**Proceedings of the 6th Annual
Federal Depository Library Conference**

April 14-17, 1997

Access and Preservation via Digital Surrogate for Spatial Data

Mary Lynette Larsgaard, University of California
Santa Barbara, CA

One of those problems of success in many spatial-data collections is how to preserve heavily used collections while at the same time making these same collections available for what appears to be an ever-increasing audience. This paper will range from work done in one specific spatial-data collection, to a consortial approach, to an overall view for georeferenced information, as seen from the viewpoint of the digital library.

Access or Preservation - or Both?

Some key thoughts in most map librarians' minds when they contemplate moving hardcopy collections into digital form are:

- Who are my library's primary users--present and future--and what are their needs?
- What purposes and benefits does this project have?
- I don't want to have to have this work done again--but how can my library financially support preservation-level scanning?
- For maps, is what is needed access to the printing separates, or is a digital picture, whose information layers cannot be manipulated digitally, acceptable for the users of this collection?

The results of having asked these questions and come up with at least some of the answers at the Map and Imagery Laboratory (MIL) are the topic of this paper.

MIL's digital efforts are in three inter-related areas:

1. Davidson Library;
2. University of California/Stanford Map Librarians Group (UCSMLG); and

3. Alexandria Digital Library (ADL).

In actuality, these three areas are interrelated, since the Davidson Library administration is a strong supporter of MIL's participation in UCSMLG and of the Alexandria Digital Library, and since UCSMLG is considered a primary source of information-intermediary audience and evaluator, data, and metadata.

Davidson Library

The library administration has supported a comparison study of scanning of air photos, by three different firms: Stokes Imaging, Luna Ltd., and TGS Technologies. The object of this study was to find out what would be an optimal way to get MIL's 3,800,000 air photos scanned, preferably as many as possible of them in time to form a part of the ADL collections. As will be noted later, MIL already had at this point a scanner which is being used mainly to scan air photos, but our primary goal at the time of purchase of that scanner was mainly to have an inexpensive data-ingest scanner for Alexandria research. Prices per scan differed considerably from firm to firm, varying directly with the size of the file generated by the scanning procedure (itself dependent on dpi and bit depth).

A few points we found out very quickly. Eight-bit depth is usually sufficient for general use of air photos (although we have experimented and scanned some at 24-bit depth, and then decided the tradeoff of having a file three times the size of an 8-bit file was probably not justified for our general users), and 24-bit depth is appropriate for air photos. Our decision as to appropriate dpi level was made by eyeball, noting that an air-photo print scanned at 600 dpi gives the general user about what is available from the actual print. Neither 150 nor 300 dpi is sufficient. 600 dpi is sufficient for many users, but how much more detail is needed for very sophisticated users and for archival purposes? Or, to put it another way, what is "enough" resolution? And the answer is, as always, that depends: who are your users and what do they need to do with the information?

Scanning to the size of the silver-halide grains (12 microns) in the film emulsion-which is appropriate for archival purposes-results in very large files. There is currently a scanner designed especially for scanning air photos (it even has a roll-film transport)-Lenzar's LENZPRO-which will scan to that level. It is also very fast, taking two minutes and eight seconds to scan a 9" x 9" air photo, resulting in an 800Mb file; it is intended for heavy-duty production scanning, and thus is ideal from MIL's point of view-but the cost of $120,000 is not!

Another decision is whether it is best to scan one's own materials, or to farm out the scanning. Once again, it depends: is this a one-shot deal, or will your library continue to do scanning over long periods of time; and moreover, will you and affiliated institutions be doing scanning of so many items that economies of scale may be realized by collaborative operations? This leads us to the next of MIL's projects for digital access and preservation.

University of California/Stanford Map Librarians Group (UCSMLG)

The UCSMLG is a close-knit group that has been in existence since the mid-1970s. On June 20, 1996, the group held a meeting at MIL. We had available to us a handout from a meeting held June 14, just a few days previous, at the University of California at Berkeley libraries, that addressed the more general question of digitization of all types of library materials. This draft has subsequently gone to a second draft and we hope to see a final version out sometime in the first half of the year.[1]

The criteria given in this draft handout, titled "Principles of Selection for Digitization," are:

- Meets current faculty and student information needs;
- Offers economies of scale by benefiting many faculty and students (locally and worldwide);
- Maintains local or consortial collection balance among disciplines, information formats, and instructional and research tools;
- Adds value over paper- or film-based copies in various ways (e.g., more timely availability, more extensive content, greater functionality, greater access, improved resource sharing due to the ubiquity of digitized resources, increasing usefulness of the total collection, etc.);
- Justifies costs of digitization, including archival maintenance and access costs for the library as well as for its users;
- Achieves the goals of conversion to digital form (e.g., publishing, archiving, replacing, preserving);
- Meets criteria of copyright, fair use, and other legal restrictions imposed on intellectual property;
- Provides orderly access and navigation to and within the item or collection;
- Is accessible from all institutionally-supported computing platforms and networked environments;
- Employs formats that follow industry standards and are fully documented;
- Platform-independent, available in a multiplicity of formats;
- Originals difficult to use;
- Commitment made by library to preserve both originals and digital files;
- Possible to capture information adequately, to enable the digital version at least to serve as a surrogate for the original, thereby reducing demand for (and thus wear and tear on) originals;
- Originals not damaged by the conversion process;

- Losses of integrity of files caused by migration of files minimal;

- Preservation problem already exists with original (e.g., risk of damage or loss);

- Security needed for original;

- Of interest to funding agencies, and

- Originals have research value; etc.

With this list in our hands, the group discussed what needed to be done, and as a result put together a first-step draft white paper[2] (goals; procedures) and a list of items that most needed to be digitized (e.g., unique, heavily used collections at each library; items each library uses heavily, such as historic U.S. Geological Survey topographic quadrangles of California; etc.). The white paper is available at:
http://alexandria.ucsb.edu/~carver/ucop3.htm.

Following is a list of some of the collections suggested for digitizing by the map libraries:

- Topographic survey [of the coasts of the United States] / U.S. Coast & Geodetic Survey. Scale 1:5,000-1:80,000. 1851? G3700 svar.U5 Case B Library has: 256 sheets, Reports T-1825-7, T-3653

  NOTE: Scan Bay area 1:10,000 sheets. First geodetic survey of the coastline.

- Pacific Aerial Surveys. [Aerial photography, Berkeley campus, 1994]. Scale [ca. 1:600] Oakland, CA : Hammon/Jensen/Wallen & Associates, 1994. Map Room G4364.B5:2U5A4 1994.P3 Case B Library has: 22 col. photos. NO copyright.

- United States. National Ocean Service. [United States nautical charts]. Scales vary ; Mercator proj. Washington, DC: U.S. Dept. of Commerce, National Oceanic and Atmospheric Administration, National Ocean Service, [18--? Map Room G3701.P5 svar.U51 Case B -

  NOTE: Scan all Bay area charts 1849-present, perhaps at 20 year intervals. Useful for shoreline changes.

- Non-copyrighted Bay-area cities. i.e., pre-1946 copyright expired, maps, especially Oakland, Berkeley, San Francisco

- Index maps for maps and air photos -- AMS/DMA index maps would be ideal, since most are small. (1K-3K sheets?)

- All CA topographic quads (or maybe those before a certain date, say 1950)(2K? sheets)

- County road maps from Caltrans. Maybe consider digitizing one set for each decade?

- Maps of historical interest, including old plat maps of Berkeley, maps that show the burned area of San Francisco after the 1906 earthquake, how the San Francisco shoreline changed, county maps showing California rancho boundaries, etc.

- Outline and base maps of all kinds (county boundaries, hydrology, main roads, etc.) useful for students and people in business

- California Forest and Range Experiment Station. Vegetation type maps of California and western Nevada. Prepared by Forest Survey Staff, A.E. Wieslander in charge ... in cooperation with the University of California. [Washington, D.C., 1932-38].

- Los Angeles City. Bureau of Engineering. Street Opening & Widening Division. Topographic Map sets of Santa Monica Mountains, Sunland-Tujunga-Verdugo Mountains, North-East Los Angeles, Sylmar-Granada Hills, Chatsworth Reservoir-Canoga-Park-Knapp Ranch, Baldwin Hills-Westchester- Playa Del Rey, San Pedro, LA Freeway Downtown Loop, Central LA, and Benedict Canyon. Los Angeles: Bureau of Engineering, 1959-75.

- Los Angeles (Calif.). Police Dept. Area boundaries of the Los Angeles Police Department [map]. Los Angeles, CA: Los Angeles Police Department, Cartography and Visual Aids Unit, [1992-96]

- Nature Conservancy (U.S.) Color infra-red aerial photos of Santa Cruz Island. Santa Barbara, CA: Pacific Western Aerial Surveys, 1985.

- United States. Bureau of the Census. 1990 County Block Maps. S.T.F. series for Los Angeles County (includes census tracts. Washington: U.S. Bureau of the Census, [1991].

- United States. Bureau of Land Management. [Township plat maps of the United States]. [1855?-

What seems to be a workable (but not easily fundable) solution is to have two scanners: one for oversize items, e.g., maps, and the other a production-level scanner for air photos, that would travel from campus to campus, starting with the items of greatest use to the largest number of University of California users (e.g., the aforementioned historic USGS topographic quadrangles), or unique items at greatest risk of being damaged or lost. In any case, the digitized items would be prime candidates for inclusion in or transparent search by the Alexandria Digital Library.

The Alexandria Digital Library

First, a few words about the Alexandria Digital Library (ADL), how it began, what it is, and its goals and accomplishments. Its overall goal is to build a distributed digital library for geographically referenced materials - maps, images, text, multimedia, and so forth. The Alexandria Digital Library is one of six Digital Library Initiative (DLI) projects funded jointly by the National Science Foundation (NSF), the Advanced Research Projects Agency (ARPA), and the National Aeronautics and Space Administration (NASA).

The six funded institutions are Carnegie-Mellon, Stanford, the University of California at Berkeley, the University of California at Santa Barbara, the University of Illinois at Urbana-Champaign, and the University of Michigan at Ann Arbor. Each of the six projects, which began in October of 1994, and run through September of 1998, has a different focus; the focus of ADL is to provide online access to georeferenced information, with an emphasis on spatial data.

Since it is estimated that about 90 percent of all spatial data is available only in hard-copy form, metadata is of the greatest importance, given that very often that is all the user will be able to find in digital form. ADL has a beta-test site up on the Web site, which we encourage you to visit, try out, and let us know how to improve. For more information on ADL, or on any of the other five DLI projects, go to the Web site: **http://alexandria.sdc.ucsb.edu**.

As a major part of the prototype, approximately one hundred items in digital form were ingested. About 60 of these were aerial photographs; only three or so were hardcopy maps, and the rest were georeferenced information already in digital form (e.g., AVHRR - Advanced Very High Resolution Radiometer; DEM - Digital Elevation Model; DLG - Digital Line Graph; TIGER files from the 1990 U.S. Census; Landsat satellite images; SPOT satellite images; a text on the Channel Islands with a link to an image of the islands). A CD-ROM was made with metadata for all the images on it, plus about 40 of the actual digital items.

Why did the prototype, in the scanning arena, focus on air photos? Several influences at the same time. Firstly, MIL has about 3.8 million air photos. The photographs of southern California, especially of the older flights (such as a 1928 flight of the coast of Santa Barbara County) are very heavily used. This very frequent pulling and refiling (some indexes are used several times in one day) - even though it is done by student assistants and not by the users - is deleterious to the indexes and the frames. It is extremely labor-intensive; one search can easily take several hours of staff time to pull and refile. It is easy to misfile; and as is true with all the "spineless" cartographic materials, when an item is misfiled, especially in such a large collection, it's gone.

The scanners MIL has are a venerable, finicky Eikonix (ca. 1987) and a Sharp JX-610. The Eikonix, which was purchased for about $60,000, takes 45 minutes to scan a color object, resulting in a 48Mb file, or about 20 minutes to scan a 15Mb black and white file; in either case the result is a 6,000 x 6,000 pixel image, no matter what the size of the object, be it 35mm slide or 5' x 4' nautical chart.

Unfortunately it is almost impossible to get a good color balance unless one is scanning a transparency on a light table (a consistent light source is essential). Maps are not transparencies, and in addition we were not interested in scanning the maps in sections, as has been done in other libraries; this meant we were effectively limited to scanning air photos.

The Sharp has a maximum size of 11" x 17," with several different dpi's possible: 150, 200, 300, 400, and 600. It takes well under five minutes to scan one item, even at the 600 dpi that MIL selected to scan its air photos (which resulted in 29Mb for black and white and 98Mb for color). Thus the Sharp (which was ordered with a special attachment so that it could scan transparencies) is ideal to scan 9" x 9" air photos, which constitute the vast majority of MIL's air photo collection; we do have about 90,000 4" x 5" obliques and perhaps 10,000 9" x 18" air photos. We did discover later on that the Sharp does introduce some distortion, in the direction of the scanning arm; this means that the scans are not appropriate for use in photogrammetry, although for general use, they are fine.

During January of 1997, MIL initiated a pilot production-scanning project, using the Sharp scanner and funded by research done by the University of Arizona. One skilled (Arc/Info, Unix, scanner) worker could scan and create metadata for three frames per hour. This worker also generated coordinates for frames off air-photo mosaics and other indexes at the rate of 400 frames per hour.

Another reason that the air photos are an excellent choice for scanning is that the size of a monitor of a computer is seldom more than twenty-one inches and for probably the majority of users it is much less. Air photos are perfect: they are 9" x 9", and thus can be displayed on many monitors at exactly the size of the original item, or even larger. While it is true that air photos are very high resolution, well beyond the 600 dpi-maximum of the Sharp scanner, for many users, the 600 dpi resolution that appears on a screen monitor seems to be acceptable.

While there are maps that are 8.5" x 11", the bulk of MIL's maps are far larger, since MIL specializes in medium-scale topographic sheets (say, 2' x 3') and nautical charts (which can easily be, as was previously mentioned, 3' x 4' and even larger). This meant that if MIL had decided to scan maps, users would first have had to view a thumbnail, and then zoom in to an area of interest. The problems of providing users with a location map, scale, north arrow, and legend that could be popped up at any time the user needed them will need to be solved over the next two years, but were certainly not anything we could deal with in the short term.

Air photos are one layer of information, which means that scanning them works extremely well as a form of delivery of information. What some users need to do is to manipulate the different layers of information that make up each map, which means that ideally the individual print separates would be scanned (although there is technology that can "scan" a printed map and separate out the layers with some level of success). On the other hand, many users just need to look at a map (sometimes slightingly called the "pretty picture" syndrome) so certainly scans of maps are by no means useless. It happens that the University of California at Santa Barbara has both a very active Geography Department that emphasizes the use of spatial data in digital form, and the National Center for Geographic Information and Analysis. Given those two points, and given that Alexandria is a research project, what would be most appropriate both for the faculty and for the ADL funding agencies would be to work with layers of information.

During a late-January 1997 meeting of the ADL Advisory Board, the main recommendation of the Board was the need for increased content (data and metadata) in Alexandria. In light of this recommendation, MIL is working on extending the pilot production-scanning project, since the nearly one gigabyte a day resulting from this work now has a place to go - disk storage - and a server capable of handling heavy traffic - a DEC AlphaServer 4100, whose system name is, appropriately enough, fat_albert.

Conclusion

Spatial-data collections find themselves in the "interesting times" of the Chinese proverb, as we simultaneously maintain our hard-copy collections while steadily and increasingly collecting data in digital form. Digitizing the hard-copy collections to keep them from damage caused by handling more and more looks like the way we will need to proceed.

[1] Draft on digitization criteria; UC selection criteria for digitization. Draft, 21 November 1996. [Oakland, CA?: University of California Office of the President?}, 1996. Message-id: Pine.ULT.3.91.970121151901.13752A-

100000@ariz.library.ucsb.edu

[2] [Carver, Larry]. Draft 1.0 (UC/S working group for distributed spatially indexed information), 8-19-96; Alexandria Digital Library for Spatially Indexed Information. Santa Barbara, CA: Map and Imagery Laboratory, Davidson Library.

http://alexandria.sdc.ucsb.edu/~carver/ucop3.htm