WEB SITE ACCESS STATISTICS AND DELIVERY OF RESEARCH RESULTS

Daniel L. Schmoldt, Matt F. Winn, and Philip A. Araman

Introduction

For the past 2-1/2 years, our Research Work Unit (RWU) has been operating a Web site (http://www.se4702.forprod.vt.edu). The site became operational in October 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 became 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 and expanded as we have seen opportunities to increase and modify services to our clientele.

Because we area research organization, our primary product is research results. The traditional mechanism for disseminating research results in our organization has been to stockpile publication reprints at our offices and at the publications office at our Research Station's headquarters. As these documents become known via formal publication and through citation in other sources, our offices receive reprint requests from interested parties. The publications center of our Station also makes lists of publications available to various library cataloging systems and specialized forestry information lists. They also, then, receive large numbers of reprint requests. Each such request, of course, requires some minimal amount of time to respond to—finding the requested document, photocopying it if necessary, and mailing it to the recipient. This process costs considerable time and money for the many thousands of publication reprints distributed each year.

Daniel L. Schmoldt Adjunct Assistant Professor of Wood Processing, Research Forest Products Technologist, USDA Forest Service, Southern Research Station, Brooks Forest Products Center, 1650 Ramble Road, Blacksburg, Virginia 24061, USA, Matt F. Winn, Forestry Technician, USDA Forest Service, Southern Research Station, Brooks Forest Products Center, 1650 Ramble Road, Blacksburg, Virginia 24061, USA; Philip A. Araman, Adjunct Senior Research Scientist and Project Leader, USDA Forest Service, Southern Research Station, Brooks Forest Products Center, 1650 Ramble Road, Blacksburg, Virginia 24061, USA.

Presented at Third IUFRO Extension Working Party Symposium, "Extension Forestry: Bridging the Gap Between Research and Application," July 19-24, 1998, Blacksburg, Virginia, USA.

One of our secondary aims as we went online, then, was to streamline this process. At first, we made these research products available as on-line lists of publications (categorized by subject area) and as corresponding on-line abstracts. Site visitors were able to order publications by filling out an electronic form. Upon receiving form requests from visitors, the requested publications were copied and surface-mailed to the customer (typically within two to three working days). This process was still very time-consuming and tedious, and reprint quality was often marginal. Nevertheless, it did make the requesting step more efficient and allowed users to read an abstract before ordering the associated document.

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. Even more importantly, 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 electronic publishing, and give some statistics regarding site usage.

Web Server Details

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 (Ethernet). Our remaining computers had 19.2K bits/see connections over telephone data lines. While the latter type of connection

[1] Adobe Systems Incorporated. Tradenames are used for informational purposes only. No endorsement by the US Department of Agriculture is implied.

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

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

Server Software

We began our Web site using a shareware version of WebStar's WWW server software called MacHTTP[3]. About one year ago we switched to a freeware version of Quid Pro Quo[4]. We find that the latter 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 files remotely and transfer them as necessary. 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 on our site 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

[1] Apple Computer Incorporated.

[3] Biap Systems Incorporated.

[4] Social Engineering Incorporated.

file to download, our Perl script automatically sends an electronic customer feedback form for them to fill out, and allows them to download the previously selected file.

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 do not 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 three steps mentioned above. Conversion 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 page's 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 document's page order (e.g. pagel.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).

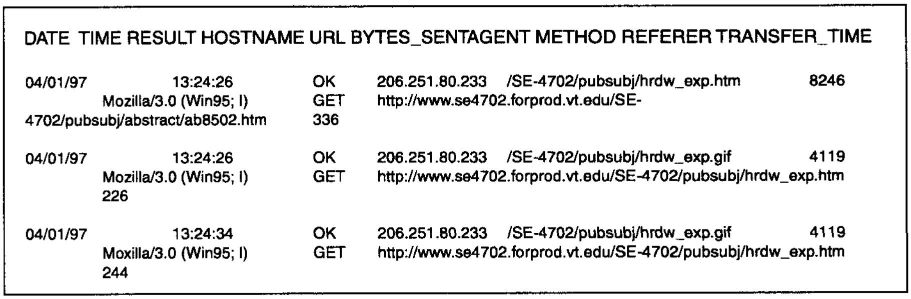
The final step in the conversion is to edit the ACD file using the Adobe Acrobat Capture Reviewer software. All suspect words found during the OCR are shown highlighted in the ACD file. Suspect words include those with a confidence level

[5] Seiko Epson, Inc.

below 95 percent, those not in the dictionary, those with uncertain fonts, and those mixed alphanumerically. 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.

Server Log File Analysis

The Web server software records each “hit” as a separate line in an ASCII log file. As noted above, each “hit” corresponds to a file requested by the client browser, which may include graphic image files (GIF, JPEG) in addition to content files, such as HTML files. Figure 1 shows several lines from one of these log files. Most of the entries here should be self-explanatory. Of particular interest for our subsequent analyses are: the amount of data transferred, the type of file transmitted, and, in the case of PDF files, which files were sent. The latter can indicate which publications were requested most often.



```
DATE  TIME RESULT HOSTNAME URL BYTES_SENTAGENT METHOD REFERER TRANSFER_TIME

04/01/97            13:24:26        OK     206.251.80.233   /SE-4702/pubsubj/hrdw_exp.htm          8246
            Mozilla/3.0 (Win95; I)  GET    http://www.se4702.forprod.vt.edu/SE-
4702/pubsubj/abstract/ab8502.htm    336

04/01/97            13:24:26        OK     206.251.80.233   /SE-4702/pubsubj/hrdw_exp.gif          4119
            Mozilla/3.0 (Win95; I)  GET    http://www.se4702.forprod.vt.edu/SE-4702/pubsubj/hrdw_exp.htm
            226

04/01/97            13:24:34        OK     206.251.80.233   /SE-4702/pubsubj/hrdw_exp.gif          4119
            Moxilla/3.0 (Win95; I)  GET    http://www.se4702.forprod.vt.edu/SE-4702/pubsubj/hrdw_exp.htm
            244
```

Figure 1. Several lines from a typical Web server log file indicate the type of information that is recorded here.

Our six-month analysis of log files (Schmoldt et al. 1997) was performed manually, which is quite time consuming. Since that time, however, we have made use of a freeware software package called *Analog,* which analyzes Web server log files. Log file analysis is performed by the software based on commands specified in a configuration file. This configuration file provides a great deal of flexibility in specifying how the log file analysis is conducted and reported.

Results

Pages accessed on our Web site have increased steadily, but have leveled off at about 6,000-8,000 per month (Figure 2). 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,” which is the traditional means to measure Web site activity. Hit counts measure Web server load, and not necessarily Web site interest. The recent low value for April and May 1998 is most likely due to the fact that we changed our directory structure slightly, albeit at the top level of the directory structure. 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 directory structure could be accessed using the old bookmarks.

The number of publications requested has also increased steadily (Figure 3). However, since PDF files became available in December 1996, the number of

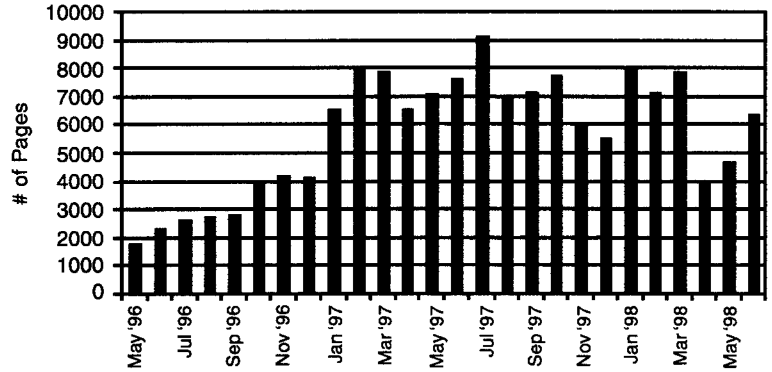


Figure 2. Pages accessed by site visitors are tracked over 24 months. Pages accessed are not synonymous with "hits," because graphic files are not included in these numbers.

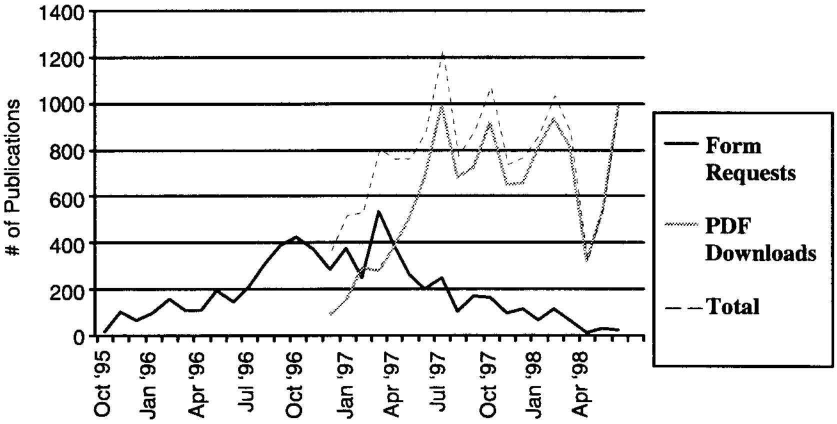


Figure 3. Publications distributed monthly are tracked for the 2-1/2 years the Web site has been operating. PDF files became available in December 1996.

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 17,000 over the past 2-1/2 years. The recent dip in publications for April/May is probably due to the bookmaking problem mentioned above.

Over the past 12-month period (from July 1997 to June 1998) an average of more than 4 MB of data have been transferred per day. Again, these numbers are reduced due to low visitation during May and June. PDF files accounted for 60 percent of the total data transferred, while representing less than 5 percent of all requests. The most frequently requested PDF file publications deal with pallet repair, recycling, and grading/sorting.

Customer satisfaction is important to our government agency. Standard customer survey questionnaires are available to Forest Service sites and are typically given to customers after they have requested services from some facility or person in the organization. Form completion is voluntary. We have modified the standard agency questionnaire to accommodate our electronic services and to provide comments to us. Results from the 300+ responses that we have received through April 1998 appear in Figure 4.

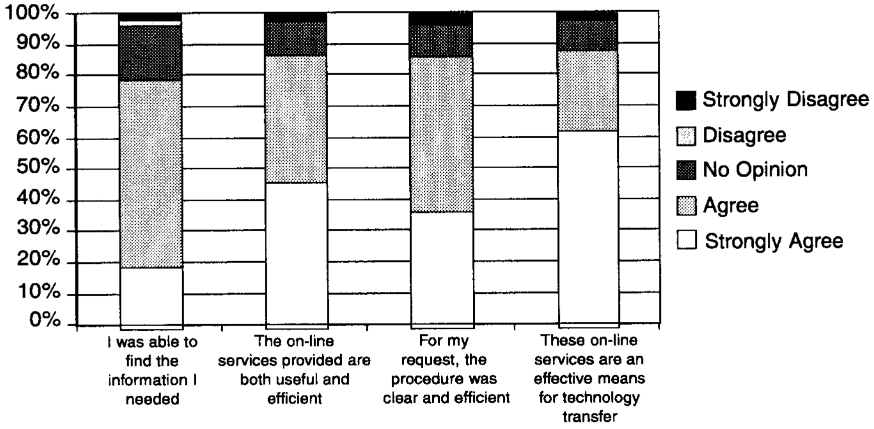


Figure 4. Our Web site is viewed quite favorably by visitors, as is evidenced by summary responses to the four questions included in our customer feedback questionnaire.

Discussion and Conclusions

Web site usage increased steadily for the first year or so of operation. Since then, site activity—both 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 feedback—in terms of questionnaire responses—this electronic means of technology transfer is well received. In fact, earlier questionnaire 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.

With the introduction of PDF file download capability to our web site we noticed an increase in the total number of publications delivered. We attribute this increase vis-à-vis mailed reprints to the immediacy of PDF file downloads. That is, site visitors can download many reprints, examine them briefly, and download others if the initial ones do not fully satisfy their interests. Requesting and receiving hard copy documents using the traditional procedure was less user interactive and immediate.

The ability to analyze Web server log files quickly and easily (via automated analysis) is very beneficial. It allows us to examine visitor usage and interests periodically, so that we can tailor the offerings of our site.

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 group—e.g. forest products, in our case—and were generally not widely available outside of the US. 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.

Literature Cited

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

Extension Working Party (S6.06-03)
Proceedings of a Symposium

Extension Forestry: Bridging The Gap Between Research and Application

July 19-24, 1998
Blacksburg, Virginia, USA

Edited by: James E. Johnson

Published by: College of Forestry and Wildlife Resources
324 Cheatham Hall, Mail Code 0324
Virginia Polytechnic Institute and State University
Blacksburg, Virginia 24061, USA
Telephone Number: (540) 231-7679
Fax Number: (540) 231-7664
E-mail: jej@vt.edu

Front Cover Photograph: Bob Veltri

Copies available at the above address for $20.00 (US). Checks payable to Virginia Tech Treasurer.

Printed in the USA by Virginia Tech Printing Services.
VT/008/0199/200/992454