DIGITIZING HISTORICAL OCEANOGRAPHIC EXPEDITION CONTENT

Peter Brueggeman
Scripps Institution of Oceanography Library

Deborah Day
Scripps Institution of Oceanography Archives

University of California San Diego
9500 Gilman Dr Dept 0219
La Jolla CA 92093-0219

ABSTRACT: The Scripps Institution of Oceanography (SIO) Archives has been digitizing a large body of archival content relating to historic SIO expeditions, through two digital library grant projects. After World War II, SIO scientists sailed the oceans of the world to study the seas, marine life and the geology of the sea floor. The SIO Archives has a rich collection of materials relating to these expeditions, which have long been of considerable interest to users, but difficult to access. These digital library projects are making these materials available as images, page scans, and encoded text, where relevant. Metadata has been developed to meet internal needs as well as external user needs, and also the needs of digital library project partners. Time-consuming reference service and access to these materials is being greatly facilitated by digitization.

KEYWORDS from ASFA: Digital records; Archives; Libraries; Imagery; Documents; Expedition reports; Audiovisual materials

Imagery, documents, and other materials relating to the history of an institution are steadily compiled by the institution, the staff, and their descendants. The Scripps Institution of Oceanography (SIO) Archives of the SIO Library was functioning as an entity long before Deborah Day, its first professional archivist, was hired in 1981. Today we find ourselves with over 80,000 images relating to the history of SIO, and project continued growth with retirements and deaths. Requests for images have changed dramatically in recent years, with demand broadening among user communities, and demand increasing for digital images. With demand increasing, reference service for materials in physical format involves an increasing percentage of repeat requests for the same materials, with the physical medium of those materials requiring time- consuming handling and procedure to process those requests. Even once digitized, constructing and maintaining a digital asset management system is imperative to gain a time savings from their digitization. Grant-funded projects, though having their imperfections as a means to an end, provide a mechanism to reshape for new demands and priorities, by establishing

infrastructure, policies and procedures, and by opening up an opportunity to capitalize on staffing turnover and change.

In addition to other resources, two-grant funded digital library projects have enabled the SIO Archives to create a digital library infrastructure on which to build its ongoing digital library program. Started with over 5,000 images, page scans, encoded texts, audio, and video, this digital library initially focused on historic SIO scientific expeditions back to 1907 due to a demonstrable high interest in them as well as the likelihood of receiving grant funding. As SIO Archives moves beyond a grant-funded digital library, digitization is now expanding out to encompass items selected by the SIO Archivist and known to be of high user interest. Scanned items include 35mm transparencies, prints, panoramas, correspondence, news clippings, and ships' logs. Encoded text encompasses expedition reports and several key texts on SIO history. Audio and video is initially modest, encompassing a news media video clip of past SIO Director Roger Revelle, as well as an oral history of Revelle, incorporating a Quicktime slide show of historic images. Scanning and image standards were set in accordance with California Digital Library Digital Image Format Standards. 35mm slides, smaller prints, and negatives were captured at 3,072 pixels on the long-side for the uncompressed TIFs. 4 x 5 to 8 x 10 inch prints were captured at 6,144 pixels on the long- side. Three derivative versions of each image were produced: a thumbnail JPEG (192 pixels long-side); a medium resolution JPEG (768 pixels long-side); and, a high resolution JPEG (1536 pixels long-side). The SIO Archives presents the user on the CONTENTdm system a thumbnail (automatically generated by CONTENTdm) and that medium resolution JPEG. Many users seeking images for PowerPoint presentations are satisfied with this medium resolution JPEG. All images are stored behind the scenes on a UCSD Libraries digital asset management service, from which staff can retrieve images for various needs. No image editing (sharpening, cropping, color correction) is done. The images are intended to be digital surrogates of the originals, having scratches, tears, stains, marks, and various flaws as in the original. During the selection process, the SIO Archives used a MS Access database to enter basic descriptive, technical, and administrative metadata for each item. Since the items were shipped in batches to an external vendor for scanning, with processing work involving several staff, it was important to create a record to track items, starting from the initial selection by the SIO Archivist. Full metadata is added when time allows, and before the images are made publicly available. The Access database is exported to ASCII-delimited text and imported into the CONTENTdm database, which is used for public display. Of the fifty-six metadata elements, only eighteen display to the public.

Fifty-six metadata elements were adopted and now serve as the standard for the description of digital objects for the SIO Archives. The metadata addressed specific needs of each of the two grants. Metadata was designed for SIO Archives' ongoing digital library needs, accommodating needs beyond the two grant projects, and relying on existing standards where applicable. The considerable experience of the Archivist in serving users was used, and common types of user requests are well known and a major consideration in designing metadata. Not all fields are mandatory, and the fields corresponding to the Dublin Core were specified. The metadata design for digital objects

at SIO Archives can inform similar digital library projects at other marine science libraries and archives so we will outline some fields and associated issues.

Object location and identity information within the SIO Archives or SIO Library collection constitutes fourteen fields of metadata. Collection Title corresponds to the Dublin Core Source field. The Title (of item) field corresponds to the Dublin Core field of the same name. Not all items have titles, yet the SIO Archivist or others know what is depicted or represented in the item, e.g. who is in the photo. Title information that is external to what is noted on the item itself is bracketed in the Title field.

The Creator field corresponds to the Dublin Core field of the same name, being the name and role of person/organization creating an object, e.g. Jane Smith, photographer.

The Notes field provides information transcribed from or supplied for the item, e.g. "Editor has used blue pencil to mark cropped edges for publication." The images are digital surrogates of the originals, having flaws as in the original.

Date (of item) field corresponds to the Dublin Core field of the same name, and is item normalized as ISO 8601: YYYY-MM-DD, YYYY-MM, or YYYY.

The Format field corresponds to the Dublin Core field of the same name, being the physical or digital manifestation of the item, e.g. jpeg, quicktime. The entry for this field is selected from the Internet Assigned Numbers Authority (IANA) official registry of Multipurpose Internet Mail Extensions (MIME) media types: see http://www.iana.org/assignments/media-types/index.html

The Source Type field corresponds to the Dublin Core Type field and is the nature or genre of the item, e.g. image, sound, text. The entry for this field is selected from the Dublin Core Metadata Initiative (DCMI) Type Vocabulary:

see http://dublincore.org/documents/dcmi-type-vocabulary/

The Document Type field records the data type of the original source: Expedition Report, Bibliography of Scientific Paper, Published Scientific Paper, Manuscript, Photograph, Track Chart, Seaman's Narrative, Ships Log, Biography of Key Scientist, Drawings of Instrument, Newspaper Clipping.

Several fields describe the physical nature or characteristic of an item. The Genre field is assigned to the item from Library of Congress Subject Headings (LCSH), describing the category or characteristic of an item, including logbooks, maps, nautical charts, clippings, drawings, caricatures and cartoons, portraits and posters. Some local genre headings were derived from the Library of Congress

Thesaurus for Graphic Materials II: Genre and Physical Characteristic Terms (TGM II), and include cruise certificates, shipboard communication, aerial photographs, underwater photographs, and trick photographs. The Physical Description field corresponds to the Dublin Core Description field and physically describes the item, e.g. Color Kodachrome slide. Other physically descriptive fields are Source Dimensions, Original Format, and Number of Pages. The Language field corresponds to the Dublin Core language field used for text objects.

See http://lcweb.loc.gov/rr/print/tgm2/

Geospatial fields are several and thorough, encompassing terrestrial locations, named undersea features, and ocean areas. The Subject Topside field describes a terrestrial geographic location for the item, using terms from a controlled vocabulary, the Alexandria Digital Library Gazetteer

See http://testbed.alexandria.ucsb.edu/gazclient/index.jsp

Incorporation of the ADL Gazetteer location terms into our metadata results in usage in accordance with US Geological Survey and US National Imagery and Mapping Agency terminology, and ensures interoperability with other University of California and National Science Foundation funded digital library endeavors.

The Location Depicted field corresponds to the Dublin Core Coverage field, utilizing LCSH terms and providing correspondence with our library cataloged materials. There can be differences between ADL/USGS/NIMA and LCSH, e.g. Samoan Islands and Samoa Islands respectively.

The IHO Number and IHO Location fields correspond to grid numbers and names from the 3rd edition (1953) of the International Hydrographic Organization Special Publication 23 Limits of Oceans and Seas, which records the designations and boundaries of oceans and seas in the world. These fields were added for geospatial cross-referencing between our archival holdings and SIO's Geological Data Center's data holdings.

See http://nsdl.sdsc.edu/

Latitude/longitude polygons for these IHO ocean areas will allow cross referencing between archives holdings and data holdings for the same ocean area. Ocean data holdings are much more granular than archival items from the same ocean area. Exact latitude and longitude is rarely known for an archival item, and typically we know only broad ocean areas for archival items. When the exact latitude and longitude is known, it is added to the Notes field. There needed to be a geospatial bridge between ocean-related data and archives holdings for SIO expeditions, and the IHO grids were selected as a standard and as being practical. For archival items associated with a terrestrial location, they cannot be assigned an exact latitude/longitude either.

The Subject GEBCO field is used for names of underwater features associated with an archival item, using the GEBCO Gazetteer of Undersea Feature Names, by the General Bathymetric Chart of the Oceans (GEBCO) of the Intergovernmental Oceanographic Commission and the International Hydrographic Organization. See http://www.ngdc.noaa.gov/mgg/gebco/underseafeatures.html

Subject fields are several. The Aquatic Sciences and Fisheries Abstracts (ASFA) Thesaurus was considered and not selected for subject indexing. Long experience with serving users was a major factor in this decision. The ASFA Thesaurus was far more precise than needed for almost all past user inquiries, and would consume more cataloging time to utilize. If we used ASFA terms to catalog our digital archival items, they would not map to the Library of Congress Subject headings used for items cataloged in our library collection. There is considerable content residing in the cataloged library collection that is of historical interest and complements archival materials. There we want to map to library catalog records from our archival metadata.

The Subject LCSH field corresponds to the Dublin Core Subject field, utilizing LCSH terms describing subject content of the item. An internal document outlines SIO Archives practice and preferences for subject cataloging its digital objects using LCSH and local subject headings. Local headings are specified where LCSH is insufficient to describe significant holdings in Archives, e.g. "Crossing the Equator" for equator crossing ceremonies; "Oceanographic research ships -- Interiors" for interior photos of research ships; a subject heading subdivision "Press conferences."

The Subject Person field is used for the name(s) of individual(s) depicted in an item, with name authority being the same as used in the library catalog.

Expeditions with multiple vessels were divided into cruises, with a cruise representing the track of each individual vessel. Lengthy cruises were divided into legs, which represented the work of the vessel between two points. Several fields parse out expedition and ship information, with their authorities being library cataloged name authority, an SIO technical report listing expeditions and cruises, a Cruise Index maintained by SIO's Geological Data Center, and the names that the University- National Oceanographic Laboratory System (UNOLS) uses to refer to its research ships, which include SIO's ships.

See http://repositories.cdlib.org/sio/reference/77-13/ See http://gdcmp1.ucsd.edu/gdc/cruises/cruise.index See http://www.unols.org/images/ships/shipimages.html

A Subject Category field is used to provide effective and simplified access for items from grant project partners. The UCSD Libraries partnered with the San Diego Historical Society in the grant funded California Explores the Ocean project. Their visual collections are made easily accessible to the general public through fourteen broad

subject heading terms which were assigned to each image: Aerial Views; Beaches; Diving; Events; Fishing; Fishing Industry; Harbors; Navigation And Communication; Ocean Life; Ocean Resources; Oceanography; People; Scientific Equipment; Vessels. This list was developed using the Library of Congress Thesaurus for Graphic Materials (TGM), using headings from TGM I and II. If TGM did not provide each archivist with what they needed, a heading was made up (for example: TGM uses the term BOATS, which was considered too narrow, thus the term VESSELS was chosen). These fourteen categories serve as a predefined search which can be used to query one collection or both simultaneously.

Seventeen fields record copyright information or technical information about the digital object and its creation, with three fields being Dublin Core: Item Rights, Item Publisher, and Filename. Item Rights is a key field because it establishes who owns the image and thus may grant permission to reproduce the image. Rights and granting permission is clear when the Archives has a recorded deed of gift for the items (or knows that rights were not granted). However items reside or arrive in the SIO Archives without deeds of gift. This is particularly a problem for early images documenting the history of SIO, i.e. studio portraits of turn of the century scientists. The descendants of the photographer cannot be located. If the Archives cannot find or determine the rightful owner, can it still digitize a slide for public display, or give a scan to the SIO Director's Office for publishing in their quarterly news magazine? For images without deed of gift, SIO Archives has to document the steps it followed in unsuccessfully locating the image owner if SIO Archives decides to grant permission to use that image. We may need to expand our metadata to better record our due diligence in determining rights for images without deed of gift. We will need our trail of effort recorded if our rights to an image are contested in the future.

SIO Archives is well poised to capitalize on its initial grant funded efforts and the supporting infrastructure of the UCSD Libraries, and to re-shape itself to better meet contemporary needs. Having a digital object management system with 5,000 images and other digital objects in place has already heightened awareness of our images among our users. Images are of highest interest to everyone, which is not a surprise. The SIO Director's Office has many images that it holds onto for its use, and that are unknown to the larger institution. When the SIO Photo Lab operation was discontinued, the Director's Office held onto the color images from the Photo Lab, and SIO Archives was given the black and white images. Thus much is hidden to the larger institution and is not readily available for use. With a digital object management system easily available to both the Director's Office and to the institution, we anticipate an increased acquisition rate of images from our institution, since we are now situating ourselves to be the best managed and logical home for institutional imagery.