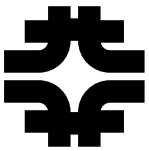


Fermi National Accelerator Laboratory

FERMILAB-TM-2056

Digitizing Legacy Documents: A Knowledge-Base Preservation Project

Elizabeth Anderson, Robert Atkinson, Cynthia Crego, Jean Slisz and Sara Tompson

Fermi National Accelerator Laboratory
P.O. Box 500, Batavia, Illinois 60510

September 1998

Operated by Universities Research Association Inc. under Contract No. DE-AC02-76CH03000 with the United States Department of Energy

Disclaimer

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Distribution

Approved for public release; further dissemination unlimited.

Fermilab-TM-2056

Digitizing Legacy Documents: A Knowledge-Base Preservation Project

by Elizabeth Anderson, Robert Atkinson, Cynthia Crego, Jean Slisz and Sara Tompson

I. Introduction

The Fermilab Library was awarded a 1998 Educate and Automate grant from the Illinois State Library in the Digitizing Illinois Collections category.[1] We applied for this grant in order to make our unique collection of scientific and technical reports publicly accessible and searchable on the Internet.

Since 1995, the Fermilab Library has been involved in the Technical Publications Fileserver Project. This project has made current scientific and technical publications available full text from our World Wide Web server and searchable via the Library's online catalog. Our Web usage statistics (**See Figure 1**) demonstrate the significant use scientists, engineers, students and others make of this Internet-accessible knowledge base. Since the project began, we have gone from about 7,000 hits in 1995, to over 400,000 hits annually. (Which means approximately 40,000 documents are served to requestors each month.)

We have a large legacy collection of scientific and technical reports that we wanted to make publicly available. These reports, authored between 1972 and 1995, are held in paper format only. The legacy collection includes a large number of "preprints," pre-publication versions of journal articles and conference presentations.

As more library customers and staff throughout the world come to rely upon rapid electronic access to fulltext documents, there is increasing demand to also make older documents electronically accessible. Illinois State Library grant funds allowed us to purchase hardware and software necessary to answer this demand. We created a production system to scan our legacy documents, convert them into Portable Document Format (PDF), save them to a server for World Wide Web access, and write them to CD discs for distribution. We hope our experience provides some guidance to others who are digitizing legacy collections.

II. Hardware and Software

Note: The authors are not endorsing any of the products mentioned in the following sections.

We already owned PC workstations and a UNIX server machine. Each workstation is a Micron Millennia Pro with a 200 Megahertz Pentium processor, a PCI 32-bit EIDE controller, 64 MB of memory (RAM), 2.1 GB EIDE hard drive, and a Micron 12x CD-ROM drive. The system board has 2 ISA slots and 4 PCI slots, 2 of which are taken by a Diamond Stealth video card and the network interface card.

The Fermilab reports are digitally archived on and served from our Web server machine, a SUN Sparc 20 with 256MB RAM, 2 1.05 GB hard drives plus an 8.4GB multidisk pack. This disk pack is configured under the Andrew File System (AFS). A number of physics laboratories share AFS space, which allows users to mount these volumes and use the documents as if they are locally accessible.

A. Hardware

We purchased the following hardware with grant funds:

- Two Microtek ScanMaker, Model E6, scanners, with Automatic Document Feeders
- Two Sony Spressa CSP-960S CD-recordable internal drives
- Two Adaptec 2910 SCSI Adapter Cards with 32-bit PCI bus (came with the Sony, but we also use for the scanner)
- TDK 74-minute recordable CD discs
- Micron Notebook PC

We purchased two scanners and CD-R drives so we could speed up production by having two people working at once, and purchased the notebook PC so we could set up mobile workstations for using Adobe Capture from our network. The Microtek ScanMaker E6 is a 30-bit, single-pass, flatbed scanner with an Automatic Document Feeder. The scanner is a SCSI device and came bundled with an ASPI-compliant SCSI interface card and the Microtek ScanWizard software. Scanners with a wide variety of features are available, but this less expensive scanner met our needs since we have same-sized documents and do not require high-resolution graphic output.

Since the SCSI host adapter card that came with the Microtek scanner was for an external device only, we used the Adaptec 2910 SCSI Adapter card for a PC workstation with a 32-bit PCI bus, which came with the SONY Spressa. Installing a SCSI host adapter is a necessity. Almost all CD-R drives are SCSI because SCSI provides the performance needed to "burn" a disc.[2]

CD recording or writing, known as CD-R, is rapidly becoming the medium of choice for storing scanned and digitized images of archival documents. [3] Others embarking upon this method of archival storage need to realize that writing to a CD-ROM is a CPU-intensive process. A recommended minimum configuration for the workstation to be used for CD-R is a 486DX2066 with 16MB RAM. [4]

The Recordable CD-ROM drive we selected was the SONY Spressa CSP-960S that came bundled with the SONY CD Right software. We purchased an internal drive since CD-R drives are not considered as robust as floppy drives or CD-ROM drives. [5]

TDK 74-minute, 650MB recordable discs were purchased for production. We chose the gold phthalocyanine discs rather than the green cyanine discs for archival purposes. The phthalocyanine dye is less sensitive to ordinary light and therefore will probably last longer and preserve information better. All phthalocyanine discs are supposed to last over 100 years.[6] It is important to consider, however, that some CD players and CD-ROM drives will read data recorded on cyanine media more readily and reliably than they will read data recorded on phthalocyanine media.

B. Software

Microtek ScanWizard software came bundled with the scanner, and SONY CD Right software came bundled with the Sony Spressa drive. A large portion of the grant funding went for one software product, Adobe Acrobat Capture. Capture was purchased to convert the scanned output files, which are in TIFF (Tagged Image File Format) graphics format, to easily accessible PDF

(Portable Document Format) files. Capture allows us to scan documents and convert the scanned TIFF files to PDF format in one package. We purchased a version of Capture with a license that allows scanning of one million pages.

C. CD-R Installation

As is the case when setting up any computer system, it was important for us to pay attention in advance to the details of the configuration. We checked available expansion slots and bays in our workstations and determined the type of interface each device required, and made sure we had enough hard disk space to hold the data. In reviewing our configuration, we determined we needed an additional SCSI cable for each machine since the provided cables did not allow for one internal and one external device.

We began by installing the SCSI card into an available PCI slot and noting its ID, or address. The regular SCSI 2 system can handle 8 devices. Each device has to be assigned a unique ID, going from 0 to 7. The host adapter is considered a device itself and typically will occupy ID 7, although no actual standards exist to assign a particular ID to a certain type of device. In addition, the Interrupt Request (IRQ) address must be unique for each device. Duplication of SCSI IDs or IRQs can cause problems – such as the device not being recognized by the host computer – so it was vital that we made sure each device had a unique ID.

We installed the CD-R into an empty bay located below our existing CD-ROM drive and secured it with the provided mounting screws. We then connected the supplied ribbon cable from the CD-R to the SCSI host adapter card. We also connected a grounding wire that came with the

CD-R. We booted the system with the Easy-SCSI device-detection software provided with the Adaptec SCSI card to insure that both devices had unique IDs and were being recognized.

Next we attached a “Mini-50 to Centron” SCSI cable from the external port on the host adapter card to the Microtek scanner. We changed the SCSI ID on the scanner, using the SCSI ID wheel, which was external. Since the scanner was the last device in the SCSI "chain," we used a terminating resistor on the cable between the cable and the scanner port.

III. Process

A. Scanning Documents

The hardware and software detailed above is necessary for the legacy collection production process of scanning, converting, Web posting and CD-ROM writing. We began our scanning and conversion process by checking that all of our original paper documents were single-sided and clean with no marks, folds, creases, or bends. Necessary corrections were made with correction fluid and double-sided copies were copied single-sided on our copier.

The first step we took prior to scanning our first document was to set up in Capture both an input and an output folder for managing our documents during the scanning and conversion process. As noted above, Capture can be used for both scanning and converting. (**See Figure 2, which illustrates both processes in the main Capture screen.)** First, we created an input folder titled “preprints” by clicking on *add input folder* under the Input menu.

We chose the following settings for the input folder by right clicking on the "preprints" folder:

- Collate all Selected Files
- Save files in c:/capture2

Next, under the Output menu, we created an output folder titled "preprints," and set the following settings by right-clicking on the "preprints" folder:

- File formats: Acrobat PDF [Image + Hidden Text]
- Processed Images: Retain in Place
- Suspect Settings: 95%
- Click all for ACD
- Under *If Word Confidence* . . . click PDF
- Under *If Word Has* . . . click PDF

Adobe Acrobat Capture gives the user the option to choose between three file formats for conversion to PDF:

- Acrobat PDF [Normal]
- Acrobat PDF [Image]
- Acrobat PDF [Image + Hidden Text]

After testing all three file formats, we found that the "Image + Hidden Text" format best suited our needs. Acrobat Capture will highlight words in both the "Normal" and "Image" formats that it does not recognize and save them as bitmap images in the PDF file. This often creates PDF files that contain bold highlighted words, increased font sizes from letter to letter in words, and

characters and fonts not identical to the original. Documents created in one of these two formats needed additional proof reading and editing. We found this unacceptable. However, these two file formats did produce PDF files significantly smaller than those formatted in the "Image + Hidden Text" option. Since we were not concerned with file size, and since "Image + Hidden Text" results in an exact copy of the original paper document with no recognition suspects indicated, we chose it.

We are scanning each page of our documents using the following settings on our Microtek scanner:

- Type: line art
- Resolution: 300 dpi
- 8.5 x 11, 100%

To begin the scanning process, we load a document into the scanner document feeder and click the Scan button. We name each document by its Fermilab-assigned document number. Scanning is automatic, and when finished, all the files (TIFF format images) are organized in the Capture "preprints" input folder.

As an extra quality control step, we have been proofing each TIFF file in Adobe Photoshop. In our proofing process, we look for incomplete scans, any severe text shifts or missing pages. We have discovered that slightly slanted or crooked text in a TIFF document will automatically be straightened by Capture when converted to PDF. Note that screen resolution of TIFF images can be poor. However, the resolution improves once the pages are converted to PDF. If we see an

error with a TIFF file, we re-scan that page(s), renumbering it in the Capture "scan filename" dialog box. Renumbering is important, as it keeps the order of the pages intact. Note that we already owned Photoshop and use the software for other applications. If you don't own this software, TIFF files can be proofed using the Capture preview option.

B. Converting Documents

Back in Capture, we set the following preferences for converting the TIFF files to PDF by right clicking on the Process button:

- Performance Preference: Most Accurate
- Primary Language: English (US)
- Page Orientations: Portrait

We then select all TIFF files and click the Process button. We name the resulting PDF document using the same Fermilab-assigned document number. During the conversion process, the "Processing Status" lights appear. These indicate the processing steps as they occur.

When the conversion process is finished, the final PDF file appears in the "preprints" output folder. We then open the PDF file using Adobe Acrobat Exchange 3.0 and proof it against the original, looking for missing pages, missing text, etc. If we find an error, we then re-scan the page and run the document through the conversion process again. If we do not find an error, we transfer the PDF file over to our Webserver. Our department maintains a Web page with links to all of our reports at http://fnalpubs.fnal.gov/techpubs/pubs_lists.html. The reports are arranged

on this page in chronological order of date posted. We use the Library online catalog as the search engine for all reports.

We use Winsock FTP (a shareware program) to transfer the file from the hard drive on our PC workstation to the archive directory, our anonymous ftp area on our Webserver. We store all our documents in this area for retrieval by our customers. This directory, and thus all documents, are accessible via anonymous ftp, AFS space, our preprint Web pages, and the Fermilab Library online catalog.

We then add the URL for the PDF file to a basic HTML Web page. Each one of our reports has a corresponding HTML page with a link to the electronic file. Once the URL is added to the HTML page, the final step we take is to test that the link is working properly by viewing the page in our Web browser, currently Netscape 3.01.

C. CD-ROM Recording

Our first priority is posting the reports on the Web, but the grant support is also allowing us to prepare CD-ROMs each containing a full year of reports. The new CD-R technology makes CD-ROM writing a viable option for archival storage and information exchange.

A debilitating problem early on in the short history of CD recording was “buffer underrun.” CD-ROM drives have an internal data buffer that stores data being transferred to the CD-R disc. When the data can't get to the CD recorder fast enough to support the continuous data stream required, a “buffer underrun” may occur. Most so-called "third generation" recorders have

sufficient buffer sizes for most recording needs, including the SONY Spressa 960S. These new recorders create discs in ISO-9660 format so it should be readable in any computer equipped with a CD-ROM drive.

Copying the files to CD-R has proven to be surprisingly simple. Once the documents are scanned and placed on our server, we use ftp to put the files on our workstation's hard drive. We launch the CD-Right software and click on the "Data CD" option from the main menu **(See Figure 3).**

From Windows NT Explorer, we then drag and drop (or cut and paste) the files to the "Recording List" area of the CD-Right window (See **Figure 4**.)

We give the volume a meaningful name and insert a blank CD-R disc. In a few seconds, the software determines whether the disc is blank or had data written to it from a previous session (See **Figure 5**.)

When the software indicates it is ready, we click on the Record button. We select the speed and then have to chose whether we want to "Test," "Test & Record if OK" or just "Record." We have decided to "Test & Record." At a 4X speed, we are writing approximately 600 kilobytes per second. Keep in mind that a faster recording speed requires that the stream of data to the disc be continuous and therefore is more demanding of the CPU and hard drive.

The software begins recording, shows its status, and alerts one with sound and a notification window when finished. To "Test & Record" 100MB with our configuration takes less than 5 minutes.

The SONY Spressa 960s and the CD-Right software both support multisession, that is, the ability to record different sets of data to the same disc at different times, or sessions. Some older CD-ROM readers cannot read discs recorded in multiple sessions. To ensure maximum readability, we chose to copy all of our files to our hard disk, then write them all to the CD in one session. Again, our workstation configuration of a 2.1 GB hard drive made this possible.

Although our recording production has been running very smoothly, we have learned a few lessons:

1. Be sure to close all other applications while recording CDs. Memory-resident programs, such as calendars or screen savers may interfere with the recording process.
2. CD Right uses the Windows Temp directory, so make sure that directory has twice as much free space as the largest file you intend to record.
3. Keep the disc and the recorder free of dust.

4. CD-ROM Organization

The TDK recordable media we purchased came with jewel boxes. We are purchasing preformatted CD-ROM label paper and will create labels for the boxes in-house. We will also include an insert with a contents note plus basic instructions on reading the CD. We will be placing a copy of Adobe's free Acrobat Reader on each CD. Acrobat Reader is necessary for

viewing and printing the PDF format reports. The CDs will be readable by all Windows/Intel-based workstations.

IV. Cataloging

The Fermilab Library began adding records for preprints and technical reports to our online catalog in 1993. The catalog is part of a Data Research Associates (DRA) integrated library system, which runs under VMS on a MicroVAX 3400 minicomputer.

In March 1994, the Library of Congress Network Development and MARC Standards Office promulgated a new edition of the USMARC Format for Bibliographic Data. In it was a specification for a new field, Electronic Location and Access, indicator 856.[7] In June of that year, we began using this field to hold URLs in some of our catalog records. In June 1995, we began including URLs for electronic fulltext reports and preprints.

This paid off when, in late 1995, we purchased the DRA Web interface to our catalog. The 856 fields we had been adding became clickable hyperlinks from our Web catalog records directly to the resources represented by the records. It therefore became possible to find a record for a preprint or report in our catalog, click on its URL link, and view or print the complete document. Some free browser helper applications are necessary in order to view and print these documents, including Adobe's Acrobat Reader for documents in PDF format, and the Ghostview Suite of applications for those documents in Postscript format (some of our reports are already on the Web in Postscript format).[8] The ability to go right to the fulltext document from our catalog has became a routine function used daily by library customers and staff.

All Fermilab scientific and technical reports and preprints are fully searchable in the online catalog. The catalog is publicly available on the Web at: http://fnlib.fnal.gov/MARION. Extensive online help is available for the catalog at: http://www-lib.fnal.gov/library/help/help.html.

V. Conclusion

A side benefit of digitizing the legacy document collection will be easier physical management of the documents. Currently we have five to six years of paper reports available in the Library for checkout. These stapled packets of paper are difficult to shelve and to keep in order, and take up many linear feet of bookcase ranges. Once we have all Fermilab reports available in two places, on the Web and on CD-ROM, we should be able to dispense with the paper copies, allowing customers to print the documents at the time they need them, rather than requiring the Library to store the reports just in case customers need them. We believe the PDF files on the CD-ROMs will be accessible for years to come, and represent a good format for archiving information. We will, however, retain one paper copy of each report, but only one.

Ultimately all Fermilab reports will be searchable and fulltext retrievable via the Fermilab Library Online Catalog. We anticipate having the documents back through 1990 available by the end of 1998, and earlier years will follow.

For those who do not have Internet access, the Fermilab Library can print, fax or e-mail documents to fulfill interlibrary loan or direct requests. We do not charge for interlibrary loans.

By the end of 1998, we will begin distributing a full year's worth of Fermilab reports on CD-ROM for a minimal charge. The $20,000 Digitizing Illinois grant we received from the Illinois State Library is allowing us to make the entire body of scientific knowledge developed at Fermilab readily available to a broad audience.

Notes

1. FY98 Educate and Automate Grant in the Digitizing Illinois Collections category: "Pioneering Science on the Energy Frontier: A Historic Knowledge-Base Preservation Project." Grant number 98-8015.

2. Miastkowski, Stan. "Install a CD-Recordable Drive." ***PC World Online*** (October 1997). (On the Web at: http://www.pcworld.com/hardware/cd-rom_drives/articles/oct97/1510p356.html)

3. Parker, Dana J. and Robert A. Starrett. ***CD-ROM Professional's CD-Recordable Handbook: The Complete Guide to Practical Desktop CD***. (Wilton, CT: Pemberton Press, 1996), p. 17.

4. Angus, Jeffrey Gordon and Carla Thornton. "Do It Yourself CD-ROMs." ***PC World Online*** (January 1996). (On the Web at: http://www.pcworld.com/hardware/cd-rom_drives/articles/jan96/jan9644.html)

5. Parker and Starrett, p. 91.

6. Parker and Starrett, p. 88.

7. The current specification is dated July 1997 and is in v.2 of the Format, which is available from the LC Cataloging Distribution Service. Related information is also available on the Web at: http://lcweb.loc.gov/marc/

8. See the Library's Web page at http://fnalpubs.fnal.gov/library/software.html for more information on helper applications.

The authors

Elizabeth Anderson, Robert Atkinson, Cynthia Crego, Jean Slisz and Sara Tompson are all members of the Information Resources Department at Fermi National Accelerator Laboratory in Batavia, Illinois. Anderson is the Systems Librarian, Atkinson is the Collection Development Coordinator, Crego is the Department Manager, Slisz is the Technical Editor and Tompson is the Library Administrator.

Figure 1

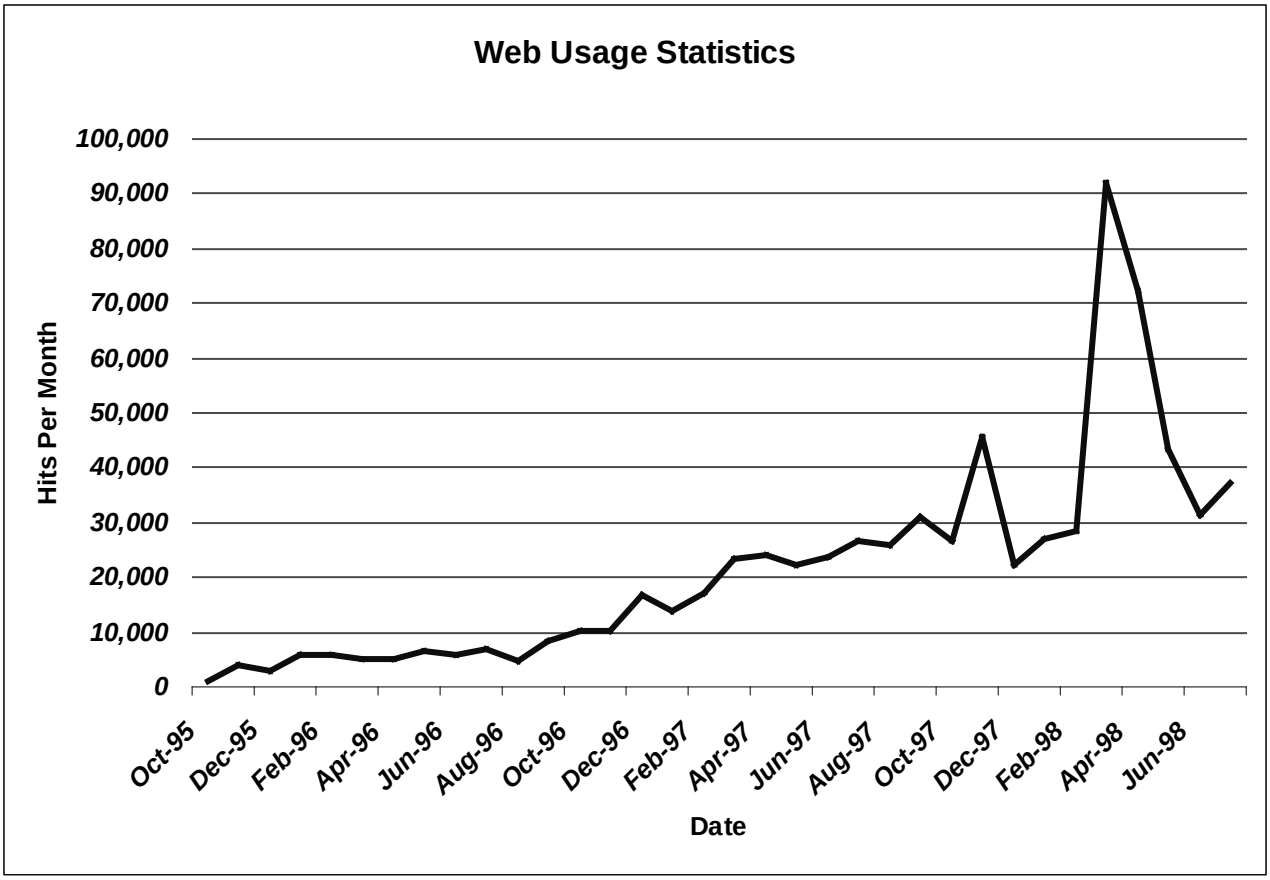

Web Usage Statistics
Hits Per Month
100,000
90,000
80,000
70,000
60,000
50,000
40,000
30,000
20,000
10,000
0
Oct-95
Dec-95
Feb-96
Apr-96
Jun-96
Aug-96
Oct-96
Dec-96
Feb-97
Apr-97
Jun-97
Aug-97
Oct-97
Dec-97
Feb-98
Apr-98
Jun-98
Date

Figure 2

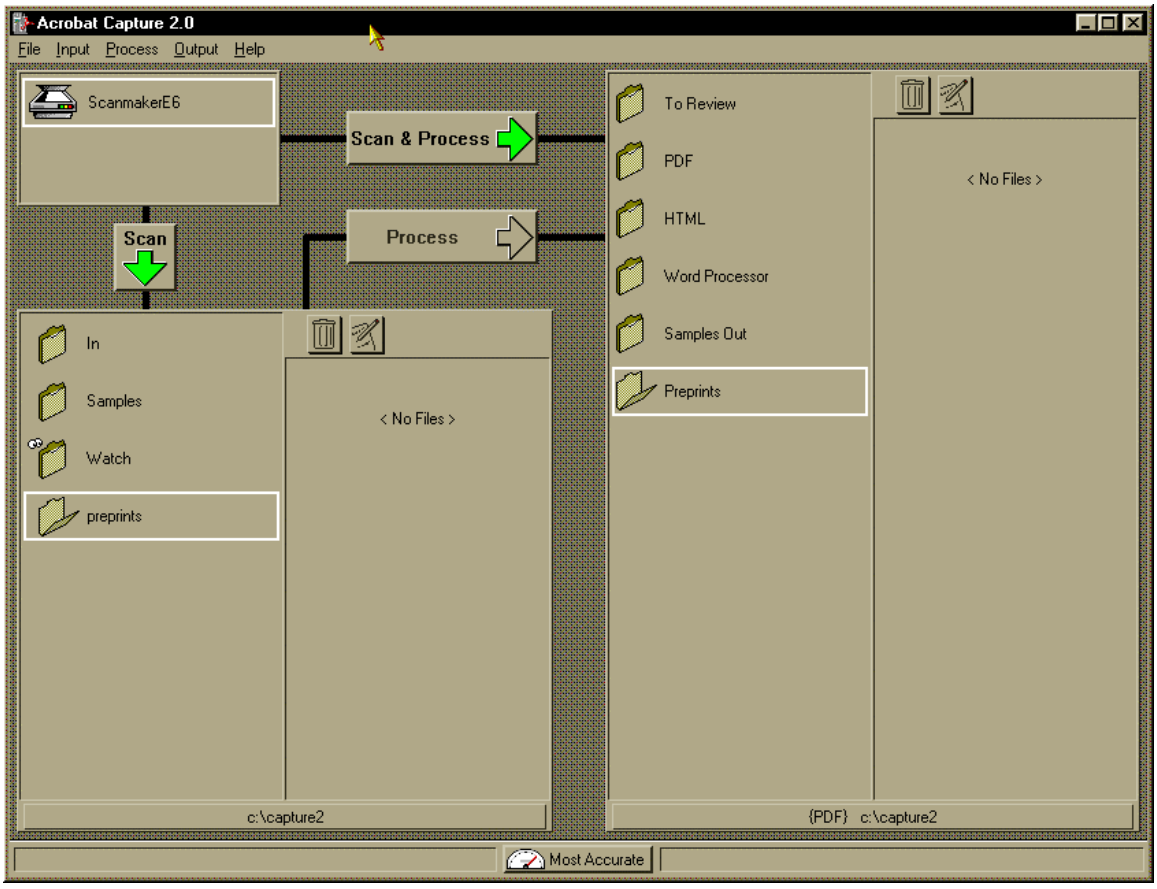

Acrobat Capture 2.0
File Input Process Output Help
ScanmakerE6
Scan & Process
Scan
Process
In
Samples
Watch
preprints
< No Files >
c:\capture2
To Review
PDF
HTML
Word Processor
Samples Out
Preprints
< No Files >
{PDF} c:\capture2
Most Accurate

Figure 3

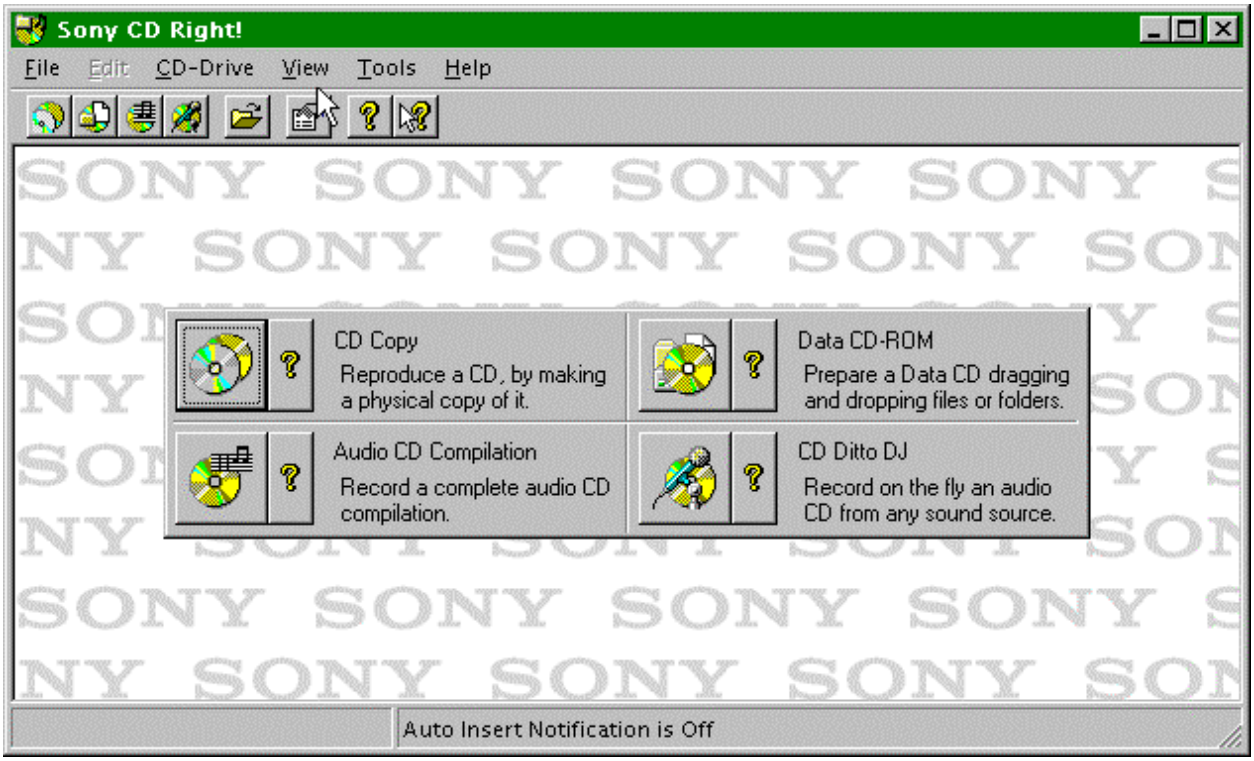

Sony CD Right!
File
Edit
CD-Drive
View
Tools
Help
CD Copy
Reproduce a CD, by making a physical copy of it.
Data CD-ROM
Prepare a Data CD dragging and dropping files or folders.
Audio CD Compilation
Record a complete audio CD compilation.
CD Ditto DJ
Record on the fly an audio CD from any sound source.
Auto Insert Notification is Off

Figure 4

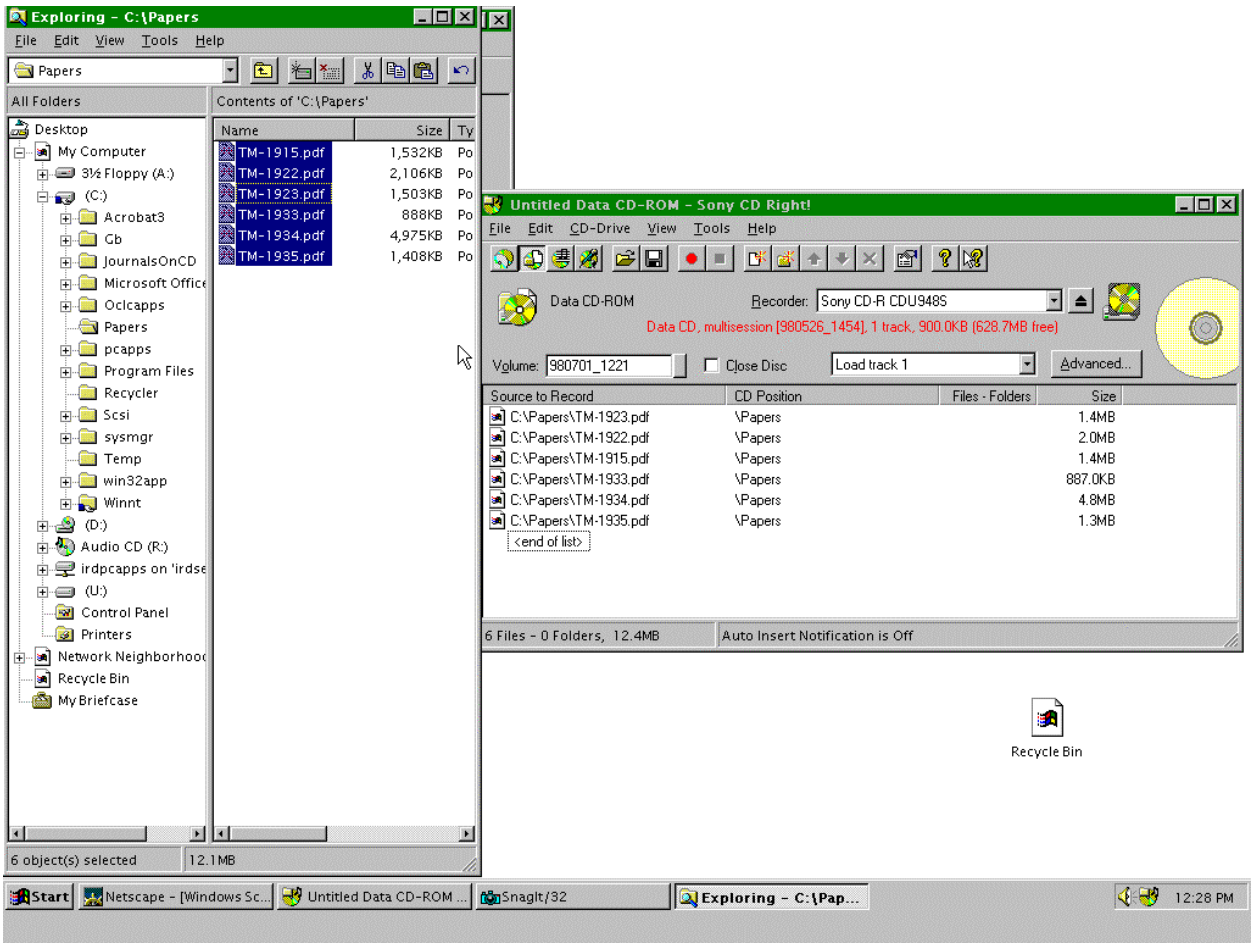
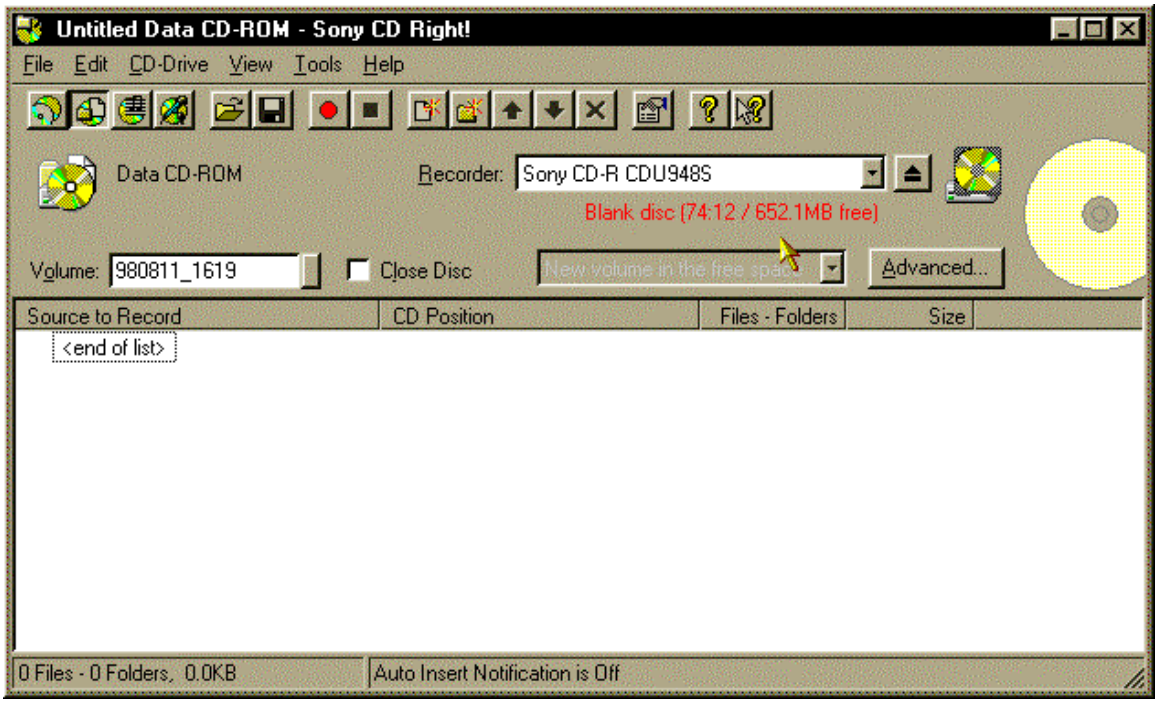

Exploring - C:\Papers
File Edit View Tools Help
Papers
All Folders
Contents of 'C:\Papers'
Name Size Ty
TM-1915.pdf 1,532KB Po
TM-1922.pdf 2,106KB Po
TM-1923.pdf 1,503KB Po
TM-1933.pdf 888KB Po
TM-1934.pdf 4,975KB Po
TM-1935.pdf 1,408KB Po
Desktop
My Computer
3½ Floppy (A:)
(C:)
Acrobat3
Gb
JournalsOnCD
Microsoft Office
Oclcapps
Papers
pcapps
Program Files
Recycler
Scsi
sysmgr
Temp
win32app
Winnt
(D:)
Audio CD (R:)
irdpcapps on 'irdse
(U:)
Control Panel
Printers
Network Neighborhoo
Recycle Bin
My Briefcase
6 object(s) selected
12.1MB
Untitled Data CD-ROM - Sony CD Right!
File Edit CD-Drive View Tools Help
Data CD-ROM
Recorder: Sony CD-R CDU948S
Data CD, multisession [980526_1454], 1 track, 900.0KB (628.7MB free)
Volume: 980701_1221
Close Disc
Load track 1
Advanced...
Source to Record CD Position Files - Folders Size
C:\Papers\TM-1923.pdf \Papers 1.4MB
C:\Papers\TM-1922.pdf \Papers 2.0MB
C:\Papers\TM-1915.pdf \Papers 1.4MB
C:\Papers\TM-1933.pdf \Papers 887.0KB
C:\Papers\TM-1934.pdf \Papers 4.8MB
C:\Papers\TM-1935.pdf \Papers 1.3MB
<end of list>
6 Files - 0 Folders, 12.4MB
Auto Insert Notification is Off
Recycle Bin
Start
Netscape - [Windows Sc...
Untitled Data CD-ROM ...
SnagIt/32
Exploring - C:\Pap...
12:28 PM

Figure 5


Untitled Data CD-ROM - Sony CD Right!
File Edit CD-Drive View Tools Help
Data CD-ROM
Recorder: Sony CD-R CDU948S
Blank disc (74:12 / 652.1MB free)
Volume: 980811_1619
Close Disc
New volume in the free space
Advanced...
Source to Record
CD Position
Files - Folders
Size
<end of list>
0 Files - 0 Folders, 0.0KB
Auto Insert Notification is Off