"Digital Library Activities at Scripps Institution of Oceanography Library"

Peter Brueggeman

IN: IAMSLIC 2002 : Bridging the Digital Divide : Proceedings of the 28th Annual Conference of the International Association of Aquatic and Marine Science Libraries and Information Centers. James W. Markham and Andrea L. Duda, eds. Fort Pierce, Fla. : IAMSLIC, 2003. pp. 113-118.

ABSTRACT: SIO Library and its SIO Archives are participating in two grant-funded digital library projects, which will bring to the Web materials in their collections, in two principal areas: historical and archival materials on Scripps oceanographic expeditions, and, a monographic series on California fisheries and marine species, including digitization of a multiyear fish landing database published therein. Scope and progress of this effort will be presented, along with some critical issues for others contemplating a similar effort.

Two digital library projects are currently underway at the Scripps Institution of Oceanography (SIO) Library and its Archives, which are referred to internally as CEO and NSDL.

CEO: With federal Library Services and Technology Act funds, the California State Library funded a University of California San Diego Libraries digital library project entitled "California Explores the Ocean" (CEO). This project involves collections at the SIO Archives, SIO Library, and San Diego Historical Society (SDHS). Utilizing a standard Web browser, California citizens can access textual, pictorial, and audio resources related to California oceans and oceanography. The goal of this project is to improve access by the public and scientists alike, to selected information relating to California oceans and oceanography.

Due to the stature of SIO as an oceanographic institution and the depth of the SIO Library's library and archival collections, the UCSD University Librarian targeted oceanography as one of his primary areas of digital library interest. CEO provides an over arching Web structure for UCSD Libraries digital library activities that are targeting oceanographic resource and partnerships, to which materials can be added in the future by other grant-funded efforts, and perhaps, through metadata, by others with significant resources relevant to California oceans and oceanography.

A graphic design consultant developed the complete CEO Web site integrating a variety of resources, which is built on a UCSD Libraries server. The site's Web page leading the user to the FB Fish Catch Statistics has links to other statistical resources relevant to California fisheries. Another Web page on the site leads users to other resources relevant to California and the ocean. The Web site provides a construct for future digital library efforts within the UCSD Libraries, as well as collaborations with other institutions, or simply providing an integrated pathfinder to the efforts of many others. All decisions made about scanning guidelines, metadata requirements and TEI encoding used throughout the project are document on the Web site. One condition of the grant is that a site evaluation be carried out, and a consultant was hired to conduct in-person usability testing, as well as an online assessment using WebSurveyor online survey software.

FISH BULLETIN: The California Department of Fish and Game's Fish Bulletin (FB) is a core resource for the study of fish and fisheries in California. Continuously published as a monographic series since 1919, but slowing down publication in recent years, FB contains in-depth monographs on a variety of topics, primarily marine, and also including some non-fish marine species. Some FB titles are of specialized interest to scientists, state officials, and those with fishery management interests. Many FB titles however are of general public interest, constituting general works on marine species. These general interest titles cover marine fish (including specific titles on sardine, grunion, halibut, tuna, etc), sharks, sea lions, clams/mussels, abalone, squid, historical shore whaling, historical commercial fishing, etc.

The vendor Pacific Data Conversion Corporation (PDCC), scanned as TIFs, OCR'd and encoded 16,500 pages of the FB from the original. A complete run was assembled through IAMSLIC listserver offerings, with strong contributions from Debbie Losey, Joan Parker, and Anne Malley. For preservation purposes, each page (including the cover, the title page, back matter, and advertisements) was scanned at 600 dpi TIFF Group IV lossless compression and copied onto CD-ROM. The FB text was encoded in SGML using Level 4, TEI-Lite with 11,165 accompanying tables, graphs, and photographs embedded throughout as 300 dpi GIF images. GIF was selected as the image format for the textual figures, since it represents tables and graphs in a sharper manner than the JPEG format, which was designed for photos. The encoded text was checked for accuracy and consistency, parsed, and enhanced. The search and retrieval tool used to deliver the Fish Bulletin is available through the University of California's California Digital Library's Online Archive of California (OAC) via Dynaweb, which will be replaced by DLXS in Fall 2002.

At the time of this writing, FB is available only for online reading, and is not available as PDFs for offline reading. Creation of PDFs from the encoded text and figure images is being investigated by OAC. FBs tend to be long monographic works with a high page count; 73% of its issues are over 50 pages in length, and 30% are over 100 pages in length. PDFs should not be created from scanned pages since it would result in immense PDF file sizes. Since the project is targeting the public and schools in addition to scientists, dial-up Internet access (low bandwidth) is an access avenue by users, who would be constrained in downloading large PDF files.

FISH CATCH STATISTICS: FB also published an important collection of fish catch "landing" statistics for California. Published under various titles as "The commercial fish catch of California for the years ....", "The marine fish catch of California for the years ...", "California marine fish catch for (year)", and "California marine fish landings for (year)", these important statistics cover 1916 through 1986, are not available online to the public, and provide a rich source of information for those who study the utilization and management of California fisheries.

At the time of this writing, over 3,000 pages of fish catch statistic tables are to be triple blind rekeyed by PDCC. Locally, each table was photocopied and magnified for easier legibility. A master codebook and spreadsheet samples were created to describe the over forty various table types. Each table was assigned a specific table type, which included the FB number, the table number, the page number, and data elements from the original source. The master codebook, spreadsheet samples, and photocopied tables were delivered to PDCC and then returned as MS Excel spreadsheets. Data from the spreadsheets were examined for accuracy and consistency, and a thesaurus of terms and a list of data available were then prepared. The spreadsheet data were converted into SAS data sets. Using tools the UCSD Libraries have developed for Web analysis of economic datasets, a user front end is being developed for the SAS data sets, so that users can locate, display, and graph data of interest and/or download the data needed in a variety of formats for use with statistical or spreadsheet software.

California catch statistics after 1986 are published in an annual publication "Final commercial fish landing tables for ...", published by Calif Dept of Fish and Game. At the time of this writing, there are no plans to capture this data due to funding limitations, though these annual publications might be added as PDFs on the CEO site.

SIO EXPEDITION REPORTS: Expedition reports commonly include the track of the vessel, list of personnel and ports of call, the expedition objective, and the scientific results of the expedition. For this CEO project, four Scripps expeditions, important for their contributions to science, were selected. Their expedition reports were used to select photographs, track charts, correspondence, cruise narratives, and other content that illustrate the expedition, scientists and work at sea.

1,000 pages of expedition reports were scanned, OCR'd and encoded by PDCC. For preservation purposes, each page (including the cover, the title page, back matter, and advertisements) was scanned at 600 dpi TIFF Group IV lossless compression and copied to CD-ROM. The text was encoded in SGML using Level 4, TEI-Lite with 218 accompanying tables, graphs, and photographs embedded throughout as 300 dpi GIF images. The encoded text was checked for accuracy and consistency, parsed, and enhanced. As with the FB, the search and retrieval tool used to deliver the Expedition Reports is OAC's Dynaweb platform, which will switch to DLXS in Fall 2002.

SIO EXPEDITION PHOTOGRAPHS & SCANNED DOCUMENTS: Thousands of images relating to oceanographic research, fish and fisheries, agar processing, whaling, coastal locations and geography from the SIO Archives and the SDHS can be accessed from the CEO site using the CONTENTdm Digital Media Management software. 5,000 black and white photographs are provided by SDHS. 1,200 black and white and color photographs and about 2,000 non-photographic items are provided by SIO Archives.

For the SIO Archives, a focus is the research ship HORIZON, one of the postwar vessels acquired by Roger Revelle, SIO Director, as he built the SIO research fleet. Scanned drawings, blueprints, diary excerpts, newspaper clippings, correspondence, and oceanographic instruments relating to HORIZON expeditions are included in addition to photographs. SIO Archives materials from historic SIO expeditions on other SIO ships are also being included, through the inclusion of digitization efforts from another digital library grant.

35mm slides and print photographs were digitized by Luna Imaging, Inc. The digital capture standards were based on the California Digital Library Digital Image Format Standards, 2001. 35mm slides were scanned as TIF images at 3,072 pixels on the long-side, resulting in 24-bit color files at 18 MB. Print photographs sized at 4 x 5 inches were scanned at 3,072 pixels on the long-side, resulting in 8-bit grayscale files at 7 MB, or in 24-bit color files at 20 MB. Print photographs sized at 4 x 5 to 8 x 10 inches were scanned at 6,144 pixels on the long-side, resulting in 8-bit grayscale files at 20 MB, or in 24-bit color files at 20 MB

SIO Archives offers one version of each TIF image for public use within CONTENTdm, a Medium resolution JPEG at 768 pixels long-side. The original TIF images are not managed within CONTENTdm, nor presented to the public, being slated for internal digital library management on a high capacity image storage system under development at the UCSD Libraries. SDHS doesn't offer the public its high resolution images either, offering for public display medium resolution JPEGs of 384 pixels long-side. CONTENTdm creates thumbnail images of these public images on-the-fly for review by users.

During the image selection process by the SIO Archivist, SIO Archives used an MS Access database to enter basic descriptive and administrative metadata for each item, using the database to track delivery and receipt of materials to the vendor. Fifty-six metadata elements were worked out for SIO Archives metadata needs, now serving as the standard for the description of digital objects for the SIO Archives. Both the SIO Archives and SDHS use sixteen metadata elements that are mapped to Dublin Core. After receiving the scanned images from Luna, the database was exported to ASCII delimited text and imported into CONTENTdm, with eighteen selected for public display. To unify SIO Archives and SDHS collections for public use, a list of fourteen broad subject heading terms was developed using the Thesaurus for Graphic Materials (TGM), using headings from TGM I and II. One or more terms were assigned to each image, and these terms seemed to cover broad categories of materials in both collections. These terms are the following: Aerial Views; Beaches; Diving; Events; Fishing; Fishing Industry; Harbors; Navigation And Communication; Ocean Life; Ocean Resources; Oceanography; People; Scientific Equipment; Vessels. If TGM did not provide what was needed to broadly categorize the content of the two respective collections, a heading was made up. For example, TGM uses the term BOATS, which was considered too narrow and not sufficient for an oceanographic research institution, thus the term VESSELS was chosen. These fourteen broad categories function as a predefined search which can be used to query one collection or both simultaneously, affording the novice user an easy entry into the body of materials.

SDHS ORAL HISTORIES: SDHS owns dozens of oral histories that document San Diego's fishing industry, particularly the rise and decline of the tuna industry. 51 oral history transcripts (1,250 pages) were scanned and converted to text using OCR, then coded into HTML.. 26 of those oral history analog tapes were digitized into WAV format for archival retention, and then reformatted to MP3 files for Web use.

NSDL: The second digital library project at SIO Library is "Bridging the Gap Between Libraries and Data Archives", which is at an earlier stage of development compared to the CEO project. "Bridging the Gap.." is a joint project between the UCSD Libraries, the SIO Geological Data Center (GDC), and the San Diego Supercomputer Center (SDSC), funded by the National Science, Mathematics, Engineering, and Technology Education (SMETE) Digital Library (NSDL) Program of the National Science Foundation (NSF). Since 1996 NSF has studied the development of a national digital library for science, mathematics, engineering and technology education. Building on work supported under the multi-agency Digital Libraries Initiative, NSF developed the NSDL program to found a national digital library that will constitute an online network of learning environments and resources for science, mathematics, engineering, and technology education at all levels.

This NSDL project involves an oceanography digital library collection providing access to many years of SIO shipboard data, historical photographs and documents, samples, selected research publications, and maps from global databases. From a global map of approximately 3000 oceanographic cruises, users will be able to identify and retrieve relevant materials, from photographs to diaries to scientific papers and data, using modern search engine tools, metadata standards, and advanced storage and computational technologies. At the same time, students and researchers are able to use the site to locate and download scientific data from SIO for further analysis and research. Requests are currently made each year to separate entities for data and historical materials to support proposal development, cruise preparation, research and publication. Currently, almost every request requires manual intervention. With these efforts, access can be made much more efficient.

Metadata is being specified to link scientific data, scientific publications and historical archival materials through a relational database. A search and delivery interface with a latitude/longitude box drawing interface, the SIOExplorer, is being developed that will present materials from data archives and libraries in an integrated format. While geospatial searching will be a key feature of this user interface, searches by subject and by format (for example, photographs or data) will also be facilitated in order to appeal to the widest possible audience. This NSDL project was funded by NSF for two years, whereas the CEO project was funded by the California State Library for one year, and is nearing completion. The NSDL project is still in development, and is more complex for the disparate resources it knits together.

At this time of writing, the SIO Archives has selected six expeditions for inclusion in the project, and is identifying 1,200 photographs as well as 600 documents and related materials for digitization. Expedition reports from these six expeditions will be produced as encoded text. Scientific publications relevant to these cruises have been identified, and publishers will be asked for permission so that this project can post them as PDFs. These SIO expedition materials will get double-duty since they are loaded on the UCSD Libraries' CONTENTdm server, and thus integrated into the UCSD Libraries' CEO initiative. Digitization and metadata issues are the same for both projects, with additional facets being developed for this project with respect to geography. Thesauri have been identified for use jointly by the GDC and the SIO Archives. For topside geographical names, we are using the Alexandria Digital Library Gazetteer. GEBCO's Gazetteer of Undersea Feature Names is used for undersea feature names, while the ocean area naming scheme from the International Hydrographic Organization's Limits of Oceans and Seas is used to name ocean areas. GDC will be acquiring the latitude/longitude polygons for the IHO ocean areas, in order to map archival and library materials with georeferenced data.

The GDC, working with SDSC, has developed techniques for inserting metadata into multibeam seafloor data files, using MB-System. Cruise-level metadata specification has been designed, and a 'gmtplus' underway processing system has been ported to work with the current GMT version 3.4.1 software toolkit. SDSC has developed an Oracle based Metadata Catalog to provide information on each digital object in the Canonical Cruise Data Structure (CCDS). The CCDS is a logical object comprised of a managed hierarchical structure that encompasses all the forms of data produced during a cruise and organizes the relationship of the shipboard data acquisition to GDC management. Currently there are 36 categories of information for a single expedition cruise leg. The data structure will be managed by the Storage Resource Broker. Additional technical details are available on request. Work is in progress at the GDC to come to terms with the idiosyncrasies of a digital collection that has evolved over 30 years. Pre-existing data structures and search mechanisms are being maintained until after a modern database approach has been implemented and tested.