AIAA 97-0467

Desktop Access to Full-Text NACA and NASA Reports: Systems Developed by NASA Langley Technical Library

Manjula Y. Ambur and David L. Adams
NASA Langley Research Center
Hampton, VA

P. Paul Trinidad
Computer Sciences Corporation
Hampton, VA

35th Aerospace Sciences
Meeting & Exhibit
January 6-10, 1997 / Reno, NV

DESKTOP ACCESS TO FULL-TEXT NACA AND NASA REPORTS: SYSTEMS DEVELOPED BY NASA LANGLEY TECHNICAL LIBRARY

Manjula Y. Ambur and David L. Adams
Information Services and Systems Division
NASA Langley Research Center
Hampton, VA

P. Paul Trinidad
Computer Sciences Corporation
Hampton, VA

ABSTRACT

NASA Langley Technical Library has been involved in developing systems for full-text information delivery of NACA/NASA technical reports since 1991. This paper will describe the two prototypes it has developed and the present production system configuration. The prototype systems are a NACA CD-ROM of thirty-three classic paper NACA reports and a network-based Full-text Electronic Reports Documents System (FEDS) constructed from both paper and electronic formats of NACA and NASA reports. The production system is the DigiDoc System (DIGItal DOCuments) presently being developed based on the experiences gained from the two prototypes. DigiDoc configuration integrates the on-line catalog database World Wide Web interface and PDF technology to provide a powerful and flexible search and retrieval system. It describes in detail significant achievements and lessons learned in terms of data conversion, storage technologies, full-text searching and retrieval, and image databases. The conclusions from the experiences of digitization and full-text access and future plans for DigiDoc system implementation are discussed.

INTRODUCTION

With the expansion of on-line catalogs to users' desktops, enhanced databases provided through local on-line systems, and the prolific use of Internet resources, scientists and researchers are increasingly using the Library from their office computers. The researchers' awareness of electronic sources is increasing along with their demands for full-text elivery of information to their desktop workstations. Users are expecting not only access to information about information (metadata or bibliographic data) but the ability to access, view, print, and manipulate the full text of information. During these exciting times of evolving information technologies, libraries are faced with opportunities and challenges of providing information services in a cost-effective manner while maximizing the use of desktop delivery.

In 1991, the Langley Technical Library developed a strategic plan with the goal of working towards building the Electronic Library by maximizing the desktop access to scientific and technical information. A major component of this plan was to explore technologies of data conversion, full-text search and retrieval software, optical storage, network access, and client-server architectures to digitize and electronically disseminate NACA and NASA technical reports.[1] The purpose was twofold: digital archiving and preservation of deteriorating documents, and enhanced access to a critical body of knowledge.

The collection of NACA documents at Langley runs from the years 1917 to 1958 and is actively used by researchers. In their paper form these documents may become less available due to aging, decaying paper, and long usage; that makes it imperative that they be archived in a longer-lasting digital format. The electronic reports system's flexibility and added functionality of searching, accessing, viewing, and manipulating the full text of documents will provide researchers with new tools to analyze information and help them to formulate new relationships and ideas. Also, the Langley researchers stressed their need and desire to have desktop access to the full text of NACA and NASA reports during Library outreach programs and focus group discussions conducted to help formulate short-term and long-term strategic plans.

Though the basic goal has been desktop access to full-text NACA and NASA reports, the following specific objectives were formulated:

* Reports must be complete, with data and format integrity including all graphs, pictures, charts, and scientific characters and equations.
* Full ASCII text of reports with searching and text-manipulation capabilities
* Hyperlink capability within a document and across documents
* Linkage of digital reports to an existing metadata database system
* A single system for all paper and electronic documents
* A client-server configuration supporting various types of clients and computer platforms

OVERVIEW

To establish the practicality and feasibility of desktop delivery of full-text technical reports to NASA Langley researchers, the Library developed two prototype systems between 1991 and 1994. The first was a CD-ROM of NACA reports and the second a Unix-based Full-text Electronic Documents System (FEDS). During 1995 and 1996, the Library has configured a production system integrating the on-line catalog, a World Wide Web (WWW) interface, and Adobe's PDF technology. The main intent of all these projects has been to design and develop a cost-effective and workable full-scale integrated information system for storage and dissemination of NACA and NASA reports to a wide audience.

NACA CD-ROM

From 1991 to 1992, the Technical Library developed a single prototype CD-ROM that contained thirty-three classic NACA reports. This prototype was funded with a research grant of $25,000 from the Center's Director's Discretionary Fund that is administered by the Chief Scientist. The main objectives were to investigate in-house development of an archival full-text NACA database using CD-ROM technology and to build the staff's expertise in optical publishing and data conversion. Both ASCII text and image databases were developed by the digitization of paper documents. The NACA CD-ROM was developed using KnowledgeSet software on a PC platform This experimental system was organized into three separate databases. The first was the full text of the thirty-three NACA reports searchable by every word in the database. It contained SGML tagging of elements and hyperlinks to references and figures. The second consisted of scanned images of the reports with a metadata searchable only by author, title, and keyword in the abstract. The third database was made up of one hundred significant graphs, tables, and line drawings extracted from the reports, searchable only by the figure title and key words from the citation.

The NACA CD-ROM was tested by users in the Library and in a research branch. They liked the power of full-text searching and possibilities of full-text desktop delivery. However, the cost of ASCII conversion was expensive, and the KnowledgeSet software was found to be unsuitable for an image database. The PC-only access was also very limiting for wide dissemination. A fuller description of this project is given below in the Technical Details section.

FULL-TEXT ELECTRONIC DOCUMENTS SYSTEM (FEDS)

With the expertise gained from the NACA CD-ROM, FEDS was developed during 1993 and 1994. This was based on a client-server architecture on a Wide Area Network. The main objective of FEDS was to develop a prototype system of approximately one hundred full-text NACA and NASA reports that exist in paper and/or electronic format and to provide Langley researchers desktop access from all three major computer platforms. This prototype was also developed with a grant from the Chief Scientist of about $80,000.

The FEDS system was developed using a Unix server and Interleaf's Worldview software. The system consisted of two databases: one of NACA documents scanned from paper and the other of NASA documents in electronic form. The paper documents were scanned as TIFF images, and ASCII text was hyperlinked to the images. The ASCII text was created by using optical character recognition (OCR) on the images. Users could access both the images and text, with the images provided for viewing and printing and the ASCII text for searching and for cut-and-paste. The electronic files came from TeX as Postscript files and were loaded into WorldView, a WYSIWIG viewer that was to provide both format and data integrity of the file. Unfortunately, problems with font incompatibilities in the conversion process made it difficult to detect integrity problems with the scientific characters and equations.

The OCRed text for full-text searching was proven to be adequate, though the ASCII file did not look good visibly. The cost of the client software for wide use seemed prohibitive. Also, linkage of full text and images of reports to a database using Interleaf's RDM software was found to be infeasible. The FEDS system was tested by researchers, and they liked the network access and full-text searching. They also wanted linkage to the Library's widely used on-line catalog system. A fuller description of this project is given below in the Technical Details section.

DIGITAL DOCUMENTS SYSTEM (DIGIDOC)

With the expertise gained from the NACA CD-ROM and FEDS, the Library proceeded to configure a production system called DigiDoc. During 1995 and 1996, the Library configured a system to take full advantage of the existing on-line catalog database, Adobe PDF technology, and the popular WWW. The vendor of the Library's STILAS on-line catalog announced their WWW interface in 1995, simplifying the project.

A high-speed microfiche digitizer was purchased with money from over-budget guidelines, with the goal of digitizing microfiche reports in lieu of sending copies of microfiche to researchers. Most of the library's NASA reports are on microfiche, and researches find the microfiche very user-unfriendly, and in some cases, completely useless. Also, during this time period, the NASA Aeronautics Centers (Ames, Lewis, Langley, Dryden) agreed to collaborate in many daily operations, including library and information services. A team chartered to look at ways to share information resources recommended that all NACA and NASA reports be digitized and provided for desktop delivery. The Langley Technical Library was designated to lead this project. As part of this recommendation, extra money was provided to invest in the necessary server and storage hardware for the system.

All these factors provided the needed impetus to develop a full-production model digital system. Phase I with one hundred reports was finished in June 1996. TIFF images from scanned paper reports and electronic files from TeX, Word, and other word processors were converted into Adobe's PDF image-plus-text format. Documents are linked to the on-line catalog by URLs in the metadata records of the on-line catalog.

Various versions of Adobe formats were experimented with and the image-plus-text format was chosen for its format and content integrity. This format provides a single file that combines the scanned images for viewing with OCRed ASCII text for searching. These and other features of PDF and Adobe's Acrobat viewer software provide the maximum text and image functionality from all computer platforms.

The prototype, or Phase I of the project, was demonstrated widely to senior managers and technical experts. The configuration and technical aspects were validated and support was given for further populating the system. Populating the system with a critical mass of documents is expensive and time-consuming, so various funding and data conversion strategies are being considered. Though the architecture of the system is open and incorporates widely used technologies, the integration of the various pieces of software and hardware is a challenge and still requires some work and fine-tuning for the necessary robustness of a large, full-text database. During 1996, the NASA STI Program Office has started to digitize NASA paper documents from the last five years. To maximize synergies and avoid duplication, we are collaborating with STI Program in building and disseminating NACA and NASA digital reports. A fuller description of this project is given below in the Technical Details section.

TECHNICAL DETAILS

This section gives the technical details of the three projects. Each project is discussed with goals and objectives, methodology used, hardware and software configurations, lessons learned, and significant successes outlined.

NACA CD-ROM

Goals and Objectives. During 1991, realizing the importance for experimenting with development of in-house, full-text databases, the Library researched the literature and decided to produce a prototype CD-ROM of selected, classic NACA reports. During this time period, CD-ROM technology was extensively used for publication and distribution of commercial databases and for archiving historical information. The American Memory project at the Library of Congress is a good example.[2] The main advantages of CD-ROMs were a high-storage capacity, its storage permanence of about 200,000 pages of text or 20,000 pages of images that cannot be accidentally erased, and its low cost and fast replication make it feasible for in-house publication and mass distribution.[3]

The main objectives that were set for this NACA CD-ROM were

* To explore optical technologies as a medium for

archiving and converting the paper-based NACA reports
* To investigate and experiment with methods for improving the availability of and access to information contained in the NACA reports
* To understand the conversion methodologies and associated technical and cost issues for conversion of paper documents to images and text
* To examine full-text indexing and retrieval software as a means of accessing text and images
* To develop in-house experience and expertise in the areas of optical disk technology, full-text database design and development, user interfaces, data conversion, and electronic publishing

Methodology The Technical Library decided to pursue funding from the Director's Discretionary Fund as a means for producing the prototype CD-ROM. This fund is administered by the Center's Chief Scientist as a source of funding for high-risk, high-payoff research that falls outside the normal research areas at the Center. In March 1991, after reviewing the literature on optical publishing and discussions with vendors, the Library prepared and submitted a proposal to Chief Scientist for $25,000 to test the feasibility of using optical storage to archive the NACA technical reports collection and make the full text of these reports more accessible to NASA Langley researchers.

The funding was approved in July 1991, and it was decided to use a small business contractor for data conversion, database build, and mastering of five copies of the CD-ROM. A team was formed in the Library to work with the contractor on all phases of the project from analysis, data conversion, database design, and user interface design to building and indexing of the database and CD-ROM production. This approach was used to take advantage of the contractor's expertise in optical publishing and Library staff's expertise in information retrieval techniques and user information needs. Based on research of the available optical publishing and full-text indexing software, it was decided to use KnowledgeSet because of its extensive full-text search capabilities, hyperlinks, and proven CD-ROM publishing record. The contractor selected to work on the project was Subsystems Technologies Inc. of Rockville, Maryland.

The Library staff selected thirty-three highly used NACA reports consisting of approximately 2,500 pages. Together, the contractor and the Library team identified all data elements, such as author, title, keywords, subject terms, table of contents, paragraph headings, etc. that were associated with this collection and created relationships for an efficient database design. Data conversion was the most critical, time-consuming, and expensive part of the project. It was decided to create both text and image databases of the reports to evaluate the costs, technology, and user-access factors associated with each option.

It was specified to the contractor that text files were to be 99.95% accurate. To create the text files, OCR was tried and found to be unsatisfactory, as the recognition rate was about sixty percent. The contractor evaluated and compared the costs of OCR and the associated editing needed to make it accurate to that of manual keying of the entire document. Based on this cost analysis, the contractor created ASCII text files with a double-keying process. In the double-keying process, each report was keyed in twice by different people. The two versions were then compared for error detection and correction, based on the assumption that the same error would not be made by both persons. Even with this approach, it was found that scientific equations and Greek characters in the reports could not be accurately represented and displayed. As a result, all the equations were scanned as images and hyperlinked from the text pages, even in Full-Text database. Images were scanned at 300 DPI resolution and saved as TIFF Group IV files.

The NACA CD-ROM was produced in July 1992, and during 1992 and 1993 it was demonstrated to a number of researchers during the Library's outreach programs. Also, it was tested and evaluated by a group of researchers in the Experimental Flow Physics Branch, with their comments and observations recorded on written evaluation forms.

Configuration The NACA CD-ROM was developed for PC platform because KnowledgeSet software was available only for the PC, with a Mac version of the software then under development. CD-ROM was mastered, and five one-offs were produced as per ISO 9660 standard.

The database design had three sub databases: Full text, Title and Key Word, and Pictures and Figures. The motivation behind these three sub databases is to provide ASCII text of reports with full-text searching and attached images, and as a separate database images of the reports, with limited metadata searching. This provided a means of evaluating the text-only and image-only approaches to see which would be better for a document archival and retrieval system. The Pictures and Figures sub database included images of one hundred of the most important photographs and line drawings, with access by words included in the captions.

The ASCII text was marked with SGML tags to create a single Full-Text database with a hierarchical structure of subject groups, reports, and table of contents. The thirty-three reports were grouped under broad subject categories such as Aerodynamics, Structures, Test Facilities, etc., and the reports' titles in each group ware linked to the text of the report. The table of contents and paragraphs were tagged to facilitate detailed searching and linking of contents to the body of the text. Hyperlinks were also created for figures and references.

Lessons Learned Data conversion of paper documents into accurate ASCII text proved to be very difficult and expensive. The cost was approximately $4 to $5 per page to create ASCII text by double-key entry. Even with this expensive method, results were not completely satisfactory because of scientific characters and equations. Some of the Greek characters and spacing of equations could not be accommodated by KnowledgeSet software.

Users found full-text searching very helpful, and the advantages of detailed searching of reports for in-depth research were evident. Based on these experiences, it was decided that accurate ASCII conversion of paper scientific and technical documents is not practical and that inexpensive OCR should be considered only for full-text searching to complement an image database.

Though converting paper documents into images was not too complicated, there were issues with image retrieval and display. Based on the power of the PC and the size and resolution of its monitor, image retrieval and display was satisfactory to poor. Printing of the scanned images provided excellent quality reprints, sometimes better quality than original paper. However, printing was slow and special accelerator boards in the printer were needed to help speed up the process. It is clear that effective use of image databases is very dependent on end-user workstation configurations.

We also gained an understanding of the differences between a structured and full-text database. The KnowledgeSet software is well-suited for full-text databases with powerful search capabilities such as boolean operators, proximity and wild card searching, and text manipulation capabilities such as ability to append notes to a paragraph or bookmark a paragraph or a page. KnowledgeSet did not, though, seem suited for fielded searching of a metadata record or for image databases with capabilities of consecutive page retrieval and display.

The PC-only access had limitations even in terms of user testing, as significant number of researchers use Macintosh and Unix workstations. It was evident for that for wide desktop delivery, access has to be made available from all platforms and the preferred access vehicle would be the Langley-wide TCP/IP network, LaRCNET.

Successes The NACA CD-ROM project has proven to be a great starting project for in-house development of full-text databases and has helped the Library staff to gain hands-on experience and expertise in the areas of CD-ROM development, data conversion, and database design. In the process of obtaining the funding and during testing, the Library also gained visibility and appreciation from management and researchers for its effort to move towards desktop delivery of full-text information. The project has also proven the power of access to digital reports and full-text searching and helped the Library to move forward as it initiated and got funding for FEDS (Full-Text Electronic Documents System) that was developed from 1993 to 1994.

FEDS

Goals and Objectives A second experimental project was undertaken in 1993. Called the Full-text Electronic Document System, or FEDS, the main objective of this project was to test a full-text document system in a networked environment, serving different types of computer platforms from a central server over a Wide Area Network. While the NACA CD-ROM project dealt with only paper documents, FEDS was a little more ambitious.

The major goals of the FEDS project were

* To determine a unified approach for displaying paper and electronic documents
* To be able to use electronic documents created in TeX by the Center's Technical Editing Branch
* To be able to use word processor documents created by the Center's researchers
* To review any new OCR technology developed since the NACA CD-ROM project

Since it had been two years since the NACA CD-ROM project, it was decided to revisit OCR to see what advances may have been made.[4] A new conversion technology, called

Intelligent Character Recognition or ICR, was promising better results than the older OCR software had given.

Methodology As with the NACA CD-ROM project, a proposal was made to the Chief Scientist's office for funding. An award of $80,000 was received in April 1993. A team, with members from the Library and Technical Publishing branches, was put together to work on the project.

A solicitation for sources was placed in Commerce Business Daily in July 1993, outlining the system's intended features and tasks that would be required of the contractor and requesting a broad description of how the contractor would propose developing the system. Approximately twenty proposals were received. The proposals were reviewed by the team, and based on technical merit and existing contracts at NASA Langley, Symbiont Inc. of Washington, D.C. was selected as the contractor.

In addition to soliciting proposals, members of the team toured various government and commercial facilities in the Washington, D.C. area to see how others were handling some of the issues they would encounter in the FEDS project. This, along with the proposals, gave the team a good overview of the state of the technology, as well as an idea of which companies were working in the field. The team also put together a matrix to be used by the contractor in the evaluation of software packages and their features. This was an all-encompassing list of required as well as desired features for the FEDS that would form the basis for selecting the software package or packages with which to implement the system.

Configuration Based on the review by Symbiont, Interleaf's document management package was selected as the software that provided the best functionality and most features for implementing FEDS. This package included a conversion program to import text and graphics files, a document indexing package called WorldView Press, and a viewing client, called WorldView, that worked on PC, Macintosh, and Unix platforms. Hyperlinks could also be placed within and between the documents themselves.

The computer platform used for this project was a Sun Microsystems SPARC 10 Model 41. This system had the ability to run either Sun OS or Sun Solaris operating systems and allowed telnet as well as X-client sessions to access the FEDS database. The computer was both a server and development system for the FEDS project.

To test the capabilities of OCR and ICR technologies, the contractor tested various commerical software packages at their offices in Washington. These were run on 486 class personal computers.

Lessons Learned. It was determined early in the project that the state of OCR technology had not changed much and that OCRed text could not be a substitute for a scanned image. Although some of the advanced ICR packages could learn or be taught to recognize certain characters, this worked only when the typefaces were all the same and paper and print quality were uniform. Depending on the quality of the paper and print, the OCR would be anywhere from seventy-five to ninety percent correct. However, even with the best recognition rates, too much information was being lost in the OCR process, so it was decided that any scanned documents would be presented as images. The OCRed ASCII text would be supplied as is only for rudimentary full-text searching within the document and for cutting and pasting small sections of text.

The use of files from electronically generated reports was not as easy as anticipated and proved to be the largest problem of the project. Even though most people write their reports and papers on a word processor, these are often not complete with all charts, tables, etc., and require editing and formatting before final publication. To get the most complete and correct version of reports for this project, PostScript files were obtained from the NASA Langley Technical Publishing Branch. These files were the same ones used to produce camera-ready copy for publication. However, scientific characters and ligatures, a typographic and printing convention where two characters are joined together, in these documents caused conversion problems when imported into the Interleaf software. With ligatures, the first letter converted correctly in most cases but the second letter in the pair would be converted into an "i." Scientific characters would be converted to random letters or symbols, depending on what font set the conversion program thought was being used. Unlike the errors from the OCR process, these conversion errors were not as obvious, and each document had to be carefully reviewed to find all of the errors. For all characters to be converted correctly, the original fonts would have to be loaded into the conversion program and sometimes hand-mapped one for one. Because of the wide variety of characters and symbols used in these equations, this proved to be very difficult. In the end, these documents were scanned from paper with the electronic text used only for rudimentary searching and cutting and pasting of text.

It was found that images of scanned documents could be converted and loaded into the InterLeaf program with relative ease. However, entering the metadata, importing the images, and linking the two took time. A database system called RDM, based on Oracle software, provided metadata searching in the Interleaf software suite. The RDM database proved to be unsuitable for variable-length metadata fields of the kind used in this project and was rather expensive. It was decided not to use RDM and to find another way of providing an index to the documents. For this, a document was created using the WorldView client as a pseudo-index for searching. This was simply a document of all metadata that were searched via keyword. Although some of these problems stemmed from the Interleaf software and others from the nature of the metadata, they reinforced the belief that any production full-text document system should utilize an existing database of metadata, not duplicate it.

Successes The FEDS project was a client-server system accessible from the three popular computer platforms that provided access to complete documents from a user's desktop computer.

This was a very good project for the understanding of multi-platform and network access issues. It also gave the Library additional experience with OCR technology and the issues of electronic documents in cross-program situations.

DIGIDOC

Goals and Objectives. The lessons learned from the two prototype projects were applied to planning of the DigiDoc system that began in 1995. DigiDoc was intended to be scalable into a production full-text document system that is usable from a researcher's desktop computer. Based on experience gained from the two experimental projects, it incorporates the latest technologies for network access and database searching, as well as being scalable to a size limited only by storage capacity and network bandwidth. The initial configuration of the system is sufficient to accommodate the roughly 100,000 NACA and NASA formal reports.

Many of the goals from the FEDS project were turned into requirements for DigiDoc, with some adjustments based on experiences from the experimental projects. There were many basic design requirements laid down along with several secondary requirements. It was felt that the ability to implement these secondary requirements would depend on the technology that was used to implement the system and that their value would have to be weighed against any possible impact on the basic requirements.

The basic requirements for the system were

* To use the existing library catalog for metadata searching
* To be accessible from all computer platforms (PC, Macintosh, and Unix)
* To be able to accept and display scanned images and electronically produced documents
* To preserve the content and format of the documents, regardless of the document's source
* To be ready for use so that the user does not have to install dedicated client software to use the system
* To allow printing of documents to a user's printer
* To have a reasonable minimum end-user computer configuration

Some secondary requirements were

* To allow full-text searching within a document
* To see from all computer platforms the same user interface
* To allow copying of text from within a document to other programs, such as a word processor
* To allow direct printing of one or more pages without downloading the entire document
* To provide a means of faxing one or more pages if a suitable printer is not available

The key objective was to use existing databases to provide the searching capability for the system. It was felt that end-users would be more likely to use the DigiDoc system if it was tied to a database they already used and which had a large amount of data. There was little incentive to search several small databases for the same information. This would allow one-stop shopping for end-users as well as reduce metadata and system maintenance that would be required for a separate database. These maintenance concerns would be critical in a large production system.

Methodology. While FEDS was under development, the popularity and use of the World Wide Web exploded. Many programs were developed that took advantage of this new Internet-based medium, and its growing popularity was creating de facto standards for the formatting of text and images that were viewable on all of the popular graphical web browsers. The browsers were also being made available for the three major computer platforms; PC, Macintosh, and Unix. At the same time SIRSI Corporation, the company that wrote the Library's on-line catalog, announced that they would develop a web-based interface to their catalog product.

This growth in the World Wide Web and the availability of a web interface to the Library's existing catalog of metadata provided a means of meeting some of the basic requirements for the DigiDoc system. The two most popular browsers, Mosaic and Netscape, were available for all three major computer platforms and were gaining wide acceptance at the Research Center.

At the time DigiDoc was being developed, there were already FTP servers and web sites that allowed documents to be downloaded, mostly in a PostScript format. This concept had been proven and worked well, both within NASA with the NASA Technical Report Servers[5] and at other government and academic sites. These servers are based on the report or papers being produced electronically. However, depending on the content of the document, local practices, and the software package used to produce it, the report may or may not be complete with all charts, graphs, tables, and pictures in their correct places.

For those who had PostScript viewers or could send the files to a PostScript printer, documents could be downloaded from a report server and either viewed on the computer or printed out. However, if the documents were large, the files would be compressed to save storage space on the server and more importantly, to take less time and bandwidth to download. These compressed files could require additional software and that several additional steps be performed before the document could be viewed or printed. This was especially true if the end-user's computer was a Macintosh or PC. Large files, even when compressed, may require many minutes to download and lots of free disk storage and memory to process and view.

The DigiDoc system had to accommodate the largest possible range of end-users and computers: not just different platforms, but ones with large amounts of memory and free disk space, as well as those with minimal memory and little free disk space. It also had to be useful to both novices and expert computer users and programmers. The NASA Langley Technical Library staff felt that while the report servers provided a good service, not all users would have the knowledge or capability to use them without some additional computer support.

At about the same time that SIRSI Corporation announced they would develop a web interface to their on-line catalog, the NASA Langley Technical Library staff saw a demo of a system called WebMan being developed at NASA Johnson Space Center. WebMan is a CGI script designed to display scanned images of documents over the web in a manner that resembles "turning" pages of the document. The script allows navigation forwards and backwards from page to page, jumping directly to other pages, thumbnail images of pages, and limited zooming. These various views of the pages are achieved by storing separate images of each page at the desired size. WebMan had the advantage of not requiring any software other than a web browser and sending only those pages that were immediately needed to the user's computer.

After reviewing the WebMan script, the Library decided to write its own script that was customized to the requirements set for DigiDoc. By this time the Langley Research Center had a site license for Netscape, and it was decided to optimize the script for that browser. The basic script was completed within a few weeks and testing was underway to fine-tune the script to deal with subtle differences in Netscape on different computer platforms.

Navigation through the document was done using a button bar displayed at the top of each page. This allowed jumping to different pages, zooming in and out, and selecting full page or thumbnails, along with several other options. Zooming or resizing of the pages was achieved by using Netscape's ability to scale images. Groups of thumbnails were displayed by using the same scaling ability. Printing was done via the browser's own printing utilities, which proved to be very slow. This was traced to a possible problem with the program the Netscape used to in their printing routine to generate PostScript.

Another development that had been gaining acceptance at that time was Adobe's Portable Document Format, or PDF. The Portable Document Format is a superset or enhancement of Adobe's PostScript language. Unlike PostScript, which is for driving printers, PDF is for cross-platform compatibility and displaying of information. At the time the DigiDoc project was started, PDF's big advancement over PostScript was the inclusion of hyperlinks within a document. However, rapid development of the PDF format caused the Library to rethink the basic design of DigiDoc, and in March 1996 the scripts that had been developed were replaced with Adobe PDF files.

Portable Document Format files require a viewer that must be installed on the user's computer. This was counter to the initial concept of not requiring the user to install specialized software. However, the PDF format had been gaining popularity, and many of the popular web sites that offered documents in multiple formats were now offering PDF as one of the options. Since PDF was becoming widely used, it would not be unreasonable to expect users to have or require them to install a PDF viewer in order to use the DigiDoc system. It was felt that the functionality gained by using PDF more than made up for any initial inconvenience of installing a viewer.

With the PDF format, a multi-page document is converted into one file, and it is possible to include thumbnail images of pages as part of that file. It is also possible to jump from page to page in the document and to include hyperlinks to points within and outside of the document. Printing, which had been a problem with the original DigiDoc script, was now much faster since it was handled by Adobe's viewing software and not the Netscape browser.

An added benefit of using PDF is the ability to do an OCR of an image and have the ASCII text included as part of the PDF file. These are formats of PDF files: Image, PDF Normal, and Image-Plus-Text. The Image format retains any images as is, without doing any OCR. The PDF Normal will OCR a document, keeping as images any characters that could not be OCRed. Image Plus Text will OCR the images but place the ASCII text behind the image. Testing showed that it was still possible to have OCR errors in the PDF Normal files, so the Image Plus Text format was selected for use in the DigiDoc system. This allowed the original format of the document to be kept without loss of information or content, while allowing text searching within the document and copying of text to a word processor or other application. The ASCII text that accompanies the image is subject to the same OCR errors as would be in the PDF Normal format, but the original content is not lost since the image is what is displayed on the screen or printout. The text is provided as is without warranty.

Shortly after PDF was adopted for use on DigiDoc, a byte-serving feature was added to the format. Byte-serving is a process where the web server will send the client only those parts of a file that are necessary to complete a screen display. This required using a server capable of doing byte-serving or installing a CGI script provided by Adobe.

Configuration DigiDoc as a total full-text document storage and retrieval system has two components: the DigiDoc server and the STILAS catalog system. The DigiDoc server stores the document files and serves them to the user on request. If a report number is known, or the user wants to browse a hierarchical listing of reports on the server, they can go directly to the DigiDoc server. The STILAS catalog system provides the capability to search the metadata record of the reports. There is not a separate database for reports in DigiDoc; rather, the library catalog and other databases that reside on STILAS, or any other system, can have pointers to documents on the DigiDoc server. The bibliographic records in the STILAS system have URLs that point to PDF or electronic files on the DigiDoc server. This allows users to search a database they are already familiar with, the library catalog system, but have the ability to bring up the full text of a document when it is available.

The DigiDoc server is a Sun Microsystems SPARCenter 1000 that drives a Disk Inc. jukebox. This jukebox is configured with 457 1.3-gigabyte read-write magnito-optical platters for a total, formatted storage capacity of just over half a terabyte. A robotics system serves eight disk drives that provide read and write capability to the platters. Magnavalt software from Tracer Technologies allows the server to control the jukebox and handle groups of the magnito-optical platters as if they were regular hard drives but without some of the size limitations of a Unix file system. An Apache web server is used to provide direct access to the documents as well as answering requests made via the hypertext links from the metadata records. Hardware and software for the DigiDoc server was acquired with funds provided by NASA Headquarters as part of Project Reliance, a cooperative and resource-sharing project between the four aeronautics centers.

The STILAS catalog is a product of SIRSI Corporation. It is an off-the-shelf software package that has become one of the most widely installed library automation systems in the past couple of years. It runs on a Sun Microsystems SPARCenter 2000.

For scanning documents, the microfiche scanner mentioned above is complemented with a paper scanner. Both scanners are driven by PCs running Turboscan software from Amitech Corporation of Springfield, VA. The scanner PCs are connected to two image conversion PCs by a Microsoft Windows for Workgroups network that allows them to share file spaces. These conversion PCs run Adobe's Capture software and convert the TIFF images created by the scanners into PDF files, OCRing the text, creating thumbnail images, and optimizing the files for byte-serving all in one process. The PDF files are placed on the Disc jukebox via PCNFS software used to mount the Unix file system on the conversion PCs. Except for putting paper or microfiche in the scanners and adjusting the scanning settings, this whole process is automatic.

For documents created electronically with a word processor or desktop publishing system, the files can be transferred via a diskette or CD-ROM, captured from the web, or FTPed onto one of the conversion PCs. The Adobe Capture software will do the conversion from Postscript, ASCII, and all of the popular word processor formats.

Lessons Learned Because of the rapid changes in technology and the popularity of the World Wide Web, the DigiDoc project went from a challenging design problem to an easily obtainable reality in about a year's time. Although these changes have solved the problems of how to serve documents to users, it only quickened the need to populate and maintain a large database and document server system.

These rapid changes have also underscored the need of having a plan in place to accommodate changes in technology that affect the file formats used to store and view the documents. To this end, it has been decided to archive the original TIFF images of any scanned documents so that these images would be available as a conversion source to whatever format may supersede PDF.

Scanning of large quantities of documents is not a trivial matter and has turned out to be a more involved process than first thought. Although the process and technologies are simple, the amount of planning and work that must go into it are very large. One additional second required for scanning each page equates to over eighty work-days when dealing with seven million pages! Maximizing the speed of the whole process takes careful workflow planning, workers who are skilled at their job, and close attention to details.

Successes Almost all of the major goals set for DigiDoc were met. In June 1996 when there were about 120 reports on the DigiDoc server, the system was demonstrated to senior management and technical staff. Their reviews of the system were very favorable and encouraging.

The images and electronic text of documents are tied to an existing large database that is widely used and that Library users are familiar with. The documents are complete with all charts, graphs, and photos in their correct places within the documents. There is also a means of doing rudimentary searching of the document's content and cut-and-paste of text to other applications.

The system has a client-server architecture that provides a single means of storing and searching for paper and electronic documents. The configuration can even be extended to encompass any type of data, not just reports. This makes maximum use of current resources and does not require users to learn how to use another system.

Although the current DigiDoc configuration does require users to install some software, that is offset by the added functionality and the software's usefulness for viewing pages from other WWW sites.

CONCLUSIONS

The NASA Langley Technical Library has made considerable progress towards the goal of digitizing NACA and NASA reports. The systems configuration to meet these objectives is completed and validated, and the necessary storage system is in place. The two import issues that are being address are the fine-tuning of the input, storage and retrieval components of the system and populating the system with a critical mass of documents.

There are approximately 160,000 NACA and NASA Reports. Current plans are to refine the methods by which documents are scanned and input into the system. This includes working with the NASA STI Program and other gvernment labs to avoid any duplication of effort in scanning the NACA and NASA collections. Digitizing legacy collections is expensive, and duplication of the work would be a tremendous waste of effort and resources.

Though there is significant activity in building digital collections, large digital databases of scholarly scientific and technical information are nonexistent. Many leading universities are involved in building digital collections of scholarly information and researching the issues of document formats, organization of digital documents, searching and navigation mechanisms, and user interfaces. Collaboration among academia, government agencies, and Industry is critical for successful development of digital libraries. Many partnerships are evident both in development of the NSF/NASA/ARPA sponsored digital libraries projects and in other projects. Both Cornell University Library's CLASS project and the electronic resources project at Indiana University-Purdue University at Indianapolis's library result from collaborations with Xerox Corporation. These projects addressed several issues we have encountered with our systems regarding scanning, file formats, searching, WWW interface and user access.[6]

The challenge for the NASA Langley Library is to keep progressing towards the goal of making digital reports accessible from users workstations while being flexible enough to incorporate new technologies and approaches as they evolve.

References

1. Ambur, M. Y.; Adkins, S. L.; and Roncaglia, G. J.: *Meeting Challenges of the Information Age: An Approach by the NASA Langley Technical Library.* AIAA Paper 94-0837, Jan. 1994

2. *The American Memory Project: Sharing Unique Collections Electronically.* LC (Library of Congress) Information Bulletin, Feb. 26, 1990

3. Hallgren, S.: *Developing Your Own CD_ROM.* CD-ROM Professional, Sept. 1990

4. Nagy, G.: *At the Frontiers of OCR.* Proceedings of the IEEE, Vol. 80, No. 7, July 1992

5. Nelson, M. L.; Gottlich, G. L.; and Bianco, D. J.: World Wide Web Implementation of the Langley Technical Report Server. NASA TM-109162, Sept. 1994

6. Crocca, W. T., and Anderson, W. L.: *Delivering Technology for Digital Libraries: Experiences as Vendors.* Second International Conference on the Therory and Pratice of Digital Libraries, June, 1995. Retrieved from the World Wide Web: http://www.csdl.tamu.edu/DL95/papers/crocca/crocca.html