

Southeast Atlantic Coastal Ocean Observation System (SEACOOS) Information Management: Evolution of a Distributed Community System

Madilyn Fletcher¹, Jesse Cleary², Jeremy Cothran¹, Dwayne Porter¹

¹University of South Carolina, Columbia, SC ; ²University of North Carolina, Chapel Hill, NC

When SEACOOS was established in September 2002, it brought together four major institutions that had extant observing systems and modeling research – the University of North Carolina at Chapel Hill (UNC-CH), Skidaway Institute of Oceanography (SkIO), University of Miami (UM), and the University of South Florida (USF) (Seim et al., 2004). A fifth partner institution, the University of South Carolina (USC), had extensive experience and infrastructure in data and information management. These partners practiced a diverse set of data collection approaches. Thus, a regional system that enabled coordination and integration of these data would require a sophisticated Information Management (IM) system, which in turn required resources and innovation. The development of the SEACOOS IM component has paralleled the evolution of the IOOS DMAC (OceanUS, 2005) conceptual development and planning, and SEACOOS has closely monitored IOOS requirements to ensure mutually supportive activities and progress. Data management efforts and the move towards interoperability has been a national community effort, with individual regions making contributions to different aspects of the whole data management enterprise. SEACOOS contributions have been particularly significant in addressing the challenge of aggregating heterogeneous data from distributed sources within an infrastructure that optimized data visualization and retrieval. This required the development of a process that served near real time data streams of in situ observations, model output, and remotely sensed imagery.

First Principles and Steps

The SEACOOS Principal Scientists fairly quickly agreed upon a vision of interoperability and a set of principles that would guide the establishment of the IM component:

1. The data observations and their ultimate applications would remain largely distributed at the participating institutions.
2. To the extent possible, the new infrastructure would be based on existing resources and practices, so that the participating institutions would maintain flexibility and autonomy where appropriate, resources would be conserved and optimized, and the rate of progress, particularly in the early stages, could be as rapid as possible.
3. The system would be developed to optimize accessibility of data and information applications by putting no embargoes on observation data and by ensuring that, to the extent possible, expensive, proprietary tools were avoided unless they were important for optimum functionality.

While establishing the fundamental principles, the partners were assessing assets and observation data to determine what each could contribute to the overall enterprise. The relevant technical personnel were identified and brought together so that they could personally get to know each other, understand their respective contributions, and build

mutual confidence. This team became the Data Management Coordinating Committee (DMCC), which has been the prime development and problem-solving mechanism for the SEACOOS IM program. During this initial assessment period, an additional important step was the conceptualization of the broad range of temporal and spatial data that would need to be accommodated by the IM infrastructure. These included – at a minimum – ocean state variables measured *in situ*, as well as remotely sensed data and model output. A conceptual framework needed to be adopted that could handle this range of data and yet could be adapted and expanded to additional types of ocean and coastal data that would become essential with the implementation of IOOS. It was decided to focus initially on near-real time physical variables. These data were common to all partner data providers, and this subset of data was a manageable size.

Components of the System

The framework that was developed has three major components; (1) the data sources, including observations, remotely sensed data, and model output; (2) the data “products”, which include the data themselves and a variety of user-defined tools and applications, as well as their documentation; and (3) a variety of functions that are applied to the incoming data to increase their information content; examples would be aggregation of data, analysis, or visualization of the data, as well as their use for prediction and assimilation into models. In order to utilize data appropriately and define their validity, it is necessary have some level of consistency through the application of specific data standards, processes, and protocols that enable quality assurance and quality control (QA/QC). The development of standards and QA/QC approaches became a 4th element of the framework. Also essential for the development of applications was identification and development of delivery interfaces, such as map-based products and their dissemination through a Web browser; these interfaces can be considered the 5th element of the framework.

The focal point for the first stages of activity became the development of the means to pull together, or aggregate, data from the various providers and provide the result in a form that had some meaning or application. The decision was made to develop a map-based product that displayed the aggregated data. The map could be relatively easily examined to assess whether the data reached some minimum level of accuracy and whether the data management processes and protocols were performing adequately. An additional advantage is that map presentations are readily appreciated by a diverse range of viewers, thus increasing the potential impact of the data aggregation product. At the same time, the processes required to achieve this defined product would address fundamental interoperability issues, so that solutions developed would have potentially broad applications in the IOOS DMAC enterprise. Application areas were identified to be addressed in phases, and these were “circulation” (Phase 1) and “waves” and “inundation” (Phase 2). The intention was to work within a diverse group of scientists and application specialists, such as federal agency staff or state resource managers, to identify and implement specific demonstrations that utilized the SEACOOS observation and IM system.

A distributed data system and aggregation process required a basic set of hardware investments. A central hub was identified for the aggregation process, and three separate servers were utilized: a web server that alternately points to one of two database servers, and two database servers that alternate roles between collection and populating the database (write operations) and servicing queries (read operations). A growing appreciation of the need for redundancy has led to the gradual establishment of parallel operations at a second site, which is near completion.

The following sections describe the processes and protocols devised and the steps taken that led to the SEACOOS IM system of aggregated data and retrievable databases.

Standards Development

To establish an efficient data sharing system, it is essential to assess and decide upon a minimal set of standards that will be used for data description, expression, and transport. As we were proposing to aggregate – and potentially integrate – data from a range of different providers, it was essential to agree upon protocols to be used, formats for recording and storing data, and vocabularies for expressing data. We needed to resolve the in-house standards already being utilized by regional data providers with standards that had been, or were being, adopted by broader data management communities working in ocean observations. This required an assessment of the variety of processes and protocols being used, their suitability for SEACOOS data (including those observations likely to be incorporated at some future time), and the changes involved in adapting to any existing observation data management systems. The SEACOOS DMCC determined that the most efficient way to conduct this assessment was in the framework of a desired application. We decided to develop a Web-based platform that enabled the following functions: (1) the aggregation of data from the primary SEACOOS partners and key federal data providers, e.g. NOAA National Data Buoy Center; (2) development of a map-based visualization of the aggregated data, which would not only demonstrate that the aggregation technologies were working but would also enable the comparison and cross-checking of actual data values and their display; and (3) utilization of the data visualization in a Web-based portal that enabled access to data and their presentation in a variety of user-targeted displays. In this way, the functionality of an end-to-end data management system could be established and verified, through the connection of raw data collection and ultimate user application.

Data Transport

It is important to identify standard mechanisms for data transport, but at the same time have the flexibility offered by a limited set of options. Selections could be made on the basis of the data source and its intended use or application. As a first step, SEACOOS IM implemented DODS/OPeNDAP (<http://www.opendap.org/>), which had been identified by OceanUS as a data transport protocol standard. With this technology, software is required at the data server end that delivers the data to the clients, which in turn can be employing a range of client applications, such as Matlab, ArcGIS or Relational Database Management Systems (RDBMS), on their servers. The SEACOOS DMCC installed DODS/OPeNDAP NetCDF servers at UNC-CH, USC, SkIO, UM, and USF and implemented data transport capabilities between the five sites. In the meantime,

the DMCC determined that raw data files could also be readily downloaded via HTTP, and this is the mechanism usually employed for data transport. In contrast, Open Geospatial Consortium (OGC) (<http://www.opengeospatial.org/>) standards make it possible for the data provider to enable a set of services, such as the Web Map Services (WMS) and Web Feature Services (WFS). With these services at the data aggregator end, the client can more readily sub-select data that suits different specific user needs or applications. The OGC standards have been particularly useful for SEACOOS contributions to national IM efforts, such as OpenIOOS (www.openioos.org), the newer OosTethys (www.oostethys.org) and the [IOOS Observation Registry](#). OGC geospatial standards utilize SOAP and REST transport protocols (a basic web service or messaging framework that exchanges information over HTTP via XML) which the IOOS DMAC recommendations follow.

Vocabulary

A key accomplishment of the DMCC was the development of a “data dictionary,” which specified the naming and reference conventions. Terms addressed the observation position, time, type of measurement, and platform that would be used by SEACOOS. This was a necessary first step before proceeding with data base development and testing of transport protocols.

Data File Format Standard

Another requirement was for a data file format standard. After a review of possibilities, NetCDF was chosen for several reasons. It was already commonly used for oceanographic data, its utility with OPeNDAP had been demonstrated, and it has the advantage that some metadata are included in the datasets. We adopted, where possible, the Climate and Forecast (CF) Metadata netCDF Conventions v1.0 and adapted the format where needed, such as through the use of standard names from the SEACOOS Data Dictionary rather than the CF Standard Name Table. The SEACOOS NetCDF specification v2.0 is available at <http://seacoos.org/documents/metadata>, while v3.0 is near completion. The convention addresses a wide range of data types, i.e. fixed point (e.g. mooring, offshore tower, and tide- and stream- gauge observations), 2D or 3D moving point (e.g. ship, surface drifter, or underwater autonomous vehicle measurements), fixed or moving profilers (e.g. bottom-mounted or ship-mounted acoustic profilers), and remotely sensed observations (e.g. high frequency radar, aircraft, satellites). As data have been used to develop a range of applications, some additional data file standards have been utilized, and they include ASCII, CSV, XML/GML, and ESRI shapefiles. An in house-developed, observations-based XML schema, which is labeled ‘[ObsKML](#),’ has also been developed for data aggregation, display, and sharing purposes.

Metadata

Although some minimal metadata are included in the netCDF file format, a robust database must include additional metadata to be able to assess the quality and characteristics of the data. Also discovery of the data by other community members is facilitated by publishing metadata via searchable clearing houses. Metadata documentation for large datasets, such as those generated in ocean observing systems,

can be a formidable task. With SEACOOS, metadata documentation has been largely carried out by the institutional program operating the observing platforms involved, except for metadata describing the aggregated datasets, which has been developed at the SEACOOS level. SEACOOS had also centralized the development of an on-line browser-based tool, called MetaDoor, that facilitated the creation and publishing of XML metadata records that are compliant with the Federal Geographic Data Commission (FGDC), a DMAC requirement. Users are guided through the web-based form, instructed on metadata entry, and checked for compliance. MetaDoor Version 2.0 includes some user management features and forms for entering basic platform and sensor metadata ([MetaDoor](#)).

As SEACOOS developed and increasingly coordinated its observing capabilities, particularly through the development of map-based applications, the need for a complete and evolving inventory of observing assets and the variables measured became apparent. The first step has been completion of a database and map application including observation platforms, sensor instrumentation, and variables measured within the SEACOOS footprint ([SEACOOS Sensor Inventory](#)). A second generation inventory will be needed as the observing system components transition into a Regional Association (RA) asset ([Regional Associations](#)). This system will need to be an ongoing inventory that is updated dynamically as the RCOOS evolves. This will require development of a new database schema for sensor metadata and the inclusion of detailed sensor metadata within observation data files. This should be an early objective for the new Regional Association IM activity.

Database Organization and Development

A basic decision to be made when first establishing the database structure is whether the data should be organized by variable or by station. If the data are stored by variable, each variable has a separate database table, and there is a data field for measurement time, station identification (ID), sensor ID, latitude, longitude, depth, and quality control flags that describe the measurement point and measured value. Accordingly, the station approach has a separate table for each station, which is populated by variable measurements for a given time. Although the variable approach can result in much replication in tables for fixed platforms, it was the better approach for complex situations created by 4-dimensional mapping. Moreover, each separate table can generate a GIS layer that can be applied to the map presentation.

Data were organized in a relational database structure, the open source PostgreSQL DB, in which data are organized according to variable type, time, location, and quality. A “data scout,” was developed to periodically retrieve the most recent data from the contributing sites and transform those netCDF files to the PostgreSQL Relational Database. Model output and references to remotely sensed image files and ancillary boundary files are also included in the database structure. Additional data were mined from federal agencies, e.g. NOS, NDBC, USGS, and NWS.

One limitation that became apparent in working with data collected to a data “logger” in tabular format was that additional work was required to move the data into more

standardized forms from which they could be more easily aggregated. This issue was particularly apparent in the development of the Carolinas Coast application ([NOAA NWS "Carolinas Coast"](#)), for which the NOAA NWS required rapid posting of new observation data as they came available. To address this need, a new general relational table schema was developed to which it would be easy to push *in situ* observation data and from which various products and web services could be pulled (Figure 1). The effort, called Xenia ([XeniaPackage](#)), is currently designed for the common datalogger/filesystem, which collects generally less than 100,000 records per hour (e.g. 10-20 observations per hour for 1-5,000 platforms). Xenia simplifies the data collection process for map visualizations of multiple layers of observational data.

Map Development

Two basic types of map-based products have been developed by SEACOOS. The first is a pregenerated, or “report-based”, map that presents the most recent information and is refreshed periodically according to pre-defined specifications. Currently these maps are accessible on the SEACOOS website and are refreshed at 1-hour intervals ([SEACOOS Observation Map](#)). The second product is an interactive map that is produced by the user selecting specific subsets of data to be included in the map visualization ([SEACOOS Interactive Map](#)) (Figure 2). In the development of these applications, we have largely used open source software: PostgreSQL for the relational database management system, PostGIS to spatially enable the PostgreSQL server, MapServer GIS for the map visualization, and Zope/Plone for the web application framework and content management. MapServer was adapted, in that it enables visualization only in two dimensions, whereas the SEACOOS application was enhanced to include time and depth. For the pre-generated maps, an automated script requests the most recent observations from the data center via an OGC web service (WMS) and saves these map images for rapid web display. These map products are archived, and both the maps and databases are accessible through the SEACOOS web portal. Specific subsets of information can be requested and specific data reports, such as table listings, maps, and graphs can be retrieved by users.

As the maps were generated from data originating at different institutions, a major difficulty was created by the different practices and protocols employed. For example, “surface temperature” could be taken from a range of depths near the surface depending on institutional practices, and a decision needed to be made about the minimum depth that represented “surface.” The same problem was encountered for “bottom” measurements. Lack of consistency also occurred with the timing of measurements, as even 1-hour interval measurements were generally taken at different specific times. Thus, a normalization process was devised and agreement was reached on the ranges that would be included under each specific category. All measurements within that range would be retrieved by the data scout at the predetermined intervals.

Archival

Generally observation data are archived with each data provider. There has also been limited central archiving of aggregated datasets of several observation types since

September 2004, as value was generated through the aggregation process, i.e. reformatting, unit conversion, QA/QC procedures. Remotely sensed imagery is maintained on the primary file system, but file space limitations may require it to be moved after a year's storage. Model output data are not archived because of their large volumes.

Visualization Demonstrations

The SEACOOS developments in database structure, data standards, and data transport are fundamental contributions to interoperability within the IOOS. A demonstration of the utility of these developments is the interactive map-based presentations of aggregated data. The three primary variables that were used for developing visualization applications for the interactive maps were sea surface temperature (SST), winds, and currents (additional variables can be accessed at the Interactive Map Page ([Interactive Map](#))). These variables all involved multiple observations and require scalar and vector processing. SST includes mooring measurements and satellite data; aggregated wind measurements are derived from moorings, ships, coastal stations, ground stations, and satellites; current data includes HF radar, moored ADCPs, and drifter data. The open source MapServer application was used, along with PostgreSQL and its extension PostGIS, as the basic mapping platform.

A data animation page was created that combines maps and graphs, with the capacity to select GIS layers, scale, platforms to graph, and time steps. Animations are powerful modes of presenting data, and they often have more appeal for general public users and educators. For these additional applications, GIFsicle and AnimationS (AniS) were used to generate and control data animations; ImageMagick was used for image manipulation and for accessing raster data. Also Gnuplot was used to generate time series graphs. Users can access animation templates ([Animation Template Log](#)), as well as instructions for how to construct individually tailored animations. The constructed animations are logged and can be removed by request.

Broader Data Delivery Issues

Quality Assurance (QA) and Quality Control (QC)

QA/QC is a complex issue that can have different requirements for different users. QA not only refers to an ability to assess the quality of the data within the data management infrastructure, but also is determined by the performance of the instrumentation and local measurement conditions. Thus, mechanisms are needed to assess the likely validity of data that are being generated at remote sites. QC, on the other hand, refers to those processes that are needed to assess the data, communicate their level of reliability, and possibly apply corrections. The SEACOOS DMCC has been largely concerned with the QC aspect, while working with the Observation Subsystem personnel to determine an understanding of their requirements and potential for assessing data (QA). The complexity of developing a QA/QC system is further complicated by the different levels of assurance required by different users. For some, near-real time data are required, and thus rapid assessment procedures are required, at the expense of a more rigorous analysis of data quality. On the other hand, when data accuracy, not time, is the primary issue,

more complete post-processing can be applied to search for spurious measurements and understand the significance of apparent anomalies.

Presently, most of the QA/QC processing within SEACOOS occurs at the institutions managing the individual observation platforms. However, at the SEACOOS level, programs are being developed to perform automated testing of data and to create data filtering methods that can be applied to data destined for mapping applications and databases. Standardization of a flagging system for suspect data is being developed, and methods for establishing spatial error estimates, which are required by specific users such as the US Coast Guard, are being analyzed.

Conveying Information to Diverse User Groups

One of the biggest challenges for a program such as SEACOOS, which represents a complicated system of data providers and data users, is how to construct a dissemination system that serves such a diverse and broad set of communities. The SEACOOS portal evolved over the five years of the program, and was constructed using a content management system to help organize and access the large information volume. This works well primarily for the information providers, as well as academics that are familiar with accessing information from such a framework. However, the system is less accessible for more specialist user groups who are interested in user-specific applications, particularly when each specialist group is likely to have a different level of knowledge or proficiency with vocabularies and concepts. On the other hand, SEACOOS and similar programs have such a broad set of potential applications, it is important to avoid limiting the front end to specific users if that deflects other possible stakeholders. One approach may be to recognize related user groups by providing links to other more tailored sites or subregional systems.

Planning for Required Capacities

As SEACOOS evolved, it successfully dealt with a number of challenges in accessing and aggregating data from multiple institutions. However, in a community in which the numbers of participating data providers is expected to grow, a significant challenge is assessing the system capacity that will be required for smooth data flow and communications. Particularly when individual institutions are making internal decisions that affect their data volumes and local infrastructure, it is hard to assess the burden placed on the overall system as new organizations are incorporated into the regional structure. An important goal, as the overall system becomes able to deal with complex biological data, is to be able to service the many biological datasets that are being developed at more remote laboratories. Historically, many have had limited communications infrastructure, and solutions are needed to establish the capacity to include these information sources into the regional system.

The Pursuit of Interoperability

A functional goal of SEACOOS, as well as the IOOS enterprise, is “interoperability,” or the ability of the various components of the systems to exchange services to enable them to work together effectively. SEACOOS achieved interoperability among key partners through the identification and implementation of a set of standards that enabled data

aggregation and exchange, as well as the generation of aggregated data products. National efforts such as [OpenIOOS](#), [OosTethys](#) and the [IOOS Observations Registry](#) are working towards standards-based operability at the IOOS level, and SEACOOS has been a major contributor to development of the necessary processes and protocols. Interoperability also requires access to relevant federal data providers, and this has largely been achieved through development of “screen-scraping” programs, which tend to need occasional script adjustments. Future adoption of XML and web services by federal data providers will facilitate access to their data. The community appears to be moving towards some consensus about standards to adopt, including OGC specifications such as [Web Mapping Service \(WMS\)](#) and [Sensor Web Enablement \(SWE\)](#). A more difficult issue will be the incorporation of state agency-based data, as there are considerable differences among state agencies with respect to data management, as well as limited resources for addressing objectives such as interoperability. However, these agency databases are often unique and contain highly valuable long-term information. Thus, identifying a way to mutually approach the standards issue should be a priority.

The Future

The SEACOOS experience has resulted in some significant “lessons learned” that can help to provide guidance for regionally based IM efforts. First, the establishment of a DMCC and strong communication and problem-solving network has been fundamental to the development of successful aggregation and interoperability approaches and technologies. This requires regular (though not necessarily frequent) meetings to maintain a strong collaborative spirit and confidence level, but day-to-day activities are easily accomplished via electronic communications. Communication tools, such as the [Twiki/Wiki](#) have been heavily used by some program personnel and provide a common platform for electronic dialogue. Second, it is essential that Information Management is recognized as a core function of an observing system and that sufficient resources are allocated at the beginning of the effort. The temptation to allocate “what’s left over” must be avoided, as that cannot sustain a fully functional integrative activity. Third, close attention must be paid to the standards issue, especially now that a number of regional systems are coalescing into a national approach, thus requiring the development of a consensus on some minimal standards. Fourth, it is important to build towards a robust system that incorporates redundancy and back-up capabilities to protect against local failures. Local storms, power failures, or infrastructure breakdowns can disrupt the regional system and a variety of users if there is too much reliance on single locations. Finally, although the initial emphasis of SEACOOS was on real time data, the value of logged data has become increasingly important and should be accommodated in regional IM infrastructures. Logged data are not only essential for assessing real time data quality, but they also provide rich information resources for assessing mid- and long-term changes. This bears on the standards issue and the difficulties involved in incorporating data archives that use a variety of formats and procedures. However, there are rich troves of environmental data in the laboratories of single investigators, state agencies, and private laboratories, and opportunities should be developed to promote their engagement with regional IM operations wherever possible.

There are significant challenges for the regional systems as they move towards the establishment of a nationally coordinated and mutually supportive system. At the same time, immediate priorities for all regional systems are to identify and develop region-specific applications that demonstrate IOOS utility in multiple sectors. The stakeholder base in the regional associations should be an important mechanism for identifying which applications to address first, but input from scientists and information managers will also be essential to establish an effective development and implementation plan. Plans must also incorporate flexibility to enable adjustment as available technologies and priorities change. The IM systems, in particular, must maintain the capacity to take advantage of new software and hardware advances, as well as participate in the evolution of open source technologies that support IOOS IM.

Acknowledgement

SEA-COOS is a collaborative, regional program sponsored by the Office of Naval Research under Award No. N00014-02-1-0972 and managed by the UNC-Office of the President. Many participants have contributed to the success of this program, and we thank particularly Jeremy Cothran, Chris Calloway, Jeff Donovan, Sara Haines, Ed Kearns, Trent Moore, Payne Seal, Vembu Subramanian, and Liz Williams.

References

Seim, H., B. Bacon, C. Barans, M. Fletcher, K. Gates, R. Jahnke, E. Kearns, R. Lea, M. Luther, C. Mooers, J. Nelson, D. Porter, L. Shay, M. Spranger, J. Thigpen, R. Weisberg and F. Werner, 2004. SEACOOS – A model for a multi-state, multi institutional regional observation system. *Mar. Technol. Soc. J.*, 37: 92-101.

OceanUS, 2005. Data Management and Communications Plan for Research and Operational Integrated Ocean Observing Systems. [DMAC-SC Implementation Plan](#)

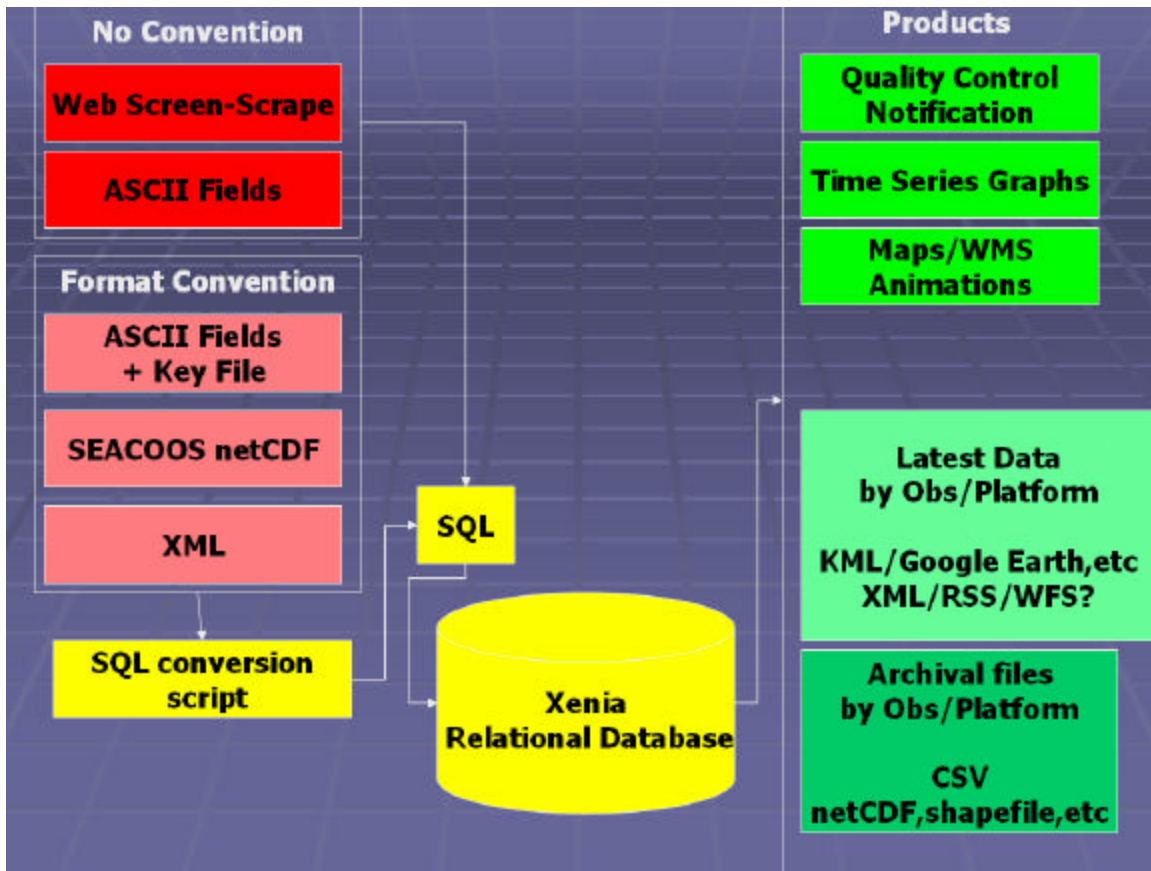


Figure 1. Schematic diagram representing data flow in the Xenia Relational Database, illustrating the multiple data formats and data products accommodated.

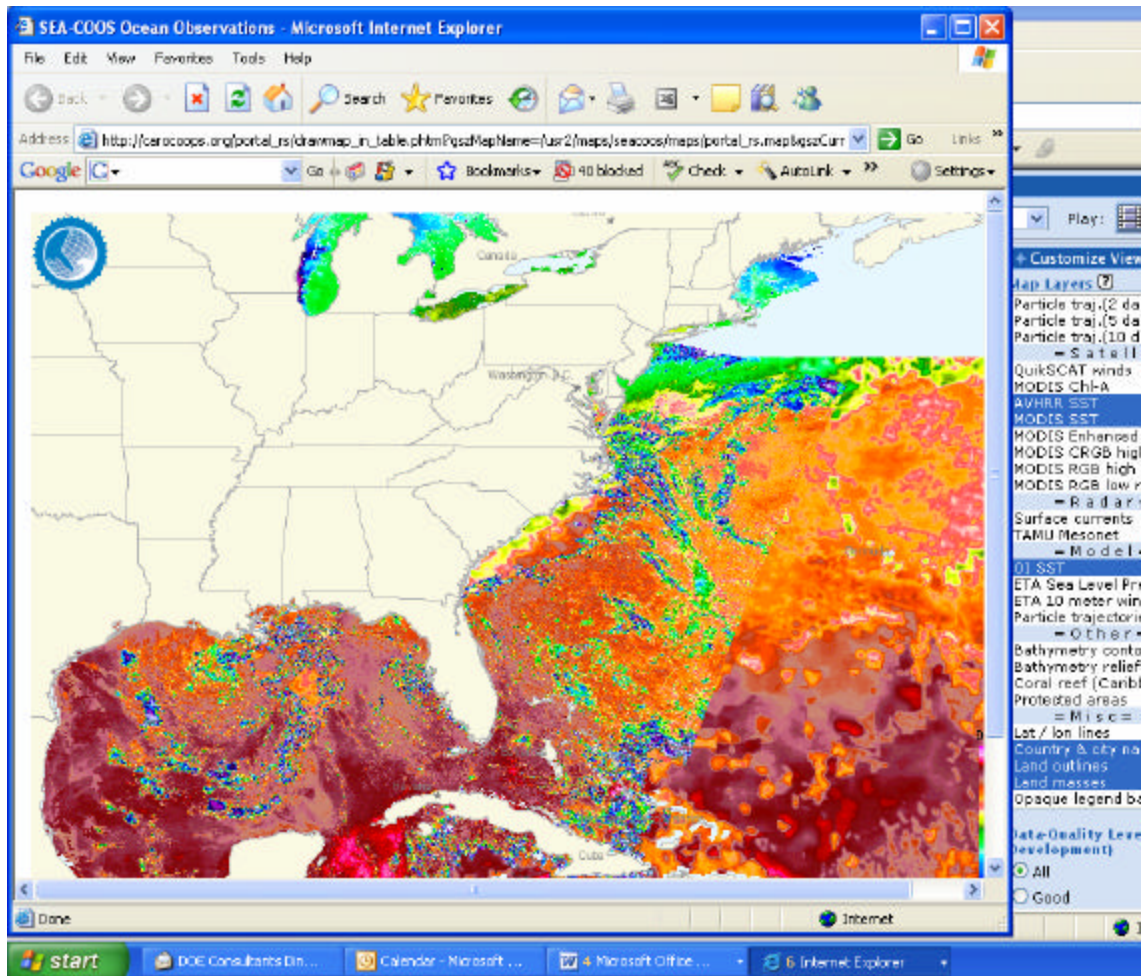


Figure 2. Screenshot of the Interactive Map from the SEACOOS web portal illustrating remote sensing images of sea surface temperature and in-situ sea surface temperature.