AC 2010-1093: MANAGING A DIGITIZATION PROJECT: ISSUES FOR STATE AGENCY PUBLICATIONS WITH FOLDED MAPS

Carol La Russa, University of California, Davis
Librarian for Environmental Engineering, Geology and Atmospheric Sciences. Physical Sciences & Engineering Library, University of California, Davis

Karen Andrews, University of California, Davis
Head, Physical Sciences & Engineering Library, University of California, Davis

Managing a Digitization Project: Issues for State Agency Publications with Folded Maps

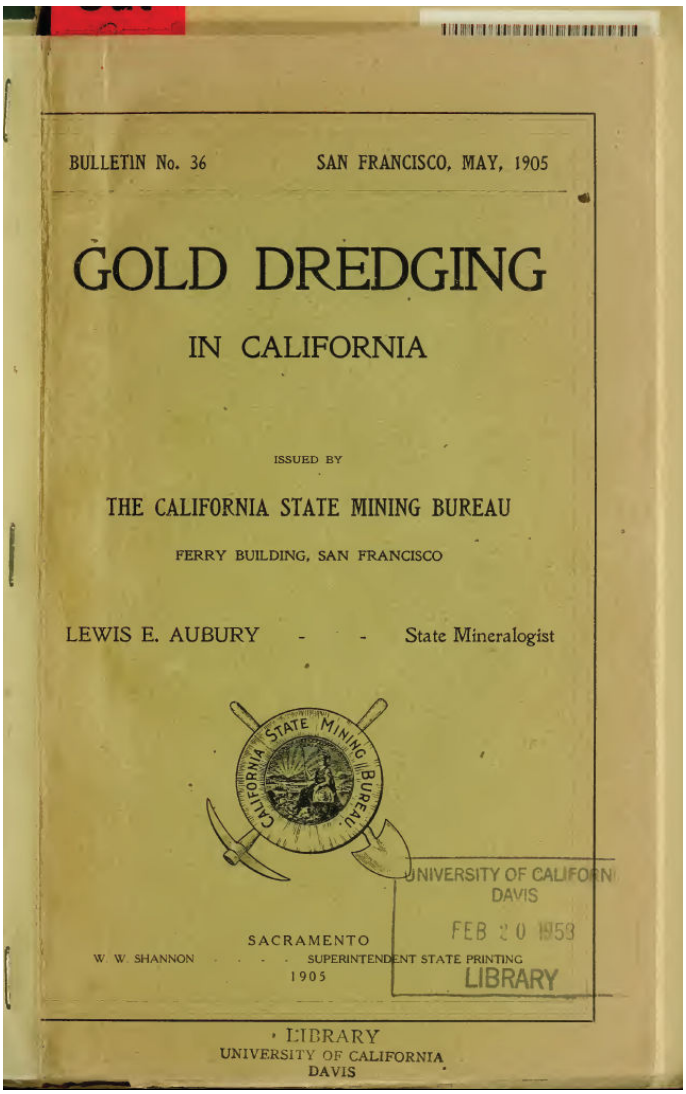
Abstract

The authors present a case study of management issues encountered when an external organization with grant money and facilities to do digitization accepted the proposal to scan and host a collection of California Division of Mines & Geology Bulletins and California Department of Water Resources Bulletins. The Bulletins are very important historically for the rich information and maps they contain about California mines and minerals, geology, water resources, highways, and weather modification projects. They were difficult to digitize because the texts were accompanied by large, folded maps, many in color. The library was responsible for gathering the materials and providing bibliographic information for each item in the series. The Water Resources Bulletins were a complex set to identify bibliographically because the title spawned many subseries over time, some with imperfect numbering and others with uncertain relationship to the original series. This project involved staff from several departments across the library and coordination with two larger organizational entities. The authors describe challenges in managing all the activities of such a project under a tight timeline and make recommendations for efficient procedures. They also identify the need to formulate better metadata extraction algorithms for use with items that are part of a series, such as government documents and technical reports, in order for these materials to be discovered by researchers. The digitized images were linked in the university online catalog, made publicly available on the Internet Archive website, and, in collaboration with the California Digital Library, deposited for preservation in the Hathi Trust. The methods developed for this pilot project will serve as a model for future collaborative endeavors involving preparations for digitizing bibliographically complex sets.

Introduction

This paper describes the management of a project to digitize and make accessible the engineering and science content of bulletins from two major state agencies. The project was complex because of two inherent challenges. The texts were accompanied by foldout illustrations and tables, and large, folded maps, many in color. The digitizing agency, the Internet Archive (IA), had not handled foldouts of this size before due to lack of appropriate scanning equipment. The second challenge was the difficulty in correctly identifying all the parts of a series due to agency changes and revised editions over time. A third challenge related to the tight timeframe during which the Internet Archive staff had funding from the Kahle/Austin Foundation and Omidyar Network to complete the project.

The Bulletin of the California Division of Mines and Geology collection is a series of publications issued from 1888 to 2001 by the Division and its predecessor agencies. The Bulletins include many California geologic maps and guidebooks and also contain information on the mines, minerals, mining law, and geologic history of California.



BULLETIN No. 36 SAN FRANCISCO, MAY, 1905

GOLD DREDGING

IN CALIFORNIA

ISSUED BY

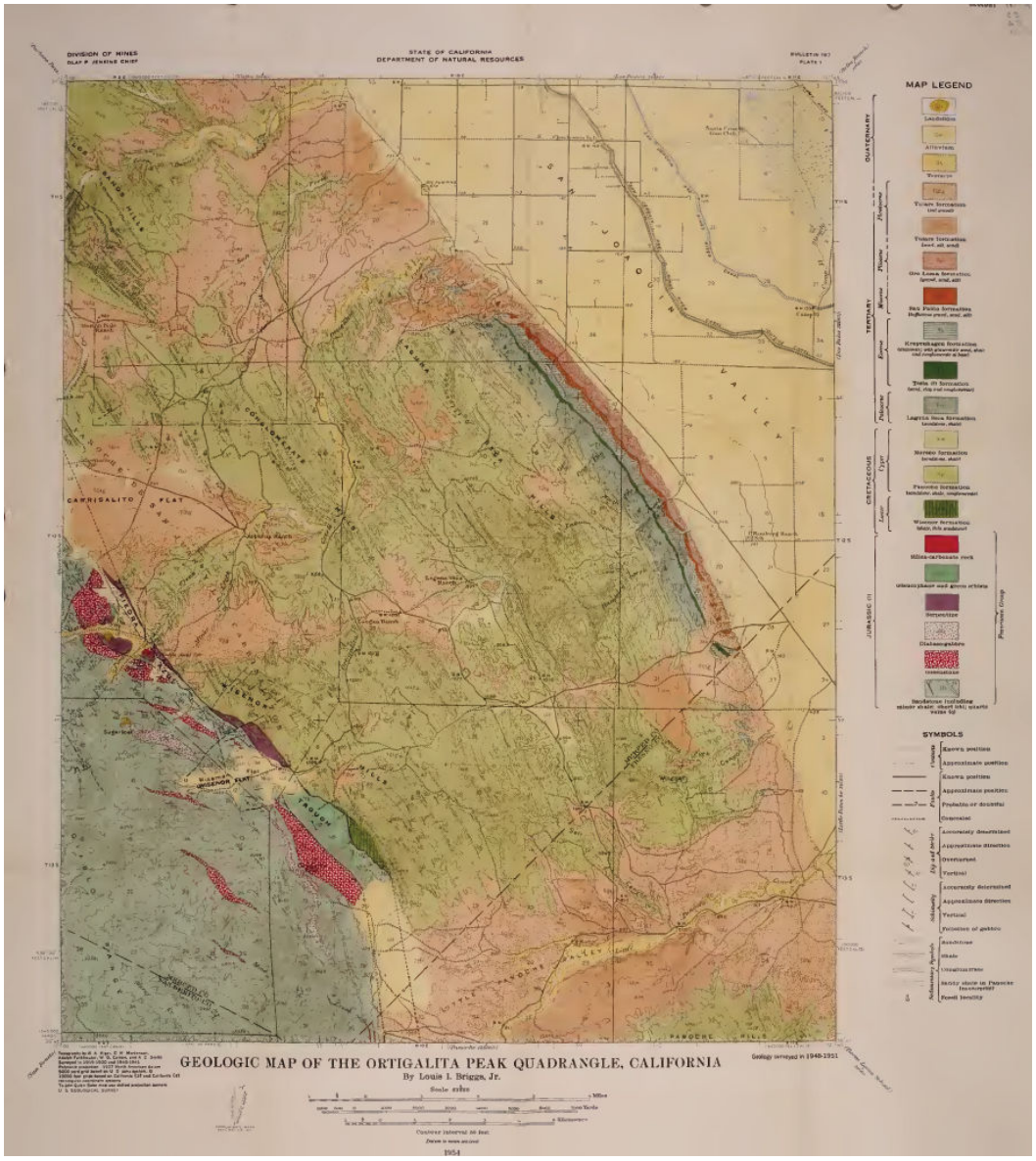
THE CALIFORNIA STATE MINING BUREAU

FERRY BUILDING, SAN FRANCISCO

LEWIS E. AUBURY - - State Mineralogist

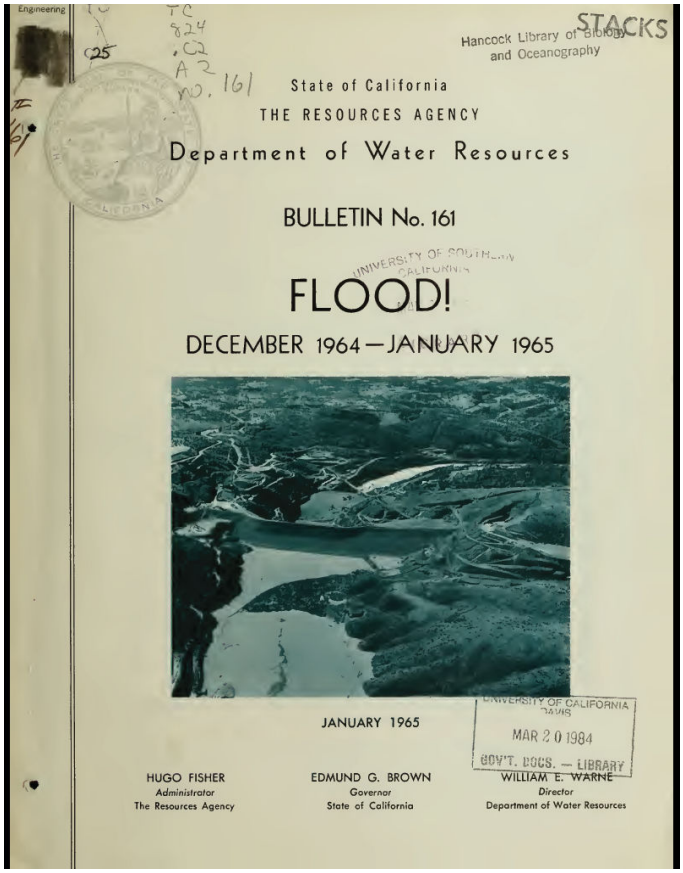
SACRAMENTO
W. W. SHANNON - - - SUPERINTENDENT STATE PRINTING
1905

A Bulletin published in 1905 of the Calif. State Mining Bureau (later Division of Mines and Geology)



Geologic map published in the Bulletin of the California Division of Mines and Geology.

The Bulletin of the California Department of Water Resources is a collection of volumes published from 1922 to 2004 by the Department and its predecessor agencies. The Bulletins have information on all aspects of water in California including water law, water projects, groundwater, water quality, flooding, water conditions, and weather modification efforts.



Bulletin published in 1965 by the California Department of Water Resources.

Timeline: In late October 2008 the Internet Archive notified the University of California of the sudden availability of grant funding for a digitization project. The University had on hand several proposals to consider. The project needed to get started as soon as possible to meet IA's ambitious goal of finishing by January 2009. Working in collaboration with the California Digital Library (CDL), the University of California, Davis (UCD) General Library agreed to move forward with its project after learning that the Internet Archive had recently acquired the equipment and capability to digitize large maps. The work of gathering the materials began in November 2008. A test batch was delivered to Internet Archive in early December, and the rest arrived at the IA site in San Francisco in mid-December. The digitization was completed by IA staff in June 2009, later than anticipated, partly due to scanning difficulties with some of the material. The books were returned in July, and in fall 2009 a quality control review was undertaken by the library. Some complications were identified and resolved.

Background

Two kinds of digitization projects are commonly seen. The first is the mass digitization project typified by the Google Books project, where minimal effort is made to ensure quality control, completeness and adequate cataloging (metadata) for individual works.[1] Mass digitization projects to date have focused on journal runs, books, and collections of

images such as for art history, photographs, or movies. Oversize maps and charts, and government documents, are typically omitted in a Google scan project.[2] The second type of digitization project is a special collection approach where a specialized work or collection is digitized with great care and attention to detail.[3]

The University of California, Davis (UCD) Library hoped with this project to find a path between these two extremes. In addition to the goal of getting these publications digitized, the CDL was also examining the process the UCD Library developed as a prototype for other campuses to use for future collaborative digitization efforts.

Goals: In early negotiations with the Internet Archive, UCD librarians established several goals for the project.

1. The titles of individual bulletins in these series should be findable.

Rationale: The librarians faced a choice in how the Bulletin series would display in the Internet Archive. The individual volume titles held valuable clues to the content, such as coverage of a specific geographic region, water feature, mountain area or river. If the Mines & Geology Bulletin were treated as a serial, then all the volumes would appear as one long, large image under that title. Researchers would have to scroll down through every issue to get to the desired volume, almost as if the digital image were a roll of microfilm. The librarians opted to have each volume appear separately. The searchable metadata should show author, title, bulletin series name and volume number.

2. The complete volume had to be digitized, especially the large maps and fold-out charts included in many of these bulletins.

Caveat: Unfortunately the Internet Archive would only be able to produce images of the maps. It did not have the ability to add the geospatial data required for geographic information system applications. Some maps were so large that they had to be digitized in slightly overlapping quadrants.

Methods for the Project

Several conference calls involving branch librarians, technical services staff, Internet Archive staff, and the California Digital Library were held to establish that it would be possible to digitize large maps and charts, and that the Archive could use library-generated cataloging for individual bulletins within these series. UCD librarians decided to provide complete cataloging (author, title, subject headings) for each volume, rather than just use the name of the Bulletin series with a volume number. The additional information, which often included geographic place names, would enhance searching capabilities and make the collection far more useful for researchers.

The same procedures were used for processing both series:

Gathering the materials: The smaller geology and mines series was done first to test the process. Volumes from each series were pulled from library shelves by branch library staff. The volumes already had barcode numbers that were used for circulation purposes. Staff sorted out the duplicates and kept the best copies for digitization. Missing volumes were noted and in several cases volumes were borrowed from another library or campus.

Obtaining catalog records for each book: Staff at the centralized cataloging department downloaded from OCLC the individual author/title records (analytics) for all volumes in the two series and added these records to the UCD local catalog. This resulted in many more records being downloaded than would be used, so codes were added to the records to enable removing unneeded records later. Branch Library Assistants matched volumes with the newly added records in our catalog by searching each volume individually. Volumes that had no matching records were turned over to the subject specialist librarian. A number of records were missing because the name of the agencies had varied and not all versions had been findable using the original search words. The librarian searched OCLC for the individual titles and noted the record numbers if found. A cataloging assistant then added these additional records to the local catalog.

Tracking the volumes: The systems department staff created a list of bar codes and call numbers downloaded from the library circulation system. This was used to create a spreadsheet for each series called a "picklist." The author, title, publication year, volume, and the local catalog system record number were added to the picklist. A shortened version is shown here, without the added columns to track when the volumes were sent out for digitization, scanner comments on any problems encountered, and return date.

Source Local Code	Barcode	Call Number	Vol	Author	Title	Pub. Date
ucd:DVXL002990953	31175010040833	TC824.C2 A2 no.45	no.45	Eckis, Rollin.	South Coastal Basin investigation : geology and ground water storage capacity of valley fill 1934/ [by Rollin Eckis].	1934
ucd:DVXL002990951	31175006446234	TC824.C2 A2 no.46	no.46	Conkling, Harold, b. 1882.	Ventura County investigation 1933/ [by Harold Conkling].	1934
ucd:DVXL002990950	31175006616935	TC824.C2 A2 no.46A	no.46A	Conkling, Harold.	Ventura county investigation. Basic data …	1934

Transportation: Branch library staff prepared the volumes for shipment to the IA digitization facility.

The process of gathering volumes, obtaining catalog records for each one, creating a tracking spreadsheet, and transporting the volumes to the scanning site, relied on the cooperation of many people from several departments. The work required one month for the geology series (202 volumes) and six weeks for the water resources series (780 volumes). Most of the work was done by a Library Assistant IV, a Library Assistant III (working his regular night and weekend shifts) plus numerous ad hoc contributions by Library Assistant Vs from other units on call for their expertise as needed. Two librarians coordinated the process: a subject expert, who resolved complex bibliographic problems and acquired loans of material for missing volumes, and a department head who coordinated communications and decision-making with the California Digital Library and the Internet Archive staff, as well as with other department heads in the library to secure assistance and staffing support.

Complications and Issues encountered

The process of gathering the right material and finding the best bibliographic record for each piece turned out to be fairly complicated. Over two-hundred and fifty emails were exchanged by the participants in a span of two months. Three- and four-way conference calls with staff from Internet Archive, California Digital Library, and several UC Davis Library units took place to coordinate aspects of the overall project. Some of the issues included:

Difficulty of Time Constraints: The engineering librarians were incredibly lucky to have the cooperation of so many library staff members who put aside their normal tasks for this project. The project could not have been done without them, but one could not ask as much again any time soon. The project was just barely doable in the time given. Compromises had to be made in order for the work to be completed on time.

Staffing Issues: At the Library, this project entailed utilizing the time and talents of staff housed in two different buildings and from several units.

a.) Workflow: Early on, the project managers faced some important issues that affected workflow. For downloading OCLC records, only cataloging staff had the authorization to work in certain sectors of the online catalog and OCLC. The question arose as to whether catalog staff should come to the library where the books were, or send the books to the building where catalog staff had their desk and workspace. Neither option was ideal. Catalog staff had a different high priority project going on during this time, and could not devote all the time needed for this project. The managers agreed to have library systems staff batch download all possible OCLC records for consideration. This resulted in local library staff less familiar with cataloging practices doing the record review. With book in hand, they spent time filtering through many possibilities to match the catalog record with the actual item. With more time and if circumstances had permitted, the most efficient way would have been for experienced catalog staff to locate the exact OCLC

record with book in hand and download only that record. This would eliminate record cleanup chores in the catalog afterward.

b.) Staff turnover and work priorities: Within two weeks of starting the project, a key systems staff person retired. This meant that the same pattern of downloading OCLC records could not be followed until the replacement person could learn the necessary programming for the second Bulletin series, which was more complicated than the first. Creative cataloging staff successfully devised a way to get the needed information without systems intervention, and although it took more time, the project stayed on pace. Meanwhile, catalog staff had another significant project to work on at the same time, of very high priority. These kinds of conflicting demands are the norm, not the exception. Fortunately, very dedicated, brilliant technical services colleagues worked very hard to meet all requests while also meeting production deadlines in their regular sphere of duties. The exceptional knowledge and guidance of a library assistant skilled in government documents cataloging was also beneficial. She came to the unit several times to help identify the best OCLC record to choose, and when to recommend original cataloging. Resorting to original cataloging was avoided if possible, because of the added staff cost.

c.) Staff capabilities and workload: Early on, the decision was made to involve only a few key staff rather than delegate to a team of the entire engineering library staff, to help maintain control of the project. This allowed some staff to focus totally on normal daily activities. Enough people had to be involved to share parts of the processing because the highly detailed nature of the work necessitated taking enough breaks to maintain accuracy. The engineering library exists as a separate branch that employs Night/Weekend Lead Assistants. These highly skilled library assistants have full responsibility for the physical facilities and safety of library users during non-weekday hours. However, beyond their nightly circulation, document delivery and ILL assignments, the night/weekend staff typically has spare time. This ensured a steady flow of book processing. One assistant developed and tested several processes to determine a best practice. He created, sorted, and merged columns from various spreadsheets with the picklists, and otherwise was an invaluable team member. The daytime library operations manager did yeoman work on this project. She personally verified many OCLC records, problem solved difficult titles, coordinated with staff from the main technical services unit, trained other library staff as needed, and kept an overview of the project that kept it on track. Without this kind of dedicated and intelligent staffing, the project leaders would not have dared to undertake a project with such a fast turnaround time and with the amount of processing and record validation that had to be completed before the books could be transported to Internet Archive. It greatly helped to have staff under direct supervision of the engineering librarians. It was critical to have staff who could devote significant time to the project without sacrificing regular work activities. And it was highly useful to have motivated, intelligent staff who communicated well and developed efficient processes with minimal direction.

Materials condition: The books to be digitized were reviewed for physical condition and ability to withstand the digitization process. Fortunately, there were duplicate copies of

many books. Staff was able to compare and choose the best one for digitization. The library had the expertise of a skilled preservation conservationist on staff at the central library. She volunteered time for the project and came to the branch library to review damaged books to recommend repairs and determine the safety in letting damaged material leave the library for scanning. When bindings had to be removed, the conservationist came and taught an engineering library assistant the best method for this task.

A tour of Internet Archive facilities developed confidence that the staff there treated all library materials with great care. Their modern scanning equipment does not damage books and staff are very careful with torn or worn pages.

Bibliographic Complexity: The water resources bulletin series contains many sub-series. Some volumes within these sub-series were cataloged individually with their own specific titles in OCLC; while others were cataloged only by the name of the sub-series. An example is this title: California High Water with volumes issued from 1964-1976. When individual title records were not available for every single volume in the sub-series, the serials cataloger and subject expert librarian made a decision to use the generic sub-series serials record for all volumes of the sub-series, adding only the year to distinguish among volumes. Some volumes of sub-series were bound individually and some were bound together, as shown by the barcodes in the example picklist.

Example of a Cataloged Sub-Series:

Author	California. Dept. of Water Resources.
Title	California high water.
Published	Sacramento, California, the Resources Agency, Dept. of Water Resources.
Description	13 v. ill., maps (part fold.) 28 cm.
Publication History	1962/63-1974/75.
Record format	<Serial>
	GV Govt Publication
	SE Serial
Check Availability	All items
Call no.	Phy Sci Engr Library TC824.C2 A2 no.69- etc.
Current frequency	Annual
Link Note	Merged with:California. Dept. of Water Resources. Water conditions in [California]. Summary report, ISSN 0163-6456, to form: California. Dept. of Water Resources. Water conditions and flood events in [California], ISSN 0163-6464.
Subject	Floods -- California -- Periodicals.
	Storms -- California -- Periodicals.

	Stream measurements -- California -- Periodicals.
Series Add.Entry	(Bulletin (California. Dept. of Water Resources))
Merged with	Water conditions in California. Summary report 0163-6456
	Water conditions and flood events in California 0163-6464 (DLC) 78648285
ISSN	0526-9873 1

Section of Picklist for Sub-Series: California High Water:

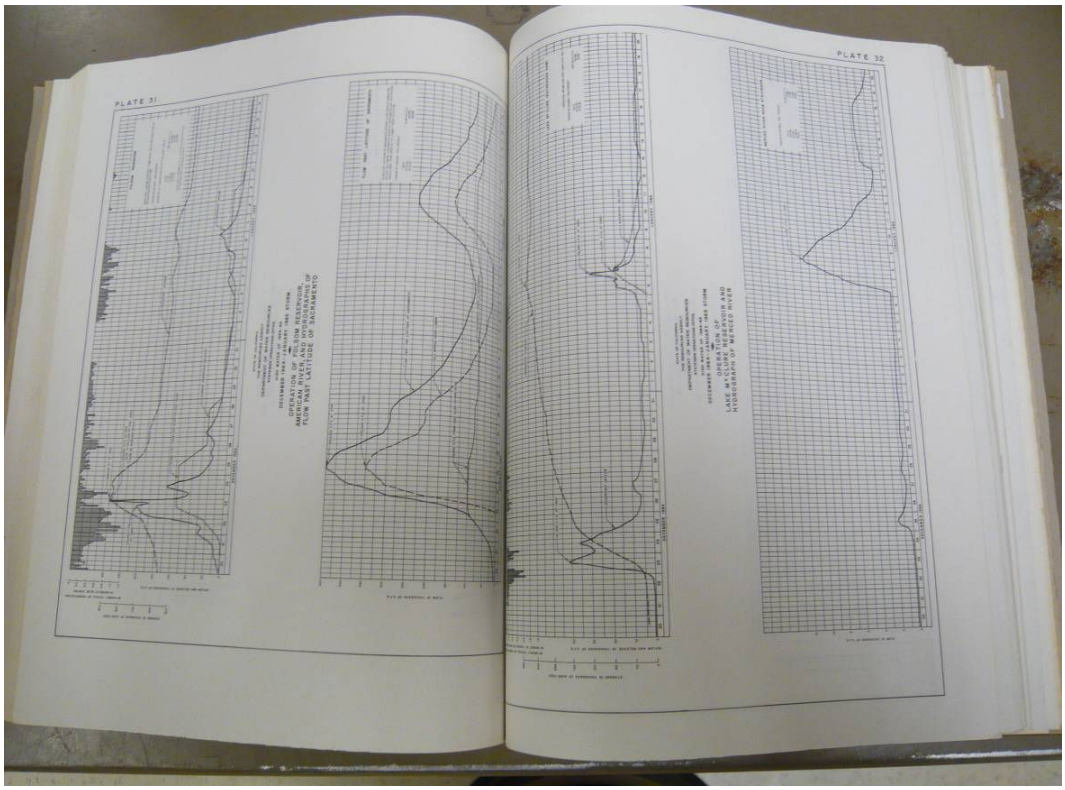
Source Local Code	Barcode	Call Number	Vol	Title
ucd:DVXL003000003	31175008236153	TC824.C2 A2 no.69:63	no.69:63	California high water : 1962-1963
ucd:DVXL003000003	31175020376987	TC824.C2 A2 no.69:64	no.69:64	California high water : 1963-1964
ucd:DVXL003000003	31175020376995	TC824.C2 A2 no.69-65	no.69-65	California high water : 1964-1965
ucd:DVXL003000003	31175020377001	TC824.C2 A2 no.69-66	no.69-66	California high water : 1965-1966
ucd:DVXL003000003	31175020377019	TC824.C2 A2 no.69-67	no.69-67	California high water : 1966-1967
ucd:DVXL003000003	31175004968437	TC824.C2 A2 no.69-68	no.69-68	California high water : 1967-1968
ucd:DVXL003000003	31175004968437	TC824.C2 A2 no.69-69	no.69-69	California high water : 1968-1969
ucd:DVXL003000003	31175004968437	TC824.C2 A2 no.69-70	no.69-70	California high water : 1969-1970

The volumes from another small series from the same water agency had mistakenly been cataloged as part of the larger series. These volumes were removed from the project. Librarians also had to make decisions regarding varying editions such as "draft" and "office version" that were found in our stacks. All of them were included for historical purposes.

Errors: Inevitably in such a rushed project there are errors. Errors were found in our picklists such as catalog record numbers not matching volume titles and at least one volume was inadvertently omitted from the list. Most errors were caught by assigning a staff person to double check the list against the books before the material was sent out.

Coordination between Archive and Digital Library: Both organizations needed a picklist but each had different specifications for that list.

Tight binding: Permission was given to Internet Archive staff to burst a book's binding when needed to obtain a good scan. However, this greatly slowed the workflow of the production scanning team. IA did not have staff to allocate to unbinding books. As a result, some books were returned to the library with tags noting that the book had not been scanned due to tight binding. This issue was later resolved by having library staff remove the tight bindings, and the books were brought back to IA for scanning.



Example of tight binding from a Bulletin of the California Department of Water Resources.

Large maps: When the UCD Library agreed to do the project, librarians were insistent that all maps be scanned or digitally photographed. This did not happen consistently. With a changeover in IA staff, some volumes from the second set (Water Resources) came back tagged "unscanned" due to oversize maps. This issue was later resolved by IA, which digitized the largest maps by dividing them into overlapping quarters.

Large map from a Bulletin of the California Department of Water Resources.

Large Map from a Bulletin of the California Division of Mines and Geology.

Transport of Materials: Arranging transport of materials was difficult to organize while keeping costs down. The library mandate was to undertake this project without incurring any actual expenses, other than in-kind contribution of staff time and expertise. However, to expedite the processing and to meet the tight deadlines, library staff personally transported the first trial batch of materials to the Internet Archive scanning quarters. One of the large maps, which was rolled into a four foot long tube, was delivered in this

manner to ensure its safety. The trip proved fruitful because the library staff were given a tour of the facilities, and saw the workflow and the equipment to be used.

Communications: Decisions made between administrators at the Archive and library staff may not always have been conveyed to the IA staff doing the scans. Volumes were skipped because they were bound too tightly even though permission was given to break the binding. There may have also been some miscommunications with scanning staff about dealing with large maps. Personnel changes at the Archive during the project also hampered communications and may have slowed the project.

Metadata: The Archive harvested their metadata from the UCD Library cataloging record. Unfortunately, from the UCD librarians' point of view, the standard algorithm IA used ignored parts of the catalog record that librarians and many researchers consider important, i.e., the title of the series with the volume number. The Archive metadata system was set up for a free-standing monograph; it only searches by author and title. This did not work well for monographs published in a numbered series. Since UCD Librarians planned to link the URLs of scanned volumes to UCD cataloging records and to OCLC records, the volumes would be accessible by series title and volume number there. However, in early 2010, the Internet Archive added the capability to upload fuller bibliographic data that will overlay the earlier incomplete data. UCD Librarians will undertake this upgrade. It will result in better search and retrieval at the Internet Archive website, matching the search capability by series title and volume number in the UCD Library catalog.

Project Wrap-up:

All volumes were returned to the library. Twenty-nine volumes were skipped because of "tight binding." And eighteen volumes with large maps were not scanned but no reason appeared on the slip in the book. Since the goal was to get as complete a set as possible, the IA staff was contacted to determine what could be done. The Internet Archive scanning facility supervisor was willing to try again to scan them, and library staff agreed to remove the tight bindings to facilitate IA's workflow. The large maps that were beyond the capability of the scanning equipment when the project began could now be done. IA had acquired a newer, more powerful overhead scanner that could divide the large map into quadrants while maintaining adequate resolution. The small shipment of volumes went back to the Archive, and library staff were grateful that IA graciously agreed to squeeze it into their current production flow so that the collection would be more complete.

For the books that were finished with scanning, the library undertook a quality control assessment. A library assistant developed a procedure for doing efficient quality assurance on the geology bulletins. So far the scans generally seem good. Occasionally a page needs rescannng, for example, when a small piece of paper, probably used as a bookmark, was left in the book and obscured information on the page. In consultation with IA staff, the practice they prefer to fix single page problems is to have library staff photocopy any problem pages. Then IA staff can scan the photocopy and replace the

faulty original image. Librarians inquired about the possibility of doing the scanning themselves and sending IA the image, but IA staff preferred the high resolution photocopy. They explained that their scanned image has corresponding metadata attached to it that they want to have, such as camera settings.

At the Internet Archive web site, several viewing options are offered. Of the current six options, the scanned image referred to as "Read Online" is done in very high resolution and meets JPEG 2000 standards. Viewing the image online is generally the best representation of the original print. The PDF versions and one called Deja Vu have compression applied to the image, so details are less crisp. This means that printouts of the PDF versions might not be acceptable substitutes for the original. One would have to save and try to print out the high resolution JP2 image for greatest accuracy.

The library catalog still needs updating. The Archive sent a list with the URL for each piece. Catalog staff at UC Davis will manually attach the URLs to the right cataloging records. Extra downloaded cataloging records, currently suppressed, that did not get used because they were not a precise match for a piece, will be removed from the catalog.

Conclusions and Recommendations:

1. Only attempt a project of rushed nature with the full support of all the personnel and departments that will need to be involved, since the project activities must have priority.

2. When dealing with government series published over long periods of time, expect bibliographic complexity and the need for librarians to make decisions about which bibliographic records to use for particular volumes, how to treat bibliographically a series within a series, which volumes are really part of the series (and not part of similarly titled series), how to deal with variant editions, etc.

3. If one cares about quality, one will need to do quality assurance.

4. The Internet Archive was advised to consider revising its use of library metadata. Its staff learned more about serials cataloging with analytics. The metadata extraction algorithm should be programmed to retain critical cataloging data like series information to ensure proper searching and retrieval of all the materials. The Library will try IA’s newly recommended software fix, which could be used up front to eliminate this problem in any future projects.

5. Map images created for this project are not really a substitute for the originals. Large maps were imaged by digitally photographing smaller sections. An overall photograph was also supposed to be taken. The resulting images are difficult to use because it is impossible to get both high resolution and continuity of the image at the same time.

6. **KEY FINDING** and **ACTION NEEDED**:

In the engineering literature, books are frequently issued as part of a series. It is crucial to locating materials to be able to search by series name and volume number, e.g. ASTM Special Technical Publication 345, or Symposium Series 102, as well as by the individual author and title of the volume. Technical reports form another very large category of material with these features. As people begin to rely more on digitized versions of books, and as retrospective collections are digitized, librarians must advocate for retention of all these search and retrieval options.

RECOMMENDATION: Metadata extraction algorithms need to increase in sophistication to take into account these added bibliographic elements of series title and volume number in addition to author, title, and subject. Standards for the bibliographic fields that metadata pull from need to include the series elements simultaneously. Current practices seem to limit the extraction to either the series title or the individual piece title, and not both concurrently.

Librarians and library organizations, working closely with mass digitization partners such as Google or Internet Archive, can educate these entities on the importance of providing complete descriptive data elements in the extraction algorithms. Disciplines such as engineering, where series are ubiquitous, will then be on a level playing field when it comes to search and retrieval of technical information.

Bibliography

1. Coyle, K. (2006). Mass Digitization of Books. Journal of Academic Librarianship, 32 (6): p. 643.

2. Coyle, K. (2006). Mass Digitization of Books. Journal of Academic Librarianship, 32 (6): p. 644.

3. Kimball, R., Weimer, K. H., & Surratt, B. (2005). Digitizing the series Geologic Atlas of the United States (1894-1945); access and preservation of older geological literature using an institutional repository. Proceedings of the Geoscience Information Society, 36: p. 109. Results of project at: http://repository.tamu.edu/handle/1969.1/2490