EVOSTC RESEARCH PLAN
EVOSTC DATA MANAGEMENT 050455

I. NEED FOR THE PROJECT

A. Statement of Problem

The Trustee Council uses goals established for the GEM program (Gem Program Document, 2002)) that make data management a top priority. Data management has been modeled on the GEM Program criteria because all aspects of the Trustee Council programs and operations depend on the data management principles established for GEM. In order for the information from Trustee Council projects to be useful, it must be made accessible through effective data management. A number of the five goals of the GEM Program rely on effective data management. For example, the "Inform" goal states that the GEM program will provide integrated and synthesized information to the public, resource managers, industry and policy makers in order for them to respond to changes in natural resources. The "Solve" goal requires developing tools, technologies, and information that can help resource managers and regulators improve management of marine resources and address problems that may arise from human activities. The "Detect" goal also has a data management and communication aspect, as GEM is asked to serve as a sentinel (early warning) system by detecting annual and long-term changes in the marine ecosystem, from coastal watersheds to the central Gulf of Alaska.

In addition to GEM other aspects of Trustee Council operations that depend on data management are peer review of proposals and work products, management and integration of project information, reporting, and office systems development and maintenance (see narrative for Project 050630 for a detailed description of these activities. Data management is thus a basic information management function that is essential to the proper functioning of all aspects of the EVOSTC staff.

During the development stages of any data system, issues are identified and need to be resolved. This requires the guidance and direction of a professional data systems manager to develop and implement a successful network and database system. Timely implementation of a networking infrastructure, database system and associated web sites requires additional technical assistance.

B. Relevance to GEM Program Goals and Scientific Priorities

In order to accomplish the Trustee Council's goals for the GEM program, management of monitoring and research data is a top priority. The purpose of this project is to provide funding to the GEM program Data Systems Manager and Analyst/Programmer in the development and implementation of a data system for GEM.

The GEM program encompasses a long term effort to monitor ecosystem dynamics in the Gulf of Alaska. In order to extract scientific understanding from these monitoring efforts, the information collected over the years must be readily accessible for analysis and synthesis. GEM data management is charged with creating the technological structure to archive and disseminate this information. The following excerpt, taken from the National Research Council's review of

the GEM Program Document, stresses the importance of concrete data and information management as it pertains to the GEM program:

> *The legacy of the GEM program will be the data it collects. Given the objective of establishing a long-term measurement program in the Gulf of Alaska and its importance to both regional and national interests, GEM must make a strong commitment to data and information management. The goals must be to facilitate data exchange among GEM scientific investigators, make data available to the public and others outside the scientific community, and archive GEM data products.*

The success of the GEM program relies heavily on the efforts of the GEM data management section. Efforts which both contribute to the construction of a robust data archiving system and guide principal investigators to produce adequate data management plans will ensure that information harvested through the GEM program will be readily available to anyone for future scientific analysis and synthesis. Internally, the GEM data management section will provide a productive technological environment for all EVOS staff through the maintenance and support of network and computing systems located within the Trustee Council Office. In this regard a high priority for the GEM data management section is the production of an automated administrative system to expedite the processes associated with the registration, documentation, and review of GEM projects.

The GEM data policies, as detailed in the GEM Program Document, incorporate ten broad elements:

1. A commitment to the maintenance and long-term availability of data.
2. Full and open sharing of data at low cost, after verification and validation.
3. Timely availability of data, depending on the type of data. Some data will be available almost immediately; other data may be available with 24 months.
4. Availability of data on the GEM public web site.
5. Identification of the origin of all data with a citation.
6. Adherence to data collection and storage standards.
7. Provision of citations to the GEM bibliography.
8. Encouragement of active participation in the GEM web site for all participants.
9. Long-term archiving of all data in a designated storage facility.
10. Acceptance of and adherence to the data policies as a condition for participation in the GEM program and receipt of funding.

II. PROJECT DESIGN

A. Objectives

Objective 1. Design, implement and manage a data and information system consistent with the provisions of the GEM Program Document that provides data, information products (maps, tables, summary reports) and documentation for scientific researchers, resource managers, policy makers, and the public.

Objective 2. Determine how best to incorporate existing and future data sets identified by the Science Director and other scientists into the data and information system.

Objective 3. Develop data management plans and work with Principal Investigators for all data gathering projects funded by the GEM program.

Objective 4. Provide for computer and network needs of office staff, including Web site.

Objective 5. Function as External Liaison: Work with and serve on regional and national coordinating committees for AOOS, Ocean.US (IOOS) and others; serve as liaison to federal/state agencies, other research entities, principal investigators, other technical support personnel, as well as stakeholders and the general public.

Objective 6. Assist EVOSTC staff in the utilization of technology to more efficiently perform their duties and to expedite the creation of the various products and assist in the administration of the events associated with the annual EVOSTC business cycle.

B. Procedural and Scientific Methods

GEM Data Management is committed to developing solutions for the management of data which use technologies that are open source. Utilizing open source technology ensures that our data management tools can be used by and distributed to other research and management entities for very low cost or free. The following seven caveats drive the GEM Data System Development effort.

1. Flexibility

For the most part, GEM data sets will be non-homogenous, independent, and unique from each other. Datasets could consist of physical measurements, taxonomic measurements, in addition to unforeseen types, or combinations of all three. The GEM data system must be able to accommodate foreseen data in addition to allowing for the absorption of unknown data and information types. The system must be able to absorb all GEM data in structured form associated with descriptive syntactic and thematic meta-data to allow facilitation of queries.

2. Scalability

Due to the nature of the GEM project, its data system must be capable of easily absorbing multiple heterogeneous datasets each year. Over the years the number of datasets could rise into

the thousands and comprise a data warehouse of a billion or more records. The data system must be inherently scalable and capable of easily absorbing new datasets into the system with minimal required maintenance. Data incorporation must be simple, automatic and straightforward.

3. Metadata

Data is useless in today's scientific world without its complementary metadata. Syntactic, semantic, and thematic metadata must be an integral part of the GEM data system and accessibility to it must exist via simple pathways. Syntactic Metadata describes programmatic/computational technical characterization of data and can include but not limited to data type, measurement units, and associated measurement error. Semantic metadata can describe contextual information about the individual data and can include descriptions like measurement type and measurement device. Thematic metadata can include descriptions which define the context of the study which produced the data and could include information detailing principal investigator, species association, study hypothesis, etc…Information describing the context of the measurement, data collection device, units, and spatial temporal relationships are just a few of the descriptive quantities which must be contained within the system. The metadata must be standardized and structured (i.e., contained in lookup tables chosen for universal usage) to assist in data extraction, data mining, and data formatting functionality. Metadata specifications must meet with Federal Government Data Committee (FGDC) requirements.

4. Transparency, Aggregation, and Data Mining

Though the GEM data system will be composed of multiple heterogeneous data sets, users of the system must interface it as if they are accessing a single dataset. The ability to generate subsets of data from both individual and multiple sets is an absolute necessity of the system. This ability to aggregate data from independent datasets into a homogenous representation must be a core property of the system. Projects will of course produce unique datasets. Many measurements of each independent dataset will be of the same semantic type but may very well be represented in differing units and data types. Structures must exist within the data system to isolate those semantic homogeneities and format and aggregate those measurements to produce a continuous transparent view of the distributed data. Users should be able to data mine the system for information which conforms to their search criteria.

5. Data Interchange Between other data warehouse systems

A paramount requirement of the GEM data system is that it be able to interact, extract, and contribute to other data systems. The facilitation of these tasks will be through the use of middleware products which must be inherently compliant with characteristics of the data system. The system should also be capable of interfacing with current oceanographic data sharing protocols such as OPENDAP.

6. GIS and WEB functionality

The system selected for the storage of GEM related data must be both WEB and GIS enabled without the application of extravagant measures to do so. Both of these technologies have become primary sources for the representation and dissemination of modern information and having a system which is conducive to the creation of ports to these technologies is a fundamental requirement of any contemporary information system.

7. GEM Data and Meta Data Archive System

The GEM data system must act as a robust and concrete data archiving system to insure backup and integrity of the data contained within it. This will include all data, metadata and computational structures.

C. Data Analysis and Statistical Methods

N/A

D. Description of Study Area

The Data Systems Manager and Analyst/Programmer will work in the *Exxon Valdez* Oil Spill Trustee Council Office in Anchorage. The Data Systems Manager will generally work under the supervision of the Science Director, although for some projects, will work under the supervision of the Executive Director. The Analyst/Programmer will work for the Data Systems Manager.

E. Coordination and Collaboration with Other Efforts

In collaboration with the staff of the National Pacific Research Board (NPRB), GEM data management staff will develop protocols for the metadata description of regional produced oceanographic datasets. The NPRB is analogous to the GEM program in that it funds projects which produce data which is of the same semantic type, i.e. physical and biological oceanographic data. The NPRB has committed staff time and monies to the purchase of Linux based server for a common data management solution to each entities (GEM and NPRB) data archival requirements. The server will be housed in the GEM office but will be utilized and maintained by both parties as outlined in the MOA entitled *Combined Linux Server Purchase and Use Stipulations* which can be requested from the internal files of the EVOSTC office.

Technical data personnel from Trustee agencies and other research entities will be invited to serve on a data advisory subcommittee. The subcommittee will assist in setting goals and policies for the GEM data system. The data subcommittee will also assist in the development of the data system and advise on how best to address the target user communities' needs and the scope of the system.

III. SCHEDULE

A. Project Milestones

The primary objective of this project is to provide an ongoing service, consequently there are few set milestone dates or endpoints.

B. Measurable Project Tasks

October 15:	PostgreSQL training will be complete and development of GEM data and metadata management system will commence.
November - December	Data Management plans issued to FY2004 projects and applicable datasets will be absorbed into the system.
October-January:	Data Systems Manager prepares data management plans for FY 05 projects approved by Trustee Council.
January:	Attend EVOS Trustee Council/GEM annual workshop
February-September:	Existing data sets identified, collected, documented and incorporated into GEM data system in an ongoing fashion

IV. RESPONSIVENESS TO KEY TRUSTEE COUNCIL STRATEGIES

A. Community Involvement and Traditional Ecological Knowledge (TEK)

N/A

B. Resource Management Applications

The GEM data system will provide a vehicle for managers, scientist, and the public to access and synthesize information collected from GEM funded projects. The utilization of this information will greatly assist resource managers in performing there duties in addition to conserving effort expended by those individuals to find the information they need to make informed decisions.

V. PUBLICATIONS AND REPORTS

GEM Data Management is not requesting funding for publication.

VI. PROFESSIONAL CONFERENCES

GEM Data Management is not requesting funding for professional conferences. GEM Data Management principal investigators will attend the annual GEM workshop.

Robert J Bochenek

Data Systems Manager - Exxon Valdez Oil Spill Trustee Council
441 West 5th Ave, Suite 500
Anchorage, Alaska 99501
(907) 278-8012
rob_bochenek@oilspill.state.ak.us

Mr. Bochenek has degrees in mathematics, physics, and aerospace engineering and has worked in scientific computer programming most of his professional life. He has been Data Systems Manager of the Trustee Council since April 2003, and prior to that, Analyst/Programmer for the Trustee Council since October of 2002.

Professional Experience:

Exxon Valdez Oil Spill Trustee Council — 2002 - present
- Data System Manager (2003 - present)
- Analyst Programmer III (2002 - 2003)

Alaska Department of Fish and Game — 2001 - 2002
- Analyst Programmer III (2002 - 2002)
- Analyst Programmer II (2001 - 2002)

Education:

Bachelor of Science Engineering in Aerospace Engineering
University of Michigan – Ann Arbor, 2001

Bachelor of Science in Mathematics
University of Michigan – Ann Arbor, 2001

Bachelor of Science in Physics
University of Michigan – Ann Arbor, 2001

Publications:

Bochenek, R. and Kelley, T. 1993. Introduction to Object Oriented Programming Methodology. Splitfire Technologies

Affiliations:

Alaska Oceanographic Observing System (AOOS) Data Management Committee (DMAC)

Michael Schlei

Analyst/Programmer – Exxon Valdez Oil Spill Trustee Council
441 W. 5th Ave., Suite 500 Anchorage, AK 99501
Voice: 907-278-8012 Fax: 907-276-7178
michael_schlei@evostc.state.ak.us

Education

B.S., Computer Science, Colorado State University – Fort Collins, CO 2002
A.S., General Sciences – Front Range Community College – Fort Collins, CO 1999

Professional Experience

November 2003 – Present: Analyst/Programmer
Exxon Valdez Oil Spill Trustee Council – Anchorage, Alaska

March 2003 – July 2003: Software Engineer
Scientific Fishery Systems – Anchorage, Alaska

March 1999 – April 2001: Computer Lab Technician
Duke Communications (Windows 2000 Magazine Branch) – Loveland, Colorado

Publications

Schlei, Michael, NTFSDOS Professional Edition, *Windows 2000 Magazine*, July 2000
Schlei, Michael, Paragon, *Windows 2000 Magazine*, February 2001

Awards and Recognitions

Inducted into the Phi Theta Kappa Honors Society – April 1998

Listed on the National Dean's List: 1997-2001

Data Management Methods/Practices

Please provide a brief description of your project's data management methods and practices relating to each category below. Please return by e-mail (Geno.Olmi@noaa.gov or James.Boyd@noaa.gov) no later than November 5, 2004.

Metadata:

The Gulf Ecosystem Monitoring (GEM) Program has chosen Ecological Metadata Language (EML) as the most appropriate solution for metadata documentation of *In Situ* style datasets being produced by GEM sponsored studies. GEM is currently sponsoring studies which are collecting primarily physical, biological, and taxonomic style measurements with a small amount sponsored projects collecting real time data being produced from ships of opportunity. GEM data management has focused much of their efforts towards creating metadata specifications for the *In Situ* type data being collected. EML has been chosen for this purpose because it is a superset of metadata protocols such as FGDC, Z39.50, and Dublin Core and provides a structure for advanced data set documentation. EML provides distinct markup language entity/attribute tags for metadata information deemed pertinent to the GEM Data Management metadata documentation effort. This metadata documentation effort is driven by two caveats: Advanced Data Discovery and Data Synthesis/Trend Analysis.

> **Data Discovery** – Proper documentation of data is critical to providing pathways for the discovery of that data by users. It is vital that users have multiple pathways for locating potential data resources which satisfy their queries. Providing individual fields for metadata descriptors which describe detailed dataset information (i.e. abstracts, measurements, sensors, units, etc…) instead of lumping these pieces of information into single text fields will greatly increase the success of data discovery and enhance interfaces to the data.
>
> **Data Synthesis/Trend Analysis** - Though data is primarily collected to prove or disprove a hypothesis put forward by a researcher, this data can serve an additional higher level purpose when combined with other data. Through the isolation of analogous data set fields, multiple data sets can be formatted to a common structure and aggregated together into a data amalgamation. This amalgamation provides a higher level data set for the synthesis of information and advanced analysis of physical and biological changes on a large temporal and geographic scope. In order to expedite this amalgamation, metadata describing datasets must exist in ways for computer systems to parse the metadata and perform the required operations for the reformatting and aggregation of fields contained within the datasets. EML, which provides a distinct recording mechanism for these fields, will suffice as a metadata container that isolates all the descriptors for this automated formatting/aggregation process.

Quality Assurance/Quality Control:

QA/QC will be carried out on multiple levels. PI's will employ their own QA/QC methods during their own personal data acquisition and will document those QA/QC methods in their metadata. GEM will then develop a secondary QA/QC methodology for assessment of collected datasets. The secondary methodology is in the works.

Storage Formats:

Metadata will initially stored in EML until sufficient EML documents are produced to model a metadata storage system using entity relational (Relational Database) methods. Metadata will be stored in a database and transferred via the EML format.

Datasets will be stored in the form in which they are received from the researcher. Utilizing the associated metadata the data contained within the dataset will be Extracted, Transformed, and Loaded (ETL) into an Online Analytical Processing (OLAP) database which is enabled for analysis, synthesis, and spatial querying/visualization. PostgreSQL, an open source scientific database, has been chosen as the relation database management system selected to host the OLAP data model.

Access/Distribution/Transport:

Access and transport will be available via multiple protocols which include HTTP, XML, Opendap, ODBC, etc…

Archival:

Due to the nature of the GEM program goals, which revolve around monitoring ecosystem change, legacy datasets will play as an important role as contemporary ones. Access to historical data will be valued as important as accessing current information.

FY 04 *EXXON VALDEZ* TRUSTEE COUNCIL PROJECT BUDGET

October 1, 2003 - September 30, 2004

Budget Category:	Authorized FY 04	Proposed FY 05
Personnel	$97.2	$103.8
Travel	$19.2	$16.2
Contractual	$8.0	$6.0
Commodities	$20.6	$15.8
Equipment	$0.0	$0.0
Subtotal	$145.0	$141.8
General Administration	$11.9	$12.8
Project Total	$156.9	$154.6
Full-time Equivalents (FTE)		1.3
Other Resources		

LONG RANGE FUNDING REQUIREMENTS

Estimated FY 05					
TBD					

Dollar amounts are shown in thousands of dollars.

Comments:

The Data Analyst Programmer III position is a fultime position, 3 months salary is to be paid out of the Data Management budget and 9 months are to be supplemented by the NOS grant and paid out of the Science Mangement budget (Project 050630A).

FY05

'Prepared: 08/4/04

Project Number: 050455
Project Title: GEM Data System
Agency: Restoration Office (ADF&G)

FY 04 *EXXON VALDEZ* TRUSTEE COUNCIL PROJECT BUDGET

October 1, 2003 - September 30, 2004

Personnel Costs:		GS/Range/	Months	Monthly		
Name	Position Description	Step	Budgeted	Costs	Overtime	
Rob Bochenek	Data Systems Manager	22D	12.0	7.3		
Michael Schlei	Analyst/Programmer III	18B	3.0	5.4		
	Subtotal		15.0	12.7	0.0	
					Personnel Total	

Travel Costs:	Ticket	Round	Total	Daily	
Description	Price	Trips	Days	Per Diem	
Anchorage to Atlanta	0.8	2	6	0.0	
Anchorage to Fairbanks	0.3	1	2	0.2	
Meeting/Conference Travel	0.7	3	9	0.2	
Data Subcommittee meeting travel					
				Travel Total	

FY05

'Prepared: 08/4/04

Project Number: 050455
Project Title: GEM Data System
Agency: Restoration Office (ADF&G)

FY 04 *EXXON VALDEZ* TRUSTEE COUNCIL PROJECT BUDGET
October 1, 2003 - September 30, 2004

Contractual Costs:	
Description	
Staff training BasicTraining	
When a non-trustee organization is used, the form 4A is required. **Contractual Total**	
Commodities Costs:	
Description	
Software upgrades and licenses Tapes for Tape Backup Laptop	
Commodities Total	

FY05

'Prepared: 08/4/04

Project Number: 050455
Project Title: GEM Data System
Agency: Restoration Office (ADF&G)

FY 04 *EXXON VALDEZ* TRUSTEE COUNCIL PROJECT BUDGET
October 1, 2003 - September 30, 2004

New Equipment Purchases:	Number of Units	Unit Price	
Description			
Those purchases associated with replacement equipment should be indicated by placement of an R.		**New Equipment Total**	

Existing Equipment Usage:	Number of Units	
Description		

FY05

Prepared: 08/4/04

Project Number: 050455
Project Title: GEM Data System
Agency: Restoration Office (ADF&G)

FY 04 *EXXON VALDEZ* TRUSTEE COUNCIL PROJECT BUDGET

October 1, 2003 - September 30, 2004

Authorized FY 03
$151.2
$7.9
$3.0
$10.0
$23.2
$195.3
$17.6
$212.9