Info

✎

The Comprehensive TEX Archive Network (CTAN)

George D. Greenwade

Abstract

This paper outlines the concept, development, and use of the Comprehensive TEX Archive Network (CTAN) — a network-accessible archive for files related to the TEX family of document processing. The CTAN is a coordinated effort among consenting well-known archive sites which provides quick identification and retrieval files in a consistent manner from hosts on different continents, thereby reducing overall network load and increasing speed of retrieval. Moreover, it provides users with a parallel archive structure between hosts with holdings which are generally synchronized to within 30 hours of one another. This is achieved by routinely mirroring one another's holdings, as well as mirroring other archives to maintain an up-to-date collection of files.

Why a Comprehensive TEX Archive Network?

Since the inception of publicly-accessible network--based archives, TEX and its related packages, macros, and utilities have been available for retrieval by users via any number of techniques. The combination of the growth of the Internet in recent years, the growth of publicly-accessible network-based archive sites, and the growth in the number of files associated with TEX and it's affiliated packages and programs, created a rather overwhelming number of files for users with network connections to sort through. In terms of "overwhelming," the number of files available has been a significant boost for users; however, in these same terms, the number of different versions available, their precise location on a given archive host, the user interface available to access these files, and the ability to efficiently identify the various pieces required to make the various iterations of TEX and its relatives work properly has evolved to be a non-trivial task.

In recognition of these problems, the then--newly-created Technical Council of the TEX Users Group formed a Technical Working Group on TEX Archive Guidelines (officially WG-92-05; informally referred to as TWG-TAG) in the latter months of 1992, with the author as its Chair.[1] While a variety of issues related to archiving have been discussed (and broader guidelines, per se, will in all likelihood be forthcoming), the concept of creating a systematically coordinated family of network-based archive sites was tacitly agreed upon as a mechanism for verifying that the ideas under discussion were workable. Also, this approach was viewed as a mechanism for creating a more efficient design, both theoretically and practically, to meet the needs of the worldwide TEX community.

General Consensus Notes From TWG-TAG. Germane to the development of a set of archives in lieu of a "hard and fast" set of guidelines for propagation to other hosts were the following concerns:

- Existing archive hosts very likely have chosen and utilize a structure which is politically and practically acceptable for their site.
- While the concept of the Internet's File Transfer Protocol (ftp) for a user interface was on everyone's mind, access via alternate means (primarily electronic mail and hard media) had to be considered.
- Very few hosts on the network possess a comprehensive archive of TEX-related materials; thus, a design guideline which includes all dimensions of TEX may not be proper for a specialized archive.
- Even if a set of guidelines were developed, there is no way to ensure that every site which possesses some aspect of TEX in its archive would follow them since archives are a function

1: I would like to take this opportunity to formally recognize the members of this Working Group and publicly thank them for their efforts on any number of topics which we have dealt with. These individuals are, in alphabetical order: Nelson Beebe, Barbara Beeton, Karl Berry, Johannes L. Braams, David Carlisle, Michael J. Ferguson, Alan J. Hoenig, Don Hosek, David M. Jones, Pierre MacKay, David Osborne, Philip Taylor, Jon Radel, Sebastian Rahtz, Rainer Schoepf, Joachim Schrod, and Elizabeth Tachikawa.

of local resources, support, and needs more than "network" demands.

- If a workable demonstration of the guidelines existed, more sites may voluntarily elect to follow the ultimate guidelines.
- No single site on the network possessed a canonical listing, much less collection, of the latest relevant files available; therefore, any guidelines developed would be hypothetical more than working.
- It is necessary to make the structure as flexible as possible, while at the same time ensuring that files may be easily located by users via some consistent and logical design.
- As much as foreseeably possible, the structure should be extensible into evolving network retrieval and browsing technologies, such as Gopher, Mosaic, and other developing utilities.
- It is essential that the archives support users from the variety of platforms under which TeX is available; while a given platform may be used for the archives themselves, it should not impose problems on any other platform which a user may wish to ultimately use the files on.
- At least initially, the concern was to provide a reliable archiving directory hierarchy from within the project, as opposed to a production system hierarchy. [2]

Genesis of the Design

Prior to defining the CTAN directory structure, extensive discussion was undertaken by TWG-TAG regarding the optimal directory hierarchy. Compared for use were the hierarchical design used on most ftp-accessible hosts now possessing a TeX archive and the flat design used by a number of mail-oriented hosts. For brevity's sake, allow it to be said that the hierarchical design was chosen. Once this decision was made, comparisons of existing archives were undertaken, primarily focusing on the holdings of three major sites — Stuttgart University's `ftp.uni-stuttgart.de`, Aston University's `TeX.ac.uk`, and Claremont College's `ymir.claremont.edu`. The end result is a hybrid of the three directory structures, focusing on top-level directories which are somewhat mnemonically-based directory names at this level which, at least in the author's view, is pleasing.

The structure is adequately diverse so that dvi-related files (such as device drivers, dvi to output programs, and dvi viewers) are distinctly different from macros, and that macros are adequately categorized into the appropriate flavor of TeX for which they are intended. The top-level directory hierarchy is presented in Appendix A.

Conceptually, every file in the archive fits into one branch of this directory hierarchy (albeit the directory mutually agreed to by the maintainers of the hosts involved in this project). Where a file conceptually fits into more than one directory, efforts are made to ensure that the file properly appears where it should.

Coincidences of Consequence. A critical dimension of this project was its timing. Two hosts were new machines; therefore, they were easily designed into whatever structure was agreed to. The main coincidences which aided the success of this project were:

- Sebastian Rahtz, who was just beginning to put together the Aston "daughter" archive at `ftp.tex.ac.uk`, began utilizing the directory design by following the very rough outline of the preliminary structure. Making the structure operational was a significant factor in more than a few subsequent decisions as it illustrated, in a close-to-production environment, the strengths and weaknesses of the then-tentative hierarchy.
- Sam Houston State University was just installing its first Unix-based host and learning its idiosyncrasies; however, its use as an archive host was established soon after its installation. This was a preferable choice for an archive host over SHSU's more established ftp host, `Niord.SHSU.edu`, as it ran a comparable software to the other two hosts, and was configurable for use as a mirror.
- David M. Jones had just released the first version of his index of TeX and LaTeX macros, providing the TWG-TAG with a relatively-comprehensive listing of files, their

2: Please note that the concerns above are those perceived by the author as the majority opinion and are not necessarily those of the TWG-TAG, nor are they necessarily those of any one given member of the TWG-TAG.

authoritative location, current version, and other critical information.

- Joachim Schrod's mirroring software for Unix hosts (which the CTAN hosts utilize) was being upgraded. Joachim was able to quickly address a few specific problems resulting from its massive use.
- Nelson Beebe was developing a few new extensions to the ftp server software which was adopted for use. Significant contributions were his definitions which allow users to retrieve recursive ZIP and ZOO archives of directory holdings.
- A bug was located in the adopted ftp server software. This led to the new release of the ftp server software, resulting in a much nicer interface, easier management, and more generic installation as compared to prior versions. In essence, while each host is unique in architecture, the configurations are virtually parallel, as are most features available to users.
- Karl Berry further extended his somewhat standardized Unix TEX directories and installation. While this has yet to play a significant role in design, it provides significant future extensions to the services of the CTAN.
- The University of Minnesota upgraded the facilities and functions of its Gopher software. This interface now serves as an alternative access method to the CTAN at Aston and SHSU.
- Significant enhancements to the ZIP utilities by the Info-ZIP team were released. This is a major item of concern as it allows for a platform-independent mechanism for the archival of files held within the CTAN collection. In essence, when combined with Nelson Beebe's extensions, this feature allows users on virtually any platform to retrieve recursive directory holdings, then convert them to the operating system specific needs which they confront.

By no means is this a complete listing of all coincidences of consequence to the project — suffice it to say that without a number of apparently disjoint and unrelated projects coming to fruition at approximately the same time, this project would likely still be in its planning stages.

The CTAN Hosts

By now, you're very likely asking where these hosts are on the network. The present CTAN hosts and their root CTAN directories are:

```
ftp.uni-stuttgart.de  /tex-archive/
ftp.tex.ac.uk         /tex-archive/
ftp.shsu.edu          /tex-archive/
```

Based at Stuttgart University (Germany),[3] Aston University (England),[4] and Sam Houston State University (United States),[5] respectively.

While the designs of the consulted archives were largely congruent, there were some modifications when compared to all existing archives. Rainier Schoepf, TEX archive manager at Stuttgart University's ftp.uni-stuttgart.de, was fully agreeable to modifying the directory structure there based on the recommendations which were developed. While it was clear that the changes would impose short run problems for users of his collection and potential problems in the management of his site's mirroring of packages, he was in agreement that the long term benefits to the TEX community of a well-known and specified structure outweighed any short term impact. Also, as the other two hosts involved were new to the network, being able to plan for the inclusion of an existing large and active archive from the start was a real benefit, especially since it had a mirroring structure already in place and the site was mirrorable by other sites.

3: In addition to the administration of the institution, thanks are extended to DANTE, the German speaking TEX Users Group, for their archival support and leadership role in maintaining this archive.

4: In addition to the administration of the institution, thanks are extended to the UK TEX Users Group for their archival support and leadership role in the development of their original archive, as well as in the creation and development of this new archive host.

5: In addition to the administration of the institution, thanks are extended to SHSU's Computer Services Division, which has been more than forthcoming in support and access as this archive has evolved. Additionally (and especially), thanks are extended to SHSU's College of Business Administration for its continued interest in and support of this project.

What's Available

As noted in the name selected for these hosts, the *Comprehensive* TEX Archive Network, the collections available are about as comprehensive and timely as humanly possible. Do these hosts have "everything"? The truthful answer is that they probably do not and probably never will — but they are and will be about as comprehensive a collection as can be arranged. As sites possessing relevant files and packages are identified, one of the CTAN hosts will include them in its mirroring passes; from there, the files will then propagate to the other CTAN hosts. Presently, there are in excess of 1 gigabyte of files available in the /tex-archive/ directory tree on ftp.shsu.edu. It would be expected that the collections in the TEX archive area on the other two hosts would be virtually identical.

How It Is Achieved. The synchronization between hosts is handled within a "mirroring" program whereby files on one host are propagated to the other hosts of the CTAN. The variant of mirroring used by the CTAN hosts is a modification of a family of perl scripts written by Joachim Schrod.[6]

Conceptually, the process entails three major points. First, hosts where files are authoritatively available for retrieval are mirrored inbound by a selected CTAN host. This CTAN host locates these files in a pre-specified area of the collection. Second, users may contribute files to a given CTAN host by uploading their files into an "incoming" area on any given CTAN host. From there, the local CTAN administrator moves these contributions into the CTAN hierarchy. Finally, the CTAN hosts mirror one another on a routine basis to collect all new files which have entered the collection at the mirrored host. In all, about one gigabit of files are referenced daily by the mirroring processes in order to maintain accuracy, timeliness, and correctness of holdings.

Existing User Interfaces

As noted throughout this paper, the Internet's File Transfer Protocol (ftp) is supported at the moment for a user interface. There are three other interfaces which should be recognized — Gopher, mail, and hard copy — as these are important to users which prefer to utilize them.

ftp. The ftp interface utilized by each CTAN host is the "wuarchive `ftpd`" program.[7] In addition to the normal ftp commands a user would expect, these hosts support a variety of additional useful functions, such as locating a file, creating `ZIP`, `ZOO`, and Unix `tar` archives on-the-fly, automatic compression of files on-the-fly, and a few other features. As these features may be modified, the best policy is to consult any files named `README` in the default directory at login.

Gopher. Two of the CTAN hosts, ftp.tex.ac.uk and ftp.shsu.edu are configured for Gopher access. The Internet Gopher, developed by the University of Minnesota, provides users with a menu-driven interface for transactions. Users may view and retrieve files in a fashion somewhat more friendly than ftp. Also, the Gopher menus presented do not necessarily represent files actually available at a site; instead, the client/server relationship utilized allows for servers to point to information stored elsewhere, which users are automatically connected to by simply selecting a menu entry.[8]

Pointing the Gopher client directly to the Gopher server on ftp.tex.ac.uk is the preferred method for access to the Aston CTAN archive; Gophering to SHSU's "front door" Gopher server on Niord.SHSU.edu is the preferred method to access SHSU's CTAN archive. In addition to the archives themselves, a variety of TEX-related files and services are available on each host. Sample menus from each are provided in Appendix B.

Mail. At the moment, no agreed-to mail interface has been installed on all hosts. At present, plans exist to eventually utilize the `ftpmail` program as

6: The program itself is available in the archive-tools/mirror/ directory of the CTAN hosts.

7: This program is available in the directory archive-tools/ftpd/mirror/ on the CTAN hosts. It is the classic ftp server/daemon utilized by Washington University in St. Louis, Missouri (USA) for their massive collection of files.

8: Gopher clients are available for a wide variety of platforms. Sources for most platforms are available in the directory archive-tools/gopher/ on the CTAN hosts.

the mail interface supporting all CTAN hosts.[9] The mail interface to the CTAN hosts will be documented in a subsequent issue of *TEX and TUG News* once the interface is installed and stable.

Hard Copy. Sebastian Rahtz has already made arrangements and compiled a hard copy of the CTAN archives for CD-ROM distribution. The distribution will be made available from Prime Time Freeware. The first version of this file set is already available and it is to be routinely updated to reflect changes made to the archives. The second release of this CD-ROM distribution is expected to be available in early 1994.

Aspects for Authors

While the process of mirroring and uploading provides a rough and ready mechanism for achieving the reliability of the CTAN, authors should be aware of a few dimensions. At the moment, no specific guidelines exist on these topics, so please view the following as personal views which will, in time, be discussed.

File Headers and Checksums. Nelson Beebe has developed a package written as an Emacs Lisp file which creates "standardized" file headers.[10] These headers, while clearly consuming space within the archive, provide valuable information which users and archive maintainers may refer to quickly. Moreover, these file headers provide valuable indexing information which may be blended into existing indices, such as David Jones' heroic effort.

Optionally included in these headers is a "checksum" option, which is of significant benefit whenever files are transferred electronically to verify correctness. The preferred is Robert Solovay's CWEB-based program which, with the assistance of Nelson Beebe, has been ported to a number of operating systems.[11]

9: Sources for the ftpmail service are available in the directory archive-tools/ftpmail/ on the CTAN hosts.

10: This package is in the archive-tools/filehdr/ directory of the CTAN hosts. [*Editorial note*: see also the text by W. Bzyl, T. Przechlewski]

11: This package is available in the archive-tools/checksum/ directory of the CTAN hosts.

Location of Files. It would behoove authors who maintain authoritative file set(s) on network hosts which are ftp-accessible to contact me so that arrangements may be made to mirror these files into the CTAN collection. Without a doubt, the ability to mirror files into the CTAN hierarchy is the least painful and most efficient method available as human intervention arises mainly in the review of the logs of the mirroring session. If an authoritative host is identified, its holdings will automatically be included when the site is referenced. In this way, additional copying to additional hosts for propagation will not be needed. Moreover, you will ensure that the latest versions are available for public consumption.

Directory Suggestions. Although it is well-specified, no official designation of the CTAN directory structure now exists. However, it is advisable for maintainers of authoritative files and packages to consider the utilization of the CTAN hierarchy, if at all possible. In this manner, the author will be able to envision how the archive is laid out for the end user. Also, if this structure is used, especially on Unix-based hosts, proper linkage to related files may be used. At the moment, one of the most wasteful uses of the CTAN directory contents are the multiple copies of files which authors elect to include. For example, as of 1 July 1993, approximately 27 copies of the LaTeX macros exist in the archive — only one of which (in macros/latex/distribs/) is authoritative. With linking, the mirror simply calls in the link and the latest versions of related files (from their well-specified directory) will be delivered to users.

Getting a File into the CTAN Without Mirroring

If mirroring is not practical (either because of inability to mirror a host, lack of a public ftp area in which to place a file, lack of ftp altogether, the contribution is a small number of files, or any other reason), authors can still easily get files included into the CTAN with electronic mail or via anonymous ftp.

Electronic Mail. To submit a contribution via electronic mail, use the address:

```
CTAN-Mgr@SHSU.edu
```

including the file in whatever manner is feasible (encoding of executables, splitting of files, indicating within a single post with multiple files where to "cut," etc.). In these cases, checksums are ideal as it provides verification that the file received for archiving purposes is indeed the same file which the author intended for inclusion. A brief note describing the contribution is very appreciated. As soon as the contribution has been processed into the CTAN, the submitting author will be notified via return electronic mail.

Anonymous ftp. To submit a contribution via ftp, connect to one of the CTAN hosts with an ftp client. When prompted for a username, type "anonymous" (all lowercase; without quotes) and use your complete electronic mail address as your password. The Aston and SHSU CTAN hosts support the /incoming/ directory for contributions (uploads) of files; the Stuttgart CTAN host supports the /incoming/tex/ directory for contributions of TEX-related files. Once connected, a typical upload session would look like:

```
cd <appropriate incoming area>  !! 1.
mkdir <your directory name>     !! 2.
cd <your directory name>        !! 3.
[binary | ascii]                !! 4.
[m]put your file(s)             !! 5.
```

ignoring everything beginning with and to the right of !! above. Each step represents:

1. cd — change directory to proper incoming area (see above).
2. mkdir — make subdirectory (optional), which is especially nice to retain together multiple files contributions which are intended to stay together; you may use any your directory name you please.
3. cd — change to the directory you just created so you may use it.
4. binary or ascii — just to be safe, it's always best to verify the intended transfer mode (binary or ascii) prior to transfer.
5. [m]put — use standard ftp put or mput command to place your files in the incoming area.

Future Directions

While the CTAN hierarchy and holdings between its hosts are relatively stable at present, this does not imply that the project is completed. Topics for potential extension are:

1. The creation of on-demand "kits". At present, the systems/ directory hierarchy includes sets of packages intended for system-specific installation. One very serious potential problem with these file sets is that they are not automatically updated when new files are introduced to the archive. For example, some of the system-specific installations include prior versions of macro sets (generally, the LATEX macros in these file sets are outdated), drivers (such as the rapidly changing dvips), or other dimensions of the included software. This implies that users retrieving these file sets, while able to install TEX, have to immediately return to the archives to upgrade their systems.

An alternative to the present process would be for servers to have the capability of creating on-demand up-to-the-minute "kits" for platform-specific installations. However, specific design features for this functionality must be developed.

2. The use of a non-ftp mirroring mechanism. At present, the mirroring mechanism used by the CTAN hosts suffice for its use. However, ftp is not an efficient mechanism for moving the quantity of files involved in this project. A much preferable solution would be the use of some alternate protocol which allows for verification and automatic updating of all hosts in the CTAN.

3. Addition of other CTAN hosts. At present, the workload of the three hosts involved in this project is non-trivial. Essentially, the eastern side of the Atlantic is served by two hosts, while North America is served by a single host. The present configuration of hosts lends itself to easy extension to another North American host to service the Americas, as well as a Pacific-based host to serve Asia, the Pacific Rim, and the Southern Hemisphere. Simply creating a mirror of one of the existing CTAN hosts in these regions, with no additional network responsibilities, would be more than acceptable.

4. Automation of archival information. While the CTAN possesses the most comprehensive collection of TEX-related materials, one very dissatisfying aspect remains. This problem area lies in the collation of information which is quickly retrievable (even by the archivists themselves) which points to related files, required files, version requirements, authoritative location of files, listing of most recently added files within an area, etc. The mirroring process is marvelous at collecting files; however, without proper and somewhat standardized documentation of each component of the archive, tracing problems may prove to be painful.

At present, David Jones' index of macros — a wholly volunteer effort — exists and serves this function admirably within the context of macros. Additionally, the "Frequently Asked Questions" (FAQ) files from the USENET newsgroup `comp.text.tex` provide more information on other selected aspects which are included in the collection. However, the fact that these are volunteer efforts, as well as the fact these only cover a microcosm at a point in time of the complete and rapidly changing archive, is troubling.

Appendix A

The Top-Level Directory Hierarchy. Once into the CTAN host's root area, the following directory hierarchy is presented:

```
biblio/
digests/
dviware/
fonts/
graphics/
help/
indexing/
info/
language/
local/
macros/
misc/
support/
systems/
tools/
web/
```

In Brief, the contents of these directories include:

`biblio/` contains bibliography-related files, such as BIBTEX.

`digests/` contains back issues of TEX-related periodicals.

`dviware/` contains the various dvi-to-whatever filters and drivers.

`fonts/` contains a collection of fonts, both sources and pre-compiled.

`graphics/` contains utilities and macros related to graphics.

`help/` contains files which provide an overview to the archive and the TEX system.

`indexing/` contains utilities and related files for indexing documents.

`info/` contains files and tutorials which document various aspect of TEX.

`language/` contains non-English related implementations of TEX.

`local/` contains local site-specific files — not of general interest.

`macros/` contains macros for TEX and its derivatives in unique subdirectories.

`misc/` contains files and programs which cannot otherwise be catalogued.

`support/` contains files and programs which can be used in support of TEX.

`systems/` contains complete system setups, organized by operating system.

`tools/` contains the various archiving tools which users may find useful.

`web/` contains WEB-related files and utilities.

Appendix B

Gopher Menus from Aston and SHSU. The menus for the CTAN archives via Gopher as of the time of this paper are provided below. For `ftp.tex.ac.uk` the Gopher menu structure appears as:

```
The menu's for the CTAN archives
via Gopher as of June 1, 1993.
Root gopher server: ftp.tex.ac.uk

1.  UK TeX Archive/
2.  Aston University/
3.  Minnesota Gopher root/
4.  Archaeology/
5.  Archie <?>
6.  Font Samples/
7.  General WAIS databases/
8.  Looking things at AMS <TEL>
```

```
    9.  Veronica (search menu items
          in most of GopherSpace)/
    10. World Wide Web
          (Hypertext fra CERN) <TEL>
```

and selecting item 1. from this menu yields:

```
UK TeX Archive

1.  Welcome.
2.  Archive directory/
3.  Indexed FTP (Directories names only)
     of UK TeX Archive.
4.  Indexed FTP of UK TeX Archive.
5.  Indexed access to archive hierarchy <?>
6.  Other Archives/
7.  TeX Font Samples/
8.  UK-TeX.
9.  WAIS database: TeX --- list of FTP
     sites with TeX material <?>
10. WAIS database: TeX Frequently
     Asked Questions (UK specific) <?>
11. WAIS database: TeX index of styles
     and macros (by David Jones) <?>
12. WAIS database: The TeX Book <?>
13. WAIS database:
     back issues of TeXhax, 1986 -- <?>
14. WAIS database:
     back issues of UKTeX newsletter <?>
```

Selecting item 2. from this menu yields the directory hierarchy described in Appendix A.

For Niord.SHSU.edu the Gopher menu structure appears as:

```
Root gopher server: Niord.SHSU.edu

1.  About the Sam Houston
     State University Gopher.
2.  Customizing the Print command at SHSU.
3.  All the Gopher Servers in the World/
4.  Chronicle of Higher Education
     "EVENTS in ACADEME"/
5.  Economics (SHSU Network Access
     Initiative Project)/
6.  Internet Information/
7.  LaTeX3 Public Document Library/
8.  Libraries, Periodicals,
     References, etc./
9.  Minnesota Gopher (Mama Gopher;
     get *your own* clients here!)/
10. SAMINFO -- Sam Houston State University
     Information System <TEL>
11. TeX-related Materials/
12. Thesaurus and Dictionaries/
13. USENET News (from Oakland University)/
14. VMS Gopher-related file library/
15. Veronica (search menu items
     in most of GopherSpace) /
16. Weather Forecasts
     (National Weather Service; US)/
17. Weather Underground
     (University of Michigan) <TEL>
18. anonymous ftp archives on ftp.shsu.edu/
```

and selecting item 11. from this menu yields:

```
TeX-related Materials

1.  Comprehensive TeX Archive Network
     (CTAN) at SHSU/
2.  UK TeX Archive/
3.  Archives of INFO-TeX/ctt-Digest
     (comp.text.tex)/
4.  EconBib (LaTeX/BibTeX styles for
     economics)/
5.  .*.*. FAQs, REFERENCE and
          PRIMERS .*.*.
6.  FAQ for comp.text.tex
     (text and WAIS indexed)/
7.  FAQ Supplement for comp.text.tex
     (text and WAIS indexed)/
8.  TeX-Index (text and WAIS indexed)/
9.  FAQ for comp.fonts
     (text and WAIS indexed)/
10. The Canonical list of MetaFont fonts.
11. A Gentle Introduction to TeX
12. Components of TeX/
13. Essential LaTeX/
14. MetaFont for Beginners.
15. NFSS in the Context of LaTeX
16. .*.*. RELATED FILES .*.*.
17. LaTeX3 Public Document Library/
18. Literate Programming Library/
```

Selecting item 1. from this menu yields the directory hierarchy described in Appendix A.