CENDI

THE IMAGING OF LEGACY COLLECTIONS AMONG THE CENDI AGENCIES

Submitted by
The Legacy Collections Task Group
CENDI Information Exchange Working Group

Prepared by
Gail Hodge, CENDI Secretariat
Information International Associates, Inc.
Oak Ridge, Tennessee

August 1997

CENDI LEGACY COLLECTIONS TASK GROUP

Barbara Bauldock (DOE OSTI), Chair
Lowell Langford (DOE OSTI)
Charlene Luther (DOE OSTI)
Gopi Nair (DTIC)
Roland Ridgeway (NASA STI Program)
Lou Knecht (NLM)

Gail Hodge (CENDI Secretariat)

CENDI is an interagency cooperative organization composed of the scientific and technical information (STI) managers from the Departments of Commerce, Energy, Defense, Health and Human Services, Interior, and the National Aeronautics and Space Administration (NASA).

CENDI's mission is to help improve the productivity of Federal science- and technology-based programs through the development and management of effective scientific and technical information support systems. In fulfilling its mission, CENDI member agencies play an important role in helping to strengthen U.S. competitiveness and address science- and technology-based national priorities.

TABLE OF CONTENTS

EXECUTIVE SUMMARY

In 1996, the CENDI Information Exchange Working Group surveyed the state-of-the-practice of image scanning and optical character recognition (OCR) technologies among the CENDI agencies. This investigation resulted in three recommendations for follow-up tasks: 1) preparation of guidelines for the exchange of images among CENDI members, 2) investigation of the job descriptions and grades for employees performing scanning and OCR functions, and 3) an investigation of the CENDI agencies' plans for imaging legacy collections, those documents held by each STI program that predate the introduction of scanning into the agency. These recommendations were presented to the CENDI members at the August 1996 meeting. They were approved and included in the Annual Work Plan for 1997.

This report highlights the discussions held by the task group on the imaging of legacy collections. The group, composed of representatives from the National Library of Medicine (NLM), the National Aeronautics and Space Administration (NASA), the Department of Energy Office of Scientific and Technical Information (DOE OSTI), and the Defense Technical Information Center (DTIC), was chaired by Barbara Bauldock (DOE OSTI). The task group met on May 28, 1997 at the DOE Forrestal Building. Each agency presented information regarding the size of its legacy collection, the amount of that collection that originated with the other CENDI agencies, the status of imaging of the legacy collection and plans for the future.

This report presents the legacy collection information by agency and discussion question topic. It is followed by a brief description of imaging activities at other organizations that may include CENDI agency technical reports, an analysis of the information, and recommendations for further study. The initial findings and recommendations of the task group were presented at the CENDI Principals' Meeting on June 24, 1997. The recommendations are:

- Create an inventory of identifiable collections for which images have been created, are in progress, or are planned. These collections could be based on report number, center, project, etc. If they are available via the WWW, the URL and a contact name should be included in the inventory. An inventory would allow the CENDI agencies to contact those organizations for the sharing of images or to point to the collection via the WWW.

- Develop procedures for periodic sharing of images among the CENDI partners to supplement the images they are creating for their legacy collections. This could be the result of massive legacy imaging or simply the images of those documents that are ordered. Determine how many images must be received by each agency to make it cost/beneficial to develop this process. Determine the policy, schedule, and procedures that would make this a cost-beneficial activity.

- Share information on new scanning equipment evaluations and purchases. This could be done by developing technology-related pages on the agencies' WWW sites and linking from the CENDI report on OCR/Scanning to these pages.

- Investigate opportunities for sharing micro-film scanning equipment, for both evaluation and production purposes.

- Discuss virtual library concepts, approaches, and plans and to determine how agencies can help each other in the further development of these concepts.

1.0 INTRODUCTION

In 1996, the CENDI Information Exchange Working Group surveyed the state-of-the-practice of image scanning and optical character recognition (OCR) technologies among the CENDI agencies. This investigation resulted in three recommendations for follow-up tasks: 1) preparation of guidelines for the exchange of images among CENDI members, 2) investigation of the job descriptions and grades for employees performing scanning and OCR functions, and 3) an investigation of the CENDI agencies' plans for imaging legacy collections, those documents held by each STI program that predate the introduction of scanning into the agency. These recommendations were presented to the CENDI members at the August 1996 meeting. They were approved and included in the Annual Work Plan for 1997.

The first two follow-up studies, being of more immediate concern to the CENDI agencies, were completed in 1996. The last follow-up study, the investigation of legacy collection activities and issues, began with the first task group meeting on May 28, 1997 at the DOE Forrestal Building, Washington, DC, moderated by chair person Barbara Bauldock (DOE). The purpose of the meeting was to share information regarding legacy collection scanning and to identify possible areas for resource sharing or further investigation among the CENDI members. The team was comprised of representatives from the National Aeronautics and Space Administration (NASA), the National Library of Medicine (NLM), the Department of Energy, Office of Scientific and Technical Information (DOE OSTI), and the Defense Technical Information Center (DTIC). The representatives from DOE OSTI in Oak Ridge also participated via videoteleconference.

The meeting was organized around a series of questions prior to the meeting prepared by the task group chair and the CENDI Secretariat Senior Analyst (see Appendix A). These questions were designed to identify the number of legacy collection documents at each agency, the degree to which these documents have been imaged, the ease with which each agency can determine if the source of a legacy item is another CENDI agency, quality issues found in the course of scanning legacy collections, and plans for the future imaging and distribution of the images.

The meeting began with presentations by each agency representative prepared in answer to the discussion questions. The Senior Analyst and Chair also presented information regarding the scanning of CENDI agency legacy collections by other government agencies.

This report presents the legacy collection information by agency and discussion question topic. It is followed by a brief description of imaging activities at other organizations that may include CENDI agency technical reports, an analysis of the information, and recommendations for further study.

2.0 DEFENSE TECHNICAL INFORMATION CENTER (DTIC)

2.1 Size of the Legacy Collection

The legacy collection is of three types: 800,000 documents on roll microfilm of which approximately 680,000 have citations on the database and 120,000 are catalogued on cards, 900,000 microfiche with citations on the database, and 11,000 hardcopy documents received from base closures without citations on the database.

2.2 Size of Legacy Collection from Other CENDI Partners

Approximately 40,000 legacy documents are from DOE and 9,000 from NASA. DOE documents can be identified by the "R" range of accession numbers. NASA documents are identified by the report number.

2.3 Imaging Activities

DTIC has been scanning images of currently received documents as part of its EDMS since October 1994. This was recently extended to include classified documents. 66,000 documents have been created with DTIC's EDMS. The hardcopy is discarded after scanning. Fewer documents than expected are being received on a daily basis, so DTIC plans to use the excess scanning capacity to convert the 11,000 hardcopy documents received from base closures. This project will take about 3 months.

DTIC's current plan is to image the legacy collection material as it is requested. However, the equipment is not yet in place, so fiche are still being duplicated. DTIC is investigating additional scanning equipment to handle both microfiche and roll film conversion. An Ameritec scanner was investigated at the recent AIIM conference.

2.4 Quality Issues

No new issues have been identified, but they are looking for software to enhance the image quality. It would also include Seaport image software with loose integration that will allow images to be exported from the EDMS server to a PC, cleaned up using the image enhancement software, and then exported back to the server.

2.5 Resources

The imaging will be performed in-house only. No resources have been allocated because there is no specific project in place.

2.6 Other Related Projects within the Agency

Scanning projects are underway at three related centers, the Naval Research Library, Redstone Scientific and Technical Information Center, and Philips Laboratory. DTIC staff plan to visit these installations in the near future to learn from these experiences and to determine if there are

images created by these centers that can be used by DTIC.

3.0 DEPARTMENT OF ENERGY, OFFICE OF SCIENTIFIC AND TECHNICAL INFORMATION (DOE OSTI)

3.1 Size of the Legacy Collection

There are approximately 1.2 million documents in the legacy collection including 80,000 classified documents. 600,000 of these documents are also in fiche.

3.2 Size of the Legacy Collection from Other CENDI Partners

Based on a study conducted by the CENDI Cataloging Working Group representatives in August 1996, DOE loaded approximately 16,000 documents from NASA and approximately 59,000 documents from NTIS during the three years from 1994-1996. Statistics for the total collection are not available. Catalog records for approximately 59,000 classified and unclassified but limited distribution documents were received from DTIC from 1994 to-date. These records are loaded into a separate database. The originating agency can be easily identified.

3.3 Imaging Activities

In 1990-1991, DOE established an SGML-based electronic format for the submission of electronic documents from DOE and its contractors. They also accept HTML files and images in PDF, Postscript and TIFF G4. The DOE F 1332.15, Announcement and Distribution of Department of Energy (DOE) Scientific and Technical Information, form is also transferred electronically with SGML coding. In 1995, OSTI established a Document Management System Team to develop a system to capture and provide information for their legacy collection following additional electronic information standards for archiving such as NARA 94-4 and 594.

DOE OSTI is currently imaging 130 new documents per day and legacy documents as they are requested. They are also going back to January 1996 as part of the InfoBridge and EnergyFiles projects. This effort will be completed in August, 1997. The programming to create an index from the pre-1975 bibliographic records on tape is to be completed in early F98. As of early May, there were 14,000 documents with images in the DOE system.

They do not plan to begin massive legacy scanning until FY 1998. The plan for FY 1998 is due on June 18, 1997.

DOE OSTI is the archival site for DOE. This requires maintenance of a permanent collection for preservation purposes.

An upcoming "Inforum Newsletter" article highlights the plans for partnering with labs and contractors to receive material electronically (see Related Activities within the Agency below). DOE also plans to identify orphan documents and scan these first. The initial date range is 1940-1975 (the years of Nuclear Science Abstracts). Orphan documents are identified as those that cannot be expected to be received from any other source electronically, such as the CENDI partners, international partners, or DOE laboratories and contractors.

3.4 Quality Issues

DOE expects many of the same quality issues as addressed by the other agencies. In addition it has had problems with some sites that provide images scanned at only 200 dots per inch (dpi). DOE's standard is 300 dpi. The quality of the paper copy is also an issue. The DOE legacy collection contains blue inked mimeograph paper, onion skin paper, old photostats with reverse images, etc. The quality of the older fiche is also suspect. There are blank pages on some of the fiche because they could not successfully image these pages from the original. Another problem is the scanning of color, particularly four-color process. The image comes out black. Some graphs and tables also have multiple colors. DOE is hoping that technology will eliminate some of these quality issues in the future.

As part of DOE's Records Management Plan, a flowchart has been developed. The flow provides for copying of fragile documents, and then scanning the copy instead of an original.

3.5 Resources

DOE estimates that with current resources they will scan 20-50 documents per week from the legacy collection, in addition to the current material being received. However, they do not know if the four people who have been detailed to the scanning for the InfoBridge project can be retained at that high a pace throughout this period.

3.6 Other Related Projects within the Agency

OSTI is partnering with Los Alamos National Lab (LANL) for receipt of electronic documents. (LANL has imaged a large number of the LANL created reports.) Five other laboratories are also scanning their documents, and OSTI will enter into agreements with them. There is also a proposal to enter into an agreement with the Information Resource Center on Plutonium. It will have scanned 1,000 DOE documents by the end of 1997, and the number of scanned documents will grow to approximately 14,000 over the next few years.

As the site of archival record, DOE OSTI must work with NARA. NARA is not yet ready to take TIFF images. If NARA agrees to the full retention plan recently submitted for approval, DOE would not send archival records to NARA. The current retention schedule requires DOE to provide the silver master, a diazo, and an electronic index to NARA for all documents dated over 25 years old.

NARA may not agree to electronic image files for another three to five years. However, DOE has negotiated with NARA for the acceptance of the ASCII SGML or HTML files that they are getting from the laboratories and contractors as part of their electronic workflow.

As of several months ago, DOE is providing only TIFF image files to NTIS and to GPO to fulfill its Federal Depository Library (FDLP) commitment. (DOE was part of a pilot project on a more electronic FDLP system.) Because the GPO system for providing full FDLP access is not in place, the FDLPs do not currently have access to DOE documents. However, the plan is to have the images loaded on the GPO system. This system will not only serve the FDLP but the public

directly. (It is not DOE's mission to directly serve the public.)

3.7 Image Distribution

DOE has a bibliographic file available on the WWW, the DOE Reports Bibliographic Database, which contains citations to documents sent to the Federal Depository Libraries since January 1994. Another Web-based product, InfoBridge, is currently being developed to deliver full text documents to the desktop. The production version of InfoBridge, scheduled for release late summer 1997, will have links to the full text of all documents received at OSTI since January 1, 1996. Eventually, InfoBridge will also link to full text documents hosted on the document originator's WWW site. The content of InfoBridge is only unlimited/unclassified material.

InfoBridge resides under the EnergyFiles virtual library concept. EnergyFiles will provide access not only to DOE products such as InfoBridge, but to other WWW sites that are selected for their relevance. Public access to the full EDB bibliographic file will continue to be provided via major vendors such as Dialog. There is limited access from OSTI to CAS, the file of limited distribution documents, and CLEO, the file of classified documents.

4.0 NATIONAL AERONAUTICS AND SPACE ADMINISTRATION (NASA)

4.1 Size of The Legacy Collection

There are a total of 743,000 documents in NASA CASI's legacy collection, including 424,000 hardcopy items (many with microfiche) and 319,000 with only microfiche available. These items date from 1975 to March 1996. There is also a collection of about 14,000 formal series reports from NACA, the agency that preceded NASA, which date from 1915-1952.

4.2 Size of Legacy Collection from Other CENDI Partners

Approximately 240,000 of the 743,000 documents are from other CENDI agencies (79,000 from DOE, 149,000 from DTIC, and 12,000 from NTIS). These can be easily identified by a code in the bibliographic record, searchable online.

4.3 Imaging Activities

NASA is currently creating TIFF images for all newly received documents using its new Electronic Document Management System (EDMS) processing. This has been in place since January 1996. Batch programs are used to OCR the image files. The result of the OCR is not checked, but could provide full text searching in the future. The hardcopy documents are retained.

In April 1996, NASA began imaging legacy documents not already imaged when the documents were requested, and using the images to fill the orders. The thought is that if one user requested a document, it is likely that work is going on in that field of study, and, therefore, someone else will also order that document. There have been some software problems and the legacy scanning

system is still under development.

NASA discovered that it takes longer to provide a requested document through the imaging process when the source is microfiche. This is because the microfiche quality is an issue (see quality issues below) and often the hardcopy must be used or a hardcopy blowback must be made from the fiche, then scanned and the image used to fulfill the request. This has caused problems in meeting the NASA-standard 3-day turn-around for request fulfillment. The NASA contractor has begun to fulfill the request via the hardcopy, if they have it, and then produce the image for legacy retention after the delivery request has been satisfied.

There are approximately 23,000 documents imaged to-date, and this is increasing every day. It is estimated that about 10-15 percent of the documents imaged are from the legacy collection activities rather than daily production scanning.

The second strategy is to begin a legacy collection imaging project involving four stages. Beginning with the 1995 accession year, NASA plans to image all material for which the NASA Center for AeroSpace Information (NASA CASI) was the original processor of the bibliographic record. This would include domestic and foreign non-partner, and NASA-sponsored material—documents that they are unlikely to get in imaged format from the originators. NASA will concentrate on those documents for which it has hardcopy first.

In the second stage of the imaging project, NASA plans to image the NACA collection. This collection of early aeronautics work is still heavily requested, but because of its age (1915-1952) the paper is deteriorating rapidly. This imaging project will focus on preservation.

The third stage is to image 1995 accessions from international partners. The international STI agencies with which NASA has agreements, such as ESA, the Israel Space Agency, and NASDA in Japan, appear to be slower to implement electronic document management systems and scanning technologies than organizations in the United States. As part of the third stage, NASA plans to work with the NASA Centers, international partners, CENDI members, and other sources of material to make agreements to receive their documents in electronic format.

The fourth stage of the legacy project would involve the processing of material from 1994 back to 1992. This procedure would be similar to that described for 1995 accessioned documents.

Once the process is completed for the five-year back-file, the scanning of other material may follow. This includes the material from the CENDI partners. This material is planned for later in the project in the hope that the majority of that material will be available from the partners in electronic form as established by the proposed CENDI image exchange guidelines (CENDI/96-3), or by simply pointing to the document on the other agency's WWW site.

There is no strategy for imaging the pre-1992 material at this time. A plan will be developed based on the experience gained from processing the 1994 to 1992 material. Perhaps only NASA-sponsored material will be scanned.
NASA is interested in making agreements for electronic documents. This may involve the exchange of images rather than the rescanning of these document by NASA. It may also involve

linking from the NASA bibliographic record for the document to the image loaded on another server on the WWW hosted by the originator or an aggregator, by retaining the URL or other address indicator in the NASA bibliographic record. NASA has been unable to move forward in this regard since agreements need to be established and programs/procedures need to be developed for modifying the bibliographic record to include the URL and for modifying the NASA WWW site. Resources are needed to complete this programming.

NASA has purchased two additional high-speed paper scanners for normal document processing and for use on the legacy imaging project.

4.4 Quality Issues

The imaging of documents from microfiche has been incorporated in NASA's new electronic document management system (EDMS). The microfiche scanning is efficient and accurate for current microfiche, and even multiple fiche can be placed in the scanner bin and run unattended, after checking the initial setting for modifications on the first fiche. However, the quality of older fiche is not as consistent. Tests showed that it takes more time and more operator intervention to scan older fiche. The setting for the scanner must be manually adjusted after each fiche, which slows the process considerably. It may be easier to create the legacy images from the hardcopy where it is available. There is also the problem of the quality of the hardcopy in the legacy collection. NASA has not done much testing on the quality issues related to legacy collection imaging, but they have identified that there are different weights and grades of paper from tissue quality to high glossy, thicker paper, all of which can cause problems in the loading of the scanner. The ink is also of various qualities. There is even some blue ink on blue paper.

NASA also identified an issue related to the scanning of blank pages. NASA did not originally plan to scan blank pages when they occur in the original documents. However, in order to hold the page order of the hardcopy blowback from the image, the blank pages must be scanned.

4.5 Resources

The resources for imaging will come from the in-house contractor staff with perhaps some outside contractor support. There have been no resources allocated to-date. NASA has a project plan only with no resources or timeline specified.

4.6 Other Related Projects within the Agency

The NASA STI Program is unaware of any official projects for the imaging of technical reports at the NASA centers. However, the STI program representative is investigating the possibility that the EOS Project is imaging documents at the Goddard Space Flight Center. If EOS is handling its documents this way, it is possible that other individual projects are also handling technical reports electronically.

Some NASA Centers are loading the full text or expanded bibliographic records for current reports. The NASA Technical Report Servers (TRS's) are linked (including the STI Program's TRS at CASI). The inclusion of reports with the TRS's has been done to different levels at

different centers; this effort is not yet well organized. A variety of formats, including SGML, PDF, and HTML have been used to-date. The NASA CIO drafted a directive to establish a standard format, but the draft was not sent out pending further attempts to achieve a consensus as to what form would best serve the NASA community and under which conditions. It has been agreed that CASI will serve as the repository if an individual TRS determines that a document will no longer be provided via its local server.

4.7 Image Distribution

NASA has a file of image documents loaded for searching via the WWW. However, the beta test has not been launched yet, because there have been problems in downloading the TIFF viewers to certain user machine configurations. It may be necessary to request online profiling information about the user's hardware/software environment, in order to help determine which viewer should be used.

NASA is still discussing how images of NASA documents should be made available to the public. All the material NASA images will be provided to NASA, NASA contractors and to other federal agencies and their contractors only through restricted access via firewalls and node addresses. The response of other agencies and foreign partners to this plan is an issue. Many of the foreign partners do not want their material distributed for free. NASA plans to work with the other CENDI members to provide NASA-imaged material to them.

5.0 NATIONAL LIBRARY OF MEDICINE (NLM)

5.1 Imaging Activities

The NLM does not routinely scan documents processed into MEDLINE because of copyright restrictions. However, there are several imaging experiments underway.

The Lister Hill Center is experimenting with scanning specific History of Medicine collections. Some of these documents may be hand written. The CENDI-proposed image exchange guidelines are being provided to the developers of this pilot.

NLM also has developed the Relais System which scans the hardcopy documents for e-mail interlibrary loan delivery. Due to copyright issues, the scanned images are not retained indefinitely.

The NLM is interested in scanning from microfilm.

5.2 Image Distribution

The NLM also has an experimental interface for MEDLINE data back to 1966 aimed at biotechnology researchers. This search interface has been developed by the National Center for Biotechnology Information (NCBI). It provides a dialog box for performing simple keyword searching. The results are ranked by relevancy and provide links to the bibliographic records of related articles based on relevancy. More advanced features allow the user to specify certain fields and additional Boolean search capabilities.

The controversial part of this project is that the system is free via the WWW. There will be a press conference later in June announcing the NLM's move to WWW access and a migration away from the traditional systems by making the access free.

An additional feature of the system includes the PubMed interconnection between the publisher WWW sites and MEDLINE citations. When the user is pointed to the publisher's WWW site, he may find the image of the journal article, the full text in one or more formats, or a subscription information or document delivery option page. There are 24 journals available through PubMed. This system can be accessed via the NLM Homepage under Databases and Electronic Resources. Cooperating journals have agreed to provide information in SGML format for the building of the bibliographic database. Two journals processed by the NLM, "Dermatology Online Journal" and "Frontiers in BioScience", are available in electronic format only.

6.0 RELATED IMAGING ACTIVITIES AT OTHER ORGANIZATIONS

Possible related activities at three organizations, GPO, NARA, and the Library of Congress, were investigated. The GPO and NARA activities with DOE are described above. No other imaging activities at GPO or NARA involving other CENDI agencies have been identified. NARA has no agency-wide imaging project or plan. The imaging activities at the Library of Congress are primarily in the area of American History.

7.0 ANALYSIS AND MAJOR RESULTS

The overlap of records among the CENDI agencies varies from agency to agency. The cost benefit of developing a system to share legacy images will need to be determined by each individual agency. The statistics on the size of legacy collections, the number imaged and the number of documents in the legacy collection received from CENDI partners is presented in the table in Appendix A.

Legacy scanning varies among the organizations. NASA images new documents and those requested, but NASA also has a multi-phase strategy for material unlikely to be available electronically from the originator. DOE has images back to January 1996 for its InfoBridge product. DOE is also working with NARA on archiving of image collections. A plan is being developed for legacy imaging in FY 1998. NLM is imaging non-copyrighted collections such as the *History of Medicine* and has a major project to link to publisher WWW sites where full text or

images are available. DTIC recently implemented EDMS for its classified collection. It plans to image legacy materials as requested and is investigating imaged collections available at related libraries

There are several key quality issues to be addressed when scanning legacy collections. All agencies face problems when scanning legacy collections due to the quality of the paper and ink of the hardcopy. The NLM History of Medicine Project will address some of these issues.

Agencies with large microfiche legacy collections, DTIC, DOE and NASA, face problems with the quality of the older microfiche. NASA has begun to address the possible impact of these problems on the procedures for scanning legacy collections. DTIC and NLM need to convert roll microfilm. DTIC is evaluating a scanner that can do roll-film, microfiche, and hardcopy.

The group discussed virtual library concepts as the technology that leap-frogs the imaging of documents. Most stated that they preferred to point to the image or full-text of the document rather than acquire the document and store it redundantly. DOE and NASA are developing virtual library concepts, which would bypass the need to exchange images for legacy collections by pointing to the electronic document on another WWW site. DOE has a working prototype in EnergyFiles, and NLM has experience in connecting their WWW site to publisher sites via the PubMed system. NASA is interested in cross-database and cross-site searching through a single search engine. However, the post-processing of the results is important, particularly the elimination of duplicates. Other CENDI agencies may need similar software to fill in the gaps in their virtual library designs. Joint development of such post-processing software was discussed. Joint discussions about the designs and concepts for virtual libraries would be beneficial to the agencies, providing new ideas and identifying areas for resource sharing.

In addition to these major results, the team collected statistics on the overlap of legacy collections, identified scanning equipment being evaluated or purchased, identified that the information collected for the OCR/Scanning Report (CENDI/96-1) should be enhanced to include new purchases, and shared information regarding the evaluation of equipment to scan roll film.

8.0 RECOMMENDATIONS

The evaluation of the CENDI agencies' legacy collections results in the following recommendations:

- Create an inventory of identifiable collections for which images have been created, are in progress, or are planned. These collections could be based on report number, center, project, etc. If they are available via the WWW, the URL and a contact name should be included in the inventory. An inventory would allow the CENDI agencies to contact those organizations for the sharing of images or to point to the collection via the WWW.

- Develop procedures for periodic sharing of images among the CENDI partners to supplement the images they are creating for their legacy collections. This could be the result of massive legacy imaging or simply the images of those documents that are ordered. Determine how many images must be received by each agency to make it cost/beneficial to develop this process. Determine the policy, schedule, and procedures that would make this a cost-beneficial activity.

- Share information on new scanning equipment evaluations and purchases. This could be done by developing technology-related pages on the agencies' WWW sites and linking from the CENDI report on OCR/Scanning to these pages.

- Investigate opportunities for sharing micro-roll film scanning equipment, for both evaluation and production purposes.

- Discuss virtual library concepts, approaches, and plans and to determine how agencies can help each other in the further development of these concepts.

Appendix A

QUESTIONS FOR LEGACY COLLECTIONS DISCUSSION

May 28, 1997

As part of the agenda for the CENDI Information Exchange Legacy Collections meeting on May 28th, each agency is asked to give a brief overview (15 minutes) of their legacy collections and the status of the imaging of that collection. For purposes of this initial discussion, legacy collection material is any material that predates the installation at your agency of an electronic document management or imaging system for current material.

Here are some questions to guide your presentation.

What is the size of your full text legacy collection: Hardcopy? Microfiche?

How many items in the legacy collection are from other CENDI agencies (estimate: check with your Cataloging Working Group representative since they gathered these statistics last year)?

Can you easily identify which items are from other agencies (particularly the CENDI partners)?

Is your agency imaging legacy collection material or does it have plans to do so?

If so, what is the strategy/ies being used or discussed (your agency versus other sources, from most recent to the oldest, certain date range only, based on customer requests, based on document type, based on medium, based on other availablility issues, etc.)?

What quality issues have been raised?

Who will perform the imaging (in-house, contractor, original centers/laboratories)?

What resources have been allocated?

Do you have a specific project, and, if so, what has been accomplished to-date and what is planned? What is the time schedule?

What projects are going on at your centers, laboratories, or libraries that might be relevant?

Appendix B

LEGACY COLLECTION STATISTICS

May 28, 1997

AGENCY	# IN LEGACY COLLECTION	# IMAGED	# IN TOTAL COLLECTION FROM CENDI PARTNERS
DTIC	800,000 roll 120,000 cards 900,000 fiche 11,000 hardcopy	66,000+ (Oct.'94 to-date)	40,000 (DOE) 9,000 (NASA)
DOE	600,000 fiche 600,000 hardcopy	14,000+ (Jan.'96 to-date)	16,000 (NASA) 59,000 (NTIS) 59,000 DTIC classified and limited) (1994-1996 stats only)
NASA	424,000 hardcopy 319,000 fiche only 14,000 NACA reports in hardcopy only	23,000+ (Jan.'96 to-date; legacy requests since Apr.'97)	79,000 (DOE) 149,000 (DTIC) 12,000 (NTIS)