Internet access to research results: an evolving effort

Schmoldt, D. L, Winn, M. F. & Araman, P.A.

USDA Forest Service, Brooks Forest Products Center, Virginia Tech, Blacksburg, VA 24061-0503 USA, schmoldt@vt.edu, mattwinn@vt.edu, paraman@vt.edu

Abstract

Since October 1995, our Research Work Unit of the US Forest Service has operated a Web site to disseminate research results. Initially, this took the form of basic information about the Unit's goals, organization, research problems, cooperators, etc. Over time the site has expanded to provide lists of available publications, abstracts of those publications, abstract searching, publication request forms, and electronic versions of publications. As part of the Forest Service's commitment to customer service, each reprint request or publication download also offers site visitors the opportunity to provide feedback to our Unit and to Forest Service headquarters using a customer survey questionnaire. The Web server's visitor log file has enabled us to analyze visitor usage and to make some valuable inferences about visitor preferences for particular subjects and for the technical content of research reports. The evolution of our Web site is described and some site-access statistics indicate how site visitors make use of the information provided there and how they respond to this mode of technology transfer. Since publications have been made available on our site more than 15,000 reprints have been distributed either in paper form or electronically, averaging approximately 700 per month currently. Many of these requests have originated internationally, mirroring the extensive reach of the Internet, and thereby including clientele that are otherwise excluded via traditional channels. Visitor response has continually encouraged the use of electronic documents (PDF files) due to their speed, quality, and ease of access.

Keywords: World Wide Web, electronic publishing, technology transfer

1 Introduction

For the past 2-1/2 years our Research Work Unit (RWU) has been operating a Web site. The site became operational in October of 1995. In the beginning, it was primarily designed to stake out a piece of the Internet with our name and our RWU's organizational mission. Quickly, it be-

came apparent to us, however, that our user community could benefit greatly if we increased the content of our site. Since that time, the Web site has continually evolved as we have seen opportunities to increase and modify services to our clientele.

Because we are a research organization, our primary product is research results. At first, we made these products available as on-line lists of publications (categorized by subject area) and as corresponding on-line abstracts. Site visitors could order publications by filling out an electronic form. Upon receiving form requests, the requested publications were copied and surface mailed to the customer. This process was very time-consuming and tedious, and reprint quality was often marginal.

Beginning in December 1996, we began to make our publications available as Portable Document Format[1] (PDF) files. Some overhead, in terms of time and expense, is incurred to produce PDF files, but the long-term savings are very attractive. PDF file availability permits visitors to download many different publications quickly without our intervention. Most notably, publication quality is equal to, or better than, reprints and photocopies.

Over time we have continued to retain and periodically analyze Web server log files. These records provide us with cursory information about users and allow us to infer certain other information. Six months into Web site operation, we performed an extensive analysis of site visitors and their use of our site's offerings (Schmoldt et al. 1997). We identified certain subject areas with high interest, and found that certain subject areas could benefit by an increased number of popular (as opposed to technical) publications. We also found use by international visitors to be quite high. This is a clientele group that we would otherwise not have been able to reach through traditional technology transfer mechanisms.

In the remainder of this report, we provide additional details about our site and give some statistics regarding its use.

2 Web server details

2.1 Server hardware

At the time that we set up our Web site there was only one computer in our unit that had a hardwire Internet connection (campus Ethernet). Our remaining computers had 19.2K bits/sec connections. While the latter type of connection may be acceptable for accessing a Web site, it does not provide the bandwidth and reliability necessary to serve text and graphics to multiple, simultaneous users. Therefore we selected the desktop computer with an Ethernet connection as our server platform. This machine also served as one scientists personal computer and, hence, our server did not use dedicated hardware. Because Web server access by users is intermittent, we did not expect that this dual use would significantly hamper the scientific use of the machine. After several months, however, we were able to move the server to another machine that was used only part-time by an adminis-

trative staff member. Finally, late in 1997 we moved the server to a dedicated machine that serves no other purpose than as a WWW and Mail server.

Our server platform is an Apple[2] Macintosh 7100/80 with a 10 Mbit Ethernet connection. We find that this provides adequate bandwidth and throughput for the load our site experiences.

2.2 Server software

We began our Web site using a shareware version of WebStar's WWW server software called MacHTTP[3]. About 1 year ago we switched to a freeware version of Quid Pro Quo[4]. We find that it is much faster than MacHTTP, is well supported, and has a compatible, high-performance commercial version available. We also operate a Mail server on the same machine. This allows us to set up mailing lists without relying on Mail servers outside of our control. In particular, we have considered creating a mailing list for our site visitors, but have not yet felt that it is warranted. We also operate an FTP server on the Web server machine, which allows us to edit and transfer files remotely. In addition to periodic backup copies of our server files, we also maintain a mirror directory structure on another machine.

Several server functions are carried out by Common Gateway Interface (CGI) plug-ins and scripts. One of these CGI's is freeware, provided by Apple Computer that allows visitors to search our site. This search plug-in is not very flexible, but provides basic functionality. The primary searchable content consists of our abstract pages. A second CGI allows us to easily mail and store information provided by user forms; this is a shareware plug-in. The third CGI is a script written by us in Perl to respond to user requests for PDF files. When a visitor selects a PDF file to download, they are automatically given a customer survey form to fill out and can then download the selected file.

2.3 Creating PDF files

Documents can be converted to PDF files in one of two ways. The first, and easiest, is to print a word processing document to a PDF file using the Adobe Acrobat PDFWriter printer driver. This method takes the least amount of time and requires the least amount of editing. Unfortunately, we dont have electronic versions of many of our older publications. Also, for those publications which we do have electronic copies of, final page layout was usually customized by the publisher. This prevents us from creating a PDF file identical to the published version. For this reason, most of our PDF files are created by scanning hard copies of the published documents and converting the scanned image to PDF format.

This method of converting documents involves three basic steps: scanning, processing and editing. The amount of time it takes to convert a paper document into a PDF file varies, but on average, it takes about 15 minutes per page. This includes all 3 steps mentioned above. Conver-

sion time depends on such factors as the quality of the original document and the page layout. Poor paper copies result in poor scanned images and thus increase editing time due to inaccuracy during processing. Tables, figures and variations in text styles also increase editing time.

The first step in creating a PDF file is to scan the paper document. Documents are scanned in black and white at 360 dpi using an Epson[5] ES-1200C flatbed scanner and the Adobe Acrobat Capture software. Each image is saved as a separate TIF file and given a unique file name. File names are given sequentially so that the processing software knows the documents page order (e.g. page1.tif, page2.tif, etc.).

After all the pages have been scanned, the image files are processed using Acrobat Capture software. The processor converts the text image into text using OCR (Optical Character Recognition) and determines its font attributes. All other parts of the image, not determined to be text, are left as bitmap images. The results of the processing are saved as an ACD file (Acrobat Capture Reviewer document).

The final step in the conversion is to edit the ACD file using the Adobe Acrobat Capture Reviewer software. All suspect words found during the OCR are shown highlighted in the ACD file. Suspect words include those with a confidence level below 95%, those not in the dictionary, those with uncertain fonts, and those mixed alpha/numerically. The file is edited by tabbing through the suspect words, comparing the processed results to the original document, and making changes as necessary. Once the file has been edited, it is saved in PDF format.

3 Results

Pages accessed on our Web site have increased steadily, but have leveled off at about 7,000–8,000 per month (Fig. 1). Our count of pages does not include graphic files (*.gif, *.jpg), but only HTML files and PDF files, so it is not equivalent to "hits". The recent low value for April 1998 is most likely due to the fact that we changed our directory structure slightly. Consequently, regular visitors that had bookmarked certain pages on our site were not able to reach them in the usual way. Once we realized this problem, we added an alias to the old directory structure, so that the new structure could be accessed using the old bookmarks.

The number of publications requested has also increased steadily (Fig. 2). However, since PDF files became available in December 1996, the number of distributed publications escalated dramatically. The number of reprints requested has continued to decrease in number as more publications have become available as PDF files. The total number of publications distributed averages between 800–1,000 per month, with the total number greater than 16,000 over the past 2–1/2 years. The recent dip in publications in April is probably due to the bookmarking problem mentioned above.

Customer satisfaction is important to our government agency. Standard customer survey question-

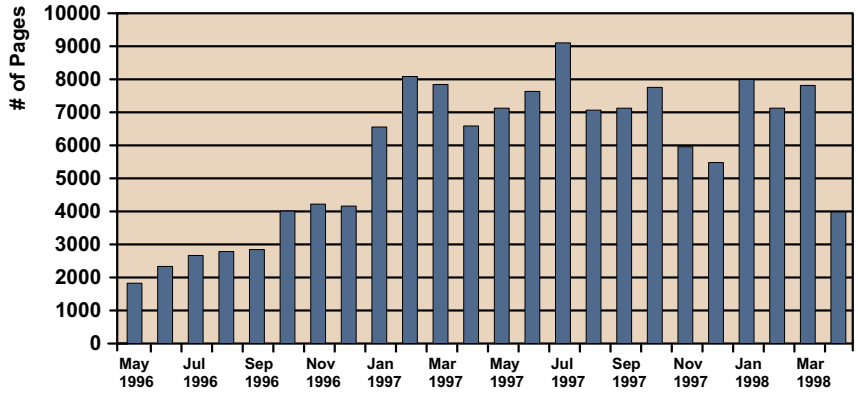

Figure 1. Pages accessed by site visitors are tracked over 24 months. Pages accessed are not synonymous with "hits"–graphic files are not included in these numbers.

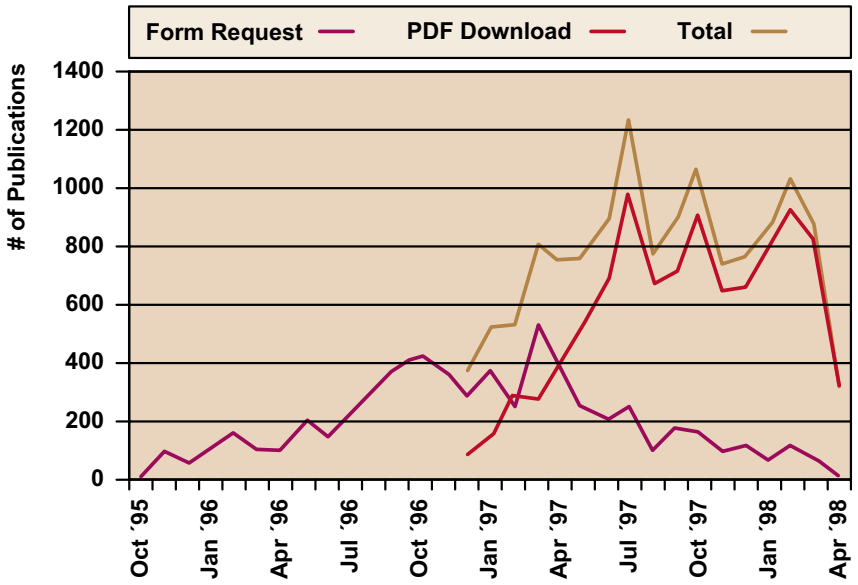

Figure 2. Publications distributed monthly are tracked for the 2-1/2 years the Web site has been operating. PDF files became available in December 1996.

naires are available and are given to customers after they have requested services from some facility or person in the organization. Form completion is entirely voluntary. We have modified the standard agency questionnaire to accommodate our electronic services. Results from the 300+ surveys that we've received to date appear in Fig. 3.

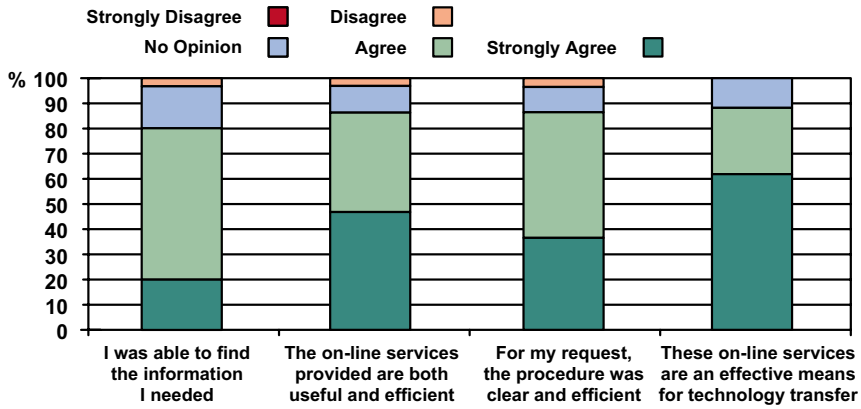

Figure 3. Our Web site is viewed quite favorably by visitors, as is evidenced by summary responses to the four questions included in our customer survey questionnaire.

4 Conclusions and discussion

Web site usage increased steadily for the first year or so of operation. Since then, site activityboth in pages accessed and publications distributed has leveled off. This equilibrium behavior is entirely expected and seems to indicate we have reached a saturation point with our clientele. Based on visitor feedbackin terms of survey responsesthis electronic means of technology transfer is well received. In fact, earlier survey responses continually encouraged us to make more publications available as PDF files. The only publications not yet converted to PDF files are very dated ones and infrequently requested ones.

As noted earlier, we have greatly expanded the customer base that we have been able to reach. Historically, research publications have been stored and distributed in paper format by our Station headquarters. Publication lists made available by them, however, were not targeted specifically to any particular customer groupe.g. forest products, in our caseand were generally not available outside of the U.S. The worldwide reach of the Internet has allowed us, instead, to distribute beyond our own shores and to target those organization and individuals most likely to benefit from our research results.

References

Schmoldt, D.L., Winn, M.F., & Araman, P.A. 1997. Wood utilization research dissemination on the World Wide Web: A case study. Forest Products Journal 47(6): 25-31.

Tradenames are used for informational purposes only. No endorsement by the U.S. Dept. of Agriculture is implied.

[1] Adobe Systems Incorporated.
[2] Apple Computer Incorporated.
[3] Biap Systems Incorporated.
[4] Social Engineering Incorporated.
[5] Seiko Epson, Inc.