Web-based document image processing

Frank L. Walker and George R. Thoma
Lister Hill National Center for Biomedical Communications
National Library of Medicine
Bethesda, Maryland 20894

ABSTRACT

Increasing numbers of research libraries are turning to the Internet for electronic interlibrary loan and for document delivery to patrons. This has been made possible through the widespread adoption of software such as Ariel and DocView. Ariel, a product of the Research Libraries Group, converts paper-based documents to monochrome bitmapped images, and delivers them over the Internet. The National Library of Medicine's DocView is primarily designed for library patrons to receive, display and manage documents received from Ariel systems. While libraries and their patrons are beginning to reap the benefits of this new technology, barriers exist, e.g., differences in image file format, that lead to difficulties in the use of library document information. To research how to overcome such barriers, the Communications Engineering Branch of the Lister Hill National Center for Biomedical Communications, an R&D division of NLM, has developed a web site called the DocMorph Server. This is part of an ongoing intramural R&D program in document imaging that has spanned many aspects of electronic document conversion and preservation, Internet document transmission and document usage. The DocMorph Server web site is designed to fill two roles. First, in a role that will benefit both libraries and their patrons, it allows Internet users to upload scanned image files for conversion to alternative formats, thereby enabling wider delivery and easier usage of library document information. Second, the DocMorph Server provides the design team an active test bed for evaluating the effectiveness and utility of new document image processing algorithms and functions, so that they may be evaluated for possible inclusion in other image processing software products being developed at NLM or elsewhere. This paper describes the design of the prototype DocMorph Server and the image processing functions being implemented on it.

Keywords: DocMorph Server, Image Processing, World Wide Web, TIFF, PDF, OCR, DocView, NLM, Ariel, Internet, Speech Synthesis

1. INTRODUCTION

Document delivery by libraries and information service providers has evolved over the past two decades. Interlibrary loan has traditionally meant photocopies of journal articles being mailed to other libraries. While photocopies are still mailed to requesters, document delivery libraries have since added facsimile transmission, and more recently, Internet document delivery. The 1990s decade has seen the arrival of Internet delivery of library documents, especially with the widespread use of the Ariel™ system developed and distributed by Research Libraries Group.[1,2] Ariel has enabled several thousand libraries to do interlibrary loan electronically via the Internet. It is a technology that is faster than mail, more reliable than fax, and offers higher resolution images than possible through conventional fax. While libraries used Ariel in the first half of this decade for interlibrary loan, the second half has seen more use of the Internet for document delivery to the patron's desktop computer. DocView, a software product developed at the National Library of Medicine, helps librarians achieve the goal of delivering library documents over the Internet to the patron's desktop.[3,4]

Running on all Windows™ operating systems, DocView is client software that enables a library patron to receive documents sent by a library's Ariel system. DocView's compatibility with Ariel enables a library or document supplier to use Ariel to scan a printed document and send the resulting images directly to a patron's computer running DocView. Ariel's scanning process produces a file of bitmapped images, which are sent via File Transfer Protocol (FTP)[5] or Multipurpose Mime Email Extensions (MIME) email.[6]

DocView is capable of displaying monochrome bitmapped images in either the Group on Electronic Document Interchange[7] (GEDI) file format used by Ariel systems, or in the Tagged Image File Format[8] (TIFF). Besides displaying the received images, DocView permits the user to zoom, scroll, pan and rotate them. In addition, a user may "bookmark" pages for easy browsing or printing, and images may be copied for insertion in word processing documents. DocView also allows the user to file and organize the received documents through a built-in document management system. Finally, DocView permits the user to forward documents over the Internet to others, using either FTP or MIME email. This last feature might be useful in an interlibrary loan service where the library receiving the document may, after completing any necessary bookkeeping, forward it to a patron.

An extensive period of beta testing that lasted 2½ years confirmed that a large majority of users found that DocView had improved the delivery of documents from their libraries.[9] DocView was released in January 1998 and is freely available. Since DocView's release, over 5,000 registered users in more than 90 countries have downloaded it. A web site established to distribute DocView includes an extensive user manual, a report on the DocView beta test, and published papers related to DocView. The software can be downloaded from the DocView home page on this web site:
http://archive.nlm.nih.gov/proj/docview/project.htm.

While libraries and their patrons are benefiting from Internet document delivery through the use of Ariel and DocView, there are still problems to be solved. These include the problems of delivering electronic documents to a wide variety of user platforms. Also, document file format often becomes an issue because library patrons often do not have the requisite software for handling bitmapped image documents. To research how to overcome potential problems such as these, the Communications Engineering Branch of the Lister Hill National Center for Biomedical Communications, an R&D division of NLM, has developed a web site termed the DocMorph Server. This is part of an ongoing R&D program in document imaging that has spanned many aspects of electronic document conversion and preservation, Internet document transmission and document usage.

2. THE DOCMORPH SERVER

The DocMorph Server is intended to serve two purposes:

1. As a resource for researching document-imaging techniques and algorithms that could find application in future versions of the DocView software and other software being developed at NLM.
2. To provide platform-independent remote document image processing capabilities to the biomedical library community via the World Wide Web.

A user may upload library document images from anywhere via the World Wide Web to the DocMorph Server to be processed. After the DocMorph Server has finished processing the images, it returns the results to the user. During the DocView beta test, some testers indicated a need for several types of document conversion. Among these were conversions from multipage TIFF to single page TIFF, from single page TIFF to multipage TIFF, or from TIFF to Adobe PDF files. There was also interest in searching documents. Files produced by the Ariel system are strictly bitmapped images; they are not text-searchable unless converted to text by optical character recognition (OCR). It is possible to include this type of document conversion in the DocMorph Server. It is also possible to include other image processing algorithms for cleaning up artifacts created by the scanning process, such as removal of image skew or borders, which would improve the accuracy of OCR conversion.

Some of these document image-processing techniques and algorithms are candidates for inclusion in future versions of the DocView software. By evaluating some of the proposed DocView capabilities first on the DocMorph Server, it is possible to bypass the costly time delays associated with beta-testing software such as DocView. With a built-in questionnaire and email feedback capability built directly into the web pages of the DocMorph Server, there is fast user feedback. Also, complete statistics on usage of the server are kept, including which functions are used, who uses them, and the processing time for each algorithm.

The first practical use of the DocMorph Server is in providing alternative document formats for users. One of the problems that document delivery librarians often experience, particularly on university campuses, is that of delivering documents to a diverse group of patrons. Convenient Internet delivery of library documents is possible when all patrons have the requisite document viewing software. However, not everybody runs Windows on his or her computer, and document delivery librarians often have to serve patrons who have UNIX systems and Macintosh computers. If a library patron receives an Ariel document, but does not have the proper software for viewing it (which may be the case for non-Windows computers), then the DocMorph Server provides a solution. The patron may send the received file to the DocMorph Server to have it converted to a suitable alternative format. Document delivery librarians may also use the Server to change a document to another format prior to delivering it to a patron.

3. DOCMORPH SERVER FUNCTIONS

The initial prototype DocMorph Server that was first made available to the public in April 1999 had three document conversion functions.

- **Concatenate one or more TIFF files to create a PDF file**. This function concatenates one or more Ariel (GEDI) files or TIFF files to create a single PDF file. This function is useful for users who wish to use Adobe Acrobat Reader™ to view a file received from a library. Librarians may also use this function to convert Ariel documents to PDF prior to delivery to patrons.
- **Concatenate TIFF files into one multipage TIFF file**. This function is used to concatenate multipage TIFF image files to form one multipage TIFF file that contains all images. This is useful for creating new single file documents. For instance, single file pages can be concatenated to form a chapter, and chapters can be concatenated to create a book.
- **Split a multipage TIFF file into individual image files**. This function is useful for document editing, especially when combined with the previous function. By using the two functions together, a user will be able to replace, delete or insert pages in a multipage TIFF file.

By the end of the first six months of use, more than 600 users registered to use the DocMorph Server. Of these, more than 300 used the system at one time or another over this period. They submitted more than 1,600 jobs containing more than 17,000 images to DocMorph. Of these, 1,400 ran successfully while 200 failed. The failures were virtually all due to file formats that were not supported by DocMorph such as GIF, JPEG and text. Of the 1,400 successful jobs, 3% were concatenation of TIFF files, 7% were splitting TIFF files, and 90% were conversions of TIFF to PDF.

A fourth conversion function was added to DocMorph in October 1999 when the system was redesigned to support compute-intensive image processing:

- **Computer-assisted reading**. This allows conversion of scanned images to synthesized speech. With this function, a user can scan a printed document, submit it to DocMorph, and receive in return an audio "document" in the form of a web page that reads the material out loud on the user's computer. This capability is intended to support an important minority in our population that often has difficulty in reading printed literature: the blind, visually impaired and physically handicapped. It has been estimated that in 1990 there were 550,000 Americans who were blind in both eyes and an additional 7.6 million who had vision impairments.[10] Blindness is defined as best-corrected visual acuity of 20/200 or worse in the better eye; visual impairment as best-corrected visual acuity of worse than 20/40 but better than 20/200 in the better eye. The computer-assisted reading function will enable these types of people to have easier access to literature. The user interface provided by DocMorph allows the user to browse the resulting audio document by using only two keyboard keys. Functionality included for advanced users permits the user to search for audio text words and to randomly access any page within the audio document.

In addition to computer-assisted reading, the TIFF to PDF algorithm was enhanced with the inclusion of automatic page rotation, a feature needed to overcome certain problems with scanning. A number of scanners on the market are not well suited for scanning bound volumes because of the way the scanner cover is hinged. For books or journals, a scanner should have its cover hinged along the short paper edge (8.5") rather than along the long edge (11"). This would permit the operator to easily move the book left

and right over the scanning area. If, however, the cover is hinged along the long edge of the paper, this often prevents both left and right pages of the book from being scanned with the same orientation. The result is that the operator may need to turn the document upside down every other page. The Ariel software does not provide any means for rotating the pages to ensure that they are all right side up. As a result, the electronic file received by the patron is difficult to use on a computer screen because every other page needs to be rotated 180^0. Like Ariel, the Adobe Acrobat Reader does not provide a function for rotating pages. If such a file is received, users have to print the document first, and rotate the pages manually. This has become an inconvenience for both document delivery librarians and their patrons, as evidenced by comments received from several DocMorph users. Our new TIFF to PDF algorithm allows the user to choose to have all pages in the document to be upright. During conversion to PDF, the software uses OCR to see if a page is upside down. Those pages found to be upside down are automatically rotated so that they are correctly oriented for reading.

4. DOCMORPH SERVER: SINGLE COMPUTER DESIGN

The DocMorph Server is designed as an application server. It may be accessed at this URL: http://docmorph.nlm.nih.gov/docmorph. While many web sites deliver information on static web pages, others are more sophisticated and allow searching, such as NLM's PubMed[12] system. PubMed is an example of an application server that provides complex searching functionality for users via the Internet. The DocMorph Server is an application server with a three-tier design. When its first version was introduced in April 1999, the DocMorph Server's single computer architecture, shown in Figure 1, consisted of a web server, a conversion server, and a database.

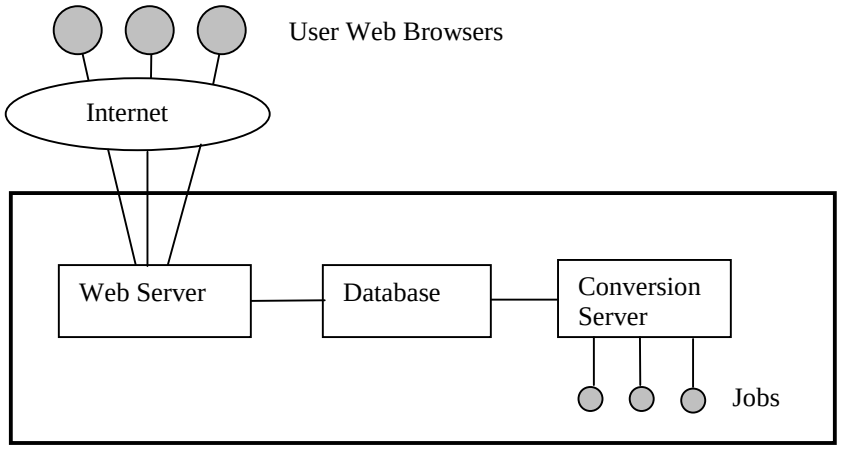

Figure 1. DocMorph Server Single Computer Configuration

The prototype DocMorph Server is designed to run on a computer hosting Microsoft's Windows NT Server operating system. Windows NT Server provides the Internet Information Server (IIS), a highly capable web server that permits processing of HTTP requests using DocMorph's Visual Basic WebClasses. WebClasses are a Windows-based alternative to CGI scripting that is commonly found on UNIX-based web servers. They use compiled Visual Basic to quickly process requests sent to the web server and return HTML-based web pages. In addition to code written in Visual Basic, other modules created for DocMorph are written in C, C++, assembly language and JavaScript.

A user of the DocMorph Server employs a web browser to access the DocMorph Server home page, which allows the user to register, log in and use the server. Once a registered user has logged in, the web browser displays four functions that permit document conversion. For each function the user needs to select one or more files to upload to the DocMorph Server. File uploading takes place using HTTP communication protocols. The Server software intercepts the uploaded file(s) and stores them temporarily on its local hard disk drive. Then it makes an entry in its database to denote the location of the file(s) to be converted, and the nature of the conversion.

A Conversion Server written in C++ periodically queries the database for work, and upon finding work to be done, the Conversion Server converts the file(s) to the requested form, and then updates the database. Meanwhile, the user's web browser queries the DocMorph Server every twenty seconds to keep track of the document conversion process. Each update request requires the WebClass running on the Web Server to query the database for job completion. When the database shows the job is completed, an HTML page is dynamically created to denote job status and allows the user to download the converted file(s). Two hours after a user has downloaded converted files, the Conversion Server removes the converted files from the computer to free up disk space. The database is used to ensure proper job selection and timing, and it also keeps track of user statistics and feedback.

5. DOCMORPH SERVER: MULTI-COMPUTER DESIGN

Initial testing and use of the DocMorph Server revealed that it ran satisfactorily for a small number of simultaneous users. On a 400 MHz processor, the average processing time for each of the initial three document conversion algorithms was 10 seconds. This included the built-in overhead for cycling between jobs, picking the next conversion job, running the job, and updating the database. Five potential bottlenecks were identified in the system design:

1. Communication speed. Users with relatively fast Internet connections (56K and above) expressed the highest satisfaction with DocMorph. Speeds lower than 33K were generally unsatisfactory for uploading and downloading document images.
2. File size. Scanned document images produce large files, even with Group 4 compression[11]. At 300 dpi resolution a typical 8.5x11 inch page can produce 100,000 bytes of compressed data, so that a typical 10-page journal article will average one megabyte.
3. Database overhead. The DocMorph Server uses a Microsoft Access database. This typically limits simultaneous access to 20 people with satisfactory response time.
4. Processing time. While DocMorph's file conversion algorithms are fast, any compute-intensive task such as optical character recognition would slow response time considerably. The average measured time to convert a typical 300 dpi biomedical journal image to text using OCR ranges from 10 to 12 seconds on a 500 MHz computer. Because the OCR process consumes virtually all the CPU time, other processes such as web page serving and database access proceed very slowly if all run on the same CPU.
5. Serving web pages. A typical Microsoft IIS web server can handle several hundred simultaneous users, but it cannot handle thousands simultaneously.

To meet our objective of implementing optical character recognition in DocMorph, it was not possible to build it into the initial single computer architecture. This is because OCR is compute-intensive and greatly slows down all the other tasks that may be running on the computer. It was decided to create a multi-computer architecture to alleviate some of the problems identified in our testing of the single computer design. The multi-computer design lessens communication slowdowns due to simultaneous file upload/download of two or more users. It also provides a web server on each computer to help spread the load created by a large number of users. Database overhead is reduced because the database is split among all computers. Finally, processing time of compute-intensive tasks such as OCR is offloaded onto separate computers.

Figure 2 illustrates the multi-computer DocMorph Server architecture that was introduced in October 1999. It consists of identical 500 MHz Windows NT servers, each running the IIS web server software.

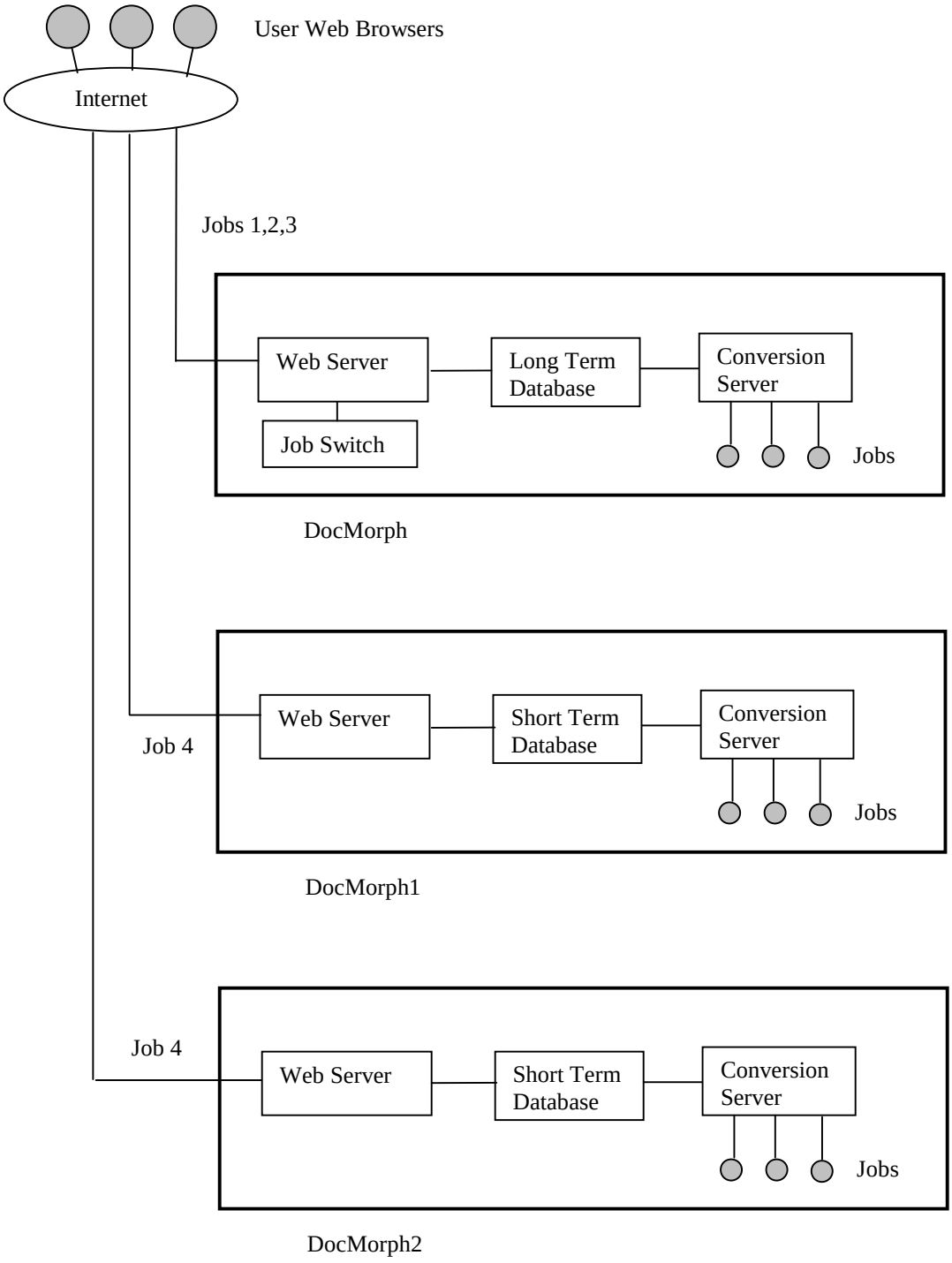

Figure 2. DocMorph Server Multi-Computer Configuration

The DocMorph Server design permits up to ten computers to run together. While the initial configuration has three computers, additional computers may be added to accommodate increasing usage. The ten computers are designated DocMorph, Docmorph1, DocMorph2, …, DocMorph9. If there is only one computer required in the architecture, the design is nearly the same as the original single computer design. The chief difference in the multi-computer architecture is the Job Switch, which controls job routing. Each time a new computer is added, all computers must be rebooted, and they begin running automatically in the newly reconfigured system.

All users initially point their web browsers to the main computer, DocMorph, which contains the Job Switch and Long Term Database. Depending on the job requested by the user, the Job Switch routes the user to a computer that handles the specific job. For example, a computer might be dedicated to the OCR function, and another to converting TIFF images to PDF. The Long Term Database on DocMorph keeps track of users and a complete history of all jobs submitted to all computers within the system. The Short Term Databases on the remaining DocMorph computers maintain data only on the specific jobs submitted to each of those computers. Each Conversion Server updates not only its Short Term Database, but also the Long Term Database on the main DocMorph computer. This way each secondary computer keeps track of its own jobs, while permitting overall system status to be obtained from the single DocMorph computer.

As illustrated in Figure 2, the DocMorph computer handles jobs 1, 2 and 3, corresponding to the three TIFF file conversion utilities. The two other computers, DocMorph1 and DocMorph2, each process only Job 4, which (as an example) is the process for converting scanned images to synthesized speech. The Job Switch in DocMorph will route all Job 1, 2 and 3 requests only to DocMorph, and all Job 4 requests to either DocMorph1 or DocMorph2. While Jobs 1, 2 and 3 each consume comparatively little CPU time, Job 4, the OCR process, is compute-intensive, requiring a separate CPU. Each computer handles its jobs in a First-In-First-Out queue that handles jobs in the order in which they arrive. As shown in Figure 3, the load-balancing Job Switch not only routes a job to the correct computer, but it ensures that jobs are distributed evenly, so that no one computer gets too many jobs. In this figure, if a newly arrived job is Job 4, the Job Switch will route this job to DocMorph 1, since it is processing one less job of type 4 than DocMorph2.

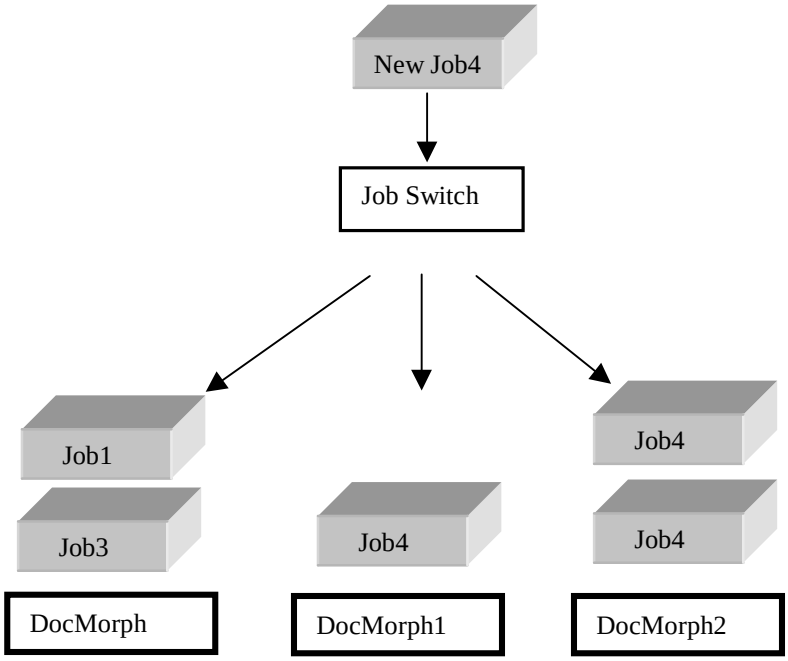

Figure 3. Load-Balancing Job Switch

Any of the DocMorph Server computers can run any job. It is up to the system administrator to configure them to maintain an even balance of computer processing, web page serving, database accessing and communications loading. The Long Term Database on the DocMorph computer maintains the entire system configuration that determines the specific jobs that will be run on each computer.

The DocMorph Server design ensures that each computer in the system can withstand outages. Although short-term power failures are adequately protected using uninterruptible power supplies for each DocMorph Server computer, if that fails and the computer goes down for a long period, then it is designed to start up automatically upon power-up. This is accomplished through an NT Service module that runs upon startup. The NT Service module reads the configuration file for the computer on which it is running, updates the system registry, and automatically starts the Conversion Server. Once the Conversion Server on each computer runs, it checks the Long Term Database on the DocMorph computer to determine the types of jobs it will run. It also updates the Long Term Database once every five minutes to let the Job Switch know that the computer on which it is running is operational. As part of the job allocation algorithm, the Job Switch checks to make sure that the computer to which the job will be routed is functional. It does this by checking the Long Term Database to ensure that the selected computer has accessed it within the past five minutes.

The DocMorph system is also designed for remote administration. For instance, it contains a built-in function available through the WWW that provides an overall status of the entire multi-computer system. The status indicates the number of jobs currently being processed by each computer, the number of jobs processed within the past two hours, the number of users currently logged into the system, and an analysis of jobs that have failed. The system administrator user interface also permits the entire system to be brought down or up remotely over the WWW, and the system to be rebooted in the same manner.

An important feature of the DocMorph system architecture is maintaining state when switching a user from one computer to another. Because the Job Switch can route a user from one computer to another and then back again, it is necessary to have a mechanism for keeping track of each user and the jobs being submitted to the overall system. This is done through the use of a cookie, a small file that can be transmitted from a web server to the user's hard disk drive via the web browser. As illustrated in Figure 4, when the user first logs into DocMorph, a cookie containing a unique identifier created using a random number generator is sent to the user's computer.

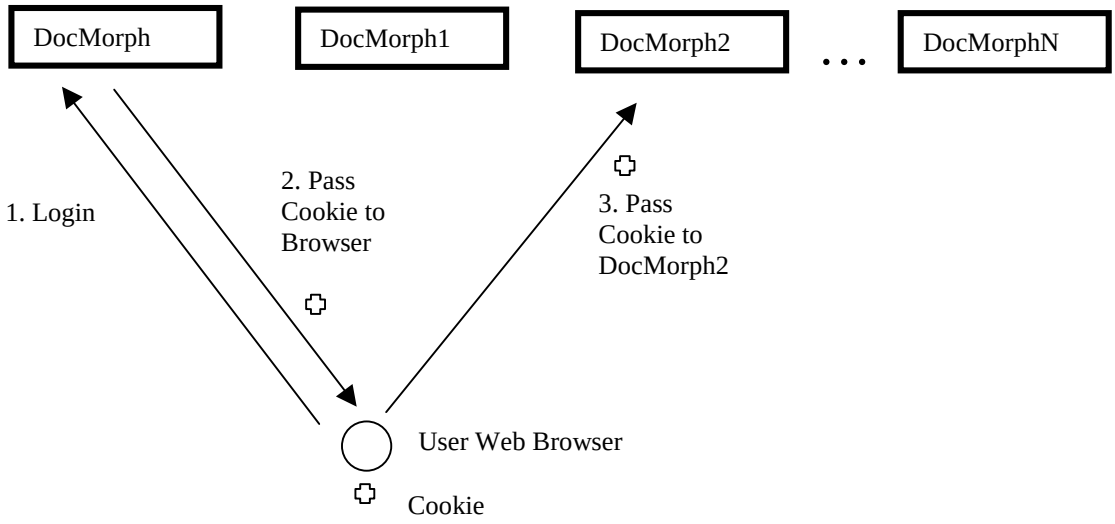

Figure 4. State Maintenance using Cookies

At the same time DocMorph sends the cookie to the user, it stores the unique identifier corresponding to the user in the Long Term Database. If the Job Switch routes the user to another computer, say DocMorph2, the user's browser sends the unique identifier from the cookie to the new computer, which in turn stores the user information in its Short Term Database. In this manner all computers to which the user is routed can keep track of that user. Each computer uses the cookie information to update system usage properly in the Long Term Database on DocMorph.

6. DOCMORPH SERVER EVALUATION

While intensive testing of the original DocMorph Server design has led to the current multi-computer architecture, evaluation of the new system has just begun. As mentioned above, it contains new and useful functions: automatic rotation of upside down images during file conversion, and computer-assisted reading. These functions will be evaluated for their usefulness and user acceptance. Based on user feedback, algorithms will be enhanced to remove bugs and overcome any performance problems.

A major innovation in the multi-computer architecture is computer-assisted reading, which converts document images to synthesized speech. To evaluate this new functionality, users who may have problems reading due to poor vision will be invited to try the system. The two areas of primary investigation will be the user interface and the reliability of conversion. There are two potential sources of error in the conversion: OCR and speech synthesis. The accuracy of OCR depends on several factors, including scan resolution, font types and sizes, text attributes (italics, bolding, etc.) and overall image quality. Speech synthesis is highly dependent on the quality of the OCR conversion as well as the language. Because the initial speech engine being used in the system processes only English, all job submissions are required to be English in origin. A user survey on the system will allow users to submit comments on their experiences with this new capability.

7. SUMMARY

The prototype DocMorph Server is based on a multi-computer architecture, and provides functions for document image processing. It is publicly available through the World Wide Web. By using a web browser, users may upload image files to DocMorph from any place on the Internet. DocMorph returns results via the web, which users may store on their computers. Current functionality includes file conversion such as TIFF to PDF, automatic detection and rotation of upside down document images, and conversion of scanned images of printed literature to synthesized speech. While the DocMorph Server is designed to provide a useful service to the public, it is also an image processing test platform that its designers are using to evaluate its algorithms and techniques for possible inclusion in other image processing software products being developed at NLM.

8. REFERENCES

1. Berger, M.A., "Ariel Document Delivery and the Small Academic Library," College & Undergraduate Libraries, Vol. 3(2). The Haworth Press, 1996; 49-56.
2. The World Wide Web address for Research Libraries Group is located at this URL: http://www.rlg.org.
3. Walker FL, Thoma GR, "DocView: Providing Access to Printed Literature through the Internet," Proceedings IOLS'95. Medford NJ: Learned Information, 1995; 165-173.
4. Walker FL, Thoma GR, "Internet Document Access and Delivery," Proceedings IOLS'96. Medford NJ: Learned Information, 1996; 107-116.
5. Postel, J. and Reynolds, J. File Transfer Protocol, Request for Comments #959, October 1985, available at URL http://sunsite.auc.dk/RFC/ .
6. Borenstein, N. and Freed, N. MIME (Multipurpose Internet Mail Extensions), Request for Comments #1341, June 1992, available at URL http://sunsite.auc.dk/RFC/

7. Information on the Group on Electronic Document Interchange (GEDI) format is available at URL http://lib-www.uia.ac.be/MAN/T02/t51.html.
8. TIFF Revision 6.0, Aldus Corporation, June 3, 1992.
9. Walker FL, Thoma GR, "Internet Document Delivery: An End User Survey," Proc. IOLS '97. Medford N.J: Information Today, 1997; 145 - 153.
10. “Trends in the Health of Older Americans: United States, 1994,” National Center for Health Statistics, Center for Disease Control and Prevention, April 1995. Available at this URL: http://www.cdc.gov/nchswww/data/sr3_30.pdf
11. Specifications of the Group 4 compression algorithm are available from the ITU Publications web site: http://www.itu.int/publications/bookstore.html.
12. PubMed is available at URL: http://www.ncbi.nlm.nih.gov/PubMed/ .