ANALYSIS OF WEB SITE ACTIVITY AND TECHNOLOGY TRANSFER ACCOMPLISHMENTS

Daniel L. Schmoldt
Research Forest Products Technologist
USDA Forest Service
Blacksburg VA USA

Matt F. Winn
Forestry Technician
USDA Forest Service
Blacksburg VA USA

Philip A. Araman
Project Leader
USDA Forest Service
Blacksburg VA USA

ABSTRACT

Government research activities are coming under increased scrutiny to justify their research direction, and to validate research project existence. One way to justify research is to pay closer attention to research clientele, their needs and their willingness and ability to adopt new technologies. Because many research products are informational rather than tangible, emerging information technologies provide a well-tailored mechanism for (1) delivering research to user groups and (2) evaluating what users find valuable in that disseminated information. During the first 6 months of activity for our Web site (Oct. 1995-April 1996) we logged and analyzed demographic information about visitors and what pages they viewed. We found that almost 1/3 of the pages accessed were publication abstract pages. By comparing the subject areas of reprints requested and the subject areas of abstract pages viewed with the number of reprints requested for those subject areas, we were able to infer how well our research products are meeting our clientele's needs. For some subject areas, the type of publications produced may not be well matched with users' interests. In general, however, we found good agreement between subject matter interest and user requests for information.

INTRODUCTION

Although the primary aim of research is to create new knowledge, the most tangible product of research is new technology, Therefore, technology transfer efforts have become a critical companion to research. These efforts can take many forms, including: workshops, training seminars, newsletters, formal and specific technology transfer arrangements, professional meeting presentations and proceedings, trade journal articles, news items, and peer-reviewed publications.

Notably, the nature of research results is most often informational rather than inventions or gizmos. Therefore, information technologies should provide an effective mechanism to transfer research results to potential users (Schmoldt 1992).

The World Wide Web (WWW) provides a useful environment for communicating information. It is two-way, interactive, and multimedia. Many different types of information can be delivered—--text, graphics, sound, images, movies, documents, and most recently interactive software applications. This distributed information environment extends the role of the computer beyond traditional data and information processing into communication technology, where networked computers form a many-to-many communication medium (Harnad 1991).

Our research work unit of the USDA Forest Service, Southern Research Station (Primary Hardwood Processing Products and Recycling), initiated a WWW site* in October 1995 to inform our clientele regarding our research objectives, accomplishments, and products. The primary goals of our research unit are: (1) to identify, evaluate, and develop new or improved automated primary hardwood processing technologies and hardwood products that make United States industry more competitive in domestic and foreign markets and (2) to develop and evaluate pallet repairs and other uses of discarded pallets to extend our timber resources. The WWW server has operated 24 hours per day since its inception, providing a large volume of information about our research unit to site visitors.

We logged and analyzed demographic and usage information about our visitors for the first 6+ months of operation. This usage information provided our research work unit with valuable feedback (albeit from a Internet savvy segment of our clientele) regarding current research direction and research products. While the types and numbers of visitors to our site will differ from visitors to other forest resources Web sites, the following description of our analysis of Web site activity can serve as a guide to others that wish to use their Web site as a report card of their technology transfer efforts.

WEB SERVER SETUP AND OPERATION

Hardware and Software

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/see 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 scientist's personal

* http://www.se4702.forprod.vt.edu

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. Since that time, however, we have moved the server to another machine that is used only part-time by a staff member.

The Hypertext Transfer Protocol (HTTP) server software, MacHTTP* version 2.2, was installed on a Power Macintosh* 7100/80. The server software was very easy to set up and to begin operation. Many different, MacHTTP server operating parameters can be set using a configuration file. These parameters include, among others: the number of simultaneous connections allowed, time delay between inactivity and connection termination, CPU cycles MacHTTP steals from other running applications, data transfer buffer size, security protection, and domain name server (DNS) lookups. These parameters allow the site administrator to tailor server operation to the needs of the users.

HTML Document Organization

The content of our Web site was organized as presented in Figure 1, for the period Oct. 1995- April 1996. That organization has changed slightly since that time. Numbers appearing in this illustration indicate the number of visitor-day accesses that those particular pages received during the period in which records were kept (see the next section for a description of "visitor-days"). At the top level resides the home page with text links to the second level of Web pages and also graphical links to other government, university, and forestry Web sites. Many of the second-level links are terminal pages. The "Cooperators" and "Scientists Involved" pages, however, link to other individuals and institutions. The "Summary of Research Projects" and "Publications" pages provide links to much of the remaining information contained at our Web site. This information includes publication lists, abstracts for all publications listed, on-line publications, and forms for requesting publications not yet available on-line.

During the 6-1/2 months that we analyzed visitor usage we had approximately 180 publications; this number, of course, was continually changing, Most publications are multiply listed within 2 or more subject areas. This makes it easier for visitors to find what they are looking for because no search facility is currently provided by our site. There are now 2 ways to obtain publication reprints (Figure 2). First, if it is available, a visitor can download a portable document format (PDF) file of the publication. Then, by using Adobe* Acrobat Reader, they can read the publication on their own computer screen and print it on their printer. If a PDF version is not yet available for a publication (which was the case for all publications initially), a visitor can fill out a simple HTML form that is easily e-mailed to our site editor. The requested publications are located, photocopied, and postage mailed to the requester.

---

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

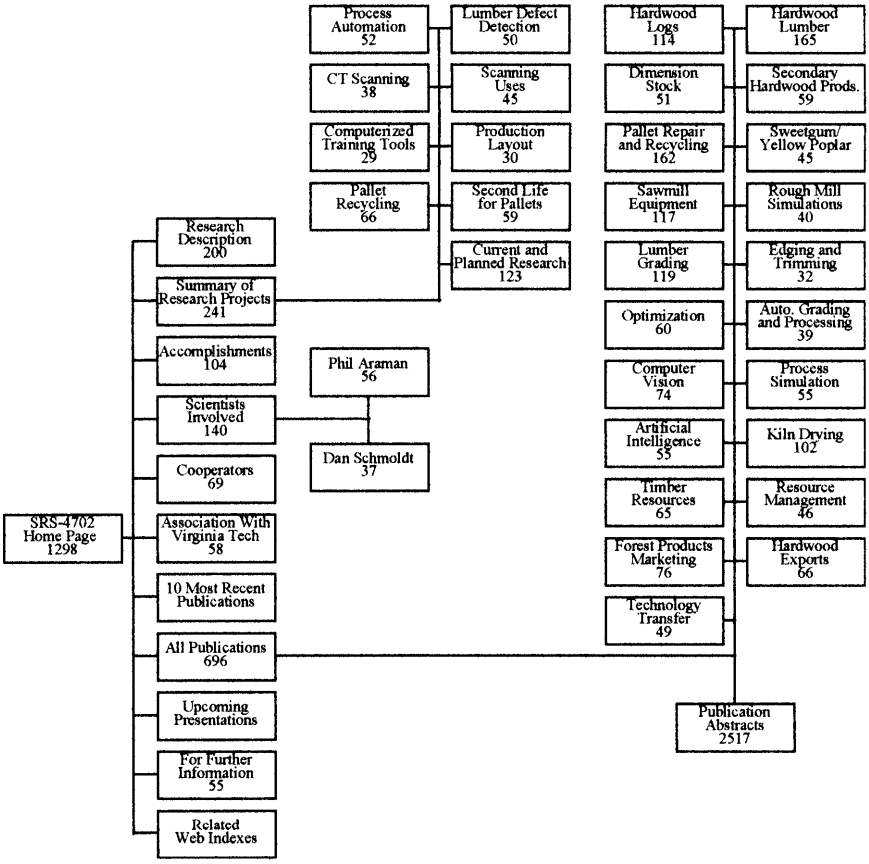

Figure l. The hierarchical organization of pages on our Web site begins with the top-most level at the left. The total number of visitor-day accesses for each page appears in each box.

HTTP Server Log File

While MacHTTP is running it records all server activity in a "log" file. Each entry in this log file records: the date and time of access, the domain name (or IP address) of the accessing machine, and the files that were downloaded. A "domain name" is a mnemonic (e.g., www.se4702.forprod.vt.edu) for an Internet Protocol (IP) address (e.g., 128.173.24 1.12). Every machine connected to the Internet has a unique IP address that identifies it, and that allows message packets to find their way from sending to receiving machine. Each HTML document (page) that is accessed by a user may result in 10-15, or more, files being immediately transferred. Each one of these file transfers is recorded as a separate entry in the log file, and constitutes what is typically referred to as a "hit". All the text in an HTML document is transferred as a single file, but many of the graphic objects that one sees on a page are separate graphic files. Consequently, the number of "hits" that a Web site receives is not a valid

indicator of site activity, but rather how “busy” (in the sense of cluttered with detail) each page is.

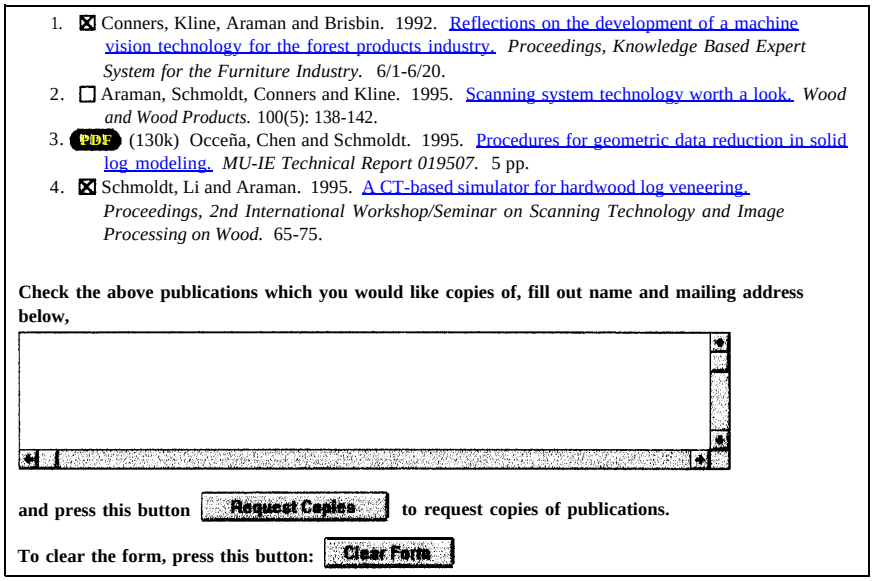

Figure 2. An example of a partial publications list illustrates: (1) on-line publications (PDF files) and (2) a reprint request form.

For the purposes of describing and analyzing visitor usage to our Web site, we have opted to define an alternative to “hits”. A single, user session is what we would ideally like to record. However, because any user access may include dozens, or hundreds, of hits over an extensive period of time, there is no way for us to reliably determine if this period of activity contains multiple sessions or a single session spread out over time. Our alternative is called a “visitor-day”. When any user accessed any one of our web pages on a given day it was considered a “visitor-day”. If the same user accessed our pages more than once in a single day it was still considered a single “visitor-day”. Unfortunately, though, because some Internet connections-in particular, modem pools—generate random IP addresses for users, a single visitor on a particular day might actually be registered as 2 or more separate visitor-days. Nevertheless, visitor-day terminology is a more realistic measure of user activity than hits counts and also enabled us to perform an analysis of visitor usage.

ANALYSIS OF USERS AND USAGE

Demographic and usage data for users was taken from the log file generated by the MacHTTP server. These ASCII log files were parsed into a spreadsheet, where extraneous hits were removed. Extraneous hits included entries representing graphic and image files served and all entries associated with access by our unit’s personnel. All pages accessed by users were retained in the

spreadsheet, but visitor-day access information was condensed out of the larger data set. Information on user type, Web pages accessed, and reprints requested was analyzed from October, 1995 through April, 1996 (approximately 6 1/2 months), Because the server was set up in October, this was considered a partial month, with consistent hits starting around October 10th.

Demographic Analysis

Domain names were examined to provide some indication of what each user's affiliation was. The last mnemonic in a domain name (e.g., "edu") is the top-level domain name (TLDN). It indicates the type of computer user (for computers in the U. S.) or it indicates the country (for computers outside the U. S.) where the computer is located (e.g., "ch" for China). This general rule has changed a little recently, there are now computers in the U.S. with the TLDN "us".

All users were placed into one of two categories: North American users or international users. North American users were subdivided into one of the following user types: government ("gov"), organization ("erg"), educational institution ("edu"), Internet access provider ("net"), commercial ("corn"), or "unknown". The government category consists of both state and federal government agencies in North America. We included in the organization category all non-profit organizations, which includes all domain names containing a "erg" TLDN, as well as those with a "us" TLDN representing a library. All user domain names ending in a "edu" TLDN were classified as educational institutions. The educational institution category also contains Canadian universities ("ca" TLDN preceded by a university name) and k- 12 schools in the U. S. ("us" TLDN preceded by a state abbreviation and "k- 12"). All domain names ending in a "net" TLDN were considered Internet access providers, Canadian users who didn't fit into the government or educational categories were also considered Internet access providers. Users connecting through an Internet access provider are usually personal accounts or small businesses.

A WHOIS lookup was done on all domain names with a "corn" TLDN to determine whether the user was commercial, educational, or an Internet access provider. If no domain name information was found, the user type was classified as "unknown". Not every machine on the Internet has a filly qualified domain name that is listed with some domain name server, therefore, not all IP addresses can be traced to a particular domain name. Also, we conjecture, without any supporting rationale, that many of our "unknown" visitors may have been search engine software robots that access and catalog Web site content.

International users were classified according to their country codes and then grouped into the continents: Europe, Asia, Africa, South America, and Australia/New Zealand.

The Internet has grown in a somewhat haphazard way over the course of several decades from a variety a separate computer networks. This has partially hindered our ability to clearly establish user demographics. Even our unit's TLDN is "edu," but we are a government site. Nevertheless, we feel that the domain-name" demographics for users that we were able to discern are adequate for this analysis, even if actual demographics are not entirely accurate.

User Activity Analysis

There were two areas of user activity that we examined, these were: (1) pages accessed and (2) publication reprints requested. We tallied the total number of accesses for each page of our Web site (Figure 1), This page-access statistic was expected to indicate the most popular or interesting or important pages across all user activity. Of special interest, we looked at which abstract pages were accessed most frequently within the various subject areas. Secondly, we also recorded information about each user request for publication reprints, In addition to the requesting user's name, postal mail address, and e-mail address, we noted the publications requested. From this information, we kept track of how often particular subject areas were requested by users, i.e. the frequency of interest in various subject areas at our site. We also kept track of how many reprints were requested from each subject area, i.e. the intensity of interest in each subject area. The combination of frequency and intensity was expected to indicate those areas of our research program that are most valuable to our Internet clientele.

There are two difficulties associated with examining activity and interest for particular subject areas, as stated above. These difficulties include an "availability bias" and a "saturation effect," and result from unequal numbers of publications within the different subject areas. Abstract page access and total numbers of reprints requested are biased by the number of publications available in each subject area. That is, subject areas with greater numbers of publications will, naturally, tend to have more reprints requested. We could normalize subject area counts by the number of publications in each subject, but this approach would ignore the fact that there is also a saturation effect. That is, someone will tire of examining abstract pages after a short time, or will only request a number of reprints that he or she can reasonably expect to read and absorb. Therefore, we need to normalize subject area counts by some value other than the total number of publications in each subject area, so that counts reconcile both the availability bias and a saturation effect.

We examined several mathematical relationships to adjust for both availability and saturation. Each relationship was a function of the number of publications in a subject area. These included, the square root, the natural logarithm, and the base 2 logarithm. They each gave relatively similar results, so we selected the base 2 logarithm of numbers of publications in each subject area to normalize subject area counts. Our intuitive rationale for this selection is that given indifferent interest in a subject area, i.e. neither strong disinterest nor strong interest, a user's curiosity will produce a binary sampling of publications on average.

RESULTS

Site visitor-day usage is presented geographically in Figure 3. By far, the greatest use was by North American users (70%), including Canada, Mexico, Central America, and the Caribbean. Users of unknown origin accounted for approximately 20% of our visitor-days. Based on the initial "unknowns' that we were eventually able to identify, we suspect that most of the remaining "unknowns" are actually of North American origin. Within North America, Internet access provider users account for more than 50% of total use. As noted above, we suspect that these user types are either personal Internet accounts or small businesses, because those are the individuals that Internet access providers typically service. Over the 6-1/2 months of initial operation, visitor-days steadily increased from 110 in the first, partial month to almost 600 in the last month.

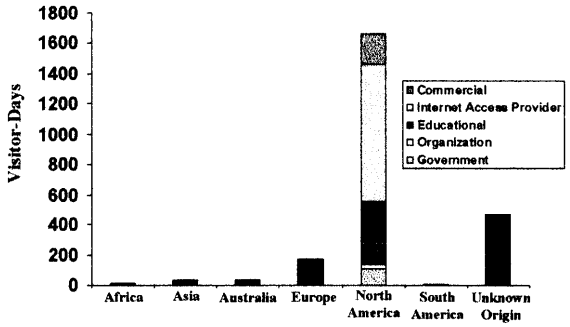

Figure 3. Visitor-days are categorized by users' continents and by access type for North America.

From Figure 1, it is apparent that publication abstract pages were viewed more than any other pages (almost 1/3 of 7768 page accesses). This is almost twice as often as the top-level home page itself was accessed (1298). This may seem somewhat counter-intuitive, but by using Internet search engines users can go directly to Web pages of information that they desire without wading down through a site's home page and other intervening pages. Also, it is possible to visit a page the first time via the site's home page, use a browser's bookmark feature, and then return to that same page repeatedly without visiting the home page first. Figure 4a shows the distribution of abstract page access by subject, where the values were normalized by the base 2 logarithm of publications per subject. "Hardwood Lumber" and "Hardwood Logs" are rather generic and can be ignored here. Then, "Computer Vision," "Marketing," "Automated Grading/Processing," and "Pallet Repair and Recycling" were the most popular subjects for abstract viewing.

During the 6-1/2 month period, 628 publication documents were requested and mailed, nearly 100 per month. These requests were made by a total of 92

users, which is approximately 7 publications per user-request. These 92 user-requests can be categorized by subject area (Figure 4b). As noted above, most of our publications are multiply categorized, so therefore, the total number of categorized user-requests exceeds the 92 actual user-requests. Aside from the pseudo-generic subject areas of "Hardwood Lumber" and "Hardwood Logs," the subjects, "Pallet Repair and Recycling" and "Lumber Grading" were requested more often than any other subject areas.

Subjects requested should not be confused with the total number of reprints requested, however. To examine total numbers of reprint requests, we can categorize the 628 reprints requested (Figure 4c), after normalizing. Here, more "Pallet Repair and Recycling" publications were requested than either "Hardwood Logs" or "Hardwood Lumber" even though the latter are relatively generic subjects. "Lumber Grading," "Computer Vision Systems," and "Sawmill Equipment" received many reprint requests also. While "Dimension Stock" and "Rough-Mill Simulations" were relatively common (ranks 6 and 9) subjects for publication requests (Figure 4b), relatively few publications (ranks 13 and 14) were actually requested (Figure 4c) for these subjects. We can also compare Figure 4a and 4c and note that subjects "Computer Vision," "Marketing," and "Automated Grading/Processing" rank high for abstracts viewed, but drop somewhat in reprints requested. "Pallet Repair and Recycling" and "Lumber Grading" shift in the opposite manner with respect to Figure 4a and 4c.

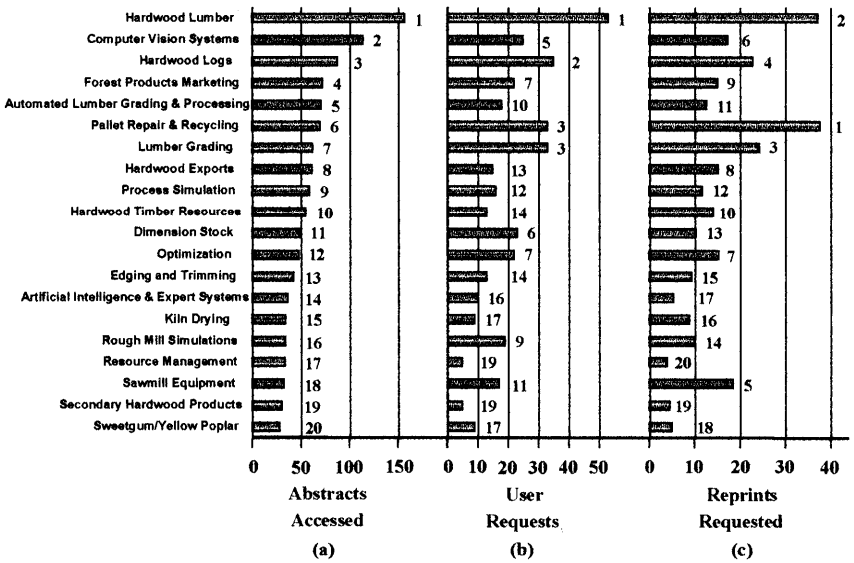

Figure 4. Abstract pages accessed (a), user requests for reprints (b), and reprints requested (c) are categorized by subject area. Values next to each bar indicate the rank of that subject's count. The values in (a) and (c) are normalized by the base 2 logarithm of publications per subject.

Requests for reprints contained more demographic information about users than was available from domain-name analysis alone. More than half of all reprint requests came from businesses or personal accesses. Although we asked that reprint requests include business information, we expect that many of the personal requests were actually unidentified business requests. Relatively few requests came from educational institutions, but a substantial number came from outside North America.

CONCLUSIONS

Visitor-day usage increased over the short time during which we've examined our Web site's activity. As more forest products professionals gain access to the Internet, we expect usage to continue rising into the near future. Eventually, though, we expect that site activity will level off at some plateau that is commensurate with both Internet access by forest products professionals and with general Internet activity.

We've found that the time required to locate, copy, and mail reprint requests has become burdensome. Consequently, we have recently begun, to make publications available electronically, as PDF files. The time required to scan and process each publication into a PDF file varies greatly. However, results are equal or better in quality to professional reprints.

We do not have detailed and accurate demographics on this Internet segment of our research clientele and how this sub-population might reflect our larger clientele population. Nevertheless, the mix of government, educational, personal, business, and international connections that we received is not dramatically out of line with what we might expect for our overall clientele group. Therefore, we feel that it is possible to draw some reasonably valid conclusions from this "cyber-survey."

By closely examining Figure 4, we can gain some insight into how well our research is being received by this Internet segment of our user community. The subject of "Pallet Repair and Recycling" is of interest to many of the users requesting reprints. In addition, many reprints were actually requested for publications in this area. This positive association seems to indicate that the topic has frequent interest (a popular subject, Figure 4b) and the research being conducted is noteworthy and valuable (high reprint popularity, Figure 4c). Other subjects, such as "Rough-Mill Simulations" and "Dimension Stock," appear to have a negative association, however, i.e. high subject interest but relatively few research reports requested. This might indicate that the listed publications do not meet clientele needs, and that we might want to revise our research direction or, alternatively, include HTML links to other Internet sources of this information. In this particular circumstance, for example, the Robert C. Byrd Technology Development Center in Princeton WV has continued and extended rough-mill simulation work begun initially by our research work unit. Because we no longer conduct much research in this area, we could expressly provide links to the Byrd Center's Web site from our simulation publications page.

In contrast, "Hardwood Exports" has an opposite association, although it is not extremely popular as a subject (13th), research reports in that subject were requested more frequently (8th) than expected given its popularity. This mix of relatively high intensity of interest despite relatively low frequency of interest for "Hardwood Exports" might indicate that those publications have done a good job of informing the users with an interest in that area.

Abstracts accessed (Figure 4a), may give some indication of users' desire to be informed about certain subjects. But, while this desire is high for "Computer Vision," "Marketing," and "Automated Grading/Processing" (Figure 4a), the technical details presented in the actual publications may not adequately satisfy that need for our user audience. Hence, the number of reprints actually requested (Figure 4c) for those subjects is lower than might be expected from abstract viewing statistics. In light of these results, we might want to re-examine the types of studies being conducted in some of these research areas, as well as, how research results are reported, e.g., the number of technical vs. nontechnical articles that are published. For some of the more technical areas, e.g. computer vision, articles that are less technical and more popular in nature may have greater interest for our clientele.

We have continued to keep track of reprint requests even after April 1996. In the 14+ months from October 1995 through December 1996, we have sent out over 2900 publications to more than 375 visitors, as a result of Web site visitation. This amounts to approximately 50 reprints per week. The total time required to locate, copy, and mail all these documents underscores our interest in making publications available on-line.

The relatively large number of reprint requests from business users is both surprising and encouraging. In the forest products community, this group constitutes the largest segment of users for our research and technology. Their interest in our research work unit's publications indicates that we are conducting research that is important to them. Also, the relatively good agreement between frequency and intensity of interest in most research subjects suggests that our studies have effectively addressed technology transfer to this user group.

REFERENCES

Harnad, S. (1991), Post-Gutenberg galaxy: the fourth revolution in the means of production of knowledge. *The Public-Access Computer Systems Review*, 2(l): 39-53.

Schmoldt, D. L. (1992). Bringing technology TO the resource manager.. and not the reverse. In ASPRS/ACSM/RT 92 Convention: Monitoring and Mapping Global Change, vol. 5. American Society for Photogrammetry and Remote Sensing, American Congress on Surveying and Mapping, Bethesda MD, pp. 62-74.

1997
ACSM/ASPRS

Annual Convention & Exposition
Technical Papers

ACSM 57th Annual Convention
ASPRS 63rd Annual Convention

Seattle, Washington
April 7-10, 1997

Volume 4
Resource Technology Institute