Electronic Distribution of Technical Reports and Working Papers: A Simple Cooperative Approach

James C. French

Technical Report No. CS-92-27

This work was supported in part by grants from Simpson Weather Associates, Virginia Center for Innovative Technology grant no. VCIT CAE-91-012, National Science Foundation grant no. CDA-8922545-02, and Department of Energy grant DEFG05-88ER25063.

Electronic Distribution of Technical Reports and Working Papers: A Simple Cooperative Approach†

James C. French
Department of Computer Science, University of Virginia

Abstract. The electronic distribution of technical reports, journal preprints and other working papers makes both economic and ecologic sense. It can be more efficient and less resource intensive than current methods of distribution. More importantly, in the case of engineering and the physical sciences the cutting edge of research is often found in technical reports and working papers. Journals have assumed an archival role because of long publication delays. This report describes the approach that we have taken at the University of Virginia and the software, techrep, that we have built to provide access to electronic archives of this sort.

1. Introduction

It is common practice for research institutions to advertise and distribute their technical reports both to disseminate results and to promote the institution. The current manual methods of distribution are cumbersome and costly, and leave room for considerable improvement. The costs incurred in connection with the distribution of technical reports — copying charges, postage, and personnel time — can be significant. In this paper we describe a specific proposal for providing an online archive of technical reports and working papers.

There are many ways one could choose to deliver technical documents. At the present time there are five ways to get technical reports from the University of Virginia Department of Computer Science. They are

- anonymous ftp
- electronic mail via mail server or human respondent
- remote access via a utility program (techrep)
- CWIS (campus wide information system)
- hardcopy by surface mail or FAX

The four online methods are simply different interfaces to the same online archive. The organization and maintenance of this archive is the subject of this paper.

2. Why Compromise is Necessary

Although the process of establishing standards is extremely important to the long-term viability of electronic document distribution, the situation demands action now. We can only hope to make choices that will not prevent the timely adoption of standards as they become available. The short-term solution must clearly be a compromise or nothing will get done at all. We recognize that each institution will have its own convention for forming technical report numbers and perhaps, multiple report series. There will be many particular details over which one could argue, but there is sufficient commonality to warrant a combined effort now.

Simply creating FTP sites is not enough. There should be some unifying concept. But, we should endeavor to unify the organization of all the online archives into a useful whole without trampling on the desire of individual sites for local autonomy in the management of their part of the global archive. The idea is to make the fewest demands on participating sites while achieving the greatest degree of utility for all participating sites.

†This work was supported in part by grants from Simpson Weather Associates, Virginia Center for Innovative Technology grant no. VCIT CAE-91-012, National Science Foundation grant no. CDA-8922545-02, and Department of Energy grant DEFG05-88ER25063

The following sections describe areas in which some compromise is necessary if this goal is to be met. The areas discussed include the document interchange format, the page ordering within documents, file naming conventions, and the choice of a Unix platform.

2.1. The Document Interchange Format

For a variety of reasons, PostScript and plain ASCII text are the obvious choices for the distribution of documents. For now we must avoid document markup because there is no universally accepted standard and every vendor is idiosyncratic. Although PostScript is a page description language not a document markup language, it has several advantages at the present time. Among them are:

- It is produced by all the wordprocessing software in common use today.
- PostScript is ASCII and easy to communicate across networks.
- It can be used to transport mathematics, graphics, and plots, all of which occur in technical papers.
- It is reasonably portable and printers are widely available.
- Online previewers are available.

PostScript also has some drawbacks:

- Because it is a page description language, all document markup is lost. This makes it difficult to create a browser or extract the text.
- It is not completely portable. At least two problems arise here: (1) differences in revision levels; and (2) some products assume that a prolog has be downloaded into the printer and therefore generate incomplete PostScript files that are not standalone.

2.2. Page Ordering

The order in which pages are generated in a PostScript file is generally determined by default settings of formatting software. These settings are often chosen so that documents collate properly when printed. Sometimes this means that the pages of an n page document are printed in reverse order, i.e., page n, $n-1$, $n-2$, etc. This makes online previewing difficult and somewhat reduces the utility of the archive. We believe that whenever possible online documents should be formatted to print in normal order.

2.3. File Naming Conventions

While it is unreasonable to expect every institution to agree on a common report numbering scheme, it is not entirely unreasonable to hope for a common file naming scheme. It would simplify matters considerably to have the base document file name correspond to the technical report

File Suffixes	
ps	PostScript file
txt	ASCII text file
Z	compressed file

Table 1

numbers. It would also be helpful to agree on file suffixes separated from the root name by a period. The common suffixes shown in Table 1 will suffice for our purposes. Thus, for example, a file named CS-90-21.ps.Z is a compressed PostScript file containing technical report CS-90-21 while the file CS-90-21.txt is a plain ASCII version of the same report.

2.4. Unix Platform

Initially it is convenient to assume that an online archive will be hosted by a machine running the Unix operating system. This is not strictly necessary, but it does provide a common framework for discussion. It may be that some of the ideas discussed below are not realizable on other platforms, but the basic common archive should not present any problem.

3. Strategy for an Online Archive

In this section we will outline an approach for the organization and maintenance of an online archive of electronic documents. This approach is based on the following three elements:

1. An online repository of PostScript and ASCII text files.
2. A file that serves as a table of contents.
3. A citation database.
4. One or more suitable user interfaces.

Elements (1) and (2) above together with a suitable user interface are sufficient to deliver documents over the Internet. The addition of a citation database facilitates the maintenance of the archive. Note that many user interfaces are possible. The four that we use are discussed in more detail below.

3.1. Organizing the Archive

The simplest organization of the archive is a flat structure, that is, put every document in a single directory. At first glance this might appear to be unduly restrictive, but in fact it poses no significant hardship while greatly simplifying the construction of user interfaces. Even though all the documents are stored in a single directory, we can impose an arbitrary degree of structure on the archive by means of file links (i.e., the Unix *ln* command). The following example illustrates some possibilities.

Figure 1 shows a hypothetical organization of an online archive for one institution. The figure depicts three separate archives — Computer Science, Electrical Engineering, and Computer Engineering — hosted on three separate machines. The organization of the Computer Science archive has been elaborated in more detail to illustrate several organizational possibilities. All the reports are stored in the directory */pub/techreports/All* on the host machine *uvacs.cs.virginia.edu*. The hierarchical appearance has been achieved by the judicious use of file links. The reports have been broken into two series, those for the Department of Computer Science (*CS*) and those for the Institute for Parallel Computation (*IPC*). Within the *CS* directory, the reports have been further subdivided into research groups while the *IPC* directory has reports organized by year. Note that a research group, say *Adams*, may have links to reports issuing from both the *CS* and *IPC* series. Furthermore, links in these directories could have different names from the canonical file naming convention suggested for the combined directory.

We have shown a file called README in each directory containing documents.[1] This file serves as a table of contents and should contain a description of the documents located in the directory. Below we will describe a particular format that can be used.

[1] The files ‘‘Directory’’ and ‘‘Contents’’ shown in the other two archives serve the same role as the file ‘‘README’’ does for the Computer Science archive.

```
uvacs.cs.virginia.edu:

       /pub/techreports

              ALL
                     README
                     CS-nn-mm.ps
                       .
                       .
              CS       .
                     README
                     CS-nn-mm.ps
              IPC      .
                       .
                       .
                     1990
                            README
                            IPC-90-01.ps
                            IPC-90-02.ps
                                .
                                .
                     1991       .
                            README
                            IPC-91-01.ps
                                .
                                .
                     1992       .
                            README
                            IPC-92-01.ps
                                .
                                .
              Adams             .
                     README
                     CS-91-38.ps
                     IPC-92-04.ps
                        .
                        .
              Cyberia
              Mentat
              Networks
              SciDB
                     README
                     CS-90-21.ps
                     CS-90-22.ps
              Suit       .
                         .
faraday.ee.virginia.edu:

       /pub/techreps
              Directory
                  .
                  . (separate archive for Electrical Engineering)
                  .

babbage.ce.virginia.edu:

       /techreports
              Contents
                  .
                  . (separate archive for Computer Engineering)
                  .
```

Figure 1: A Possible Organization of Archives

Note that the unifying concept to which we referred earlier is simply that each directory containing documents should have a text file serving as the table of contents. The particular directory structure is not important. In this way, it is possible to remotely access archives and conveniently determine the content of the stored documents. As shown below, this convention is sufficiently useful to build a common interface to archives without unduly constraining the local sites' ability to structure their archives.

3.2. User Interfaces

To maximize the utility of an online archive, there should be several ways for interested parties to acquire the documents. The following sections describe four possibilities that are widely available and there are others.

3.2.1. ftp

The most common method used today to download reports over the network is by *anonymous ftp*. A user simply starts an ftp session to a known host to get reports. One often employs a two step procedure, first getting some text file describing the reports and then actually getting particular reports.

An ftp session is a cumbersome way to browse an archive, but effective to pick up a known file or set of files.

3.2.2. Electronic Mail

This is often the most straightforward way to acquire a report if its identifier is known. Although this could be handled manually, it is easily automated. For example, sending electronic mail to *techrep@uvacs.cs.virginia.edu* containing the message

```
send CS-90-21
```

will result in the UVa mail server sending a copy of technical report CS-90-21 to the requestor by return electronic mail.

The code for mail servers of this sort is widely available in source code archives on the Internet. One example is the *netlib* server described by Dongarra and Grosse [DONG85]. This server is used by Argonne National Laboratory for the distribution of mathematical software.

The installation of the mail server is made a little simpler if it can look for the documents in a single directory, but this is not essential. The server can manage details like splitting a large file into smaller pieces and mailing a multipart response.

3.2.3. Gopher Internet Protocol

Increasingly, many institutions are creating campus wide information systems (CWIS), many based on the popular gopher internet protocol developed at the University of Minnesota [TOME92]. The gopher protocol provides the appearance of a hierarchically organized global information space with a menu driven interface. The structure of Figure 1 can easily be accommodated by the gopher methodology.

3.2.4. UVa Technical Report Utility: techrep

The foregoing three interfaces all offer varying degrees of functionality for extracting a document from an electronic document archive, but no support for putting the documents online in the first place. This is a completely manual process. The technical report utility, techrep, was developed to provide both convenient access to online reports and to assist in the maintenance of the online archive. Note that fully automating every aspect of the process (e.g., assignment of report numbers, support for arbitrary structuring of the archive, etc.) was not an objective.

4. techrep

The following sections describe some of the design decisions and operational details of the techrep software.

4.1. Document Capture and Archive Maintenance

Document capture has been simplified so that authors can put documents online. The utility uses a graphical user interface to simplify entry of the necessary details (author name(s), title, report number, etc.) and the name of the file containing the PostScript document. It creates an optional title page, concatenates it to the original document and puts the PostScript file in the common archive. An entry is made into a citation database and the online table of contents is updated. This process ensures that the table of contents always agrees with the archive.

A limited update capability is provided and may be used by the user who originally posted the document or by a distinguished user, the technical report librarian. An update is treated as a delete/insert pair. This simplifies the maintenance of the archive.

At the present time the utility only maintains the common archive. The placement of links in any additional directories is very much like the assignment of subject descriptors to the document and is now a manual process. Similarly the assignment of report numbers is not done automatically. This could be automated but for now it is handled by the technical report librarian.

One of the more troublesome aspects of archive maintenance is the provision for a user-defined cover page. Our current version provides a generic cover page containing basic information. We expect to improve on this in later releases of the software.

4.2. Format of the Citation Database

The format of the citation database is shown in Figure 2. The entries are in *refer* (also more recently called *bib*) format. Each entry is composed of one or more authors (%A), a title (%T), the report number (%F), and report date (%D). In addition the user id of the posting user is retained with the entry (%Z). This is used to enforce the update policy. Entries are separated from one another by blank lines.

```
       .
       .
       .
%A J. C. French
%A A. K. Jones
%A J. L. Pfaltz
%T Scientific Database Management (Final Report)
%F CS-90-21
%Z techrep
%D August 1990

%A J. C. French
%A A. K. Jones
%A J. L. Pfaltz
%T Scientific Database Management (Panel Reports and Supporting Material)
%F CS-90-22
%Z techrep
%D August 1990
       .
       .
       .
```

Figure 2: Format of the Citation Database

```
                                    .
                                    .
                                    .
    [CS-90-21]
        J. C. French, A. K. Jones and J. L. Pfaltz, Scientific Database
        Management (Final Report), August 1990.

    [CS-90-22]
        J. C. French, A. K. Jones and J. L. Pfaltz, Scientific Database
        Management (Panel Reports and Supporting Material), August 1990.
                                    .
                                    .
                                    .
```

Figure 3: Fragment of a Table of Contents File

4.3. Format of the Table of Contents File

The format of the README file corresponding to the citation database fragment of Figure 2 is shown in Figure 3. The essential features to note are: (1) the entries are separated by blank lines; and (2) the report number is delimited by brackets so that it can be easily found.

The actual format of the entries is immaterial. The software simply displays the list for user selection. The two conditions above enable us to support selection by mouse click. Note that techrep could still be used if the conditions above are relaxed. That is because a report number can be specified directly on the selection screen.

4.4. Remote Site Access and Registration

Since we would like to regard the distributed collection of online archives as a shared resource, it is necessary to make some provision for remote access to nonlocal archives. techrep provides this facility by presenting a list of registered sites to the user for selection. After the user selects a remote site, techrep makes a connection and the user proceeds as if the archive were local. The only distinction is that write operations are disabled. That is, a user cannot insert, update, or delete documents at a remote archive.

We are maintaining a registry of remote sites on our FTP server. This registry is an ASCII file containing all the information necessary to make an ftp connection with the remote sites. techrep can be used to fetch this file on demand so that all users can have the most up-to-date list.

Figure 4 shows a fragment of the remote site file containing the entries corresponding to the archive sites depicted in Figure 1. Each entry of the file is a colon (:) delimited line of four fields. The fields contain the following information.

```
                                .
                                .
                                .
UVA Computer Engineering:babbage.ce.virginia.edu:/techreports:Contents
UVA Computer Science:uvacs.cs.virginia.edu:/pub/techreports/All:README
UVA Electrical Engineering:faraday.ee.virginia.edu:/pub/techreps:Directory
                                .
                                .
                                .
```

Figure 4: Directory of Archives

1. Name displayed by techrep for user selections.
2. Complete domain name of the host machine supporting the archive.
3. Complete path to the directory containing all the reports.
4. Name of the file containing the table of contents.

At this time the remote site registration is manual. Sites wishing to be added to the registry can simply send mail to *techrep @uvacs.cs.virginia.edu* containing these four items.

4.5. Managing Multiple Report Series

techrep can be used to support inter and intra departmental technical report series by several strategies. Intradepartmental series are best handled by a hierarchical structure on a single host, but multiple archives will also work. Separate departmental report series are best handled by multiple hosts each supporting a single departmental archive, but could use a single host and a hierarchical structure.

In both cases, it is possible to impose further structure by employing the directory structure as a pseudo-menu. The actual choice will be influenced more by local resource constraints than by software restrictions. techrep will support multiple registrations for a single machine or many machines at one site or any other combination that seems appropriate. For example, Figure 1 shows three separately hosted archives with the Computer Science archive organized as two formally distinct report series and several informal collections organized by research group. This organization is supported by the three entries in the directory of archives shown in Figure 4.

4.6. A More General Browsing Approach

In a future release, techrep will be enhanced to incorporate more extensive browsing facilities. The specific facilities will include:

1. the addition of string searching to the selection screens;

2. a file selection algorithm for traversing the directory structure; and

3. a file view selection mechanism to customize the reading of files.

When a long list is presented to a user for scanning, it can be very tedious to move about even when scrolling is available. The addition of string searching will help alleviate this situation. The latter two features are discussed below.

4.6.1. File Selection Algorithm

When techrep is enhance to include arbitrary browsing of online archives it will employ a particular convention to select files for viewing. The basic file selection algorithm is given in the pseudo-code of Figure 5.

The idea is based on the following structuring conventions. The remote site registration contains the root directory of the online archive and a distinguished file name denoting the table of contents file. techrep will read a directory and if it contains the distinguished file name, the contents of the file are presented to the user as a table of contents for selection. If the distinguished file name is not present, then techrep will display the directory entries as a menu.

If a directory entry is selected from this menu, techrep will descend the hierarchy otherwise it will assume that the selected file is to be viewed and use the viewing conventions described in the next section. Whenever the table of contents file is displayed for user selection the behavior of techrep will be the default technical report selection described earlier.

```
select_file(directory_name, distinguished_file_name)
    get directory listing of directory_name
    if distinguished_file_name occurs in directory listing
        display contents of distinguished_file_name as menu
        await user selection of entry
        construct actual file_name from selected entry
    else
        display directory listing as menu
        await user selection of file_name
        if  file_name is a directory
            file_name = select_file(file_name, distinguished_file_name)
    return (file_name)
```

Figure 5: Selecting a File for Viewing or Transfer

4.6.2. File Viewer Selection

When a file is selected for viewing, techrep will choose a viewer according to a set of user supplied rules. The default rules are specified in Table 2. The rules will be applied recursively as suffixes are examined from the right. For example, to view a file named *prog.src.tar.Z* the following statements would be invoked

```
uncompress prog.src.tar.Z; tar -tvf prog.src.tar | more
```

resulting in a directory listing of the compressed *tar* file.

5. Summary

This paper describes several ways to distribute technical reports online. They all rely on a single electronic repository and a small amount of local information, principally a text file that serves as a table of contents to the archive. The combined approach is effective and based on relatively few compromises. We are continuing to monitor the project to gauge its success and to determine if additional functionality is warranted.

The chief advantage to this comprehensive approach is that institutions can form and share large online technical libraries with an absolute minimum of overhead and compromise. All the user interfaces described earlier provide the essential functionality to distribute technical documents. The techrep software[2] provides some support specifically tailored to document distribution and facilitates remote access to the available archives.

Suffix Translation Table

Suffix	Viewer
ps	gs $f
txt	more
tar	tar -tvf $f | more
gif,tif,jpg	xv $f
Z	uncompress $f

Table 2

[2] The software is available by anonymous ftp from uvacs.cs.virginia.edu in the directory /pub/techreports. The file techrep.tar.Z should be downloaded using the binary mode of ftp.

While the technology is not perfect, it is sufficiently developed to be useful. Because the overhead and resource requirements are low, we can begin wholesale distribution of electronic documents now without waiting for a perfect solution. Nothing has been suggested here to preclude the introduction of new developments, say in the document interchange format, as they become available. We are proposing a simple open design that is amenable to technological evolution.

Many institutions already distribute at least some of their technical reports electronically in PostScript or other forms. We hope that many others can be induced to do so. The economic advantages are tangible; the potential for increased research productivity is manifest.

Acknowledgements

I would like to thank J. Terry who implemented the first version of techrep and C. Viles who made significant contributions to later versions. D. Wrege and T. Zhang have also contributed to the project. In addition, several individuals from other institutions were kind enough to test the software and offer constructive comments.

References:

[DONG85] J. J. Dongarra and E. Grosse, ‘‘Distribution of Mathematical Software Via Electronic Mail’’, draft document, Argonne National Laboratory, Nov. 1985.

[TOME92] C. Tomer, ‘‘Information Technology Standards for Libraries’’, *JASIS 43*, 8 (Sep. 1992), 566-570.