis/game/, Accessed 7/2011.