TOPS On-Line - Automating the Construction and Maintenance of HTML Pages

Kennie H. Jones

NASA Langley Research Center

Abstract

After the Technology Opportunities Showcase (TOPS), in October, 1993, Langley Research Center's (LaRC) Information Systems Division (ISD) accepted the challenge to preserve the investment in information assembled in the TOPS exhibits by establishing a data base. Following the lead of several people at LaRC and others around the world, the HyperText Transport Protocol (HTTP) server and Mosaic were the obvious tools of choice for implementation. Initially, some TOPS exhibitors began the conventional approach of constructing HyperText Markup Language (HTML) pages of their exhibits as input to Mosaic. Considering the number of pages to construct, a better approach was conceived that would automate the construction of pages. This approach allowed completion of the data base construction in a shorter period of time using fewer resources than would have been possible with the conventional approach. It also provided flexibility for the maintenance and enhancement of the data base. Since that time, this approach has been used to automate construction of other HTML data bases. Through these experiences, it is concluded that the most effective use of the HTTP/Mosaic technology will require better tools and techniques for creating, maintaining and managing the HTML pages. The development and use of these tools and techniques are the subject of this document.

Background

In October of 1993, NASA's Langley Research Center (LaRC) presented its first Technology Opportunities Showcase (TOPS). TOPS was designed to present LaRC's developing technologies to industry and to initiate mutually beneficial partnerships between LaRC and industrial enterprises. Special emphasis was given to strengthening existing and cultivating new strategic partnerships. There were 185 exhibits of technologies which included medical applications, structures, materials, remote sensing, aeronautical systems, sensors and non-destructive evaluation. Although over 850 industrial representatives representing about 400 different organizations attended TOPS, LaRC's Center Director wanted to preserve the investment in information assembled in the TOPS exhibits by establishing a data base. The data base would allow those who could not attend TOPS to benefit from this information.

LaRC's Information Systems Division (ISD) accepted the challenge to construct this data base. The HyperText Transport Protocol (HTTP) server and Mosaic were the obvious tools of choice for

implementation, for two reasons: they provide free public access to the data base and they support multimedia displays. Research revealed that some TOPS exhibitors had begun creation of HyperText Markup Language (HTML) pages describing their TOPS exhibits intending to include these pages as links from developing home pages for their organizations. One exhibitor had created a TOPS Home Page as a suggested framework for accessing the exhibit pages as they evolved. As there were 181 TOPS exhibits, ISD recognized that depending upon exhibitors to develop HTML expertise independently and to construct their own HTML pages for their exhibits would not produce a complete data base in a reasonable time. It was also deemed impractical to assemble a team of HTML experts to construct HTML pages for all 181 TOPS exhibits by hand. Instead, an automated approach was designed that eliminated the need for hand-editing the pages, while simplifying modifications and enhancements. An HTTP compliant TOPS data base was successfully implemented by a few part-time volunteers in less than a month.

Following the completion of the TOPS data base, the techniques and tools developed for that application were also used to automate the construction of HTML pages presenting the current organization of LaRC and to facilitate the automatic conversion of binary document enclosures in e-mail messages to a variety of presentation formats. These applications are also described.

Constructing the TOPS Data Base

Each TOPS Exhibitor had prepared a Technology Information Sheet, a one page description of each exhibit that had been presented at TOPS. These sheets contained descriptions, lists of potential commercial uses, technical references, and responsible points of contact. Constructing an on-line data base was not under consideration at the time the Technology Information Sheets were produced and consequently, the only record of the sheets was in paper form. The paper sheets were assembled, digitized and processed with optical character recognition (OCR) software to produce text files. A processor was developed that automatically converted these text files to HTML pages. This guaranteed a consistency in the format of the HTML version of the Technology Information Sheets (hand editing would likely not result in consistency). Also, since the Technology Information Sheets are now maintained as text files, any modifications required to the pages can be made using any text editor without the need for knowledge of HTML syntax. Following any modifications to a page, the processor is re-executed for that exhibit and the modifications appear in the HTML version.

The Technology Information Sheets provided a minimum set of consistent information on each exhibit. Other efforts at LaRC to create HTML data bases on large data sets by hand have followed an "add a page when I get around to it" philosophy. The authors create the pages on each subset and, when ready, add them to the data base. This methodology can result in data bases that are never complete. The TOPS approach avoided this pitfall by selecting a minimal subset of information on each exhibit (the Technology Information Sheet) and, using the processor, adding all exhibits to the data base initially. The processor also added to the end of each Technology Information Sheet HTML page a section entitled "More Information." This section contains links to other pages if available:

- HTML pages further describing the exhibit - If anyone has created HTML pages providing more information on an exhibit, a link is added to access the Universal Resource Locator (URL) for those pages. Exhibitors are then able to add multimedia displays of their exhibit to the data base. Pages added thus far are well done including charts, photographs, videos sequences, and audio segments. These "exhibit tours" are made available to the processor by simply adding the exhibit identification number and the URL to a text file.

- HTML pages describing the LaRC organizations that supported the exhibit - If anyone has created HTML pages providing more information on the organization that created the exhibit, a link is added to access the URL for those pages. These multimedia displays provide the reader with links to other activities at LaRC. These "organization tours" are made available to the processor by simply adding the exhibit identification number and the URL to a text file.
- The on-line Langley Technical Reports Server (LTRS) - All exhibits have a list of technical references on the Technology Information Sheet and this list is available as a link from the Technology Information Sheet HTML page. However, if the author has contributed the publication to the LTRS, a direct link is included to that service rendering the complete publication available on-line
- Up-to-date information on the point of contact - Address, telephone number, e-mail address, etc. on the point of contact are provided directly from the LaRC personnel data base. This is an important point. Many HTML pages are appearing around LaRC that include this information embedded in the HTML file. Often the author obtains the information by executing the LaRC "whois" server and adds the output (via a text editor) to the HTML page. Representing a "snapshot in time," the information is accurate at the time the page is created. However, it is likely that the information will not be updated should it change at a later time.
- An automated "Request for Information" form - One of the objectives for constructing the TOPS data base is to identify more contacts in industry interested in LaRC's technology. Prior to the on-line data base, these contacts were made by telephone or paper mail. A Request for Information form was included for each exhibit that allows an interested viewer to request more information about that exhibit. Once the form is submitted, the information is e-mailed to a central repository and to the point of contact for the exhibit. A reply to the requester also echoes back to the central repository. This method not only makes contact easier, but opens the door for automated metrics on the contacts made.

The processor also automatically constructs several indices to access the exhibits. The TOPS administrators originally organized the exhibits into groups related by discipline. Each exhibit has a unique identification number that specifies a group number and an exhibit number. An index is assembled that lists all exhibits in sequential order. Another index is assembled that lists all groups. Each group is linked to a sequential list of exhibits belonging to that group. The TOPS administrators also placed the exhibits into technology categories (aerospace, medical, transportation, etc.) using relevance of the exhibit to the technology as the criterion. An index is assembled of all technical categories. Each category lists all exhibits in that category. Exhibits may be added by providing a text file of the Technology Information Sheet and adding the exhibit to the text files describing the indices. Re-executing the processor will reconstruct the indices as well as creating the HTML version of the Technology Information Sheet.

The processor has been modified to support the next TOPS in 1995 whereby the exhibitors can submit their exhibit's Technology Information Sheet, exhibit tour, and organization tour via e-mail as text files or word processor documents and the processor automatically posts the information to the TOPS data base in HTML format. This will eliminate the major effort in the previous data base generation of collecting and digitizing the Technology Information Sheets.

Cognizant of the rapidly evolving future of digital imaging and the need for a data base of photographic information, LaRC's Photographic Section, had constructed a data base on a Macintosh computer for keeping track of photographs stored on a video disk system. The data base defined "folders" that represented a request for a photographer resulting in one or more photographs. In addition to the

photographs, they added textual information about the photographic session (date, requester, title, description, people in the photograph, etc.). All of the photographs taken at TOPS had been placed in the data base. This system, however, was designed for internal use and was not scaleable for on-line access by the LaRC researchers. To provide the information on-line, the photographs were digitized. The TOPS processor extracts the folder data from a dump of the data base and constructs an HTML page for each folder. The processor also constructs an HTML index of the folder titles linked to the folder pages and a sequential list of photographs in HTML format. The indices and folder pages link to the photographs. Using this approach, the Photographic Section remains the maintainer of the data base and modifications can be automatically transferred to the HTML data base.

Langley’s Organizational Browser

As the popularity of Mosaic spread across LaRC, individuals began to construct "home pages" providing an overview of their organization. Although a valuable effort to provide information about LaRC, this was proceeding in the conventional manner. Only those organizations having someone that, through their own initiative, learned about Mosaic and HTML had a developing home page. The URL’s for these pages were added to the LaRC home page as they became available. Thus, the LaRC home page displayed a subset of LaRC total capability.

Applying the same technique used for TOPS, a processor has been developed that extracts information from the LaRC personnel data base and constructs an "Organizational Browser." The browser presents a hierarchical map of LaRC organizations that links to pages on each organization. Each organizational page contains a list of managers, secretaries, and other personnel in the organization. Every person’s name appearing in the browser links to LaRC’s "whois" server that provides up-to-date information on that employee (address, telephone number, e-mail address, etc.). Because all of this data, organizational structure and employee data, come directly from the LaRC personnel data base through execution of the processor, updates to the data are automatic.

URL’s and descriptions (for links) for those home pages of organizations that already existed were listed in a text file. The processor reads this file during execution and adds links to these pages on the browser page for those organizations. Additional pages can be added by members of an organization by e-mailing information to another automated processor in the form of text files, word processor documents, or HTML pages. This processor:

- Reads the e-mail.
- Extracts the link description.
- Extracts a URL or, if text files or word processor documents are provided, constructs an HTML file.
- Updates the text file to include the new URL and link description.
- Re-executes the browser construction processor for the subject organization to replace its page in the browser.

The "whois" server was modified to read a directory in the selected personnel’s directory on the LaRC Postoffice computer, if available. In this directory, an employee can add personal information about themselves such as a picture, biographical information, skills descriptions, etc. Again, the approach is to provide a minimal set of information on each employee while allowing custom data to be added.

For both the organizations and the personnel, WAIS indices are created as the data are updated that

allow keyword searching of the subject areas.

As with TOPS, this approach guarantees a minimal set of up-to-date information on every organization and person at LaRC while providing a means for individuals to customize the information for their organization or themselves. This customization requires no intervention by developers or maintainers of the organizational browser

Conversion and Distribution of E-mail Enclosures

When the current ISD Chief came to ISD, he was used to generating e-mail messages on a Macintosh that included enclosures of word processing documents, charts, spreadsheets, etc. However, when he e-mailed these to the ISD staff, the Macintosh binary files were converted to an ASCII representation that was unreadable to the UNIX users (the majority of the staff). A processor was developed that:

- Reads the e-mail and, if an enclosure is present, extracts the enclosure.
- Extracts the data from the Macintosh file.
- Determines the source of enclosure (word processor, spreadsheet, etc.).
- Executes a Windows emulation on UNIX that executes the proper application which reads the enclosed document and constructs text, Postscript, and HTML files of the document.
- Posts the HTML, Postscript, and document files to a directory available via an HTTP server.
- Reconstructs the e-mail message to include the original message, the URL for access to the HTML, Postscript, and document files, and the text version of the document.
- Forwards the new message to the original recipients.

Using this approach, the ISD Chief does not change his way of doing business, but the ISD staff can now access his document in a variety of ways. If the document is all text, they can simply read their e-mail. If the document contains formatting or graphics, they can open the URL specified in the message. From the URL, they can view the HTML file or view or print the Postscript file. If they define a MIME type in Mosaic for the document type, they can automatically launch the application to view or edit the original document, regardless of their computing platform.

Details of the Processors

The TOPS processor is a UNIX shell script that calls three separate FORTRAN executable programs. The first program reads text files extracted from spreadsheets created and maintained by the TOPS administrators that describe the exhibit identification numbers, exhibit descriptions, exhibit groups, and technical categories. Using this information, the program creates a directory structure reflecting the exhibit organization to contain the HTML files. This program also creates the exhibit indices in HTML format. The second FORTRAN executable program reads the text files containing the Technical Information Sheets and the text files describing the exhibit tours and organization tours. This program creates the exhibit pages. It can be executed to create an individual exhibit or all exhibits at once. The third FORTRAN executable program reads the photographic data base dump and constructs the HTML pages providing access to the TOPS photographs.

The subroutines developed for these processors were generalized and reused in the organizational browser processor. Other routines were developed to extract and manage organizational and personnel data from the LaRC personnel data base. This processor is also a FORTRAN executable program. The

library of subroutines were also used to create two FORTRAN executable programs that establish links to pages submitted by users via e-mail. Through these programs, users can add a functional statement, a home page, or pages describing current activities for their organization.

The library of subroutines was also used to create a FORTRAN executable program that converts the e-mail enclosures to various formats. Additional subroutines were developed that perform the document conversion. The document conversion subroutines were added to the organizational browser program allowing users to submit word processing enclosures as well as text files for added pages.

FORTRAN was selected as the implementation language for these utilities by the author simply because of his expertise on this language. C or Perl would have been better choices for portability of the libraries but due to relative inexperience in the use of these languages, this choice would have significantly increased the development time. Recognizing the need for portability if these utilities are to be used for other applications, the libraries will be ported to either C or Perl.

Conclusion

The approach used for each data base described provides a means of quickly presenting a minimal set of information on each data set, consistency in the "look and feel" of the pages, and a means of automatically modifying the pages (change the text files and/or processor and regenerate the pages. There are three major advantages of the techniques and tools described.

- Quickly providing a minimal subset of data on each section of a large data set initially provides a consistency and completeness to the presentation of data that is preferable to completing one section at a time until all sections are done.
- Automating the production of HTML pages is advantageous in both the construction and maintenance of large HTML data bases.
- Linking to a single source for data is essential for practical maintenance of the data bases. If this is not controlled, a few years from now, the world will have data distributed around many data bases that are obsolete.

Although the use of HTML pages displayed with Mosaic provides an efficient and effective means of presenting information, construction of these pages is commonly a labor-intensive activity requiring special skills not commonly available (knowledge of HTML, UNIX files systems, HTTP servers, etc.). The most effective use of this technology will require better tools and techniques for creating, maintaining and managing the HTML pages. This requirement is necessary because:

- Many people with data to share are not and are not likely to become fluent in the construction of HTML pages.
- Those fluent in the use of HTML can make better use of their time than constructing and maintaining HTML pages (processes that can be easily automated).
- Large data sets exist already in other formats and conversion of these to HTML is better accomplished through an automated processor.
- Data sets can be maintained more easily in formats other than HTML (word processors, data base management systems, etc.) and automatically converted to HTML for display.
- Consistency in display of large uniform data sets is important for effective use and consistency is more easily enforced through automated processors than construction by hand.

Further development of these tools and techniques will benefit future use of the World Wide Web.

URL's of interest:

TOPS: http://www.larc.nasa.gov/tops/tops.html

LaRC's Organizational Browser: http://www.larc.nasa.gov/orgs/orgs.html

LaRC's Home Page: http://www.larc.nasa.gov

About the Author

Kennie Jones is the Assistant Head of the Scientific Applications Branch of the Information Systems Division of NASA Langley Research Center. In that capacity, his major responsibility is support for scientific information management systems. He has a M.S. degree in Computer Science and a B.S. degree in Biology.

Telephone: (804) 864-6720

E-mail: k.h.jones@larc.nasa.gov


LaRC Home Page

Last Updated Sep 15, 1994
by

Kennie Jones