Cyberinfrastructure and Collaborative Synthesis and Analysis of the Long-term effects of the Exxon Valdez oil spill

The Exxon Valdez Oil Spill Trustees Council (EVOSTC) faces challenges in ensuring that critical data and products of long-term monitoring and herring population studies are available to the general science and policy community, both now and into the future, and that these products are useful for gaining an effective understanding of the effects of the spill on Prince William Sound and the recovery status of affected ecosystems. The AOOS Long-Term Monitoring and the PWSSC Herring Research and Monitoring projects propose an ambitious monitoring agenda that could facilitate a thorough understanding of the effects of the oil spill, but only if these new monitoring products are effectively managed and collated along with historical data on these systems, and then used in a comprehensive synthesis effort. We propose to collaborate with the AOOS LTM and PWSSC Herring teams to build an effective data management cyberinfrastructure for current monitoring efforts, to organize these data with historical data to prepare it for synthesis, and then to convene two cross-cutting synthesis working groups to do a full-systems analysis of the state of recovery of the ecosystem.

National Center for Ecological Analysis and Synthesis

Since 1995, the National Center for Ecological Analysis and Synthesis (NCEAS)[1] has been advancing the state of ecological and environmental knowledge through synthetic and collaborative research that aims to discover general patterns and principles based on existing data. The premise at NCEAS is that many decades of data have been collected that can be synthesized to produce novel insights into important scientific and societal issues, and that the expertise and information resources necessary to accomplish these syntheses are latent but distributed throughout the science community. To promote effective synthesis of environmental data, NCEAS has sponsored and executed more than 450 working groups over 15 years, many of which have had major scientific and policy impacts (e.g., changes in habitat conservation plans for endangered species, and creation of marine reserve initiatives based on scientific principles). Sociological studies of the working groups in action at NCEAS have demonstrated major shifts in the culture of synthesis in ecology and gains in collaborative productivity via the working group model at NCEAS (Hackett et al. 2008).

NCEAS personnel have extensive expertise in cyberinfrastructure systems for synthetic environmental science (c.f., Reichman et al. 2011, Jones and Gries 2010, Jones et al. 2006). The Knowledge Network for Biocomplexity (KNB)[2] data repository network links more than 250 field stations and marine labs in a cohesive data sharing federation that is built to handle highly heterogeneous science data generated from independent research projects

[1] http://www.nceas.ucsb.edu
[2] http://knb.ecoinformatics.org

and investigators. NCEAS has developed software systems supporting long-term data preservation and sharing, is a leader in metadata systems for science data, and is a progenitor of the DataONE[3] interoperability framework to create a global data federation for open access to scientific data. Our focus on project-level data management for highly heterogenous data has allowed us to manage current and legacy data that are critical to synthesis but that often are not captured by large scale agency data systems, such as the EOSDIS program or the IOOS program. Thus, the initiatives at NCEAS for capturing complex but smaller-scale biological and physical data will be an effective complement to the ocean observatory data management systems that are provided by AOOS.

We propose a two phase project: **Phase I**, during the first two years, will focus on data federation and cyberinfrastructure development to create the necessary data resources for synthesis; **Phase II**, during years 3-5, will comprise a multi-year working group effort using internationally renowned researchers to synthesize what is known about spill effects and recovery of ecosystems, and about herring population dynamics.

Phase I: CyberInfrastructure for Long-Term Monitoring and Herring Population Monitoring

Managing heterogeneity

Data collected prior to and in-response to the Exxon Valdez oil spill are profoundly heterogeneous. They range from long-term, automated sensing of oceanographic and atmospheric conditions, to short-term, experimental, monitoring, and behavioral studies of biological components of the system. The scientific data to be collected in these studies includes data on population trends, behavior, physiology, disease, and genetics of many species, as well as oceanographic and meteorological data at both regional and local scales. This diversity of data and data collection protocols substantially complicates data management by EVOSTC long-term monitoring projects. In addition, investigators on both the long-term monitoring and herring population studies are affiliated with many different institutions and agencies, each currently collecting data from many sites within the spill region and managing it within the frameworks dispersed among these agencies. Any data management system will necessarily need to accommodate this heterogeneity and dispersion by preserving the original data and providing mechanisms to access, integrate, and analyze the data for crosscutting synthesis.

Cyberinfrastructure design and development

The AOOS data management team is planning to use an established suite of open services and standards to expose data from both the LTM and Herring projects. This suite is well suited to handling high volume time-series data that is collected on shipboard cruises and on semi-automated sensing platforms. Much of the biological data collection, however, is managed on a project-by-project basis, with correspondingly higher complexity for each unit of data. We propose to collaborate with the AOOS data management team in order to design and implement a system to capture and manage these more complex data while simultaneously preparing the data for use in future synthesis activities. There will be five

[3] http://dataone.org

principal deliverables in Phase I: cyberinfrastructure design; semi-automated data population; services supporting the full data lifecycle; semantic data discovery services; and data organization in preparation of synthesis.

Cyberinfrastructure design. NCEAS engineers will work with the AOOS data team to extend the AOOS data management plan to incorporate metadata tools and catalogs that are customized for project-based data management for biological data. The design will include both tools for data access and for data contribution and management by the participating scientific staff. The planned AOOS software stack based on THREDDS and OPeNDAP will be supplemented with more biologically-oriented metadata management tools like Morpho in order to enable individuals to describe and deposit all of their heterogeneous data in a uniform data repository. We will also evaluate the extent to which AOOS would benefit from the data preservation and replication services that are part of the DataONE architecture, and design the system accordingly.

Semi-automated data population. Once these complex, heterogeneous data are in a data repository, they still will not be directly queryable. We propose to extend existing software (the KNB Data Manager Library), which we have written for automatically creating relational databases out of heterogeneous data sources, for use within the AOOS system. This would allow the AOOS data catalog to expose even complex, project-specific data sets as relational databases, providing all of the power of both SQL and spatial database systems but without custom development for each new study. This system would be incorporated into the Tier 1 AOOS services that provide basic data management for data and models.

Full lifecycle services. The AOOS Tier 2 services are designed to query, extract, and visualize data from Tier 1, and utilize several community standard services that are in common use (e.g., Web Mapping Service). These services mainly cover the read and access portion of the data lifecycle. We propose to extend these services by also incorporating the full lifecycle services that are part of the DataONE data federation. These include open services for writing data and metadata, controlling access to data products as they are populated in the system, and services for replication and preservation of data. By adopting the DataONE service framework, this will also link the AOOS system to the broader DataONE federation, which includes partners such as the USGS, the KNB, and NASA Distributed Active Archive Centers. This broader federation will be critical in other stages of the project, e.g., for access to MODIS data during Phase II synthesis and analysis activities.

Semantic data discovery services. Data discovery is difficult for complex, multidimensional and cross-disciplinary data that will be collected by the long-term monitoring team. The AOOS system includes a comprehensive Asset Catalog in their Tier 3 services. This catalog provides an index of all project data, but does not directly index other non-project data resources that may be needed for analysis and synthesis. We propose to work with the AOOS team to develop their asset catalog, in particular focusing on linking it to the DataONE federated catalog and providing powerful semantic search capabilities for the project. The NCEAS team has pioneered a semantic scientific observations model that allows scientists to precisely discover measurements of interest and subset data to only include observations relevant to their studies. We developed the

Extensible Observations Ontology (OBOE; Madin et al. 2008) to enable semantic search and access services that facilitate much higher precision and recall than have been possible with traditional metadata-driven systems. We will incorporate these semantic search services into the Tier 3 asset catalog, and help to develop the catalog so that semantic markup of data on ingest is easily accomplished.

Data organization for synthesis. The current AOOS plan is to emphasize the capture of historical data from previous studies related to the Exxon Valdez oil spill during the first two years of the project, as well as to prepare the system to receive the monitoring data generated during this project. We will collaborate with the AOOS team in order to collate summarize, visualize, and integrate these historical data in order to prepare them for the synthesis and analysis to be conducted in Phase II. Our experience at NCEAS is that, even with current information management advances, there is significant labor and time needed to prepare data for any given synthetic study—often a year or two is required for even small-scale studies. NCEAS has developed a group of scientific programmers who specialize in assisting in cross-cutting analysis and modeling, and we will employ one of these scientific programming specialists to collate, standardize, integrate, summarize, and visualize the data needed for synthesis activities.

Phase 2: Synthesis and Analysis

Despite decades of monitoring and analysis of EVOS-affected systems, there is still a major lack of understanding of oil spill impacts and recovery at a holistic level. Many of the studies to date have been at the single species level, and recovery status is tracked on a case-by-case basis. In addition, because all of the historical data have never been fully integrated, it has been impossible to conduct a holistic analysis of the effects of the oil spill and recovery of impacted regions. Such a holistic view is critical to guide future monitoring and recovery initiatives, which are expected to continue for decades. In Phase II, we propose to conduct two holistic synthesis activities aimed at understanding the long-term consequences of EVOS and the effectiveness of recovery initiatives:

- Synthesis Working Group: Long-Term Effects and Recovery from the Exxon Valdez Oil Spill
- Synthesis Working Group: Understanding Dynamics of Herring Population s following the Exxon Valdez Oil Spill Dynamics.

The first will address system-wide impacts of and recovery from EVOS, and the second will specifically focus on the herring project in an effort to assess reasons for the lack of recovery of herring and develop models that will assist in future recovery initiatives.

NCEAS has an extensive history of convening highly productive synthesis activities through its use of a working group model, involving face-to-face meetings and ongoing virtual collaboration supported by the Center (Hackett et al. 2008). Under this successful NCEAS model, committed working group participants conduct relevant analysis and modeling on a continuous basis for approximately two years, punctuated by periodic working meetings to come to consensus and drive further work by participants. The momentum of the group is maintained by postdoctoral fellows that reside at NCEAS, working on the group's analysis,

modeling, and other synthesis tasks while being able to take advantage of the computational and analytical support services available at NCEAS. Working Groups are composed to represent a wide variety of scientific expertise, including both scientists that are closely involved in the problem at hand, as well as researchers from adjoining disciplines that help broaden the scientific perspective of the group. In addition, Working Groups typically include a mix of more senior scientists and younger scientists that are eager to dive into the required analysis and modeling activities.

To initiate these Working Group activities, NCEAS will organize and constitute the groups during year two. Working Group leaders will be selected for their knowledge of the issues at hand as well as their ability to effectively motivate a group of up to 14 other working group participants. We would expect that multiple of the PIs from the LTM and Herring projects would be participants in the synthesis working groups along with nationally renowned experts in population and community modeling, ecosystem modeling, and coupled whole-system analysis. In addition, because NCEAS is already running a working group on ecotoxicology associated with the BP Deepwater Horizon spill[4], we would expect significant coordination and cross-pollination with these new EVOS synthesis groups.

References

Hackett EJ, Parker JN, Conz D, Rhoten D, Parker A. 2008. Ecology transformed: The National Center for Ecological Analysis and Synthesis and the changing patterns of ecological research. Pages 277–296 in Olson GM, Zimmerman A, Bos N, eds. Scientific Collaboration on the Internet. MIT Press.

Jones MB and Gries C. 2010. Advances in environmental information management. Ecological Informatics 5: 1–2. doi:10.1016/j.ecoinf.2010.01.001

Jones MB, Schildhauer M, Reichman OJ, and Bowers S. 2006. The new bioinformatics: integrating ecological data from the gene to the biosphere. Annual Review of Ecology, Evolution, and Systematics. 2006. 37:519–544.

Madin JS, Bowers S, Schildhauer M, and Jones MB. 2008. Advancing ecological research with ontologies. Trends in Ecology and Evolution 23 (3): 159-168. doi:10.1016/j.tree.2007.11.007

Reichman, OJ, Jones MB, and Schildhauer MP. 2011. Challenges and Opportunities of Open Data in Ecology. Science 11 February 2011: 703-705. doi:10.1126/science.1197962

[4] Anderson, Cherr, and Peterson; Ecotoxicology of the Gulf Oil Spill: A holistic Framework for Assessing Impacts