Exxon Valdez Oil Spill
Restoration Project Annual Report

Evaluation of the Data System for the EVOS Long-Term Monitoring Program

Restoration Project 00455
Annual Report

Following the death of Charles Falkenberg, the principal investigator, this report was compiled by Trustee Council staff to create a permanent record of the preliminary results from this project. This report was not submitted for peer review.

Charles Falkenberg

ECOlogic Corporation
19 Eye Street, NW
Washington, DC 2001

September 2001

[NOTE: Report compiled by Sandra Schubert, Trustee Council Staff]

Evaluation of the Data System for the EVOS Long-Term Monitoring Program
Restoration Project 00455
Annual Report

Study History: This project was funded for one year only (FY 00) due to the untimely death of the principal investigator, Mr. Charles Falkenberg. Following Mr. Falkenberg's death, this report was compiled by Trustee Council staff to create a permanent record of the information generated under this project. This report was not submitted for peer review.

Abstract: This project funded a technical consultant to outline the key data and user issues that the Trustee Council should consider in designing a data system for GEM (Gulf Ecosystem Monitoring and Research Program), the Council's long-term monitoring and research program. In addition, a report analyzing existing systems that deliver similar data was to be prepared and strawman proposals developed for a range of data systems that could meet the needs of GEM. The consultant, Mr. Charles Falkenberg, died before completing that report. This report – which consists of a series of advisory memos from Mr. Falkenberg and language he drafted for a chapter on data management in the GEM Program Document – has been compiled by Trustee Council staff to create a permanent record of this information.

Key Words: Data management, data system, GEM (Gulf Ecosystem Monitoring and Research Program), *Exxon Valdez* oil spill

Project Data: All information developed under this project is included in this report.

Citation: Falkenberg, C. 2001. Evaluation of the data system for the EVOS long-term monitoring program, *Exxon Valdez* Oil Spill Restoration Project Annual Report (Restoration Project 00455), ECOlogic Corp., Washington, D.C.

INTRODUCTION

This project funded a technical consultant to outline the key data and user issues that the Trustee Council should consider in designing a data system for GEM (Gulf Ecosystem Monitoring and Research Program), the Council's long-term monitoring and research program. In addition, a report analyzing existing systems that deliver similar data was to be prepared and strawman proposals developed for a range of data systems that could meet the needs of GEM. The consultant, Mr. Charles Falkenberg, died before completing that report. This report – which consists of a series of advisory memos from Mr. Falkenberg and language he drafted for a chapter on data management in the GEM Program Document – has been compiled by Trustee Council staff to create a permanent record of this information.

The documents comprising this report are:

To: Phil Mundy

From: Charles Falkenberg

CC: Molly McCammon, Bob Spies

Date: July 31, 2000

Re: Role of data coordinator for GEM

Overview

As I look at the success factors for the various projects that I have been involved in, the most important factor has always been the person who understands the fundamental goal of the system and can represent that to the users, suppliers, and the system developers. For GEM, this would be the director of data management or the data coordinator. Because this is the most important aspect of my recommendation I thought I would start with a memo describing this role to explain why I think it is so important.

I would like to see this description become more refined with input from yourself and any anyone else you would like to include. The result could be part of a final report for this project or it could just be a job description of this position for GEM. My own role is to respond to your concerns and tailor this description to meet the needs of GEM.

In summary, the role of the data coordinator will include:

- Articulating the vision of the data management system and data policies for GEM.
- Be the liaison between the GEM board and the data suppliers, the system users, and the developers of the system.
- Participate in the review of proposals for on-going monitoring projects as well as targeted research in order to evaluate the compliance with the GEM data policy. Approve the final delivery of the data from the project leaders.
- Oversee a staff that is responsible for maintaining the system, entering new projects, ingesting new data, and responding to needs of the user community.

The data coordinator will need to understand the issues that arise when data providers create data products and documentation. He will need to understand the user community and how they will want to access and use the data. He will be the point person for the development of any system to manage the GEM archive or generate data products for the user community and he will advise the board on what to expect from theses groups. The best candidate will be a data oriented professional with the credentials to bridge these groups, articulate their needs to the board and articulate the board's goals to the users and suppliers of the system.

Keeper of the vision

The vision of the GEM archive needs to be included in any description of the GEM program. The vision for the archive is closely tied to the vision of GEM as a whole, but the means and methods for building a long-term archive cannot be found unless the mission of that archive is stated explicitly. The current description of GEM presents the program as a resource for managers and policy makers, however the vision of how the data system will meet those goals is not included. Although the board will establish the vision for the archive, the data coordinator will advise the board during this process and articulate the vision to the data suppliers, the system developers, and the community of users.

The vision begins with a description of what it will provide today's user and, for GEM, the user in the future. Today, managers and policy makers will likely need tailored output products that will require several different datasets as input. Future research and management applications may be counting on the time series provided by GEM and therefor represent a different type of demand. A statement about what the GEM archive will provide to these future users is also critical to the mission of GEM.

The GEM data policy is another part of the archive vision. The data policy will govern how, when, and in what format, data will be provided to GEM. GEM may also wish to set fair use policies or provide the necessary documentation to govern how the community can or should use GEM data. The policies may be different for on-going monitoring and targeted research. It may be different for different types of data (e.g., manually collected data can take longer to compile than data collected by an instrument). The policy will also change over time as collection and storage technologies change.

Supporting the user community

A basic archive will store the data and allow the user to search by project or by variable (parameter) on order to locate and download data files. Even with a basic archive there will need to be on-going changes to the web site and issues and questions will come up on a daily basis. The data coordinator will manage the operational support of the user community and oversee a user support representative who would address the daily issues.

Supporting the target user community will require more than operational support. An active participation in the relevant meetings within this community and a working relationship with the key representatives of this community will be the coordinators responsibility. As GEM evolves the requirements of the target user community will change. The data coordinator will need to capture these requirements on an on-going basis and provide the board with updates on how GEM is meeting this community's need.

Making raw data available over the web will be the minimum goal of the data system. However, if the success of GEM is based on supporting the resource management and policy communities, it will need to create several products tailored to particular parts of this community. Custom data products, resource maps, graphs, and graphics are all possibilities. GEM could standardize these products and produce them on a regular basis with the most current data available. The analysis and definition of the applications to produce these products will require the data coordinator as well as an application developer. The data

coordinator will be in the unique position of understanding of the needs of the user community, the data available as input, and the technology available to create the product

Supporting the data supplier

The target users are not the only community that will have questions about how to interact with GEM. The data suppliers will also need to have a contact person at GEM that will provide information about the data format and data delivery requirements. My experience is that adding data to an archive can be a time intensive process. Modifications may be needed to the format of the data or the values of the meta-data elements. Often several iterations are needed to insure that valid data and meta-data are received. A project database will also need to be maintained to track the projects funded by GEM. A supplier support person that would answer to the data coordinator will likely be responsible for maintain the projects and ingesting new data.

In addition to the data delivery tools, most archives have software tools for adding new data to the system. The meta-data entry tools and automated ingest tools will reduce the load on a supplier support person and help ensure consistent input data. The data coordinator will need to be involved in the definition and specification of these tools and oversee the development of the tools by internal or external software developers.

GEM will also interact with data suppliers from several outside agencies that archive relevant data. The data coordinator will need to understand these resources, monitor the changes taking place in these other archives, and maintain a working relationship with the various agencies. Data on which standard products rely may need to be archived by GEM if alternative sources stop collecting or archiving them. The data coordinator will need to attend meetings and stay active within this community.

Support for the technologists

It is not clear how much of the data archive will be built by GEM and how much GEM can rely on existing data management technology. However, there will be several tools that will require some software development. It is very likely that GEM will be developing tools for ingesting data, managing the archive, or creating custom products. This development could be done by internal staff at GEM or by outside software developers. In either case, the data coordinator will play an integral part in defining the requirements for these tools and ensuring that the delivered products meet those requirements.

In my experience as a software developer, the coordinator role is the most important ingredient in the development of a successful system. There are a great many issues that arise during the software development process and the resolution requires an understanding of the user community, the scope of the solutions, and the ramifications of the compromises. The person who holds the vision and the goals for the system is the only one who has this understanding.

Support for the board

The data manager will oversee the ongoing operation of the data archive. This will include the management the project database, the ingestion of new data, and the decimation of data to the user community. Although the data manager cannot take on responsibility of monitoring the complete status of funded projects, he can report to the board on the status of the data related aspects.

The data policy will be an ongoing issue with the suppliers and the users. The data coordinator can help refine these policies and help the board enforce them across all projects. The data coordinator might also chair a data committee made of the representatives from several stakeholders that could help draft and communicate the data policies.

The most important part of acquiring consistent data in a timely manner is to include a requirement in the RFP. This means that each proposal must include a section describing what data will be delivered, when it will be delivered, and how it will be delivered. The data coordinator can assist the board in evaluating the proposals to ensure that they will comply with the GEM data policy. This will put the coordinator in a unique position to verify compliance as the project proceeds.

One of the stated goals of GEM is to identify the data needs of the management community and then perform a gap analysis of the data that not currently available to meet this community's need. This will be an on-going process and will require a broad understanding of the other data that is being collected in the Northern Gulf of Alaska. The data coordinator could assist in this periodic gap analysis and provide a unique perspective that includes both the user needs and the specifics of existing data.

As envisioned, the role of the data coordinator is quite large. Once GEM is in full operation, the role will probably require a staff of 2 to 3 people supporting the web site and the needs of both users and suppliers. The demands of maintaining contacts with the suppliers, users, developers and the board may prove to be too much. If so a logical split would be between the user-related functions and the data supplier-related functions. However, with a complete and competent staff a single person could perform these tasks in the foreseeable future.

To: Phil Mundy

From: Charles Falkenberg

CC: Bob Spies, Molly McCammon

Date: September 4, 2000

Re: A brief classification of web-based scientific data systems

---

The evolution of science data systems

The advent of the World Wide Web revolutionized the delivery of scientific data. Data centers began delivering data over the web and individual projects used the web to make non-proprietary data available to the public. As a result, the initial data systems were supply or mission driven. The groups that collected or archived data made them available to all interested parties. Generally, these data were collected by a single type of instrument or by a single research group and were homogeneous. The sites were quite useful to someone who was familiar with the project, the instrument or the data itself.

A large number of data-oriented web sites were built, thereby creating a need for a single location from which several sites with similar datasets could be accessed. Data clearinghouses, a second stage in the evolution of data systems, were developed to address this problem. Currently, these clearinghouse web sites store little, or no, data themselves but provide access to several other locations that do archive the data and meta-data. A clearinghouse will use summary meta-data to build a list of relevant datasets and provide a link to the site that contains the full meta-data and the sometimes the actual data. The Federal Geographic Data Committee (FGDC), for example supports the development of FGDC clearinghouses that use the FGDC meta-data standard. The Cook Inlet Information Management and Monitoring System (CIIMMS) is another sophisticated example of a clearinghouse.

In a third, and recent, stage of development, web-based data systems have been designed to meet the needs of a specific user community. These sites are not geared to a particular set, or type of data but rather toward a particular set of users. They represent a demand driven approach as opposed to the original supply driven approach. These sites provide data and, more importantly, data services that are tailored to the target audience. Examples of the target communities include K-12 educational support or particular aspects of the earth science or space science research community. The GEM system could well fall into this category because it has identified sections of the resource management community as a target user community.

This is a somewhat simplified description of data systems. Although these three types of web sites evolved in stages, they can also be thought of as different species rather than the evolution of a single species. All three of these types will have a place in the future landscape. Groups that collect data will continue to provide them over the web along with some simple data services. Data from these sites will be indexed by a clearinghouse and accessed over the web by a user-oriented site. The end user will come to the site that

is gear to his specific needs. That site will use the clearinghouse functionality to locate data at the collector's site or at other large digital libraries. The data will then be downloaded by the user's site, where his discipline specific data services can be applied.

A classification of web-based science data systems

When evaluating the complexity of a new data delivery site, it is important to be precise about the specific features it will provide. Common features emerge from data-oriented web sites and these provide a metric for classification. As an example, some sites provide on-line searching but the criteria may include keywords only. Other sites may also include spatial and temporal searches. The value of the classification is the understanding that is gained about the effort needed to provide the suite of features. Below is a summary of the relevant features.

- **Meta-data access**: Retrieval of the full collection or granule level meta-data
- **Data access**: Retrieval of the actual data files that were submitted to the archive
- **Access to remote data**: Access to other sites with relevant data
- **Directory-oriented navigation**: Hierarchical navigation through a fixed set of web pages
- **Search-oriented (spatial and non-spatial)**: Data selection that matches user entered values
- **Data subsetting**: Extraction of specific data values by space or time
- **Data reformatting**: Creation of an output format tailored to the user's application.
- **Data regridding or reprojection**: Re-sampling data for modeling and visualization
- **Generation of graphs**: Creation of graphs with selected data
- **Generation of maps**: Creation of maps and overlays with data values or locations
- **User registration**: Maintenance of a customer information and history
- **User profile**: Maintenance of customer preferences

Complexity is introduced when the number of possible services and features is increased but complex systems are not always an improvement. Simple systems can be inexpensive and be very useful to particular users. In addition, science data sites that offer complex services to a small user community can also be built at a reasonable cost. The best way to limit the overall expense is to have a clear idea of the target community and the specific needs that the system will address for that community

Systems combine a subset of the features described above and, therefore I have created a broader set of categories. A summary of these categories is provided below. The following sections describe each class in more detail with links to representative sites. These sites illustrate the features of that class of systems.

Directory-oriented sites provide access through predetermined lists of data files and directories. These sites can include FTP access or simple lists of files in an archive directory. Data files are found by navigating a hierarchy of directories that narrow the search for the specific files. These sites are inexpensive to maintain and can provide simple and easy access to data.

Search-oriented sites allow the user to type or select some essential parameters of interest and then perform a search for results that meet these criteria. Search criteria include meta-data keywords such as variable and collection name, or spatial domain and temporal range. The result is a list of data files in the archive that meet the criteria. Sites that offer additional data services on the results of the search are considered analysis-oriented and described in the next category.

Analysis-oriented sites allow the user to tailor the output after a search has been performed. They often support the creation of graphs or other visualization utilities over the selected data. These sites are frequently more sophisticated and include spatial or temporal search. However, it is also possible to provide analysis tools over a simple directory-oriented archive.

Clearinghouses are sites that do not archive data but catalog data that is stored at other sites in order to gather together sites with similar data. The FGDC supports the development of individual data sites that allow meta-data to be accessed from a central FGDC clearinghouse. This FGDC network is called National Spatial Data Infrastructure (NSDI) but there are many clearinghouses that are not part of the NSDI.

Image archives are built around the growing number of remote sensing images that are collected each day. These sites are often tailored to a specific satellite or mission and provide a wide range of functionality over the imagery. They can be an extension of the directory-oriented sites, providing simple static lists of images that can be traversed like a directory structure. More recent sites, however, provide additional services like zooming or image merging.

GIS-oriented sites are currently quite rare. These sites provide a map based interface to all data traversal and allow layers to be selected or deselected. The look and feel of these sites is similar to a GIS like Arc/View.

Directory-oriented Access

Directory-oriented sites are characterized by cascading lists of links to subdirectories or files. These lists represent a directory like hierarchy where subdirectories contain all data for a particular data type or time period. The meta-data may be available in one of the files but the hierarchy itself is a model of the meta-data.

Data can be added to these sites with little effort. Placing a data file in the right directory is often all that needs to be done. The files containing the meta-data and the data do not need to be parsed and therefore the format of these files does not have to be controlled in order to ingest the data. (Data and meta-data that are not in a standard format will cause difficulty to the user in the long run however.)

These sites are easy to maintain and update. They can be quite inexpensive and provide a great deal of functionality without a much complexity. However, they do not provide easy integration across datasets and often require the user to have an understanding of the projects and data.

- CMDL Data Archive
  - Parent site http://www.cmdl.noaa.gov/
  - FTP display and access
  - Data access http://www.cmdl.noaa.gov/info/ftpdata.html

- NCAR data support section
  - Parent site http://www.scd.ucar.edu/dss/
  - Hierarchical break down of disciplines and datasets
  - http://www.scd.ucar.edu/dss/catalogs/index.html
  - Limited search offered over dataset

- Space Environment Center
  - Parent site http://www.sec.noaa.gov/
  - Gopher style retrieval
  - Data access http://www.sec.noaa.gov/Data/index.html

Search-oriented Access

The distinguishing characteristic of a search-oriented data site is the ability to enter a search term, date range, or a spatial range in order to get a list of relevant datasets. This makes it easier for a user without preexisting knowledge to find relevant data (e.g. precipitation data). Search-oriented sites also allow data to be integrated across discipline. As an example, it was important for the Sound Ecosystem Assessment (SEA) to access fisheries data and oceanography data for the same region and time period. This can be quite cumbersome with a directory-oriented site.

This search functionality is usually implemented with a database of meta-data that can be queried with the search criteria that is entered by the user. The meta-data for any dataset submitted to the archive must therefore be well formatted so it can be parsed and ingested into the database. In most cases, however, the data files that are downloaded to the user are the same files that were submitted to the archive. These files are not necessarily read during the ingest process and formatting discrepancies can be overlooked at this point.

Keyword search is the most basic type of search. Sometimes these terms are project specific (e.g. station name or cruise id) but often general terms are available (e.g. precipitation, temperature). Spatial and temporal search capabilities are common in a more specialized group of search-oriented sites. Most Database Management Systems (DBMS) offer spatial search capability but it is often an added feature. Spatial search also requires a consistent spatial reference system (or map projection) and this introduces some additional complexity. However, spatial and temporal criteria are often the most important for locating relevant data.

- NOAA site for Oceanographic and Meteorological data (AOML)
  - Access to several tailored sites
  - Parent site http://db.aoml.noaa.gov/dbweb/RetrvData.html

- Applet based map
- Example site http://db.aoml.noaa.gov/dbweb/CTD/CTDGUI.html

- Ocean Chemistry data (AOML)
  - Parameters, time, no spatial selection
  - http://www.neptune.aoml.noaa.gov/sk.html

Analysis-oriented sites

Analysis-oriented sites focus on the data more than the meta-data. Although some level of searching is often supported the distinguishing characteristic is the ability to generate graphs or maps of the data, or the ability to reformat or subset the data. Graphs are specific to the type of data. Examples include hydro-graphs, cross sections, or iso-line graphs.

Directory-oriented and most search-oriented sites return the data files that were submitted to the archive. This means that the data returned to the user would overlap, but might extend beyond, the time range or spatial extent that was used as the selection criteria to locate the data files. A site that offers data subsetting, will clip the data to the requested range or extent. These sites also allow the user to choose different output formats, which may include graphs or maps.

In order to implement this functionality, the analysis-oriented site must be able to read and parse each data file submitted to the archive. This requires additional steps during data ingestion and the data submitted to the archive must meet the expected format. Therefore, most current analysis-oriented sites supply homogeneous data from single type of instrument and produce graphs that are standard for those types of data.

Some of these sites create custom data products the meet the user's specifications. These products may be shipped to the user on tape or CD or they are stored at the site and made available for a limited amount of time. These sites usually require the user to register by entering a name, address, and email and selecting a password. Maintaining customer information adds to the complexity of the system. The customer database must be backed up regularly and the system must enforce some level of security and privacy. However, it also allows for closer ties to the customer base and a better understanding of whom is using the data and services.

- Radiosonde real-time data archive (NOAA FSL)
  - Access to wind data by time and space with plotting capabilities
  - http://raob.fsl.noaa.gov/

- River discharge hydrographs
  - Access to graphs and data from river gauging stations around the world
  - Parent site http://pyramid.sr.unh.edu/csrc/hydro/welcome.html
  - One of the data access sites http://www.rivdis.sr.unh.edu/maps/

- Modeling and Monitoring of the Neuse River (UNC)
  - Nice web site with access to results data
  - http://www.marine.unc.edu/neuse/modmon/

- Web-Based system for Terrestrial Ecosystem Research (WEBSTER)
  - Provides spatial and temporal subsetting and several output formats
  - http://webster.sr.unh.edu/ (follow links to "Data")

Meta-data Clearinghouses

With the proliferation of scientific data sites there became a strong need to provide a single point of entry for a set of distributed data resources. Clearinghouse sites are quite varied in what they offer and how they are built. However, the distinguishing characteristic is that they do not archive data locally but maintain a catalog of the data at several other sites. If the user finds data through the clearinghouse, he must follow a link to the actual archive site in order to access the data. Umbrella projects (e.g. Long Term Ecosystem Research and PICES) often maintain a clearinghouse that includes a catalog of the data at the individual project sites.

In some cases, these sites maintain a database of meta-data that can be search locally. These meta-data must be submitted to the clearinghouse or retrieved from the remote sites automatically. In other cases, meta-data is also stored at the distributed sites and a meta-data server is provided that can be queried from the clearinghouse. This second option is slower but it provides the most up to date description of the data at the distributed site. In either case the clearinghouse produces a list of the datasets or data granules at several distributed sites that meet the criteria. Each entry in the list can include a link to the distributed site to access the data.

The FGDC has created a content standard for meta-data that describes spatially coordinated data. All of the clearinghouses that are part of the NSDI support this meta-data content standard and several other scientific sites support meta-data in this standard. This standard is well suited to GIS data and is being reviewed by the International Standards Organization (ISO) as an international spatial meta-data standard (ISO211).

- Global Change Master Directory (GCMD)
  - Access to a huge number of datasets related to global change
  - Caches meta-data locally for quick search
  - Parent site: http://gcmd.nasa.gov/
  - Data site requires Java, text only is also available
  - http://gcmd.nasa.gov/cgi-bin/md/zgate?SERVICE=INIT&FORM_HOST_PORT=FREETEXT

- Alaska Geographic Data Committee (AGDC)
  - Queries remote sites for the meta-data
  - Allows a subset of the remote sites to be searched
  - Parent site: http://agdc.usgs.gov/
  - FGDC home page: http://www.fgdc.gov/

- Default user interface: http://agdc.usgs.gov/cgi-bin/searchgate
- Provides the ability to tailor the search interface

- Cook Inlet Information Management/Monitoring System (CIIMMS)
  - Caches some meta-data locally for quick search.
  - Allows the selection of remote databases as well
  - Parent page: http://www.dec.state.ak.us/ciimms/

Image Archives

Image archives provide access to a homogeneous set of images, often collected by satellite. Most of the current image-related sites could be classified as directory-oriented sites. They provide a series of pages with predetermined lists of images cataloged by day or by region. However, because the data are all images some value-added functionality can be offered. In some cases these sites are updated in real time with the most recent image from a particular platform. Some sites provide some image analysis tools as well with pan and zoom capabilities.

- GOES and AVHRR images (ETL http://www.etl.noaa.gov/)
  - Directory-oriented
  - Parent page: http://www1.etl.noaa.gov/climsat/
  - Sample page: http://www1.etl.noaa.gov/climsat/realtime.html
  - Sample page: http://www1.etl.noaa.gov/climsat/goes9.html

- Minnesota Department of Natural Resources - ForNet
  - Parent site: http://www.ra.dnr.state.mn.us/
  - Image archive provides spatial search and zoom and pan
  - Requires Java (HTML only version available)
  - http://www.ra.dnr.state.mn.us/imageview/tm/ivtmJava.html

GIS-oriented

GIS has been around for many years, but has been slow coming to the web. Although this is changing and more web-based GIS tools are available, most of them are geared toward traditional GIS data and not time-coordinated science data.

GIS functionality can be quite complex to implement over the web. It often uses some Java applet at the client and/or special servers to generate maps. ESRI is working on several new technologies for delivering maps over the web. In addition, the Open GIS Consortium is developing an open standard for web-based map delivery called the Web Mapping Testbed.

- NOAA data centers and archives (e.g. NODC, NGDC, NCDC)
  - Interesting GIS-oriented interface using HTML
  - http://gis.ncdc.nndc.noaa.gov/atlas.htm

- The USGS GEODE server
  - Access to several types of data through a GIS interface
  - Requires special Java tools and access
  - http://dss1.er.usgs.gov/

- Open GIS Web Mapping Testbed
  - Open standard for maps on the web
  - Parent site: http://www.opengis.org/wmt/
  - Example implemented by NASA Digital earth: http://www.digitalearth.gov/
  - http://viewer.digitalearth.gov/

Where GEM fits

The details of the GEM system are not complete and the role of the data system is evolving. However, there are a couple of important observations. At its most basic, GEM is a data collection and archiving program. It could well provide a data archive and only basic data services over that archive. User oriented sites that make use of GEM data would access data from the GEM archive and provide services on those data that are geared toward its own community. Sites like CIIMMS could provide an intermediate level of searching and download that GEM could use as a more general front end.

However, GEM can play a stronger role. It has positioned itself as a system to support resource managers in the Gulf of Alaska region. In order to achieve this, GEM will need to act more like a site tailored to this community, pulling data from several other sources and generating data products on a regular basis that meet this communities needs.

These are two very different views of the GEM data system. Indeed, there are other possibilities as well, but the ramifications of this choice and role of the GEM system in the larger data infrastructure will be described in another memo.

To: Phil Mundy

From: Charles Falkenberg

CC: Bob Spies, Molly McCammon

Date: December 29, 2000

Re: A "strawman" proposal for a GEM data system

Background

At the October workshop there was a public consensus that consideration should be given to data management issues up front, and that these issues were going to present an ongoing challenge to GEM. The data managers that attended the breakout session were in general agreement that GEM would need to exercise a good deal of control over the data. This included centralizing anything that was not being well maintained at an external site. There were some dissenting opinions from some researchers and this will no doubt be an ongoing debate.

Several people at the meeting suggested that I put up a "strawman" that would allow all parties to have something to focus on (and throw stones at). I have been reticent to do this because of the potential controversy but the GEM plan presented at the October workshop generated a healthy debate and I am sure this will as well. The following proposal breaks out the critical components, assumptions, and priorities related to a GEM data system.

Guiding Assumptions

To begin I would like to outline the assumptions that I have guided my design. These are the salient points that I think will have the greatest impact on the final GEM system. These include well-known procedural and policy issues followed by data storage and data retrieval issues.

1. GEM will need to adopt an evolving set of data policies and standards that are appropriate for each discipline. Standards emerge from practical data application and no single set of standards is comprehensive enough to anticipate the requirements of the different disciplines involved in GEM.

2. It will be necessary for GEM to be able to hold projects accountable for the submission of data that is promised under a GEM contract. Providing a high degree of visibility into the promised deliverables from each project may be the best do this.

3. GEM will need to provide an archive to store at least a subset of the GEM data. Even if some GEM data is stored at other sites, GEM will need to provide an archive for many GEM datasets and any other relevant data that is at risk of being lost.

4. GEM will need to have a flexible set of data retrieval and synthesis options to support a changing set of user requests. The user community will likely be quite diverse and need different types of data retrieval tools.

The differences between data storage issues and data retrieval issues are important to make clear. Although these two are not completely unrelated, the drivers are quite distinct. The user community, which has not yet been finalized by GEM, will drive the data retrieval and analysis issues. The data retrieval systems will have to evolve and expand along with this user community. The data storage decisions, however, will be driven more by the characteristics of the data. These storage issues are also more immediate (a sentiment that was expressed at the October workshop). There are already data that is at risk of being lost and need to be archived and GEM will begin collecting more data soon.

Data Policy

The GEM data policy will set out important guidelines for the smooth operation of the over all data system. Separate from this will be the standards that are used to validate the data submitted to GEM. These standards should include format and content standards such as the FGDC standard with a set of "valids" allowing the meta-data to be searchable. Examples of "valids" include species codes, station names, and data units. GEM can adopt existing standards from the EVOS ecosystem projects and from the individual disciplines from which the data are collected.

Data Submission

GEM will need to pay close attention to the process by which data are submitted. Data submission is often a problem and a smooth procedure will be critical if GEM is to build a long-term archive. The data policy needs to address this procedure directly. Proposals need to include the list of datasets that will be submitted along with the formats and content of those datasets. The proposal and the data need to be linked together and available to the public to provide visible accountability.

Only the person who is responsibility for fulfilling the contract, can insure the scientific merit of the data. However, GEM will need to check the format of the data and meta-data. For example, GEM will need to insure that the dates are in a valid format valid and the values like station ids or species codes are consistent across the archive. If problems are found, GEM can pass data back to the supplier and ask for corrections. This will prevent GEM from making any updates to the data that are submitted to the archive.

Data Storage

GEM will need to maintain an archive over which it has basic control. Much of the new data will need a home and, as GEM evolves, data that is at risk of being lost will need to be rescued. GEM could contract out these services or it could maintain its own set of hardware and software systems.

The archive can be separate from the end user search and retrieve tools. The archive will need to store meta-data about each dataset but the searchable database could reside elsewhere with pointers into the archive. The archiving plan will need to address how the hardware and software will be upgraded to the new technologies that will become available over the next century. These issues can be separated from the retrieval decisions. GEM will have some time to refine the tools used for retrieving data but the data will need a home soon after it is collected.

Data Retrieval

GEM will need to provide for basic access to the data that is in the GEM archive and the critical data that is located in an external repository. These external repositories may include data that was collected by GEM as well as data that GEM has identified as critical but that was collected by others. Current web technology makes this task quite tractable.

GEM will also need to augment a general system if it is to meet the needs of particular user communities. This would include customizing the data in the archive and creating a data product that is tailored to a particular community.

Strawman proposal

From these assumptions, I have developed an architecture, which consists of components to address the areas of concern. The following list of steps is a summary of the system requirements:

Establish an ongoing policy and standards group

- Refine a data policy that includes data guidelines for proposals and for submission
- Write a standards document that includes recommended formats and list of "valids"
- Establish a process by which the standards can evolve over the lifetime of GEM

Provide a connection between administrative systems and data archives

- Link all documentation about a project together with pointers to the data.
- Provide web based access to the public portions of these data
- Provide GEM staff with the tools to add and modify the administrative data

Establish a GEM archive with basic ingest and retrieval

- Build systems to check the input data to insure that it meets the format standards and establish a procedure for iterating with the provider if any discrepancies are found.
- Select a basic archiving strategy and implement a system

Provide a general access interface to GEM data

- Use CIIMMS or another clearinghouse system to provide immediate access to all GEM data in its original form
- Provide the GEM staff with the tools to add data to this system

Support custom data product generation

- Use GEM staff to create custom data products from GEM and non-GEM data for targeted user communities.
- Support the automation of these systems so that the custom products can be produced on a regular schedule without staff intervention.

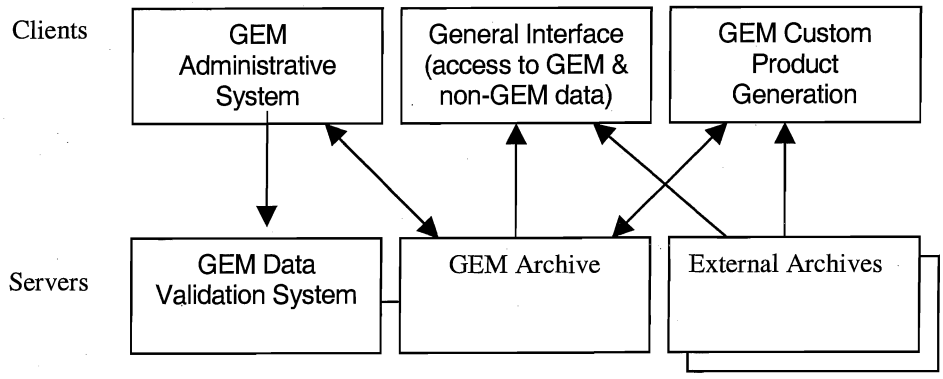


Figure 1. GEM data system proposal

Figure 1 shows the components described above and how the data will pass between them. Each component is explained more fully below.

GEM Administrative System

The GEM staff will use the databases and tools in this system to administer GEM projects. EVOS has several existing databases that are used solely for administration. The GEM system should tie project proposals and year-end reports to the actual data submitted by the project. This will insure that the project is accountable for the data submitted to GEM. In addition, future researchers can locate the project proposals and reports that are associated with the data and meta-data. Some of these data will be public and therefore this system may need simple web interface that allows a user to search by project id or text string. The database will need to maintain a spatial reference to support ongoing gap analysis by GEM.

GEM Data Validation System

The data validation system will include a variety of programs and scripts that check the format of the data that is submitted to GEM. Species codes, date formats, and required meta-data are examples of the kinds of data elements that will need to be checked before they are inserted into the database. This system will enforce the standards that are established by the standards body, it cannot insure the scientific value of the data. GEM will need to control the implementation and the execution of these validation routines. GEM may provide the routines to its suppliers so they can check data before submitting them, but GEM will be ultimately responsible for insuring the data it archive meet the GEM standards

GEM Archive

This is the storage component of the GEM system. It will support the basic storage and retrieval of a large number of data and meta-data files. This system will assign a unique identifier to both the data files and meta-data files. Most complicated searching will be done by other, more user-oriented, systems and the GEM archive will only be accessed once the user has identified the necessary data. As examples, the administrative system will be interested in tracking the project id, a general interface will search FGDC meta-data, and the custom product system may have several different spatial and non-spatial search fields. In order to minimize cost, an established archiving center or Distributed Active Archive Center (DAAC) could operate this component under contract to GEM.

External Archives

The gap analysis done by GEM has uncovered the relevant data that are currently being collected and archived by other projects. These data may be in databases or in ASCII files and available over the web. Indeed, GEM may fund projects that have internal archiving capability that will be used to store the GEM data. GEM will need to insure that these data meet the standards established by GEM but there may be no reason why these data cannot continue to reside in an external archive. However, over the next century some of these external archives will need to be brought into the GEM archive for long-term storage and it will be too late to check for non-standard data at that point.

General Access Interface

GEM will need to make all of its data public within a reasonable period after it has been collected. Although the user community is not yet finalized, GEM data will be useful to a large number of yet undefined applications. A general access system, such as CIIMMS, is ideal as a basic interface to GEM. It is not geared to any particular user and has the flexibility to deliver all of the GEM data for the near future. In addition, these clearinghouse systems will allow GEM data to be integrated with external data sources. GEM could make data available through one of these systems within days after it was archived. GEM could use CIIMMS or another existing system and perhaps even add a GEM nameplate.

Custom Data Product Generation

The benefit of a general access system is that it is not tailored to a specific user community. The search criteria are general and a user can download the data as it was submitted to GEM. As GEM grows it will want to produce products that are designed to meet the specific needs smaller groups of users. Examples include resource management applications that use GEM data to produce monthly reports of conditions of the Gulf or predictions of physical or biological trends. These applications will need to be built for the specific communities and the results may need to be produced by hand before being automated. GEM staff could work with the specific user group and refine the requirements for these products before any system is developed to generate them automatically. As GEM evolves, the number of these small, targeted applications will grow and GEM will become vital to large number of smaller groups.

Summary

This "strawman" proposal has been designed to minimize cost were possible. A report circulated to the data committee in October suggested that cost of managing data for a large project like GEM should be between 10% and 20% of the overall project budget. Existing retrieval systems and archiving centers can be used offset the cost of the overall GEM system. This proposal focuses on the administrative systems and the archiving components first. Sophisticated data retrieval systems are postponed in favor of existing systems that are general and easily supported. As the user community is identified, custom products can be produced manually at first and then automated after they have stabilized. These points are summarized below:

- The administrative systems used by EVOS would be combined and extended to support the mission of GEM and access to the data submitted by each project.

- GEM would establish data standards and build a system to check the data submitted to GEM.

- A storage management group in one of the federal or state agencies could operate the archive system component. GEM would build or extend the software needed for the archiving system.

- The CIIMMS project could supply a good deal of the technology and functionality for a general access interface. GEM would hire data support people to add data to the gateway and to manually produce custom products.

The interaction between the components of this proposal has not been included in this discussion. There will be several issues about how meta-data and data are packaged and labeled; how the archive will be accessed over the Web; and how the validation system will interact with the administration system and with the archive itself. The value of this proposal is the partitioning of the components and the basic relationship between them.

My hope is that this "strawman" will generate debate that will refine the requirements for these components. Future design documents will include increasing levels of detail on what each of these components include and how they fit together.

To: Phil Mundy

From: Charles Falkenberg

CC: Molly McCammon; Bob Spies; Andy Gunther

Date: June 26, 2001

Re: Background for GEM data policy

Phil,

This memo is designed to provide the background needed for the GEM data policy. It includes a summary of several data policy statements that could be used as a starting point by GEM. These include an early policy statement by the U.S. Global Change Research Program (GCRP) data policy, the data policy for the GLOBEC program and a data policy from the Environmental Monitoring and Assessment program (EMAP) at the EPA. The GLOBEC policy was built from the GCRP and the EMAP policy was built from both, and therefore, these represent somewhat of a policy evolution.

Each policy is summarized and discussed in the body of this memo and the last section is a set of general recommendations for a GEM policy. There is an appendix for each data policy that contains excerpts from each policy and a final appendix that has as short list of data policies from other projects.

Thanks,

Charles Falkenberg
Director of Research
ECOlogic Corp.

US Global Change Research Data Policy

In July of 1991 the U.S. Global Change Research Program (GCRP) established a data policy to facilitate full and open access to quality data for global change research. This policy represents the position of the U.S. Government on accessing GCRP data.

The policy is made up of seven statements and an annex that describes each statement in detail. The policy statements establish:

1. A commitment to maintenance and accessibility of long-term data sets.
2. Full and open sharing of the data for global change researchers
3. Preservation of the data over the long term at a designated archive
4. Easily accessible meta-data about the data and aids for obtaining the data
5. The use of national and international standards
6. Minimal cost for providing the data

7. Initial periods of exclusive access to the data which are set by the funding agencies

This data policy introduces the basics and establishes a foundation for the policy statements of groups within or associated with the GCRP. It sets out a commitment to store meta-data as well as data and make both "*easily accessible*". It also establishes the objective of "*full and open sharing*" and gives the funding agency the authority to define the duration of any exclusive data use period.

The full text of the statements are included in appendix A, and the full text of the letter and a more complete description of each of the above points can be found at the following web site: http://globalchange.gov/policies/dmwg/dmwg-gcp.html

GLOBEC Data policy

The GLOBEC data policy was drafted in 1994 and is still in effect. GLOBEC is part of the GCRP and uses the GCRP data policy as a foundation. The GLOBEC policy consists of 12 policy statements that cover many of the details of the program, including the cooperation between disciplines and the support for interdisciplinary research. The 12 statements are summarized below and bullets have been used to summarize the text associated with the statements.

Philosophy and Motivation

- To enhance the value of the data by providing a set of guidelines for the collection and storage of the data sets

Objectives

- To establish a data management office (DMO)
- To increase the value of the data by supporting interdisciplinary research
- To make plans for data collection and storage prior to the field experiment
- To share data with the scientific community to maximize the value

Quality and Methodology

1. Investigators must use methods that are adequate to insure data quality
2. Documentation of the collection techniques must be submitted with the data
3. The investigator is responsible for estimating accuracy and precision and recording it
4. Physical data must be collected with biological data
5. The investigator is responsible for insuring the quality of the data is as high as possible

Data Exchange and Archival – Methods and Schedule

- It is not ethical to publish data without prior attribution or co-authorship
- The investigators are entitled to the fundamental benefits of the data set
- The purpose of the archive is to facilitate collaboration between scientists
- Any substantial use should include collaboration with the data collectors
- The DMO will keep a list of all data access

- The data policy will be supplied with any data

6. An inventory of measurements will be submitted within 3 months of collection
7. Measurements without manual analysis will be submitted within 6 months
8. All measurements and standard analysis will be available within 1 year
9. Data will be submitted to DMO or be available online as a GLOBEC database
10. The data will eventually transfer to NODC

Sample Preservation

11. Biological samples will be preserved for later analysis

Modifications to the policy

12. Requests for exemptions will be submitted to the GLOBEC steering committee

Relevance of GLOBEC Policy to GEM

The GLOBEC data policy is quite specific and details the protection of the investigators, the data and meta-data that will be submitted, the responsibilities of the investigators and the data management office (DMO), and specific time frames for the submission of data. GEM will be similar to GLOBEC and therefore, this policy, and the specifics included in it, have a great deal of relevance.

The data collection rules outline several requirements before a cruise begins. The investigator must submit a description of the data that will be collected and the DMO will use this to create a data plan before cruises begin. Anticipated precision and accuracy, and a complete description of the collection methodologies are part of the data plan, which is posted on the GLOBEC site for review by participating scientists.

The data submission rules describe the specific data and meta-data that will be submitted and timetables for different types of data. An alternative to submitting data is provided, if the investigator is willing and able to maintain a GLOBEC standard distributed database. This ensures that even if the data is not archived by the DMO, the data meet the GLOBEC standards for both format and content.

The GLOBEC policy also provides specific protections for the intellectual investment of the scientist and makes a strong ethical stand for the use of the data by other researchers. In addition, the policy states that the DMO will maintain a list of all data requests, by dataset, and include the policy when those requests are filled.

The policy outlines the responsibilities of the DMO during the collection, storage and archival stages. The DMO produces a data plan for the cruise, accepts the data and reports on data submissions to the program manager and the steering committee.

The one missing element from this policy is the specifics of QA/QC. Although some of this could be covered in a data management plan, the basic philosophy belongs in the data policy. The best-case scenario would allow time for the DMO to check the submitted

data against the data standards and report any inconsistencies back to the investigators. These errors might include invalid dates, units of measure, or unknown station or species identifiers. The investigator will then have an opportunity to make the necessary corrections and view the data on a limited access server before giving the management office a final signoff to general access to the data.

Excerpts of the policy are included in appendix B, and the full policy is available at http://cbl.umces.edu/fogarty/usglobec/reports/datapol/datapol.contents.html

Additional documentation about data management at GLOBEC can be found at the following URL: http://globec.whoi.edu/globec-dir/globec-doc.html

EPA EMAP data policy

The Environmental Protection Agency (EPA) is conducting an Environmental Monitoring and Assessment Program (EMAP), with slightly different goals than GEM. The data policy for EMAP was drawn from the GLOBEC policy and reflects certain additional federal requirements as well. A short policy statement was written up for a pilot study conducted by EMAP in the western states and it is included in appendix C. The statements are summarized below

This statement is shorter than the GLOBEC policy and less specific but it has some addition points that are worth noting. The policy includes:

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. Different types of data should be available within the study within 6 or 15 months
4. All data should be available publicly within 24 months
5. All data will be identified with a citation
6. All data will be made available on the EMAP public web site
7. Participants will adhere to the EMAP standards
8. Citations will be provided to the EMAP Bibliography
9. Active participation in the EMAP web site is encouraged for all participants
10. Stream and coastal data will copied to STORET for long-term archival
11. The text of a data use statement to be included in proposals.

Relevance of EMAP policy to GEM

The EMAP policy draws from the GLOBEC policy and also includes specific parameters for how soon after collection data must be submitted to the archive. It makes a stronger statement about data distribution, stating that all submitted data will be available through the EMAP web site. Although the policy mandates a data citation it does not provide the same protections for the investigators from whom the data originated.

The STORET data management system is identified as the long-term storage site for the some of the EMAP data. Unlike GLOBEC however, the policy does not indicate that it will be removed from the EMAP web site once it is submitted to STORET.

This policy does address the issue of quality assurance (QA/QC). The second rule in the policy stresses full and open sharing as a fundamental objective and states that all data will be publicly available *"following verification and validation"*. The process of verification and validation is spelled out in the EMAP Information Management Plan: 1998 - 2000 and in other EMAP document on the web. These documents compare the process to the publication of scientific results, including 4 steps that are accomplished the data task group and the Information Management (IM) staff. A procedure similar to this would be appropriate for GEM. The steps are transcribed below from the EMAP site (http://www.epa.gov/emap/html/pubs/docs/imdocs/guiman.html):

- ***Submission:*** *The Task Group submits data and its associated metadata, along with appropriate publication designations, instructions, authorization, and approvals. This will normally occur through FTP transfer to a restricted access directory on the USEPA Internet server.*

- ***Editorial Processing:*** *The IM Staff logs the submission and conducts appropriate quality assurance (QA) checks and reviews. If necessary, the Task Group revises the submission to meet publication standards. The IM Staff will not alter the content of any submission.*

- ***Formatting:*** *The IM Staff posts the submission to the USEPA limited access server, including making any format conversions or inserting links (pointers) to related data sets on the server. The Task Group reviews this "proof" version of the publication, within a specified time frame, for errors.*

- ***Public Release:*** *The IM Staff posts the submission to the EPA Public Access Server*

Quality Assurance and Control refers to the data as well as the meta-data. For data, QA/QC refers to the assurance that the instruments are returning and recording accurate data. This is true for the meta-data as well but the meta-data must also adhere to the standards outlined for the project. These include the format of the data and the sets of valid values that are required for many of the meta-data elements. This type of quality control will be important to GEM and procedure will be needed to insure it is done properly.

Excerpts from the policy for the EMAP pilot study are included in appendix C. The complete document can be found at the following URL:
http://www.epa.gov/emap/html/pubs/docs/imdocs/wpdatapol.pdf

Additional information about EMAP is available at the web sites below:
http://www.epa.gov/emap/
http://www.epa.gov/emap/html/pubs/docs/imdocs/index.html
http://www.epa.gov/emap/html/pubs/docs/imdocs/impover.html

Recommendations for a GEM data policy

There are several similarities in these data policy statements. First is the stated goal for each policy that the data are to be freely available to the community with a specific set of deadlines. In addition, the GLOBEC policy contains specific statements that protect to interests of the researcher. These protections include the proper data citation and an ongoing list of who has downloaded the data. GEM could increase the value to the investigators of providing public access to the data by crediting these citations to the data supplier and considering this credit during the evaluation of new proposals.

The EMAP policy is concise and includes the philosophy for QA/QC. The process of QA/QC can be drawn out and, for GEM; it could consume a considerable amount of a person's time. A comprehensive procedure and a clear statement of policy are the best way to make this a regular and efficient part of the data submission activity. The EMAP procedure includes posting the data to a "limited access server" for review by the data submitter in order to achieve a final signoff.

Steven Hale, the information manager for the EMAP confirmed that the EMAP policy had worked well and that he would not make additional changes for the future. He also said that the QA/QC process was drawn out for some of the participants (and datasets) in the EMAP program but that the process they put into place had worked well.

Robert Gorman who manages the GLOBEC DMO also confirmed that their policy had worked well. He had a very practical opinion on the problem of accepting and validating data. Like the SEA program, he placed as few restrictions on the data providers as possible in order to reduce the obstacles during data submission. The experience of EMAP is encouraging, however, that a clear policy and procedure can improve this process.

The wording of the GEM data policy will need to come from the chief scientist and several committees will review and refine it. The GEM data policy should, however, include the following key components:

- A description of the long term goal of creating a GEM data archive
- A commitment to full and open sharing
- A summary of the data and meta-data that are expected to be submitted
- An outline of when the data and data descriptions must be submitted (This should include an planning document before the data is collected and the length of the exclusive access period.)
- The commitment that GEM will make to protect the interests of the data source
- A statement describing the process of QA/QC and the standards that will be applied during this process

- A description of how and when the data will move to NODC
- How GEM will support distributed access to data maintained by the investigator

GEM will need a full data management plan which will be much more comprehensive than the policy. The plan will describe the standards and procedures in detail and therefore the policy can be more concise. The policy, however, will convey the intent of the program and provide the key drivers for the data management plan.

Appendix A: The data policy of the U.S. Global Change Research Program

Below are the data policy statements for the Global Change Research Program (GCRP) from the GCRP web site http://globalchange.gov/. These were drafted in 1991 and represent the U.S. government's commitment to full and open sharing of data within the GCRP.

The full text of the letter and a more complete description of each of the above points can be found at the following web site:

http://globalchange.gov/policies/dmwg/dmwg-gcp.html

The overall purpose of these policy statements is to facilitate full and open access to quality data for global change research. They were prepared in consonance with the goal of the U.S. Global Change Research Program and represent the U.S. Government's position on the access to global change research data.

- *The U.S. Global Change Research Program requires an early and continuing commitment to the establishment, maintenance, validation, description, accessibility, and distribution of high-quality, long-term data sets.*

- *Full and open sharing of the full suite of global data sets for all global change researchers is a fundamental objective.*

- *Preservation of all data needed for long-term global change research is required. For each and every global change data parameter, there should be at least one explicitly designated archive. Procedures and criteria for setting priorities for data acquisition, retention, and purging should be developed by participating agencies, both nationally and internationally. A clearinghouse process should be established to prevent the purging and loss of important data sets.*

- *Data archives must include easily accessible information about the data holdings, including quality assessments, supporting ancillary information, and guidance and aids for locating and obtaining the data.*

- *National and international standards should be used to the greatest extent possible for media and for processing and communication of global data sets.*

- *Data should be provided at the lowest possible cost to global change researchers in the interest of full and open access to data. This cost should, as a first principle, be no more than the marginal cost of filling a specific user request. Agencies should act to streamline administrative arrangements for exchanging data among researchers.*

- *For those programs in which selected principal investigators have initial periods of exclusive data use, data should be made openly available as soon as they*

become widely useful. In each case the funding agency should define the duration of any exclusive use period.

Appendix B: The data policy of the GLOBEC program

Below are several summary paragraphs and 12 data policy statements that were excerpted from the GLOBEC data policy. This summary is included as an overview. The complete policy is available at this web site:

http://cbl.umces.edu/fogarty/usglobec/reports/datapol/datapol.contents.html

Philosophy and Motivation

The fundamental objectives of U.S. GLOBEC are dependent upon the cooperation of scientists from several disciplines. Physicists, biologists, and chemists must make use of data collected during U.S. GLOBEC field programs to further our understanding of the interplay of physics, biology, and chemistry. Our objectives require quantitative analysis of interdisciplinary data sets and therefore data must be exchanged between researchers. To extract the full scientific value, data must be made available to the scientific community on a timely basis.

Precedent and perception have resulted in a disparity of data collection, storage, and archival methods. This makes the exchange of data difficult and may suppress dissemination of data. The U.S. GLOBEC Scientific Steering Committee seeks to enhance the value of data collected within the U.S. GLOBEC program by providing a set of guidelines for the collection, storage, and archival of these data sets.

The policy detailed below applies to all U.S. GLOBEC investigators. Field data, retrospective data sets, and numerical experiments must all be included in the U.S. GLOBEC database.

Objectives of the U.S. GLOBEC Data Policy

The U.S. GLOBEC data policy consists of twelve concise statements addressing the collection, sharing, and archival of data within U.S. GLOBEC programs. Preceding these statements is text which seeks to provide some details and the motivation behind the specific policy statements. In setting forth these statements and establishing a data management office, the Steering Committee intends to increase the value of data collected in support of our mutual scientific objectives. The Steering Committee will not attempt to force an investigator to comply with these policy statements, but does wish to encourage and organize full and accurate communication. Plans for data collection must be communicated prior to execution of a field experiment to insure that all necessary data are collected. Data collected during field programs or in laboratory experiments, organized for retrospective studies, or produced from a model must be shared with the scientific community to maximize the scientific value of the data.

Quality and Methodology

1. *Investigators must select methods and equipment which are adequate to insure that data quality is sufficient for the objectives of the U.S. Global Change Research Program and U.S. GLOBEC.*

2. *Documentation of the measurement and analysis techniques used to produce the data set must be submitted with the data to the U.S. GLOBEC Data Management Office.*

3. *The investigator is responsible for estimating the accuracy and precision of each measurement and recording this information in the database.*

4. *The overall objectives of the USGCRP and U.S. GLOBEC demand knowledge of the physical setting of the ecosystem. To this end, physical data must be acquired with biological measurements.*

5. *The investigator is responsible for insuring that the quality of the data available to the community is of as high a standard as possible.*

Data Exchange and Archival - Methods and Schedule

A data system must facilitate the exchange of data and insure the long-term existence of the data set. National requirements for submission of data to the National Oceanographic Data Center (NODC) must be satisfied either by the investigator or by a data management office. Because both submission and retrieval of interdisciplinary data sets, a data management office is needed to facilitate exchange and cooperate with NODC on establishing a national capacity for the exchange of interdisciplinary data sets. While this office will accept responsibility for submitting data to NODC, the primary objective of this office is to provide a mechanism for the exchange of interdisciplinary data sets.

The reader is reminded that it is not ethical to publish data without proper attribution or coauthorship. ***Beyond this, the U.S. GLOBEC Scientific Steering Committee believes that the intellectual investment and time committed to the collection of a data set entitles the investigator to the fundamental benefits of the data set. Therefore, publication of descriptive or interpretive results derived immediately and directly from the data is the privilege and responsibility of the investigators who collect the data. The purpose of a data archive is to facilitate collaboration between scientists, the combination of multiple data sets for interdisciplinary and comparative studies, and the development and testing of new theories. Any scientist making substantial use of a data set should communicate with the investigators who acquired the data prior to publication and anticipate that the data collectors will be co-authors of published results.*** *This extends to model results and to data organized for retrospective studies. As possible, the U.S. GLOBEC Data Management Office will encourage and*

facilitate the ethical and courteous use of data within the archive. In particular, the U.S. GLOBEC DMO will maintain a list of all data access and will notify those who access the data of our commitment to the principle that data is the intellectual property of the collecting scientists.

6. *Within three (3) months after collection, a detailed inventory of measurements made during the cruise or field season must be submitted to the U.S. GLOBEC DMO by the chief scientist of the experiment in cooperation with the participating principal investigators.*

7. *Measurements which do not involve manual analysis and which would be useful to the science community must be submitted by the principal investigator within six (6) months after collection.*

8. *All other measurements and any standard analyses of these measurements must be available to the community within one year after collection.*

9. *Investigators will either submit data to the Data Management Office or place it on-line as a U.S. GLOBEC distributed database.*

10. *The DMO will serve as an intermediate archival location and data source, will transfer data to the NODC, and will prepare the necessary documentation for data collected in foreign waters.*

Sample Preservation

11. *Biological samples will be preserved following currently accepted practice for the particular contents. Sub-samples of a representative subset of the samples must be preserved in reagent grade alcohol for later genetic analysis.*

Modification of Policy

12. *Requests for exemption from the data policy should be submitted to the Program Manager and the U.S. GLOBEC Steering Committee.*

Appendix C: The data policy of the EMAP program

The following data policy is from the Environmental Protection Agencies (EPA) EMAP program. The core policy statements are included but the complete document can be found at the following URL:

http://www.epa.gov/emap/html/pubs/docs/imdocs/wpdatapol.pdf

These policy statements represent a commitment of the Western Pilot Study Steering Committee. All participants are expected to comply unless there is a good reason otherwise reported to the Committee.

- *The Environmental Monitoring and Assessment Program requires a continuing commitment to the establishment, maintenance, description, accessibility, and long-term availability of high-quality data and information.*

- *Full and open sharing of the full suite of data and published information produced by the Study is a fundamental objective. Data and information will be available without restriction for no more than the cost of reproduction and distribution. Where possible, the access to the data will be via the World Wide Web to keep the cost of delivery to a minimum and to allow distribution to be as wide as possible. All data collected by this Study will be publicly available following verification and validation of the datasets.*

- *Organizations and individuals participating in the Study should make measurements that do not involve manual analysis available to other Study participants within 6 months after collection. All other measurements should be made available to Study participants within 15 months after collection. Data and metadata should be publicly available on the EMAP web site within 24 months after field collection. Advise the Chair of the Western Pilot Study Steering Committee if these schedules cannot be met.*

- *All data sets and published information used in the Study will be identified with a citation; for data sets an indication of how the data may be accessed will be provided.*

- *All data sets generated as part of the Study will be made available on the EMAP public web site. These data sets must be described and a quality assessment provided. All such data set descriptions will be made available for inclusion in the EMAP Data Directory/Data Catalog, accessible on the EMAP web site. In addition, steps will be taken to assure their continuing availability.*

- *Participants will adhere to the 'Core Information Management Standards for the EMAP Western Pilot Study'. National and international standards will be used to the greatest extent possible.*

- *Citation information for all the Study's published reports will be provided to the EMAP Bibliography, accessible on the EMAP web site.*

- *Organizations and individuals participating in the Study should actively participate in the EMAP Western Pilot Study web site to share information and coordinate the Study's disparate activities.*

- *To the extent feasible, data from the Streams and the Coastal groups will be copied to STORET for long-term archival and use.*

- *Suggested Data Product Requirement for Grants, Cooperative Agreements, and Contracts:*

  Describe the plan to make available the data products produced, whether from observations or analyses, that contribute significantly to the <grant's> results. The data products will be made available to the <grant official/contracting officer> without restriction and be accompanied by comprehensive metadata documentation adequate for specialists and non-specialists alike to be able to not only understand both how and where the data products were obtained but adequate for them to be used with confidence for generations. The data products and their metadata will be provided in a <standard> exchange format no later than the <grant's> final report or the publication of the data product's associated results, whichever comes first.

Appendix D: Other data policies

There are a many other policies from various groups and agencies in the scientific community. Below are a few other data policies that are available on over the web. These are included to demonstrate how different agencies are addressing similar problems.

The NSF policy for ocean data includes explicit cut off dates for submission
http://www.geo.nsf.gov/oce/programs/oceandat.htm

The NASA planetary data system includes a peer review process for data.
http://pds.jpl.nasa.gov/qs/

The Coastal Ocean Processes program CoOP policy
http://www.skio.peachnet.edu/coop/datapol2.html

Smithsonian Environmental Research Center
http://www.serc.si.edu/datamgmt/policy1.htm

USGS surface water information policy
http://water.usgs.gov/osw/pubs/ofr92-56/ofr92-56.html

NASA EOS validation data
http://www.daac.ornl.gov/eos_land_val/policy.html

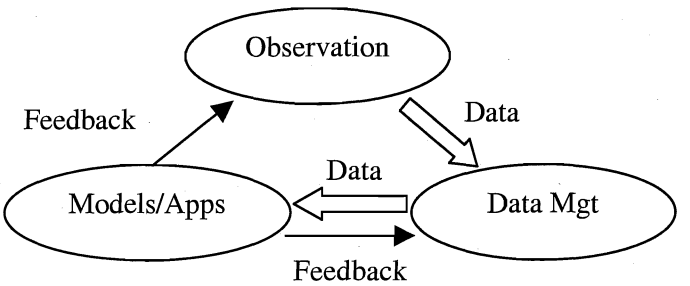
NOAA Atmospheric Chemistry Experiment (ACE) program
http://saga.pmel.noaa.gov/aceasia/info_participants/datapolicy.html

Editorial note: References GOOS document, a NASA document, and several web sites. The web sites are included inline but may need to be moved to a bibliography.

Chapter 13: Data Management and Information Transfer

The Role of Data Management

The data management component of GEM will receive the data and meta-data from the field, provide quality control of the meta-data, store and manage the data, and provide mechanisms for retrieving those data. It will include the systems necessary to automate as much of that procedure as possible and the programs needed to create the custom data products that will be provided to the modeling and applications components. As such the data management system for GEM fits well into the definition established by C-GOOS (GOOS 2000).



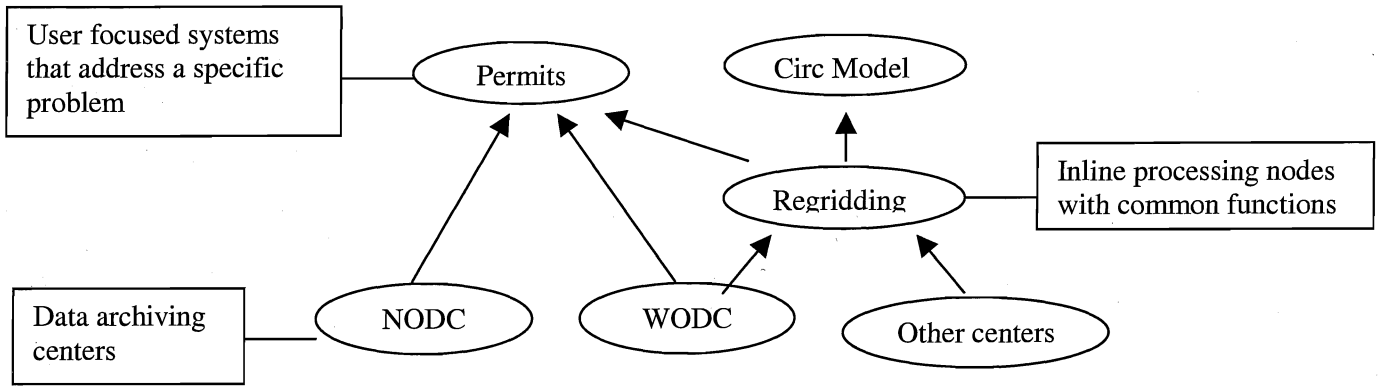
The GOOS model is a general description of an end-to-end system that is based on the tripod of observation, data management, and models and applications with the data management component acting as the intermediary between the observational component and the applications. Data flows from observation through the management system to the modeling and application component. In turn, the applications component informs and refines the both the design of the observational component and the design of the data management system. The monitoring plan may be altered to include new data and/or regions that are identified during the modeling phase as key to understanding the natural system. The interfaces and data products distributed by the data management system will also be refined with feedback from the applications.

Scientific data management systems have grown rapidly since the advent of the World Wide Web. Initially, projects or groups that collected or archived data made those available over the web through a simple interfaces based on the navigation of links. These supply-oriented systems reflect the structure of the data that was made available by providing links to lists of datasets by years, dataset name, or variable name. Many of these are still in wide use although newer systems include more sophisticated search options such as spatial and temporal selection. However, these systems make few assumptions about the intended user community and it becomes

the users responsibility to locate, evaluate, integrate, and preprocess the data into a form that is suitable for the target application.

As the applications that use scientific data become more sophisticated, and the community is able to access and integrate large amounts of data to address a single problem, new data systems will be built that are address the data needs of specific user applications. The output of these systems will be higher order products such as maps, graphs, visualizations, and data in interoperable formats. NASA has funded some projects with a demand-oriented focus (ESIP NRA) and in the future more user communities will find ways to build these types of targeted systems.

The landscape of data product delivery will likely include large archives that supply data in a raw or partially pre-processed form. Application oriented sites will accesses data from these archive sites through a high bandwidth connection and may use intermediate sites, which provide value added services that are not available from the originating archive (see diagram). Common data services available at the archive or through intermediate sites will include subsetting, reformatting, reprojection, regridding or aggregation.



Although predicting the evolution and the impact of the web on scientific data delivery is speculative at best, the landscape of future data systems needs to be evaluated in order to understand the role of the data management component during the extended life span of GEM. Initially, GEM will act as both a data archive and a user focused delivery system, accepting and archiving data from the observational component and creating products that are customized to meet the needs of the habitat specific applications. During this phase, GEM will establish the procedures for assuring the quality of the data that is submitted to the archive as well as the operational details of ingesting data and making it available. As the archive grows, older datasets will be moved to the National Ocean Data Center (NODC) for permanent storage. GEM will continue to maintain a meta-database that provides a data search interface to locate and access GEM data that is maintained by the originating project, the GEM archive, or the data archive at NODC.

Over the long term, however, GEM will likely turn over the entire archiving task to a center such as NODC that is better equipped to maintain the data for extended periods of time. This is only possible after the data flow between the observational component and the applications component has been established and the tools and structures are in place to build the custom data products from a distributed set of data archives. GEM will retain the meta-database and continue to provide custom data products and services to set of targeted users.

Characterizing the data within GEM

Within the data management component, data is classified by the operations that must be applied to it during the archive and retrieval cycle. This classification often cuts across the content-based classifications used during data analysis. While biologic data is more often collected by observation or laboratory work and physical data is frequently measured by instrument there are significant exceptions. A satellite image of ocean color that contains biologic variables will have more in common, in a data management context, with the physical variables in a Synthetic Aperture Radar image than to the phytoplankton results collected from the settled volume of a bottle sample. The settle volume could include both physical and biologic results but be retained by the data management system as a single data holding. The meta-data and processing that is associated with the chemical and biologic data from the bottle sample will be nearly identical, as will the processing and meta-data associated with both types of satellite imagery.

GEM will be collecting and processing a wide range of data from different collection and recording techniques that place different challenges to quality control and assurance. In order to classify these differences for the data management component, data can be separated into broad categories that reflect the handling and storage requirements. These data categories include:

- **Observational** data collected or recorded by an individual
- **Measured** data collected by an instrument and stored in formatted files
- **Modeled** data generated by a running computer model
- **Geographic** or reference data used by a Geographic Information System
- **Remotely Sensed** image data taken from a satellite or aerial platform

The criteria used to characterize these data types are:

- **Interoperability:** how easily the data can be used in alternate applications
- **Consistency:** the degree of similarity between the data for different points
- **Size of file:** the size of the data for a single instance
- **Number of files:** the number of instances that make up the dataset
- **Repeatability:** whether or not the same data can be re-sampled
- **Lag time:** the length of time needed between collection and submission
- **Alternate sources:** whether the data is maintained at multiple sites

- **Meta-data:** The content and/or format of the meta-data

Observational

Observational data are collected by human observation, lab results, and manual data entry. These data include species counts and locations, and can include a large number of ad hoc observations of conditions or unrelated sightings. These data are manually entered and capture a person's observations or calculations, which makes them less consistent, often complex, generally low volume, and occasionally error prone. The observations are not repeatable and the formats are not customarily interoperable. The lag time between collection and submission can be long if extensive lab or manual work is involved. The meta-data describe the collection and or processing location and sometimes the conditions. These data are often in a database managements system (DBMS) or a spreadsheet, which forces a level of consistency that allows automated processing upon retrieval. Examples of observational datasets from the GEM habitat themes (see chapter 10) include:

Wetlands

- Lab results for stream chemistry
- Plant and animal observations from field study
- Isotopes of N and levels of P, Si, Fe from lab

InterTidal/SubTidal

- Species counts for substrate classification
- Lab results for chemical/biological oceanography

Alaska Coastal Current

- Lab results for chemical/biological oceanography
- Species counts for zooplankton
- Diet composition for nekton
- Nekton measurements from net tows
- Bird surveys

OCS/Alaska Gyre

- Lab results for chemical/biological oceanography
- Species counts for zooplankton
- Bird and Mammal surveys

Measured

These data are mostly measurements of physical variables such as air temperature or salinity but they may also include biologic variables as in the case of the acoustic measurements of the biomass of nekton or zooplankton. These data are usually stored in files with formats that are set by the collection instrument. The data files are consistent across the dataset but have a low level of interoperability with other systems. The fact that data collection is automated means that size of the files and the number of the files can be large. Little special processing is involved, usually, so the lag time between collection and submission does not need to be long. The meta-data

includes instrument details and conditions and the data formats are standard enough to allow customized processing during retrieval. Example from the GEM themes include:

InterTidal/SubTidal

- Physical oceanographic variables

Alaska Coastal Current

- Lidar measurements
- Hydro-acoustic plankton or nekton surveys
- Fluorescence measurements

OCS/Alaska Gyre

- Physical oceanography
- Hydro-acoustic plankton or nekton surveys
- Fluorescence measurements

Modeled data

Numeric, and to some degree statistical models, can generate a significant amount of data. As an example the circulation model can provide a snapshot of ocean current vectors across the GEM region, at many depths, for time steps as small as 10 minutes. Other models produce smaller result sets but often these results are used by other models as input and must be cataloged and delivered by the data management component. However, unlike most other datasets these data can be recreated and often are as the model matures. These data are consistent across the data set, can represent a high volume of data, and are not generally interoperable. The lag time between data generation and data submission (and even use) can be very short. The meta-data needs to describe the classification and version of the model and may need to include relevant input parameters. The meta-data may be used to track the lineage of the output data including the references to the input data and, if relevant, the models that created those input data. The modeled output data for GEM is not yet defined.

Geographic

These data are the reference data used by Geographic Information Systems (GIS) and include base layers such as elevation (bathymetry) and shorelines but can also include soil types or habitat characterization. These data formats are rarely used to store data collected by a project but are frequently employed to display the information in the spatial context of a map. These data are usually interoperable across different systems and may be stored at several different locations. The meta-data is focused on the spatial definition and may include information about the resolution or precision of the data. GEM will not generally be ingesting these data from projects but it may store reference information in this format, which is also a prime candidate as a format for custom data products created by the data management component.

Remotely Sensed

Remotely sensed imagery can come from satellite or aerial platforms. These are generally large files and may be used on a regular based by the analysis being conducted by GEM but images from NASA or NOAA may not need to be archived if they can be retrieved again from the source. Arial photography has also been used by EVOS projects to capture the spatial distribution of nekton in Prince William Sound. These images along with satellite images may in some cases be archived by GEM and provided to the application component. These data will require a large amount of storage and are quite interoperable with GIS and image analysis tools. The meta-data describe the instrument and platform and often include details of the image quality and the spatial reference system. Examples in the GEM themes could include:

Wetlands

- LandSat images of watersheds
- MODIS imagery
- Aerial photography

InterTidal/SubTidal

- Ocean color imagery from SeaWiFS
- Aerial photography

Alaska Coastal Current

- Ocean color imagery from SeaWiFS
- MODIS ocean products

OCS/Alaska Gyre

- Ocean color imagery from SeaWiFS
- MODIS ocean products

Impact on GEM

Although the data standards set by GEM will be similar across the datasets in a given type, each dataset will have its own set of standards and QC and ingest processing. As the GEM data management component becomes active, new datasets will be added to the archive. For each new dataset, GEM will set data standards and create the software to perform the QC against those standards. The data management plan will outline what needs to be in place before a new dataset can be added to the GEM archive and the GEM data manager will oversee the process of adding a new data

As each collection effort is funded and organized, a plan that outlines the data inventory and its submission schedule will be established. In addition, the plan will include the procedures for performing the QC process and how discrepancies will be resolved.

Characterizing the GEM user community

Over its lifetime, GEM will serve a large and diverse user community with needs that will vary from simple data download to the creation of tailored data products. In most cases meeting the requirements of particular user groups will require detailed analysis and the creation of tailored products but generalizations can be made about the types of applications that GEM will provide data for.

The user groups interested in each application will have different levels of data analysis and data reduction capabilities and each will need to search for GEM data with different criteria. Some applications require regular or periodic access to GEM data and others are irregular or sporadic. The largest discriminator between the applications, however, is the type of data products that GEM will create them and the level of processing that will go into creating those products. These applications of GEM are relevant for all four of the main GEM themes: watersheds, intertidal, Alaska coastal current, and the Alaska gyre.

1. **Basic research and analysis** is perhaps the most fundamental application of GEM data. This will be done by researchers who are collecting data for GEM and by other researchers that are investigating the GEM region. In general this community will have a good understanding of GEM data and will be searching for specific variables within a region of interest. Access is less likely to be irregular but research applications expect access to data as soon as it can be made available and so FTP or file-download of the original data will generally be sufficient.

2. **Modeling** is also a critical application of GEM data. Verbal and visual models will be drawn from research applications but statistical and numeric models will require access to customized data products that are tailored to meet the needs of the model as closely as possible. Most of the search criteria may be saved by the system and may be reused on a regular basis in order to execute the model with the most recent set of parameters. The types of preprocessing could include reformatting, spatial or temporal aggregation, regridding, and re-projection.

3. **Resource management** applications will increase in number over time and may become a common use of GEM data. These applications will require a separate set of product than the modeling applications. Management applications will be both periodic and sporadic and the product may include reports, graphs or maps. Examples include regular stock analysis reports that are used by fisheries managers set catch limits, or irregular access to watershed data that would be relevant to permit requests.

4. **Public outreach** encompasses several different applications that GEM will be supporting to varying degrees over its life span. These include providing public information about the state of the ecosystems that are being studied by GEM as well as supplying visibility into the general administration of the GEM program. Other outreach activities will include supporting educational programs and possibly emergency response. These applications can be supported with maps and graphs that describe various aspects of the central GEM themes. Access is likely to be quite irregular and may be accomplish through the creation of a few standard maps and graphs on a regular basis.

Supporting GEM applications with user interfaces

In order to support these applications, GEM will initially provide three different modes of access. Although this will change over time the design will include basic search and download, tailored product creation and display, and open map access. For the most part, basic search and download will support research applications, tailored products will be used by both modeling and management applications, and open map access will support public outreach applications. Together these three modes of access characterize many of the scientific data delivery systems available on the web.

Basic search and download is currently the most common method of accessing data on the web. Many projects have an interface that makes some level of search available and then allows data to be downloaded by clicking through to an ftp site or a web page containing data links. Examples include CIIMMS (http://info.dec.state.ak.us/ciimms/), which been used successfully to provide basic access to meta-data and data relating to Cook Inlet and other systems such as GLIMPSE (http://lternet.edu/data/), EMAP (http://www.epa.gov/emap/index.html), and Beija-flor (http://beija-flor.ornl.gov/lba/) which provide basic access for the NSF Long Term Ecological Research program, the EPA Environmental Monitoring and Assessment Program, and the Large Scale Biosphere-Atmosphere Experiment in Amazonia sponsored in part by NASA. In addition the GLOBEC program provides basic data download through its own database (http://globec.whoi.edu/globec-dir/data-access.html).

Although these systems provide different types of search criteria and each has a different orientation they all provide access to meta-data and, in most cases, the actual data collected by the program. GEM can use one of these systems or something very similar to provide access to data soon after it is submitted to GEM. Research applications are often focused on specific variables and regions and these basic systems meet the majority of those needs. In addition, a basic search and download tool will provide the minimum access to GEM data and may support the other applications including modeling, resource management, and public outreach. Although budgetary constraints may require that the creation of custom map and data

products be cut back, the basic search and download functions will be supported as long as data is collected and archived by GEM.

The meta-database maintained in order to support the basic search and download functions would also support access to remote database services that are funded by or relevant to GEM. Remote databases like the EVOS hydrocarbon database and other databases maintained by the group that is conducting the data collection effort will be included in the GEM meta-database for searching purposes. The data will then be available through the remote web site set up to support those data.

Map creation systems such as the Open GIS Consortium's Web Mapping Server (WMS) (http://www.opengis.org/techno/specs/01-047r2.pdf) and the ArcIMS system (http://www.esri.com/software/arcims/index.html) from the Environmental Systems Research Institute (ESRI) make preprocessed maps available to users over the web. Both of these systems provide maps to web browsers and to freely available viewers. Because the WMS protocol is not tied to any particular vendor it has been enjoying rapid acceptance and deployment in a wide range of applications and in the future, the use of WMS in educational and outreach applications is likely to be very large.

Once GEM has identified a set of standard map products that would be useful to the public or to particular educational programs, they will be available through one of these Internet map protocols. These products will likely include base maps and general information maps but might also include regular maps of the Alaska gyre or currents that affect the GEM habitats. Web sites designed to support the educational program or the public interests will display these maps and may, over time, support more complicated map viewers that can access and overlay maps from other sites that are relevant to the goal of the web site.

Data products tailored to specific modeling and resource management applications will be the most useful facet of the GEM data distribution and also the most expensive to create. It is not possible to create a single data distribution system that meets the wide range of user needs in modeling and resource management. Therefore, GEM will prioritize the products that are needed by particular groups and create them in sequence. These products will be designed with the close involvement of the specific user community to which they are targeted and initially they may need to be created with a significant amount of manual effort. However, once automated, a separate web-based interface can be created that will be used by the target user group to create and download these products on a regular (or irregular) basis. Over time, after many of these products have been designed and the distribution of them automated, certain common functions will emerge and GEM will begin to build a library of data processing utilities.

Examples of modeling products include the reformatting and regridding of data to match the execution grid and time steps of the model. Non-GEM data may be pulled from another site and integrated into data product and several different products may be generated at a time to meet the needs of a single modeling application. The

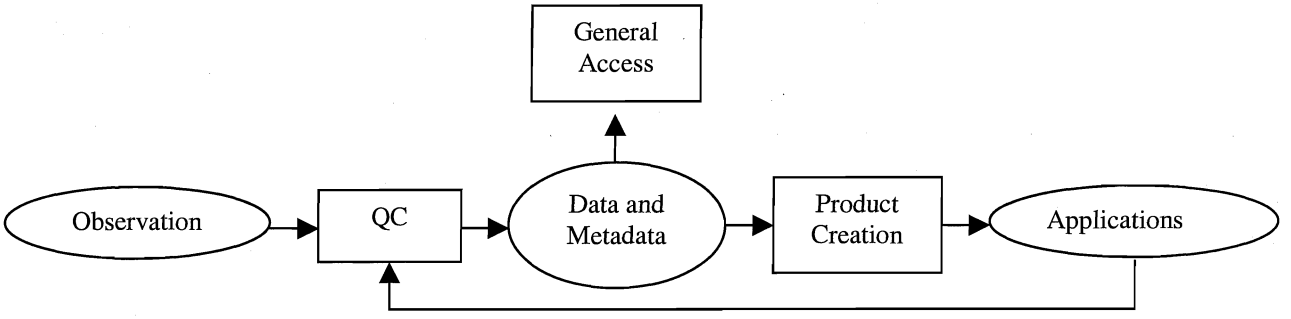
creation of a suite of products may be done by hand and it may require that GEM start with algorithms that were written by the modeling group itself. However, after the modeling group has used the products successfully several times, the process of creating the products could be automated and a simple interface built to allow the group to create and download the product. If the requirements for the product were clear enough, the manual step may be bypassed.

For resource management applications, a report or spreadsheet used to manage fish stocks may require access to several different datasets and the extraction and integration of different variables. Unless the report is already in existence it may require several attempts before a truly useful product can be created. Once this is accomplished, the process could be automated and the resource management office could trigger the report through a simple interface created for that product. In this way the application component of GEM will feedback information and tailor the design of the data management component.

Over time, GEM will create a wide range of products to meet the specific needs of the GEM modeling and resource management communities. The creation of each product will involve GEM staff and the interaction with the target user group. Depending upon the scope of the effort for each product, several tailored products could be created for the modeling and resource management community each year. These products coupled with the basic search and download and with the web-based map delivery services will support a wide range of both specific and general data distribution needs.

The structure of the GEM data system

The GEM data management system will address the issues related to the data types supplied by the observational component and the demand placed by the applications component. As such the data management system is positioned in-between the other two components and must develop and maintain an interface to both. In addition, modeling and map creation applications will generate new data that will also be archived and delivered by the GEM data system.



Supply side support

In order to support the ingestion of data from the observational component of GEM, the data management system must provide quality control (QC) of the meta-data (and to some degree the data) and quality assurance of the data and the meta-data. Quality control will ensure that the meta-data comply with GEM standards and that valid values are supplied in format that can be used to store that data in the GEM archive. Values such as station identifier, date, and latitude and longitude need to be valid or fall within a reasonable range. In general, each data type will have unique issues and GEM will create new QC procedures and programs, although over time some of the QC algorithms can be shared across data types. GEM will also need to provide quality control on some of the data values such as species identification, but the submitter will do the most of the quality control for the data itself. The validation provided by the data management component is done to ensure that data can be found and retrieved using an accepted set of search criteria.

Quality assurance includes the design of the quality control processes and documentation of the quality control activity. The data management component of GEM will not be able to provide the quality control over the most of the data but it can ensure that the documentation of the submitters' quality control is available along with the data. The data management system will also provide quality assurance of the meta-data.

Demand side support

On the applications side of the data management system, software modules will create the custom data products and standard maps. These routines will not be developed all at once when the system is deployed but over time as the archive is populated with data and the user demands become clear. Custom routines will integrate third party software where possible. These external routines may be Commercial Off The Shelf (COTS) software or they come from the growing library of free software available over the web. These custom routines will pull datasets from the GEM archive and from other relevant data sources and provide preprocessing. Examples of the types of operations include:

- **Reformatting:** Often, raw data may need to be reorganized in order to be usable by an application. As an example, an application may need multiple observations pulled into a single output file containing only those variables of interest from a subset of stations. This file may also need to be ordered by date or species and written out in a comma-separated file that can be manipulated by a spreadsheet. Other output formats may include GIS, image analysis formats or special binary formats for visualization applications.

- **Aggregation or subsetting:** Modeling applications often need summary or averaged data and so datasets may need to be merged or clipped to capture the temporal or spatial region of interest completely. Some file formats support clipping but many of these routines will be tailored to the input data. Aggregation routines may come from the application space or they may simple average or sum calculations.

- **Projection:** Data is usually collected with latitude/longitude coordinates and some regional models use a map projection that preserves spatial relationships more accurately for the region. Satellite data and other data may need to be projected or reprojected into a specific map projection for the application. Software is available to perform some of these reprojection operations from both commercial and freeware sources.

- **Map creation and visualization:** Some data products may be best represented in the spatial context of a map or a graph. The generation of these maps or the creation of a multidimensional or graph oriented visualization require data extraction reduction and rendering. There are a large number of software utilities available to assist in this process and they will be integrated into single utility to create the custom product.

Most custom data product will require a user interface to allow the entry of parameters and to trigger the creation of the product. In most cases these will be simple web pages that support various pull down menus to select input or display parameters. Simple interfaces that are designed to support one or two data products are easier to use and maintain. Although over time GEM will support a large number of custom products, and interfaces may need to be merged to reduce the overall maintenance load.

Meta-database support

The core of the data system will be the meta-database and a data storage component. The meta-database contains the descriptive information and is used to integrate the access to the data by supporting cross dataset searching. The ability to search for all datasets within a given spatial or temporal range or all datasets containing particular variables requires a single meta-database. The QC routines will ensure that the meta-data submitted to GEM meets the standards necessary to support cross dataset search. No dataset will be added to the system unless it can be located using a search of this meta-database.

The meta-database maintained by GEM will also support access to remote GEM archives that are maintained by individual researchers. GEM will also evaluate whether to ingest meta-data about datasets that are relevant to the GEM system but are not directly supported by GEM. The ongoing gap analysis conducted by GEM will continue to reveal datasets and data collection activities that compliment the GEM mission and one of the GEM goals is to integrate with those projects. The data

management system will reflect this integration by allowing users to locate relevant data that may not be archived by GEM.

Most search and download systems include some level of meta-database support. GEM will evaluate the use of these existing systems and the evaluation criteria will include the structure of the meta-database. Although an existing meta-database structure may be found to suite the needs of GEM, the population and use of the meta-database will be the central activity of the GEM data system and any existing system will need to be modified.

Data storage

The storage of the data in files or in another storage mechanism is a separate function of the data system that in time will require a significant amount of storage space. The meta-database will contain pointers to the data itself, which may physically be in a separate storage facility. The evolution of large archive technology has been rapid over the last few years but GEM will be able to postpone the use of tape or optical media for several years until the space requirements demand it. GEM will evaluate the use of an external site to store the data as well as the use of GEM computing hardware. Unlike the search of the meta-database that places a heavy computational burden on resources while returning a small amount of data, accessing the data itself requires no significant computation but can return a large amount of data. Therefore the network connectivity is also an evaluation criterion for the data storage subsystem.

The format of the data files will be defined by the GEM data management plan and become a GEM standard. Although the QC procedures will not validate the scientific quality of the data, these programs will need to validate the format of the data. Data product creation routines require that input data files are in a recognizable format and contain data in a format that can be processed automatically.

GEM administrative support

Managing the projects funded by and associated with GEM requires a project-oriented database. The administrative information includes the original proposal, comments submitted by the review panel, status reports and notes, and the final report. This information will be valuable over the long term as the data collected by the project is evaluated in retrospect. The proposals and reports will contain the original hypotheses as well as the problems that were encountered during data collection. Future researchers will use this project genesis to understand the original goals of the project and issues that might affect data quality.

Much of these administrative data are in the public record and will be made available over the web. The GEM meta-database will include the project specification so that the data submitted by the project can be displayed along with the administrative details. This link between the administration of the project and the data submitted

would also allow GEM to evaluate whether all the data for a given project has been submitted.

Building a data system for GEM

The GEM data system will grow in phases as the data holdings are expanded and new users are served. However, during the first few phases a system foundation will be defined and created which includes the storage system, the meta-database and some sort of basic data delivery mechanism. After a foundation or core system is in place, progress on the components to support the data suppliers, the data users, and the administrative systems can be made in parallel.

The creation of the core system components will also need to be done in phases. The initial phase will include the definition of purpose, a prototype of the basic system, and possibly some data rescue. A second phase will include a more comprehensive description of the intended user interactions followed by detailed specification, design. A third phase will include the implementation of the system foundation, which will support the functionality of the existing prototype and the framework for the subsequent phases. The fourth and subsequent phases will include the creation of new ingest and delivery modules that allow for new data types to be ingested and new data products to be created and delivered. Each phase will be completed within 6 to 12 months so that feedback may be incorporated into the design of each phase.

Phase I

The first phase includes the creation of the basic documents that define the GEM program and data system. In addition, the first phase will include an operational prototype that will be used to refine the requirements for a larger system and deliver a small amount of data over the short term. The documents defining the system will be created with community feedback and include this GEM plan and the following data related documents.

- **GEM Data Policy:** The data policy sets out the GEM philosophy of data sharing within the GEM program. It describes the responsibilities GEM has toward the data suppliers and the responsibilities the suppliers have to GEM. The GLOBEC data policy offers a good starting point for the GEM data policy. (http://cbl.umces.edu/fogarty/usglobec/reports/datapol/datapol.contents.html)

- **GEM Data Management Plan:** This is the overall plan for how data will move through the system, how standards will be checked, how data can be retrieved and how data will migrate out of the system. It will describe the procedures for submitting data to GEM and how GEM will manage data in remote databases. It will also outline the quality assurance (QA) procedures for the data in GEM including the quality control (QC) of the data done by the

supplier, and the QC of the meta-data, which will be done by suppliers and by GEM.

These documents publicize the objectives and plan for the GEM data system and are intended for the GEM stakeholders. These two documents will be the initial documents describing the data system and will be the basis from which the technical specifications created in subsequent phases will be drawn.

The data management plan will include a description of how the system will be used and by whom. From this framework a prototype of the system can be built using GEM data or possible data from previous EVOS projects. A data rescue effort will bring datasets from previous EVOS projects into the GEM framework and provide an opportunity to prototype different storage options and meta-databases. In addition, several basic delivery mechanisms will be tried and the key user groups will provide feedback.

The prototyping effort will include establishing a basic system that supports the storage and delivery of a small amount of data. GEM may store the data or it may be stored remotely at the supplier's site. The prototype will build upon other EVOS related system such as CIIMMS and SEA and provide basic services that are maintained by a small staff at GEM.

The result of phase one will be a clearer understanding of how the system will be used and who represents the key user communities that will interact with the GEM data system. In addition, a basic framework of hardware and software services will be in place to support the delivery of a small amount of data. The lessons from this phase will educate phase II, which will include a detailed specification of the system.

Phase II

Phase II will build upon the lessons learned in phase I and result in a detailed set of specifications of a more comprehensive foundation. These specifications will expand upon the phase I solution or define a new organization that will support larger data volumes and more complex data access and information delivery. The technical specifications will include the following documents:

- **GEM System Operations Concept:** The primary purpose of this document is to provide a technical of description the planned functionality and operations of the GEM Data System. The document describes the functionality based upon community input and derived requirements from science coordinator and the data manager. The document will include a description of all functions associated with the GEM data system and related interfaces to external entities that directly provide information to, or receive information from the system. Internal process flows, both normal and contingent, are also described.

- **GEM System Requirements Specification:** This document will establish the requirements baseline for the GEM data system. It will include a conceptual and functional architecture, interface, functional, performance and design requirements.

- **GEM System External Interface Requirements Specifications:** This series of documents (1 per external interface) defines the data exchanges between GEM data system and external entities. It will identify data flows, performance constraints, and implementation responsibility.

An approach to system specification that is very effective when the provider and user communities are diverse is the development of representative user scenarios through Use Case Analysis. This is a process by which the users of the system and the process workflows through which they interact with the system are identified. The output of Use Case Analysis is a set of functions and interfaces that the data system needs to support. The set of users will have been refined in phase I, and a representative will have been identified to help formulate the use cases. The general classes of users include:

- **Data suppliers:** who collect, analyze, and submit data within a GEM project.
- **Program administrators:** who monitor and report the status of GEM projects.
- **Data access users:** who access GEM data through a basic user interface.
- **Applications users:** who access GEM data through a tailored application.
- **System operators:** who support data ingestion and access functions.

The scenarios of how each user will act on the system will clarify the functionality that the end system needs to have (requirements), and external entities with which the system needs to interact (interfaces). The use case diagrams will be created in cooperation with the user representatives who participated in phase I.

From the user scenarios a Conceptual Architecture that identifies the main components of the data systems can be developed. The architecture and the user scenarios will then be folded into the Operations Concept, which will specify the data and information flows through the system. This includes the identification of external interfaces, the specification of Operational Scenarios, and the development of a Functional Architecture that supports those scenarios. Lastly, a formal System Requirements Specification is developed that together with the Operations Concept will form the technical backbone of the data system specification. These technical documents, once completed, will be the supporting information necessary to outline the development necessary. GEM may undertake the development or GEM may release a public Request For Proposal (RFP) for some or all of the system components identified in the technical specifications.

Although there will be several documents produced in this phase some can be produced quite quickly. Each will capture an important specification necessary to

define the system for a developer but some will be brief. The overall time needed for this phase will be between four and six months.

Phase III

Phase II will produce all the necessary specifications to carry out the development in phase III and so the actual steps involved in creating the GEM data system will not be known until that phase is complete. However, the development will likely include the creation of custom software components including the meta-database and the integration of existing software components such as CIIMMS. The storage of the data may be distributed across GEM facilities, the supplier's facilities, or perhaps the facilities of an external partner that can provide long-term data storage. Finally, large portions of the GEM system requirements may be satisfied by existing systems (e.g. STORET) and the implementation process may mostly consist of integrating with those systems.

Depending upon the requirements that emerge from phase II, the development process may be undertaken by GEM directly or be completed under a proposal submitted to GEM. In either case some new phases might be identified and the development might be broken up. However, because the prototype will already have been in operation the development of the new system will face fewer uncertainties and therefore fewer problems.

Enhancement phases

The ongoing enhancement of the data system will build off of the foundation by adding new data types and new custom products for specific user groups. These enhancements will be identified by the GEM program committee and grouped together in phases. The operation of the system will also be an ongoing effort, as will the effort of rescuing data and migrating data to new forms of storage media, but these issues are separate from the enhancement of the system.

Each new dataset collected by GEM will require a QC analysis, a data definition document, and a suite of programs to check the QC standards outlined in the document. The data definition document describes the format of the data, the QC procedures and any other collection details that accompany the dataset. This document will be created in advance of the data collection and may cover the details of several years of data collection.

The GEM system architecture will outline how these QC programs are integrated into the system foundation and into the procedures for submitting data. Each new QC program will check a new data file against the data standards and report any inconsistencies. Once the core system is in place, these programs can be created as new data holdings are collected and brought into the data system. Periodically these programs may be consolidated or refined in order to reduce the ongoing software maintenance.

The creation of custom data products that are tailored to specific user communities will also be an ongoing enhancement to the foundation. The needs of each new group will be analyzed and a custom interface or report will be created. The product will be refined with feedback from the user group as it is used. The product will stabilize after several iterations and the needs of a new user group can be addressed.

As with the QC programs the custom data product generation programs can be consolidated and refined periodically. The development of a custom product will initially be undertaken as an ad hoc product but once it has stabilized it may be integrated with a similar product. Periodically the processes needed to create several products may be merged and refined into a single process that is easier to maintain. The change will likely be done only in the process that creates the product and not in the interface that the user has come to rely upon.

The addition of new holdings and the creation of new data product will be phased in over several years following the creation of the core system foundation. Each phase will result in the consolidation and enhancement of the GEM data system as new holdings and new data products are made available. The creation of the GEM data system will therefore not be a monolithic process but an ongoing process that interacts with the data suppliers and users of the system. By iterating through the development process, GEM will be able to increase the functionality of the system and keep costs to a minimum.

Budgeting and Personnel

The GEM data system will be created in phases and enhanced over several years but the anticipated budget will be set early in the process. Although establishing the correct budget for data management will be difficult, experience from other projects suggest that the data system cost should be between 10% and 20% of the overall project budget (Hale 1999). This budget will cover the cost of the hardware needed to house the archive (or the expense of contracting this service), software development costs for ingest or application components, commercial or other software costs, and the staff necessary to administer and operate the system, support the data suppliers, and support the application users.

The initial computing hardware needed to support the GEM data system does not need to be very extensive or powerful. Judging from prior EVOS projects, the volume of GEM data for the first few years will not require specialized storage system such as tape or optical storage and the initial demand will be is also not likely to be extensive. If this assumption holds true after the data management system has been analyzed, GEM would need hardware to support the following servers functions:

- A web server to provide access to meta-data and possibly data
- A meta-database of GEM and possibly non-GEM holdings
- An archival server for data storage

- A testing and development environment

Over time the archival functionality may be contracted to a third party data center that is able to provide large volume storage, 24-hour access, and migration facilities to move data onto the most current storage media.

GEM staff will need to maintain relationships and support both the data suppliers and the user community. There is not much overlap between these two and in the long run two staff persons may be necessary. The first will maintain the supplier relationships and support the process and program needed to ingest of new datasets. The second will provide customer support and development for the custom data products. Additionally, a data manager will oversee the data system overall, participate in the broader scientific data management community, and meet with the GEM program committee on the data related issues of the program and of individual projects. The data manager will also participate in the review of GEM proposals by evaluating the issues related to the data system.

Conclusion

The GEM data system acts as the intermediary between the GEM observational component and the modeling and applications component and it will respond to the needs of both. The data system itself is partitioned by the needs of the data suppliers who make up the observational component and the needs of users of the applications.

- The analysis of the system is based both upon the expected types of data and the anticipated user community.

- The structure of the system includes both the QC procedures for adding data received from the observational component and the generation of custom products needed to feed the modeling and end user applications.

- The data management staff will be partitioned by the support provided to the data suppliers and the customer support provided to the application users and modelers.

These two sides are brought together by the shared meta-database and the data storage architecture that make up the foundation of the system.

The phased approach to development provides for the creation of the shared foundation in order to support ongoing enhancement of the support for the observational component and the modeling and applications component. Although the initial phases will need to focus on the foundation, it is the phasing in of these enhancements that will lead over time to a full and robust data system with the flexibility to adapt to the changes in the all of the components of GEM.

References

Hale, S. S. 1999. How to Manage Data Badly (Part 1). Bulletin of the Ecological Society of America 80(4): 265-268.

Hale, S. S. 2000. How to Manage Data Badly (Part 2). Bulletin of the Ecological Society of America 81(1): 101-103.

Intergovernmental Oceanographic Commission. 2000. Strategic Design Plan for the Coastal Component of the Global Ocean Observing System (GOOS). UNESCO, Paris.