AC 2009-925: ROUNDING UP THE COLLECTION: THE STORY OF TRAIL DIGITAL CONTENT COLLECTION

Patricia Kirkwood, University of Arkansas

Patricia is the Engineering and Mathematics Librarian at the University of Arkansas. A member of the Greater Western Library Alliance (GWLA) TRAIL project since 2006. Currently she is the chairperson of the Collections Group.

Michael Culbertson, Colorado State University

Mike is the Engineering College Liaison Librarian at Colorado State University's Morgan Library. its implications for libraries." with Allison Mike is currently developing a study to look at how diverse populations use virtual reference services and developing a project to digitize the Colorado State University Civil Engineering Reports. He has been a member of the GWLA TRAIL project since 2008.

Esther Crawford, Rice University

Esther is the Head of Kelley Center for Government Information and Micro forms. She has recently added civil engineering subject specialist to her duties at Rice University and has been a member of the GWLA TRAIL project since 2008.

ROUNDING UP THE COLLECTION -- THE STORY OF TRAIL DIGITAL CONTENT COLLECTION

Abstract

Recognizing the importance of technical reports and the challenges to access presented by this material, GWLA (Greater Western Library Alliance), in collaboration with the Center for Research Libraries, developed a project to identify, digitize, archive, and provide persistent and unrestricted access to federal technical reports issued prior to 1976. Significant progress has been made since the project began in 2005. The first collection digitized, the "NBS monograph series", was relatively easy and, along with a small subset Atomic Energy Commission series, is available at the TRAIL (Technical Report Archive and Image Library). The Bulletin series of the US Bureau of Mines found its way to the group, but provided challenges in digitization technique due to maps, foldouts, and other illustrations. Defining what is a federal technical report; determining what agencies (existing or defunct) are appropriate for inclusion; and finding paper copies of the reports of interest has been a more complicated task than expected. This paper will describe the efforts of the taskforce of engineering and government documents librarians to define, collect, and digitize one of the largest bodies of grey literature in science and engineering.

Introduction

For decades librarians and researchers in science and engineering have discussed how to provide greater access and bibliographic control to the "grey literature": unpublished reports, pre-prints and similar documents. A significant portion of the grey literature consists of technical reports commissioned by the federal government. Technical reports are a means of communicating the progress of research in fields of technology and science. Federal technical reports have an additional attribute; they are produced using public funds, they are meant to be widely accessible. These reports are highly detailed and contain valuable information serving specialized audiences of researchers. While availability and access to the more recent (1994-current) technical report literature has greatly improved with delivery via the Internet, legacy technical report documents remain elusive to researchers. Many large research libraries across the country have sizeable collections of federally funded technical research reports, frequently a million or more reports ranging from several pages to several hundred pages. However, these collections, particularly legacy collections, are often difficult to identify and locate for several reasons:

- Dissemination to libraries has occurred through a variety of agencies and organizations over many years; dissemination was often based on institution profiles creating incomplete sets of reports.
- There is limited bibliographic access and control in science, technology and medicine indexing sources and often more than one index must be consulted to retrieve a report.

- Collections are usually available in some combination of print and/or microfiche and are difficult to access without known citations and mediation to navigate through the various collections and organization strategies.

- Depending on institution preferences and availability some collections of reports were produced and distributed using poor quality media resulting in disintegrating and unusable pieces of collections at many institutions.

- Most library catalogs and bibliographic utilities only include access points at a broad series level and even fewer records for individual technical reports in their online cataloging systems making it difficult for users to determine the availability of reports in local library collections.

- Most legacy reports are not accessible in electronic format and are difficult to acquire via Interlibrary Loan, compounding the difficulties experienced by end users in accessing the research reports.[1]

In 2005 librarians from libraries comprising the Greater Western Library Alliance (GWLA), recognizing the serious preservation and accessibility issues surrounding federal technical reports, proposed that GWLA provide seed money for an effort designed to digitize these documents and place them in an open source repository on the Web. The GWLA directors agreed and the Technical Report Archive and Image Library (TRAIL) project was established. Led by the University of Arizona and in collaboration with the Center for Research Libraries TRAIL has been charged by the GWLA directors with identifying, digitizing, archiving and providing persistent and unrestricted access to federal technical reports issued prior to 1976. A timeline covering the history of the project can be found at: http://sites.google.com/a/gwla.org/trail/about-the-trail-project. In addition to digitizing the reports TRAIL intends to leave a print archive through either: a) creating complete print runs of each series being digitized, or b) identifying and supplementing/completing existing print collections that will serve as print repository copies of the digitized content.

Literature Review

The 2006 conference, *Scholarship and Libraries in Transition: a Dialogue about the Impacts of Mass Digitization Projects*, gave an imperative for TRAIL and other efforts to digitize unique documents by stating, “In fact, evidence is mounting that any material that is not available in digital form does not get used.”[2] A select number of articles discuss the digitization of government documents in general. Hartman (2001) describes efforts by the ALA Government Documents Round Table (GODORT), beginning in 2000, to coordinate digitization of government documents and to avoid the possibility that "... multiple institutions would digitize the same publication, unnecessarily duplicating costs and efforts."[3] Sleeman argues that digitization of federal documents is more about access than preservation. He also points out successful projects that have digitized federal documents in specific series, such as the effort by the University of North Texas to load the publications of the Advisory Commission on Intergovernmental Relations.[4] Hartman (2000) provides a framework for preserving documents from federal agencies that no longer exist.[5] Of particular interest to the TRAIL effort, Anderson, et al., describe a project to digitize the legacy publications of the Fermi National Accelerator Laboratory. They report that even in the early days of the Web their collection received an average of 40,000 hits a month.[6]

Defining a collection

The first part of our work was to define what we intended to digitize. The idea was that we would work with legacy federal technical reports that were in most danger of becoming lost. The trick was to set up parameters for that population. One suggested parameter was that the reports should have appeared before 1976. One of the rationales for this was that in 1976 government documents began to receive MARC records through the MARCHIVE project. Thus limited cataloging information existed for federal technical reports that appeared before the mid-seventies. We reasoned that this factor might lead to a greater likelihood that the reports would be discarded. We were also concerned about the physical condition of technical reports that appeared before 1976. Many were printed on poor quality paper with high acidic content. The participants in TRAIL also decided to make reports from “dead agencies” a priority. As we learned from the documents librarians on the project the federal government has gone through frequent reorganizations. Reports from agencies that no longer exist may be more likely to be discarded, for various reasons. Another factor considered is the copyright status of the items being digitized. Changing "ownership" status of government publications, including those done "for" the agency by an outside organization meant working with materials published prior to 1976 is cleaner and less likely to run into copyright related issues. This is especially critical as we move into the mass digitization stream. A strong agreement by the group was to avoid digitizing anything that has not be cleared of security status that would make the information sensitive. Lastly, we needed to establish a scale by which the resources of the project would be used most effectively. As we need to have a content stream of 18 shipping boxes per month (1/2 pallet) for the digitization stream, it was important that we are actively processing at least 2 collections at a time of more than 5000 items. These items could be the result of multiple series (such as the various smaller U.S. Bureau of Mines materials) or large series like the *Reports of Investigations* (also from the U.S. Bureau of Mines). With a document stream guaranteed, we could work with smaller collections (such as the Saline Waters) as they were identified, reviewed, and collected. For this reason, the Bureau of Mines series as well as the National Bureau of Standards series are currently our "large" collections.

Procuring Collections: Happenstance or planning?

As we have moved into production mode, the quantity of materials needed to have a successful materials stream has expanded dramatically. Though we have been working diligently to identify collections for digitization, sometimes we must simply accept series as we become aware of materials that fit our parameters and are readily available. The Bureau of Mines Bulletins as well as the other Bureau of Mines materials are examples of this phenomena. In a survey of interested librarians the members of TRAIL did at the beginning of the project (see *Ranked report sets for digitization project*) the Bureau of Mines reports showed up in 8th place. However, in 2006 librarians involved in TRAIL were approached by a librarian at a library in Alaska that was closing and had many Bureau of Mines materials available for our project. One of the nodes in TRAIL agreed to short term housing for these materials while they were evaluated for the project. GWLA, incidentally, paid all the shipping costs as they have done with most of the TRAIL collections. When we were ready for our third series (we had done National Bureau of Standards Monographs and a sampling of Atomic Energy Commission reports as a trial) and wanted to test the Node/Central process as a method of preparing materials, we decided to use the Bureau of Mines Bulletins that had been housed waiting digitization. The series was over

80% complete. It also had unique characteristics – large fold outs as well as a few color plates. It was also accessible and allowed us to develop node procedures easily. This series was easily completed and shipped for processing.

Unfortunately, the other series offered to the project were not as complete. They are fascinating, useful resources and are being digitized. However, because they were not complete, processing these series became (and remains) a serious issue. This has caused roadblocks as we have moved into production mode and are concentrating on a limited number of series at a given time. We have learned a lesson. Just because materials are available right now, doesn't mean we can use them now. However, we will continue to pursue the discard lists to gather material as appropriate. And when we can (a good example is a set of NUREG's TRAIL recently accepted), we will house the material temporarily until decisions are made as to how to collect that collection.

On the other hand, planning has also been involved. Members of the project have utilized the *U.S. Government Manual*, *WorldCAT* and their own catalogs and holdings to come up with agencies and series that would be appropriate candidates for digitized collections in TRAIL. During Phase II of the Project we decided to focus on technical publications from agencies which no longer existed, often referred to as "dead" agencies. Using the *United States Government Manual, 2007-2008*, *Appendix B, Federal Executive Agencies Terminated, Transferred, or Changed in Name*, to produce the list of "dead" agencies, the government documents librarians in the Task Force researched each agency and identified a list of possible technical publication series published by the agencies. The list was further refined to include the following agencies as possible candidates for digitization:

- Atomic Energy Commission
- Bureau of Mines
- Bureau of Plant Industry
- Bureau of Plant Industry, Soils, and Agricultural Engineering
- Department of Medicine and Surgery
- Maritime Administration
- National Council on Marine Resources and Engineering Development
- National Park Service
- Office of Oil and Gas
- Federal Radiation Council
- Federal Radio Commission
- Bureau of Public Roads
- Office of Public Roads and Rural Engineering
- Rubber Producing Facilities Disposal Commission
- Rural Business and Cooperative Development Service
- Rural Electrification Administration
- Federal Council for Science and Technology
- Office of Scientific Research and Development
- Shipping Board
- Shipping Board Bureau
- Shipping Board Emergency Fleet Corporation
- Shipping Board Merchant Fleet Corporation

- Ships, Bureau of
- Soil Conservation Service
- Soil Erosion Service
- Bureau of Soils

These agencies were chosen, mainly, on the basis of the existence of appropriate technical reports series and the lack, at least to our knowledge, of digitization efforts covering these series.

Throughout the project finding a balance between planning what agencies and report series to include and scrambling to evaluate and determine whether to accept unexpected offers has proven to be a challenge. We have also discovered that other projects are digitizing reports of specific federal agencies (NACA documents are a good example) and we are trying not to duplicate their efforts.

Inventory Management

Once report series have been accepted as collections for digitization they are sent to TRAIL sites for processing. Inventory control in the course of such processing has become one of the more interesting features of this distributed collection effort. Initially the design of the TRAIL project allowed for a given library, defined as a node, to assemble two copies of the collection; send one of the copies to the central processing unit for digitizing and addition of metadata; then make the necessary arrangements to archive the second copy. At the central processing unit student assistants would work under the supervision of an engineering librarian and a cataloging librarian. The actual scanning is done by Google through the Google Books project; special handling (large foldout especially) scanning goes to a vendor. All scanning is done at a level to be compliant with GPO requirements, thus helping insure an excellent image for display and download.

As we worked with collections we learned several things that have changed the process. First, waiting for collections to be assembled in their entirety takes more time than can always be allowed. With the move towards a more robust document stream, the simple number of items needed to keep the operation running at a level our partners require has meant that some of the materials bypass the node and are sent directly to the central processing unit. The advantages of this are twofold – reduced shipping costs and quicker turn around. The disadvantage is the loss of inventory control. Until the central processing unit has created the metadata and placed a received inventory list, the node personnel don't know if the donated materials meet the specifications for the project or if all the materials expected have indeed arrived. We have worked to minimize this issue with an Access™ database designed and hosted by the central processing unit. It is now being used to track shipments and keep records straight. The database relies on access to spreadsheets with inventories and other collection information that are available through a Google Documents site. Though this work flow has improved matters, the more distributed model of assembling a collection it has created will not be appropriate for every series or even every agency. It is likely that modified approaches will be used by different nodes depending on the expected access to materials and the librarian’s comfort with ambiguity during this process. Assembling a collection for digitization is much different than assembling a collection that will reside on your library’s shelf, especially when so many participants are

involved in the process. In short, what has evolved is that once a node site has assembled a collection, most of the processing is done by the central processing unit.

Meshing the talents of government documents and engineering librarians

The project has benefited from both the talents and perspectives of government documents and engineering librarians. The TRAIL members with government documents backgrounds have been able to provide insights into how publishing works in the federal government and, crucially in some cases, how it worked in the past. The documents librarians also know about the life history of government agencies, how agency names and responsibilities changed over time and which agencies were most likely to publish technical reports. The members of TRAIL who manage regional depository libraries have provided particularly important knowledge. They sometimes learn of depository libraries that are interested in offering up collections of technical reports. They also know whether such donations are possible under the regulations that govern depository library collections. The engineering librarians, on the other hand, have been able to make judgments as to how well collections of federal technical reports contribute to the literature of engineering and scientific knowledge. They have looked at the standards of scholarship in report collections. They have considered whether the mechanics of scanning might compromise the preservation of the reports. For example, reports with foldout maps, errata, supplements etc. have required special scanning treatment as have similar materials in pockets. The additional cost, extra time and, occasionally, technical difficulty involved in scanning these items has had to be weighed against the necessity of scanning a report in its entirety and in its proper context.

Issues surrounding the removal of items from depository libraries

Depository status has also had an effect on what technical report series could be included in the TRAIL project. More often than not, publications chosen for digitization and inclusion in TRAIL were distributed through the Federal Depository Library Program (FDLP) and, therefore, remain property of the United States Government.[7] Before these materials can be donated for digitization, a process known as "needs and offers" (http://www.fdlp.gov/collections/collection-maintenance/144-needs-and-offers-nao) must be followed. Regional depository libraries (those who receive everything offered through the FDLP) administer the needs and offers process for the selective depository libraries (libraries that receive only a selected portion of depository materials) in their respective regions.

End Product

Currently TRAIL employs a pilot Web site hosted by the University of Hawaii (http://digicoll.manoa.hawaii.edu/techreports/index.php). A more permanent Web site is in development and can be viewed at: http://sites.google.com/a/gwla.org/trail/Home. The site includes extensive information about the TRAIL project along with a searchable database of the reports that have been digitized. One feature of this site is the "Status of Collections" page (http://sites.google.com/a/gwla.org/trail/trail-project-status) which shows project participants and visitors the status of collections in process for inclusion into TRAIL.

Conclusion

To date TRAIL has digitized National Bureau of Standards reports and placed them on the project's pilot Web site. Atomic Energy Commission reports are currently being digitized. The Saline Water Transport and Use Office reports, the Bureau of Mines Reports of Investigations and the Bureau of Mines Information Circulars are in processing. We have learned that federal technical reports issued before 1976 are large in number and hard to pin down. We are actively seeking donations of reports that fit the parameters of the project. We are also cultivating contacts in federal agencies and in the documents librarian community.

Ranked report sets requested for digitization project:

Requests	Agency or subject area
18	DOE: Department of Energy
16	EPA: Environmental Protection Agency
14	NASA: National Aeronautics and Space Administration
11	USDA: United States Department of Agriculture, including 7 Forest service requests
10	NBS: National Bureau of Standards
9	USGS: United States Geological Service
8	AEC: Atomic Energy Commission
6	NACA: National Advisory Committee for Aeronautics
6	US Army research, technology, and engineering reports
6	USBM: United States Bureau of Mines
5	NTIS: National Technical Information Service
3	DOD: Department of Defense
3	NRC: Nuclear Regulatory Commission
2	‘Aerospace’
2	DOI: Department of Interior
2	National Labs; e.g. Argonne, Los Alamos, Oak Ridge, Sandia
2	NOAA: National Oceanic and Atmospheric Administration
1	BAE: Bureau of American Ethnology
1	BIA: Bureau of Indian Affairs
1	BLM: Bureau of Land Management
1	Defense Research Laboratory
1	DOT: Department of Transportation
1	‘Electrical Engineering’
1	‘Environmental Impact Statements (on Idaho)
1	Environmental issues and studies
1	‘Fire, Safety, automobiles’
1	Highway Research Record/Transportation Research Record
1	LA-UR’s: Los Alamos Unlimited Release
1	‘Military agency scientific and technical reports’
1	NIOSH: National Institute for Occupational Safety and Health
1	Ocean engineering
1	Optics and lasers
1	OSRD: Office of Scientific Research and Development
1	PHS: Public Health Service

1	Post WWII, including the BIOS, CIOS, FIAT & JCIA titles, as described on CRL's webpages
1	“Pre 1975 government documents—high demand, high use categories done first”
1	Simulation and Training.
1	IBP: International Biological Program 1964-1979 National Biological Information Infrastructure via USGS
1	The Yearbook of Agriculture
1	‘Water/ocean/atmosphere’

1 GWLA, Technical Report and Image Archives, *Project Rational.* http://www.gwla.org/ (accessed 4 February 2009)

2 U.S. National Commission on Libraries and Information Science. *Mass Digitization: Implications for Information Policy: Report from “Scholarship and Libraries in Transition: A Dialogue about the Impacts of Mass Digitization Projects”* Symposium held on March 10-11, 2006, University of Michigan, Ann Arbor MI.

3 Hartman, Cathy Nelson. "Ad hoc Committee on Digitization of Government Information." *DttP* 29, no. 3 (Fall 2001): 17-21.

4 Sleeman, William. "It's Not All on the Net: Identifying, Preserving and Protecting Rare and Unique Federal Documents." *Government Information Quarterly* 19 (2002): 87-97.

5 Hartman, Cathy Nelson. "Storage of Electronic Files of Federal agencies That Have Ceased Operation: a Partnership for Permanent Access." *Government Information Quarterly* 17, no. 3 (2000): 299-308.

6 Anderson, Elizabeth, et al. "Digitizing Lagacy Documents: a Knowledge-Base Preservation Project." *Illinois Libraries* 80, no. 4 (Fall 1998): 211-219.

7 *United States Code. 44, Section 1912.* Viewed on LexisNexis Academic http://www.lexisnexis.com/us/lnacademic/search/loadForm.do?formID=AC07STFedStCodesSrch&random=0.487266893811688 (accessed 4 February 2009).