CENDI / 1998-4

CENDI

CENDI DIGITAL LIBRARY INITIATIVES:
Toward a Digital Future

Report of the CENDI Digital Library Initiatives Task Group Meeting of June 30, 1998
Department of Energy, Forrestal Building, Washington, DC

Sponsored by the CENDI Information Exchange Working Group

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

October 1998

TABLE OF CONTENTS

CENDI DIGITAL LIBRARY INITIATIVES TASK GROUP

Valerie Allen (DOE/OSTI)
Barbara Bauldock, Chair (USGS/BRD)
Robert Bunge (NTIS)
Simon Chung (NASA LARC)
Anne Frondorf (USGS/BRD)
Marcia Hanna (DTIC)
Gail Hodge (CENDI Secretariat)
Michael Hoegler (NAIC)
Sam Kosecki (USDA/NAL)
Alexa McCray (NIH/NLM)
Kris Vajs (NTIS)

EXECUTIVE SUMMARY

At the February1998 CENDI Meeting, the CENDI members approved a proposal by the CENDI Information Exchange Working Group to review the projects and plans related to the development of digital libraries among the agencies. The information gathering meeting was held on June 30, 1998 with eight of the nine CENDI agencies in attendance.

The agencies have different definitions of digital libraries. In some cases, the libraries are strictly digital versions of textual material that was previously dealt with on paper. In other cases, the emphasis is on a virtual library, which is a homepage organized to link to a variety of collections. In several cases, the agencies were providing both digital libraries of their own science and technology materials, and creating virtual libraries based on information from other sources ranging from other federal agencies, to academia, to international organizations, to industry.

The agencies are in differing stages of development. Some are in the strategic planning stage (NASA). Others are implementing their infrastructures (NAIC). Some are conducting pilot programs (DTIC, NTIS). Others have full scale operational systems (NLM, USGS/BRD, DOE/OSTI and NAL) which continue to evolve. Most of the high-end evolution involves the integration of existing resources (through embedded links) and the incorporation of additional services such as digital reference, post-processing of downloaded material, and re-use of material through remote experimentation and modeling.

A digital library has many components -- collection development, metadata, a search engine, post-processing functions, digital reference and other services, and community building. While most of the digital-virtual libraries include these components, the CENDI agencies emphasized different aspects of their projects during their brief presentations. This suggests the strengths that each agency has that could be used to forward the digital future of all (see section 3.0).

The resource types and formats included range from the traditional bibliographic databases accessible via the Web to data sets, gene sequences, and satellite imagery. Most agencies have moved away from the early emphasis on full text or images of textual documents (extensions to their bibliographic databases) to multimedia, sound, numeric data, related web sites, etc. DTIC is working with the Library of Congress on better guidelines for metadata to support photos, sound bites, and multimedia. The agencies are devising strategies for integrating the bibliographic databases with these new resources.

Some agencies are adding digital library services that go beyond collection and access. Digital reference (or AskA service) provides access to reference staff or scientists. NAL provides such a service from its Land Grant Universities. DOE and NAIC provide real-time machine translation. Downloading with special emphasis on manipulating the resulting file is available through NTIS's Encapsulator program. USGS/BRD is working on modeling software and distributed vocabulary support for searching. The National Library of Medicine (NLM) is using its Unified Medical Language System to provide vocabulary support in a middleware layer. NLM is also at the forefront of connections between bibliographic systems and online journals.

With the advent of digital-virtual libraries, their integration with legacy systems, and the advent of new types of services, the agencies reported an unexpected challenge. Digital and virtual libraries have sociological implications that were not anticipated by the majority of the agencies. These included the training needed to move legacy staffs into this new environment, the personnel needed to handle increased in collaborative efforts (some of which involves complex agreements and licensing issues), the limited technical resources available to support digital library development, and the possible need to reorganize the company for new ways of doing business.

In addition to reporting on the current status of digital-virtual libraries and future plans, the agencies highlighted challenges that they had encountered. These are indicated as research needs:

- Economics and funding models for digital-virtual libraries
- Distributed searching
- Metadata crosswalks
- Distributed vocabularies that can be integrated
- Tools for profiling and customizing collections
- User usability and evaluation methodologies/metrics
- Sociology of digital libraries (among collaborators, users, and agency staff)

Recommendations

Based on the discussions, the Digital Library Initiatives Task Group recommends the following actions:

- Through CENDI or its Working Groups, prioritize and promote one or more of the research needs identified above.
- Update the scanning/OCR technology tables that were done for the Scanning and OCR report to reflect upgrades in equipment as a means of sharing information about these rapidly changing technologies.
- Produce an “areas of expertise” inventory (specific to DL), as an extension of the highlights presented in section 3.0 of this report.
- DTIC will distribute its guidelines for image and sound metadata.
- Consider how the CENDI agencies might contribute testbed material or otherwise be involved in the February 1999 round of digital library research initiatives.
- NAIC invited other agencies to attend its DLIPS demo at DTIC on July 30.

1.0 BACKGROUND

At the CENDI Meeting on February 2, 1998, the Information Exchange Working Group proposed an investigation of CENDI agency digital library initiatives. The aim was to review the projects and plans currently underway, to identify the state-of-the-practice among the agencies, to identify the issues and common concerns, and to identify research projects (both technology and policy) that could be done collaboratively and result in advancing the state of federal digital libraries.

The task group met on June 30, 1998, at the Department of Energy's Forrestal Building. Representatives from the National Air Intelligence Center (NAIC); the Defense Technical Information Center (DTIC); the Department of Energy, Office of Scientific and Technical Information (DOE/OSTI); the National Technical information Service (NTIS); the US Geological Survey/Biological Resources Division (USGS/BRD); the National Institute of Health, National Library of Medicine (NLM/NIH); the US Department of Agriculture's National Agricultural Library (NAL); and the National Aeronautics and Space Administration (NASA) were present.

Prior to the meeting, a series of questions was developed by the CENDI Secretariat to be used as discussion points by the presenters. Those questions are provided as Appendix A.

2.0 DESCRIPTION OF AGENCY INITIATIVES

2.1 Department of Energy/Office of Scientific and Technical Information (DOE/OSTI)

EnergyFiles and DOE Information Bridge

Two major independent but complementary DOE/OSTI projects fall under the broad mantle of digital library initiatives. One has a digitized text emphasis and the other a virtual library perspective. The DOE Information Bridge builds on the digitization of OSTI's traditional material, DOE technical reports. It is a major component of the second initiative, EnergyFiles, which is a web-based virtual library environment and subject pathway system for energy-related information both from within DOE and from other sources.

The DOE Information Bridge (http://www.doe.gov/bridge; http://apollo.osti.gov/dds) makes the full text of DOE technical reports available electronically on the web. There are currently nearly 30,000 reports available dating from January 1996. The full text of older reports residing in the OSTI repository is scanned and provided as those reports are requested. The DOE-generated information included in the DOE Information Bridge is contributed by the Scientific and Technical Information Program (STIP) partners at the various DOE and contractor research and development (R&D) sites throughout the DOE complex. The DOE Information Bridge is now available via the GPO Access gateway for public use with no password requirement. It is also available on a separate system to DOE and its contractors via registration/password and provides additional features such as the online ordering of paper documents and bibliographic data for journal items and foreign literature acquired through international exchange agreements. In both cases, the full text and bibliographic records are fully searchable and may be downloaded by the user. It is estimated that combined downloads from the DOE and Public "Bridges" will total about 50,000 reports in FY 1998. The public help desk support for DOE Information Bridge is provided by GPO, while OSTI provides the DOE and DOE contractor support.

The workflow processes used to populate the DOE Information Bridge are quite different from OSTI's traditional information processing activities. In the past, many manual processes were used and non-electronic products were produced. Currently, the laboratories and facilities submit their documents and the corresponding metadata electronically or in paper, and OSTI makes the full text and metadata available through diverse channels by various means. The image or hardcopy is OCR'ed to create the full text. This full text is stored, indexed, and searched using OpenText. The PDF version is stored on an optical jukebox. TIFF, GIFF, and PDF formats are retained and accessible on the DOE and contractor site; however, only the PDF version is available on the public site. In the current model, the full text and metadata records must be integrated centrally at OSTI. However, the plan is to continue moving away from centralized processing and toward linking centralized metadata files with distributed full text files residing at the sites. This will change the laboratories' STI submission requirements as they will simply submit metadata records with the corresponding URL's for the full text.

Data within the DOE Information Bridge is indexed using the 38 subject categories OSTI has used for many years. However, the metadata record structure is being modified to a Dublin Core format with administrative and optional fields added to support OSTI processing.

The second initiative is the EnergyFiles virtual library (http://www.doe.gov/EnergyFiles). It is a very heterogeneous system emphasizing information resource collection and accessibility and it includes more than 400 collections from OSTI, other DOE programs and offices, industry, and academia. The collections are grouped under 14 subject categories which are consolidated from the 38 categories that are used in the DOE Information Bridge. There are multiple resource types, including electronic journals, databases, regulations, technical reports in full text when available, and conference proceedings. There are also multiple document types such as HTML and PDF. An EnergyFiles collection development plan that outlines the scope and source criteria in some detail is available on the EnergyFiles Home page at More About EnergyFiles.

Utilizing the dynamic nature of the Internet, EnergyFiles provides access to both information resources and a "virtual workspace" that allows users to customize services through workspace tools. In essence, the site goes beyond being a library to being an energy portal, presenting information in addition to providing capabilities to make it more useful through user manipulation. The customized services include such capabilities as machine translation in real time (using Systran, developed by NAIC) and a push technology application which allows the user to establish notification mechanisms based on information profiles. The virtual workspace also provides access to remote experimentation laboratories and a link to the Argonne National Laboratory "Ask-a-Scientist" service, and it continues to grow in capabilities and functionality. Some of the workspace resources have limited access due to licensing agreements. EnergyFiles has involved a multitude of collaborative relationships on various levels.

The primary audience for EnergyFiles is scientists, engineers, program and project managers, both within the DOE R&D community and among U.S. industry and academia who need STI to facilitate R&D efforts. It also has significant value for the general public. An analysis of user segments indicates that over 24 percent are from the US Commercial sector and more than 20 percent are from the US Government sector. More than 8 percent of users are affiliated with the US Education sector.

Due to the great number of collections and resources on EnergyFiles, a key issue for further development has been searching across heterogeneous document types. The EnergyFiles virtual library currently relies on links to integrate the material, but for the future, the approach being considered is a modular one. The incorporation of a search engine is the next important development effort which will make the location of the material more transparent to the user. Users will have the capability to choose specific databases and collections, to limit a search to subject areas, or search “All” if preferred. The Z39.50 heterogeneous file search protocol is not currently used.

Major issues for OSTI based on development to date include funding and staffing resources, the accelerated development rate of web products, and heterogeneous document type searching. Current efforts are focused on the migration of OSTI production systems to ORACLE software and the redesign of the database management system upon which the redesigned electronic infrastructure will be built.

EnergyFiles, the DOE Information Bridge, and many other websites developed by the Department of Energy’s programs provide useful collections of scientific, technical, public affairs, and educational information. These successes lay the foundation for the development of a National Library of Energy Science and Technology. In addition to technical report literature and other information resources, access to electronic journals in scientific disciplines will be a key component of such a National Library. Site licenses are being procured so that greater access to electronic journals may be offered in the near future. In recognition of the fact that journal literature is a very critical element in the scientific community, OSTI is further exploring the concept of a web-based journals product using existing technology to search and link journal article citations with fulltext articles, to include links to fulltext of references within the article. This product, focused on the physical sciences (and similar to the National Library of Medicine’s PubMed which is focused on the life sciences) would complement the report literature already available electronically to the scientific community through the DOE Information Bridge and other OSTI products.

2.2 Defense Technical Information Center (DTIC)

Defense Digital Library

The DTIC hosts several digital collections and sponsors a number of ongoing digitization efforts aimed toward the defense Scientific/Technical information community. Although not specifically referred to as a digital library, DTIC makes these materials available through its own Internet-based web pages and hosts over 90 more Internet sites in addition to providing numerous links to other pertinent sites.

DTIC's primary mission is to collect and provide information produced by, and of interest to, the defense community. The greater part of the organization's current digital content may be found under DTIC's STINET web site. Through STINET, users can search DTIC's three main databases, plus others, in addition to linking to those produced by other organizations. STINET offers multi-database searching with retrievals ranked within each database, including several non-DTIC databases. Included are: DTIC U2 Technical Report Database, DTIC Full Text Technical Report Database, and DTIC Thesaurus. *Non-DTIC databases include the DoD Index of Specifications and Standards, DOE OPENNET Database, DOE Reports Bibliographic Database, NASA Technical Reports Database, NASA Open Literature Database, NASA NACA Technical Reports Database, and the NASA Goddard Technical Reports Database.* In the near future, DTIC's How To Get It Directory will be added. Documents retrieved from the DOE or NASA databases may be ordered directly from the respective agencies.

Examples of other searchable DTIC databases and resources under the STINET umbrella include the Air University Index to Military Periodicals (AULIMP) and the Research & Development Descriptive Summaries (RDDS) database. In many cases, the full-text documents, articles, or summaries are available either in HTML or PDF formats. STINET also provides the capability to search several specialized research tools such as Technology Navigator, a web-based collaborative tool that enables national security community technologists, project, and resource managers to quickly track rapid changes in products and services available from industry and academia. STINET offers a large number of these resources through its public site. Registered users with access to Secure-STINET will also find expanded and additional collections of controlled materials and tools, plus links to other resources. Recently added is STINET On SIPRNET (SIPRNET is DOD's Secret, US-only network), a new service which is freely available to all SIPRNET users.

From the DTIC home page, users will find and be able to view and/or search a variety of reference tools such as the Dissemination Authority List (DAL), the Subject Categorization Guide for Defense Science and Technology, and an acronym list. Several current awareness products are described and in many cases are fully searchable. One recently added example is TopicLINKS, which directs users to locations and sources of information available from the Internet on topics that categorize the areas of scientific and technical interest upon which DTIC's field and group structure is based. DTIC's Manpower and Training Research Information System (MATRIS) publishes and provides full-text access to a number of directories, in addition to hosting several specialized collections.

DTIC hosts LabLINK, the website for the U.S. Department of Defense Laboratory System. DoD Laboratories are owned and operated by the Military Departments: Army, Navy, Air Force. This Home Page serves as a vehicle for interaction and coordination among the DoD labs, and also serves to let outside organizations find out about the DoD labs. LabLINK is sponsored by

DDDR&E Office of Laboratory Management & Technology Transition. Through LabLINK users may access TechTRANSIT, a website that provides the technology transfer activities of the DoD laboratory community.

A number of other specialized programs, such as the Dual Use/Technology Transfer and the DoD Small Business Innovation Research (SBIR), are directly accessible from DTIC's home page, as

are the Information Analysis Centers (IAC), which are sources of highly specific scientific and technical information. Many of the IACs host their own databases and digital collections.

Surveys of DTIC's broad user community and research performed by experts from both within and outside the organization have determined that DTIC's digital content is of high caliber in terms of quality and quantity, and is appropriate for the community it serves. Most comments focused not on what is available, but rather on the way it is presented, the difficulty in finding and accessing it, and the limited scope of digital materials included. In response to this, DTIC is currently supporting two major efforts. The first will pull together and reorganize existing collections, materials, and resources in addition to identifying other pertinent resources aimed specifically at the scientific and technical information (STI) community. Added functionality, such as providing a more sophisticated search capability, commands high priority.

The second effort, the Defense Virtual Library (DVL), is a collaborative partnership among DTIC, the Defense Advanced Research Projects Agency (DARPA), and the Corporation for National Research Initiatives (CNRI). The goals of the DVL are to (1) use the prototype experience to learn lessons and expose challenges that require further research and development, (2) provide Internet access to a variety of DTIC-hosted materials, (3) provide access control as appropriate, (4) provide virtual library connectivity to diverse digital libraries, and (5) provide a means for management and rapid resolution of digital object names (Universal Resource Names) and locations on the Internet.

CNRI's Handle System for location independent identification of web accessible digital objects was adopted. (The Handle System is the basis for the Digital Object Identifier [DOI] which is being developed by a consortium of commercial publishers.) If the user has a Handle, he can search the Handle System directly for the item. Handles are becoming more widespread; e.g., LC has adopted them, and they should become more available as part of bibliographic citations in the future.

The DVL repository includes textual materials in TIFF and PDF format drawn from the DTIC digital document collection. Similarly, the citations for these text documents are extracts from the regular DROLS production. In addition, photographs and sound files are being added to the repository as test materials. Some 300 photographs in JPEG and GIF format on the early development of the atomic bomb were obtained from Los Alamos National Laboratory. Sound files consisting of extracts of oral history interviews and military music are currently being digitized for addition to the collection.

In the future, DTIC proposes to add more materials, and different material types including video, executable computer programs, simulations, maps, spatial data, art and graphics. Medical imagery, aerial photographs, and educational materials with a DoD emphasis also would be interesting. DTIC would like to see the Handle system more widely used.

DTIC views its digital library efforts as an excellent opportunity to explore the challenges associated with providing access for all of the DoD to a diverse collection of digital materials.

2.3 National Aeronautics and Space Administration (NASA)

Expert Center for the NASA Electronic Library

The NASA STI Program is in the midst of strategic development of the Expert Center for the NASA Electronic Library. This is an outgrowth of the NASA reorganization concept of a lead center. The agency has in its strategic plan the need to generate and communicate the knowledge it funds. The vision is that by 2001 there will be a virtual library available at each employee's desktop. This includes NASA and non-NASA-produced material. The document types include traditional technical report and text, as well as engineering drawings, bibliographic records, full text, images, electronic, and photographs. Although currently there are no plans to put the full NASA database on the Web, this is a desire for the future.

NASA STI Program is interested in using the digital-virtual library concept to manage the STI system more proactively. Given the fact that there is no additional funding for implementation of such projects by lead centers, in this case NASA Langley Research Center (NASA LARC), and that electronic projects as well as web sites abound within NASA, the key is a distributed, but collaborative effort. They want to be in a position to recommend standards across the agency for publications, viewing, and manipulating. Although they are focused on internal tools, they will also be satisfying public information needs, when possible.

Currently, the NASA STI Program is evaluating implementation approaches, taking into account the legacy of electronic efforts that already exist within NASA. Information life cycle management is another key concern. They are working with the CIO to establish the policies that would support the virtual library, and its archiving function.

The NASA STI Program has determined two main groups of users. The first is the primary market which is the strategic enterprises determined by NASA, such as aeronautics and earth science. Then there are also the NASA cross-cutting processes, which are not directly pushing NASA's agendas but are needed as support. For example, information about cost center management cuts across all processes. There is a need to satisfy both groups.

The cultural aspect is very important. The strategy for developing this expert center is to mobilize within LARC and then to spread out to the other centers. There are many other projects going on that are not connected to this effort at the STI Program and it is important to connect them. Some of the centers are involved with NRL, NIST and NSF on the "Web of Science". Others have local projects.

NASA identified the following major challenges:

- creating the expert center without additional funding
- modifying the independent culture of NASA centers to include the Electronic Library in its information life cycle management or having policies from Headquarters that support the project
- technologies and infrastructure to maintain the links of the virtual library

On the latter issue, LARC would like to use best practices and have looked at Michigan's Internet Public Library project.

2.4 National Air Intelligence Center (NAIC)

Digital Library Input Processing System and InfoSpace

NAIC's digital library emphasis is on the workflow and technologies for processing digital information. The DLIPS (Digital Library Input Processing System) was designed in 1996 to be the system to meet the future information technology push and the need for organizational downsizing. Large budget cuts have pushed the installation ahead of its scheduled full-scale implementation. The system's goal is to achieve maximum processed input with minimal human intervention and maximum retrieval.

The challenge was to develop a digital library capable of operating on a processing continuum from manual to fully automatic. This is necessary because NAIC deals with a wide range of materials with various source qualities, including foreign languages and very poor quality textual originals. The system has been shown to be able to scan and OCR up to 6,000 pages per day of good quality English text.

DLIPS includes document scanning (currently, an EK500 scanner with BTG software), machine translation (Systran), entity tagging, a record authoring tool, etc. Workstations also include simple provisions for gisting.

Tools include the Systran Machine Translation software. Russian, German and French tools are well developed and they are looking to implement 11 languages in the short term. The future hope is that any person in any language could read any document.

NAIC is moving away from their traditional people, facilities, and nomenclature (PFN) indexing sets. While manual indexing is supported by the system, resource limitations require that the PFNs be created automatically, or the need for them be eliminated or reduced through more sophisticated retrieval algorithms. RetrievalWare is at the core of the system. As part of the testing phase, products are being extracted using RetrievalWare on a daily basis.

Much time has been spent on the technologies related to the system, but just as importantly, teams have been working to ascertain the most effective and productive workflow. Workflow determination testing went through alpha testing from July through October 1998, and the system successfully met IOC in October 1998. FOC is scheduled for March 1999. Key aims during these testing periods are to smoothly integrate these new technologies among the legacy labor force through operational training and their efforts to establish new workflows. The alpha testing in this development approach occurs significantly before the traditional acceptance testing phases of product acceptance, and the full involvement of the alpha test development team provides important feedback to the code writers and process developers. A new branch organization was designated in January 1999 to provide DLIPS processing.

Continuing NAIC efforts will be focused on maximizing process effectiveness. Early questions regarding the numbers and types of processing personnel have essentially been answered. Now the question seeking answers is determining effective throughputs and impacts on information quality and usability.

NAIC indicated that associated technology improvements constantly challenge system developments, and that changes are always being considered. A change in the scanning software is underway, from BTG to KOFAX. NT boxes with 233mHZ dual Pentium servers are currently being used, and changes in hardware will obviously follow evolving demands for improved user interfaces and processing throughputs. The key to the success to date of the DLIPS and other information processing systems is the modular design approach which permits the incorporation of improved technology through hardware or software replacements (scanning, OCR, translation, automatic indexing, and retrieval) without needing to make major changes in the underlying processes and their products. Other challenges include the effect that the digital library environment has on outsourcing requirements, contracting policies and conflicts with rapidly changing needs, and pressures poses by ever shorter processing turn-around requirements.

NAIC identified several key challenges during DLIPS development:

- implementation of a complex process in a reduced time frame
- the lack of metrics from outside sources, and the need to develop their own
- retraining personnel and managing the impact of the changes on the workforce

NAIC expects a continuous expansion of data sources. Organizations are increasing their overall analytical requirements for textual and imagery information. NAIC has enjoyed significant improvements in its communication of imagery information, improving turnaround from months to days, and expanding its distribution to hundreds of additional customers. A similar improvement is underway in their textual world, and although the current development emphasis is on converting paper documents, the hooks are built in for future electronic sources, including CD-ROM, Digital Video Disc (DVD) and others.

2.5 National Technical Information Services (NTIS)

SpecFinder

At NTIS, the digital library is defined very broadly as “marketing entry points” for sales opportunities. NTIS has a number of files that are currently available and are being integrated. These include the NTIS Web Site, the International Trade Book Store (20,000 records with online ordering); industry, federal and military standards; depository library information; an Open Source project for the intelligence community; and Specfinder for US business access to DoD technical solicitations.

The latter was the focus of the presentation by NTIS. Specfinder, which became operational on May 18, has as its audience US businesses interested in technical information for DoD solicitations. The content is current active solicitations. The industry standards (some of which are contained in the military standards file on NTIS) are referenced. The specifications and standards may also include engineering drawings.

The acquisition of data includes many partners, including Intessera Technologies, Data Marketing of Virginia, and various DoD organizations. DoD organizations supply the drawings and the solicitations. Agreements had to be struck with industry standards organizations, since many of the standards are copyrighted. NTIS currently has 20-30 partners in this project.

A team approach is used in developing this digital-virtual library product. The teams include the database creation staff as well as business and industry specialists at NTIS. The aim is to tightly integrate these virtual collections with NTIS’ legacy ordering and bibliographic systems. These systems are needed to support the EDI (electronic data interchange) transactions. In order to support the location and ordering of the standards, the bibliographic system now includes the standards documents. This means that they can be sold through the regular NTIS channels.

The architecture for this system is a distributed one. Oracle databases are used for the bibliographic and order systems. The document is stored in Adstar, NTIS’ image-based storage system. The web site is operated by Intessera. Verity is used as the front-end search tool, with hooks to Oracle databases. Access is only through standard Web protocols, not through Z39.50. Intessera provides a parser that scans the solicitation text document for the standards that are available in the database. This is done without manual effort and there may be standards referenced in the solicitations that are not linked.

When the user accesses the system, he can search the active solicitations by keyword, where they came from, solicitation number, etc. The resulting document includes embedded hyperlinks to the standards and drawings. Through a Java applet the user can display all the standards connected to that specification. The items on the standards list can be moved over one at a time for ordering, or all at once. The user can also preview the first page of each standard to ensure that it is what is wanted. Another Java applet allows the user to download one of the resulting standards and save the rest for a later download.

Manipulation of the downloaded document is supported by the Encapsulator, a Windows client program that allows the management of the document, including PDF. The user can select whether to have the paper copy sent or to download. The drawings themselves are of large size, but they result in relatively small files. There has been no Internet bandwidth problem to date.

A cash register function (credit card or NTIS deposit account) is available to support the ordering system. A copyright clearance screen is also included, where the user must acknowledge understanding and agreeing to the copyright notice provided.

SpecFinder has proven to be a good pilot project, because the standards are generally smaller files than technical report files. Future plans include continued emphasis on technical information but with a new line of products. NTIS is looking for niche audiences that can be supported by their traditional systems, and new applications. In addition to new variations on the Specfinder model, Specfinder itself will expand. Currently, at the Richmond, VA DoD procurement center alone, they are providing approximately 500-1000 solicitations per day. Other centers are expected to join in the near future.

NTIS also plans to integrate its database backfiles into the same system. A new Oracle data has been built with hooks to the other programs, and ten years of the NTIS database are being moved to the new system. This will allow more flexibility in product design.

2.6 National Agricultural Library (NAL)

Electronic Information Initiative

NAL, the largest agricultural library in the world, has been serving agriculture since 1862. Established by Congress as the primary agricultural information resource of the United States, NAL is part of the Agricultural Research Service of the U.S. Department of Agriculture (USDA). In its dual role, NAL serves as a national and departmental library and as the U.S. point of contact for international agricultural information efforts with the Food and Agricultural Organization (FAO) of the United Nations.

NAL collects materials in more than 75 different languages. The collection contains over 3.5 million items and over 23,000 journal titles, received annually. NAL's AGRICOLA database is its major product. In July 1998, AGRICOLA became available free on the World Wide Web at http://www.nal.usda.gov.

The underlying foundation for digital library initiatives was laid in December 1992 when NAL launched its Electronic Information Initiative (EII). Early recognition of imminent changes in information production and delivery systems, combined with the adoption of an EII vision, positioned NAL to play a leadership role in the creation and development of the emerging “electronic library.” NAL's commitment to digital information was emphasized with the January 1, 1995 declaration that electronic information would be the preferred medium. To fulfill its mission to “ensure and enhance access to agricultural information for a better quality of life” and to further the digital library cause, NAL is actively working within the USDA, with other federal entities, and with national and international partners to explore new methods and technologies that advance access to agricultural information.

NAL is seeking to develop a comprehensive agricultural information network — a one-stop web environment for agriculture. In doing so, NAL, in collaboration with land-grant universities, established the Agriculture Network Information Center (AgNIC). AgNIC is a discipline-specific distributed virtual library with the goal of helping users find quality information on the Internet. Key to this project is the development of specialized areas of expertise. There are currently 20 participants, primarily land-grant universities, with virtual knowledge centers. They provide collection development support, create products, and handle digital reference inquiries. NAL is contributing in seven subject-specific areas, including collection development and reference. AgNIC is one of the first discipline-focused distributed networks on the Internet. It is a model that can be applied to other disciplines. It is among the first discipline-specific distributed networks to use librarians and subject specialists to provide online reference assistance.

NAL is taking a lead role in the effort to preserve USDA's electronic and print literature. Because of the transitory nature of digital information, a long-term strategy is being put into place to ensure that the growing body of USDA digital publications being produced is systematically identified, prioritized, preserved, and archived. The basis for this strategy is the report, “Framework for the Preservation of and Public Access to USDA Digital Publications.” A national steering committee, chaired by NAL Director Pamela Q. J. André and composed of representatives from within USDA and other key external stakeholders, is being organized to oversee the implementation of this plan.   NAL also has been actively participating in the National Preservation Program for Agricultural Literature within the framework of the United States Agricultural Information Network (USAIN). This national preservation program for agricultural literature is a discipline-based cooperative plan involving NAL and land-grant universities. One major activity emanating from the program is, “Preserving the History of United States Agriculture and Rural Life: State and Local Literature, 1820-1945,” funded by the National Endowment for the Humanities and for which NAL is serving as the national depository for state-preserved microfilm. NAL's own contributions to the USAIN effort are in the digital arena. To date, more than 25,000 text and image pages of USDA embrittled journal and monograph materials have been digitized, representing the first publications in the library's digital archive collection. SGML files have been created for the digital products and additional efforts by the library will be directed toward Internet access to these images. NAL is committed to digitizing the entire journal run of the Journal of Agricultural Research and continues to move forward with this project as funds become available.

NAL continues to work toward becoming a virtual library. NAL’s goal is to deliver to its customers a seamless electronic infrastructure with access to information resources that are available through computer networks. A major step was taken with the establishment of an Electronic Media Center. Emphasizing information's modularity and portability, NAL intends to create an adaptive user interface. NAL has been addressing needs to build bridges between the secondary services and primary literature sources. Developmental work continues in providing Internet links from the AGRICOLA database to agricultural full-text electronic journals. NAL is working on a mechanism to simplify access to all USDA publications available on the web. Parallel efforts are directed toward agricultural electronic publications from research facilities worldwide. In developing the virtual library, NAL is recognizing the need to emphasize to end-users that information is not free and that there is substantial value-added in what NAL is providing. This is especially important to reiterate to users when, through partnerships, links are created directly to commercial sites and the role of the library is less visible.

NAL provides leadership in the identification and implementation of new methods and technologies to improve access to, and management of, agricultural information. The Library's information access and management efforts focus on five major goals:

- **Information services**: create conditions by which NAL's diverse customers can efficiently and cost effectively identify, locate, and obtain desired information on agricultural topics.

- **Electronic access**: enhance access by contributing to the content, organization, and management of the information superhighway, especially as it relates to agricultural topics.

- **Information products**: create products that support information needs in a changing agricultural environment, and make them widely available through electronic publishing, Internet access, and state-of-the-art storage and retrieval methods.

- **Outreach**: promote the availability and use of NAL's resources and information products.

- **Training**: develop and implement programs that enable customers and staff to take full advantage of current and emerging technologies and information systems.

NAL supports the development of new knowledge and technology, and disseminates information that is essential to solving critical agricultural problems facing the nation and the world. The information and database services of the National Agricultural Library are vital to the nation's ability to sustain a viable agricultural economy, maintain a healthy environment, and ensure safe, abundant, affordable and high-quality agricultural products for consumers.

2.7 National Library of Medicine (NLM)

<u>PubMed</u>

The NLM, although best known for the MEDLINE database, hosts over 40 online database under its MEDLARS system, containing over 18M records of bibliographic and factual information. Half of the MEDLARS records are from MEDLINE. Even though it is not a full text database, it is a very rich database because of the carefully controlled scope and the use of MeSH (18,000 terms in 15 hierarchies and 40,000 entry terms). Other databases include AIDSLINE (which is a subset of MEDLINE), BIOETHICSLINE (in collaboration with Georgetown University), SPACELINE (in collaboration with NASA) and TOXNET (which brings together not only MEDLINE records, but those of other secondary publishers in the area of toxicology).

Based on the experience from this history, NLM believes that Digital Libraries are larger than just a digitization project. It is the provision of new services in an electronic environment.

The decision was made last year to make MEDLINE free on the web. This is because the NLM was able to charge for distribution but not for database creation. The distribution costs via the web are negligible, and so putting it out for free made sense. In addition, NLM introduced Internet GratefulMed, which brought the GratefulMed search software from diskette to the web; PubMed, that provides links to publishers full-text material; and a specialized, custom-designed search engine called Entrez.

NLM has had several digitization and digital collection projects. These include the Relais interlibrary loan system, which uses scanning and fax technologies to provide interlibrary loans from its collection. Copyright issues will not allow a database to be built, so the records must be located every time and rescanned. The images from the History of Medicine collection (60,000 photographs and artwork) have been digitized. In support of the Regional Medical Programs, 1,500 documents or about 40,000 pages have been scanned. The newest project is a collection of papers from eminent biomedical scientists. The initial collections will be introduced in September. Many of the projects have included historical material that can provide valuable information to the scholars of medical and biological history. They help students gain an appreciation for the methods and successes of science. These collections include text, audio, still images, and video.

In addition to the collections, NLM has emphasized the development of tools. Creation of a workable search engine was key to NLM's efforts. They wanted a well published and documented set of Application Programmer Interfaces (APIs). It was also imperative that the vendors demonstrate their systems with NLM data. They bought Excalibur (now RetrievalWare). However, because of the lack of a robust API and licensing issues in a networked environment, NLM is using Excalibur only for data entry and web publishing. Capitalizing on the efforts of the National Center for Biotechnology Information at Lister Hill, NLM incorporated its Entrez search engine into PubMed. Entre includes sophisticated statistically based neighboring algorithms (conceptually related neighbors) developed by John Wilbur.

Another tool is the Unified Medical Language System (UMLS). This system includes a metathesaurus which incorporates over 40 other thesauri. The 1998 release has over 400,000 concepts and over 1 million natural language strings. The Semantic Network includes 132 semantic types and 53 relationships which identify the strings. There is also a lexicon and lexical processing tools, and an Information Sources Map.

PubMed currently is using parts of the UMLS as middleware to provide search help based on semantics. It can help users when they want to broaden or narrow a search. It is particularly helpful when a search retrieves zero hits. The UMLS could be particularly valuable when intelligent agents are developed to identify relevant databases and to translate the user's request from natural language to the language of the information resource.

In support of its own digital library research needs and to advance the future of digital libraries, NLM is involved in the NSF Digital Library-2 Initiatives (DLI-2). This initiative is co-sponsored by a number of federal agencies including the Library of Congress, NLM, National Endowment for the Humanities (NEH), and NSF. It is more applications oriented than the first DL initiative which set the stage with technology and infrastructure development. The main research areas for DLI-2 include human-centered DL research, content and collections, systems-centered research areas, and testbed and applications.

Human-centered research includes research into information discovery and retrieval methods. The design of more intelligent user interfaces will be studied. This may include information visualization. User and usability studies, including metrics will be studied. The social implications of digital libraries also is included.

In the area of content and collections, the funding agencies are interested in methods and technologies for data capture, representation and preservation. Metadata issues will be explored, particularly for domain specific information objects. Intellectual property rights and new economic and business models to support digital libraries will be investigated.

The DLI-2's systems-centered research agenda include open, networked architectures, system scalability, intelligent agents, systems evaluation and performance studies, data compression, and authentication. Testbed and applications are aimed at specialized tools, such as those for mark-up and semantic encoding, specialized applications for specific domains, and new methods for establishing relationships among networked knowledge sources (beyond the current pointer/link mechanism).

The first round of DLI-2 projects must be submitted by July 1998. Awards will be given soon thereafter. Several CENDI agencies are involved in projects, either by providing content for testbeds or for applications. NLM’s digital library challenges are well reflected in the DLI-2 agenda. NLM is providing Visible Human data and the UMLS Knowledge Sources. NLM is particularly interested in research agendas that center on requirements of health care digital libraries, including information for health care consumers. Another round of funding is expected in February 1999.

2.8 US Geological Survey/Biological Resources Division (USGS/BRD)

National Biological Information Infrastructure

The focus at the USGS/BRD is not on creation of digital collection so much as it is on collaboration to make a digital-virtual library for the communication of biological information available. As the lead federal agency on the National Biological Information Infrastructure, USGS/BRD is chartered to provide the means for collecting, accessing and utilizing the biological resources of the nation. Collaborators include other federal agencies, universities, and state and local governments.

To achieve this, USGS/BRD has collaborative agreements with a variety of sources —federal, academic and commercial. Using a web-based site, the NBII brings together disparate resources. Key to the ability of this site to integrate well is the requirement to use the NBII specified metadata standard. The NBII is intended to be a complement and an extension of the National Spatial Data Infrastructure.

The NBII metadata standard is an extension of the Federal Geographic Data Committee (FGDC) standard for geospatial referenced information. This is a federal standard required for use by any federal agency for data that can be referenced geospatially. Much of the NBII information can be so referenced, but there are laboratory-only research reports and data that do not fit this scheme well. To resolve this, the USGS/BRD wrote guidelines for the use of this geospatial scheme for non-geospatial data. In addition, there was the need to extend the standard through a biological profile. Elements added to the FGDC standard include the genus-species name (Latin and common) and the classification for that genus-species.

In addition to the development of standards, BRD is providing tools to support the creation of the metadata to that standard. Metamaker is available as both Windows-based software on diskette and a web-based form. The goal is to create metadata as the report is generated.

The architecture of the NBII is a distributed network of web-based resources (a virtual library). These resources include legacy data sets with metadata added after the fact, and new sets that are built with the metadata included. The metadata is searchable via a clearing house mechanism (CHM) based on the NSDI CHM which is searchable via the Z39.50 protocol.

BRD is now working on building bibliographic information for publications into the CHM. Bibliographic cites are extracted from various repositories and then converted to the standard format based on a subset of the NBII metadata standard appropriate for publications. At some science centers, the metadata for publications points to the full text of the publication if it is online.

In addition to the development of the metadata, USGS/BRD supports the development of vocabulary tools to aid in the indexing and retrieval of NBII information. This vocabulary term will be used as controlled keywords in the metadata for publications, data sets, and web sites. The development of the vocabulary is a collaborative effort with the California Environmental Resources Evaluation System (CERES) of the California Resources Agency. The goal is to develop a very shallow vocabulary at the national level that can house the more detailed vocabulary needed by the individual states, or other organizations, that will supply information through the NBII. This will allow the maintenance of the most changeable part of the vocabulary to remain with the owners of the vocabulary. The vocabulary also may be extended to accommodate international efforts such as the European Environment Agency's GEMET Thesaurus and the International Council for Scientific and Technical Information (ICSTI) Working Group on Life Science Vocabulary, pointer web-site to life science databases. In addition, BRD has taken the lead in the Integrated Taxonomic Information System Project, which is a collaborative project among several federal agencies to create a taxonomic authority file. The terms from ITIS will be used in the appropriate taxonomic field in the metadata.

Key challenges for USGS/BRD have been maintaining connections to standards for both geospatial and non-geospatial information and developing vocabulary aids that can be navigated directly over the Internet (not just work behind the scenes). BRD recently co-sponsored an NSF workshop on Taxonomic Authority Files and a workshop at the Digital Libraries '98 on Networked Thesauri.

2.9 National Library of Education (NLE)

The NLE is the largest federally funded library in the world devoted solely to education and the federal government's principal center for one-stop information and referral on education. The NLE began in March 1994 as a continuation of the Education Research Library which started a century ago with the private collection of American schoolbooks from Henry Barnard, the first commissioner of the Office of Education.

The Library is charged with several key functions. These include the acquisition of education information, the development of national and international education information services, and the establishment of resource sharing networks among a variety of information providers. Today the NLE houses more than 200,000 books and about 750 periodical subscriptions in addition to studies, reports, ERIC microfiche, a rare book collection, and CD-ROM databases. An online integrated library system provides access to the current collection via author, title, and subject searching.

From the Library's inception, the NLE has been striving to become a virtual library rather than focus on more traditional library activities such as acquisitions, cataloging, and reference services. Projects such as the ERIC system, the Department's World Wide Web site, the Gateway to Education Materials, and the Virtual Reference Desk are examples of the library's attempts to create a library that is customer driven and multi-media in nature.

The NLE also maintains the U.S. Department of Education's Online Library. Individuals with access to the Internet can tap a rich collection of education-related information, including:

> Information about key Department of Education initiatives, such as GOALS 2000, Technology, and School-to-Work Programs
> Full-text publications for teachers, parents, and researchers
> Directories of effective programs, exemplary schools, information centers, and sources of assistance
> Student financial aid information, and more

The National Library of Education serves the educational community through three areas: the Reference and Information Services, Collections and Technical Services, and Resource Sharing and Cooperation.

Reference and Information Services – a one-stop-shop that responds to telephone, mail, electronic, and other inquiries for education information:

Provides service to all customers – government, public, and international.

Specializes in search and retrieval of electronic databases; document delivery by mail and fax; research counseling, bibliographic instruction; interlibrary loan services; legislative reference services; and selective searches.

Answers the 1-800 toll free service, serves walk-in patrons, responds to mail, e-mail, and facsimiles. Typically, this division receives between 400-500 phone calls a month and 200 to 300 letters a day.
Mails out education publications published by all components of the U.S. Department of Education upon request. The department's publications are prepared in different formats for better accessibility. There are bilingual versions, alternate formats for people with disabilities, and also online accessibility for most publications.

Collection Development and Technical Division directs the acquisition, preparation, and assessment of all collections in all formats. This division within NLE prepares several hundred thousand monographs, a thousand journal titles, many databases, legislative reference materials, textbooks, and rare materials. The U.S. Department of Education staff may check out materials except rare books and archive collections from the Library. The loan period is 30 days and renewal is for 14 days.

Resource Sharing and Cooperation Division develops and maintains a network of national education resources. Major activities include: OERI Toll-free Electronic Bulletin Board System, the award winning U.S. Department of Education's web site which contains more than 21,000 files and receives approximately 5 million hits each month. This web site offers cross site indexing across approximately 170 department of Education sponsored web sites and management of Educational Resources Information Center (ERIC).

ERIC

The ERIC system encompasses the world's largest and most frequently used education database as well as a network of 16 subject-specific clearinghouses, 11 adjunct clearinghouses, one affiliate clearinghouse, and three supporting service components. ERIC is sponsored by the U.S. Department of Education, Office of Educational Research and Improvement, and is administered by the National Library of Education. ERIC has been an important component of the national education dissemination system for more than 30 years, ensuring that education information reaches those who need it, including teachers, administrators, parents, and students.

The ERIC database is the world's largest education database. Created in 1966 to capture and make available the "fugitive" education research, the database now includes nearly one million records. Each year, ERIC adds more than 30,000 records to the database. ERIC now has acquisition arrangements with more than 2,100 organizations that submit documents for the database.

The database is available in print, online, and on CD-ROM. There are now five online and six CD-ROM vendors who offer access to the entire ERIC database or portions of it. More than 1,000 institutions in 27 countries around the world provide access to the microfiche collection of full-text ERCI documents; electronic document delivery also is available for many of the more recent documents.

AskERIC

AskERIC is a personalized, Internet-based service that provides education information to teachers, librarians, counselors, administrators, parents, and others throughout the United States and the world. AskERIC began in 1992 as a project of the ERIC Clearinghouse on Information & Technology at Syracuse University. AskERIC answers approximately 1,400 questions each week and draws from the resources of the entire ERIC system and many other sources. Anyone needing the latest information on special education, curriculum development, or other education-related topics can simply "AskERIC" by sending a request to askeric@askeric.org. Information specialists send personal E-mail responses to questions within two working days. Responses include a list of ERIC citations that deal with the topic, relevant full-text materials, and referrals to organizations and other Internet resources for additional information.

Anyone wishing to search for answers to education questions will discover an abundance of electronic resources at the AskERIC Virtual Library (http//www.askeric.org/virtual). These resources include lesson plans, AskERIC InfoGuides, ERIC Digests, education listserv archives, and much more. The ERIC database also can be searched online from the AskERIC Web site.

AskERIC Listserv Archive – ERIC fosters dialog and information exchange through the creation and administration of electronic discussion groups. More than 40 listservs currently are managed by ERIC Clearinghouses. A list of ERIC-sponsored listservs with links to subscription information is available on the ERIC systemwide Web site.

The AskERIC R&D team includes professional staff members and Syracuse University graduate and undergraduate students who use cutting-edge technology to help AskERIC bring high-quality information services to the education community. Current R&D projects include developing real-time distance education over the Internet as well as developing customized searching tools and streaming audio and video.

Online ERIC Document Delivery. The ERIC Document Reproduction Service (EDRS) Web site (http//www.edrs.com) allows Internet users to search the ERIC database of recent, copyright-cleared documents.

Listservs. ERIC serves as a catalyst in fostering dialog and information exchange through the creation and administration of electronic discussion groups. More than 40 listservs are currently managed by ERIC Clearinghouses, including early childhood education, elementary and secondary school administration, and school library and media services. The ERICNews listserv provides subscribers with bimonthly updates on new ERIC publications and services.

Special Projects

ERIC Clearinghouses and support components bring creativity to the ERIC system through a number of special projects.

Education Resource Organizations Directory. ACCESS ERIC assists the U.S. Department of Education by maintaining the database for the Education Resource Organizations Directory located on the Department's Web site. The Directory enables Internet users to search more than 2,100 national, regional, and state organizations, including information centers; comprehensive and technical assistance centers; and many other types of programs, services, and organizations.

ERIC Search Wizard and Expert Searches. This state-of-art search engine developed by the ERIC Clearinghouse on Assessment and Evaluation allows users to select terms from the Thesaurus of ERIC Descriptors to build effective, high quality searches. The Wizard features seamless online ordering and readily available information on journal and document sources.

Gateway to Educational Materials (GEM). The NLE is spearheading a consortium effort, GEM, which is a special project of the ERIC Clearinghouse on Information and Technology. The goal of GEM is to create an operational framework that will provide the key to "one-stop, any-stop" access to the thousands of lesson plans, curriculum units, and other educational materials on the Internet. To accomplish this, GEM created the K-12 metadata standard for describing educational resources. GEM also provides software, training, and support so that GEM consortium members with Internet-based collections can easily use GEM to describe their resources. These descriptions are assembled in the Gateway Catalog which went online in February 1998, and currently includes more than 2,000 records. The catalog contains links to the materials and so creates easy access regardless of where the materials reside on the Internet.

Virtual Reference Desk (VRD) The VRD is a new project creating the foundations for a national cooperative digital reference service. The project is sponsored by the National Library of Education and the ERIC Clearinghouse on Information and technology, with support from the Office of Science and Technology Policy.

- **Resources of the Virtual Reference Desk:** AskA+ Locator. The AskA+ Locator contains over 70 quality online expert services that answer the questions of the K-12 community. Digital reference services, also called "Ask-An-Expert" services, are Internet-based question and answer services that connect users with individuals who posses specialized subject or skill expertise. As opposed to static Web pages, digital reference services use the Internet to place people in contact with people who can answer specific questions and instruct users on developing certain skills.

- **Software Development and Certification**: The Knowledge Base is Internet software that allows students, teachers, experts, and others to search across AskA services archives for answers to previously-asked questions.
- **AskA Starter Kit**: This instructional resource guides organizations in the development of new AskA services in their areas of expertise by providing how-to-advice and methods based on experiences of exemplary services and in-depth empirical research.

- **Organization**: The Virtual Reference Desk seeks to identify and provide the resources necessary to link all K-12 community members to necessary expertise in order to satisfy information needs.
- **AskA Consortium and Standards Development**: The Virtual Reference Desk seeks to aid organizations of all types in the creation of a AskA services and offers resources and guidance to those interested in building and maintaining quality digital reference services.

Test Locator. Test Locator (http//www.ericae.net/testcol.htm) describes more than 11,000 assessment instruments and their availability.

National Parent Information Network (NPIN). NPIN is a special project of the ERIC Clearinghouse on Elementary and Early Childhood Education and the ERIC Clearinghouse on Urban Education. Begun in 1993 as an Internet site for parents (http//www.npoin.org), NPIN services have since expanded to provide information via E-mail with the AskERIC program, through workshops, and through its toll-free telephone number (1-800-583-4135).

Native Languages Project. The ERIC Clearinghouse on Rural Education and Small Schools conducted a special project to develop and publish a list of courses offered in American Indian and Alaska Native languages at postsecondary institutions. More than 50 different American Indian and Alaska Native languages are listed, along with contact information for the colleges and universities where each language is taught.

Virtual Libraries. Several ERIC Clearinghouse Web sites offer virtual libraries of full-text documents in their topic areas. For example, The ERIC Clearinghouse on Counseling and Student Services has developed 11 Web-based virtual libraries on career development, cultural diversity, and many more educational topics.

United States Network for Education Information (USNEI). USNEI was created in 1996 to meet the information needs of all who are involved in international educational mobility, including parents, students, educators, advisers, institution, and organizations. USNEI provides a central reference point for information on the United States and foreign education and also refer persons to the correct authority to help with the specific questions and assistance.

USNEI also serves as the official information service for the United States under the terms of international agreements involving the provisions of education information. USNEI is the National Education Information Center under the terms of the 1996 Convention of the Recognition of Qualifications Concerning Higher Education in the European Region.

While the management center for USNEI is housed at the NLE, USNEI is actually both an Internet presence and a distributed referral service spread across the United States and with contacts around the world in the United States embassies and consulates.

The NLE is creating a library that is virtual and collaborative, proactive and responsive, and creative and visionary while maintaining the highest standards of customer service.

3.0 ANALYSIS OF THE INITIATIVES

The definition of a digital library differs from agency to agency. It was necessary to begin, if not explicitly then implicitly, to identify a digital library versus a virtual library. A digital library is defined as taking a collection and making it digital. A virtual library means bringing digital collections from different distributed sources under a central umbrella. The main finding of the review of the initiatives is that there are different approaches to the concept of a digital library. Some agencies are focusing on the internal creation of electronic information (including scanning and OCR technologies); other are focused on the sociological aspects of developing collaborative communities. While there are technologies involved, the emphasis tends to be on the provision of tools to allow others to contribute or to provide tools for users to access the information provided by others.

The digital library includes several key components:

collection development
search engines
indexing/categorization
metadata
profiling/user push technologies
archiving and preservation
digitizing technologies and the workflow to support them
personnel issues

Virtual libraries also add the following components:

preservation of URLs and persistence
metadata standards
distributed search engines
business and intellectual property concerns
sociological and cultural issues

While most of the digital-virtual libraries include these components, the CENDI agencies emphasized different aspects of their projects during their brief presentations. For example, NAIC emphasized its input processing system involving scanning, OCR, and machine translation. DOE emphasized the move into broader dissemination of DOE information via its Information Bridge and EnergyFiles projects. NASA highlighted the vast number of currently undocumented NASA web sites and the plans to bring them under one umbrella, acknowledging the cultural difficulties in doing so. NLM emphasized its infrastructure (the Entre search engine, the Unified Medical Language System, the MeSH thesaurus, and its relationships with publishers), and its involvement in the DL-2 digital library research initiatives. NAL focused on community building and on its role in digital archiving discussions. USGS/BRD also emphasized community building and the provision of tools to support the creation, dissemination, access, re-use, and exchange of biological information. DTIC emphasized its repository architecture, particularly the development of “handles” for persistent location and the security aspects needed to provide a restricted system.

The agencies are in differing stages of development. The NASA STI program has a strategic plan for bringing order to the multitude of Web sites created independently by NASA centers and their contractors. NAIC is weeks away from initial implementation of its Digital Library Input Processing System (DLIPS) which support full text capture, editing, translation, automatic indexing, and product development with a minimum of human intervention. DOE has a Virtual Library (EnergyFiles). A major component, or the public reading room for DOE information within the library, is the DOE Information Bridge, the digitized full text collection of technical reports. NTIS is integrating selected niche markets for technical and business information that will integrate with its document ordering capabilities. DTIC is working with several collaborators on the Defense Virtual Library. This is a continuously evolving system. NLM has advanced in a particular resource type area, integrating its databases (both bibliographic and data) with the full text at publisher sites. USGS/BRD has the National Biological Information Infrastructure (NBII) under its belt and is now seeking to build a more robust virtual library with the NBII-2. NAL has the AgNIC system and a ever growing number of expert centers developing to support community development, extending the resources well beyond its own ability to produce them.

The resource types and formats that these efforts address range from the traditional bibliographic citations accessible via the Web to data sets, lesson plans, gene sequences and satellite imagery. Most have moved forward from the early emphasis on full text or images of textual documents to multimedia, sound, and numeric data. DTIC is working with the Library of Congress on better guidelines for metadata to support photos, sound bites, and multimedia.

Some agencies are adding digital library services that go beyond the collection and access. NAL provides a digital reference (or AskA service) to provide users access to reference staff or scientists. DOE and NAIC provide real-time machine translation. Downloading with special emphasis on manipulating the resulting file is available through NTIS's Encapsulator program. USGS/BRD is working on modeling software and distributed vocabulary support for searching. Through its use of the Federal Geographic Data Committee and National Spatial Data Infrastructure standards and practices, BRD is able to provide geographic information system (GIS) searching through its clearing house mechanism. NLM is using its Unified Medical Language System to provide vocabulary support in a middleware layer.

With the exception of NAIC and DOE, which have implemented or are implementing wholesale replacements of their input processing system for text in the DL environment, the DL efforts among the agencies are outside the mainstream production system. Part of this is because they are pilot projects or projects done in cooperation with others. However, the bibliographic databases are not being left behind in this environment. USGS/BRD, which has no legacy system for publications or a bibliographic database, is using a distributed system to create a bibliographic database (metadata) that can front-end its publications system. In most cases, the bibliographic databases are being moved to the web (most recently AGRICOLA from NAL). The agencies are devising strategies for integrating the bibliographic databases with these new environments. The legacy database and ordering systems of NTIS are key to their strategy of connecting new document types, like standards and solicitations, to other types of material and services. NLM has successfully integrated its MEDLINE database with gene sequence data and with links to publisher information.

Based on the varying approaches, status of development and research interests, certain agencies have strengths in certain areas. These strengths can be shared with others.

- Scanning and OCR - NAIC, NLM, DTIC
- Collaborations within the agency - DOE, DTIC
- Collaborations outside the agency - NLM, BRD, NTIS, DTIC
- Metadata guidelines for non-textual material - DTIC, BRD
- Metadata for textual information - NAL, DOE
- Geospatial information - BRD
- Preservation and archiving - NAL, NLM
- URL persistence - DTIC
- Digital reference - NAL
- Document ordering and electronic commerce - NTIS, DOE
- Remote experimentation/modeling - DOE, BRD
- Machine translation - NAIC, DOE implementation

In addition to reporting on the current status of digital-virtual libraries and future plans, the agencies highlighted challenges that they had encountered. These are indicated as research needs:

- Economics and funding models for digital-virtual libraries
- Distributed searching
- Metadata crosswalks
- Distributed vocabularies that can be integrated
- Tools for profiling and customizing collections
- User usability and evaluation methodologies/metrics
- Cultural change (among collaborators, users, and agency staff)

With the advent of digital-virtual libraries, their integration with legacy systems, and the advent of new types of services, the agencies reported an unexpected challenge. Digital and virtual libraries have sociological and cultural implications that were not anticipated by the majority of the agencies. These include the training needed to move legacy staffs into this new environment, the personnel needed to handle increased collaborative efforts (some of which involve complex agreements and licensing issues), the limited technical resources available to support digital library development, and the possible need to reorganize for new ways of doing business.

4.0 RECOMMENDATIONS

Based on the discussions, the Digital Library Initiatives Task Group recommends the following actions:

- Through CENDI or its Working Groups, promote one or more of the research needs identified in Section 3.0.

- Update the scanning/OCR technology tables that were done for the Scanning and OCR report to reflect upgrades in equipment as a means of sharing information about these rapidly changing technologies.

- Produce an “areas of expertise” inventory (specific to DL), as an extension of the highlights presented in section 3.0 above.

- DTIC to distribute its guidelines for image and sound metadata.

- Consider how the CENDI agencies might contribute testbed material or otherwise be involved in the February 1999 round of digital library research initiatives

- NAIC invited other agencies to attend its DLIPS demo at DTIC on July 30

APPENDIX A

DISCUSSION QUESTIONS FOR THE DIGITAL LIBRARY INITIATIVES WORKSHOP

Questions for CENDI Digital Library Presentations
June 30, 1998

Please prepare a brief description, with particular attention to the architecture and organization aspects of the DL project. We discussed the metadata in detail at the meeting last year. You don't need to answer all these questions. Remember that we are looking for lessons learned and where we can collaborate and cooperate with one another.

GENERAL DESCRIPTION

What digital library initiatives do you have underway at this time?

Who is the audience?

What kinds of materials are provided: bibliographic records, full text, images, sounds, full motions, etc.?

What is the scope? How is content acquired? Is content all from within the STI program, your agency, others?

How is the DL maintained? How is new information added? How is old information updated?

ORGANIZATION OF THE PROJECT

Is this a collaborative effort with others?

Who was involved in the development of the DL? What kinds of skills were included in the team?

SYSTEM ARCHITECTURE

What is the architecture of the system? What is centralized and what is distributed? Where do the indexes reside? Where do the resource files reside?

How is the content accessed? A central search engine? Web browser access only? Distributed search engines for each resource or connected server? Is it based on a search protocol such as Z39.50?

What kinds of tools are provided? Is there consistent subject indexing or a categorization scheme to organize the library?

What kinds of standards are you using? Is there metadata involved? (Don't spend much time on this since we did a review of the metadata formats last year.)

What kinds of services are provided?

What kind of support/help is provided to the user?

CHALLENGES/ISSUES AND THE FUTURE

What have been the challenges and issues involved in doing the project?

What is your vision for the future and how will it grow from where it is now to what it will become?

Where could you use help in this endeavor? What could CENDI do to support your effort?