A World Wide Web Interface to CTAN

Norman Walsh
O'Reilly and Associates, 90 Sherman Street, Cambridge, MA 02140, U.S.A.
norm@ora.com

Abstract

There are a lot of different software packages, style files, fonts, etc., in the CTAN archives. Finding the things you need in a timely fashion can be difficult, as I found out while writing *Making TEX Work*. The ability to combine descriptions of packages with the directory listings from CTAN could help alleviate some of the difficulty. The HyperText Markup Language (HTML) is the document structuring language of the World Wide Web and it provides one possible means of combining different views of the archive into a single vision. The CTAN-Web project is my attempt to provide this vision.

Introduction

A functioning TEX system is really a large collection of programs that interact in subtle ways. Processing even a relatively simple document like this one requires several programs (TEX, a previewer, and a printer driver at the very least), most of which read input files or can be configured in other ways. It was this complexity that lead me to start writing *Making TEX Work* (Walsh 1994), a book I hoped would unravel many of these intricacies (end of plug ;-).

In the process of writing *Making TEX Work*, I looked at a lot of the software packages, style files, fonts, etc., in the CTAN archives. It really made me appreciate how much stuff the TEX user community has made freely available. By my estimates there are more than 31,000 files in more than 2,300 directories in `/tex-archive` on `ftp.shsu.edu`.

My first challenge was to find the things that I wanted to write about. This was a long process that involved coordinating (at least mentally) the lists of files in the upper-level CTAN directories, entries from David Jones' `TeX-index`, descriptions maintained by the CTAN archivists, my own intuitions about what was available, and the tidbits that I had collected over the years from `Info-TeX` postings. It was occasionally tedious, but it was never really difficult (at least technically).

When the book was beginning to fall into place and I was starting to try to track down all the loose ends, I came to a realization: in the early days, finding things had been an end as well as a means. Now, with pressure mounting on an almost daily basis to finish, I discovered just how hard it was to find things on CTAN. This is not a criticism of the CTAN archivists in *any* way. Without their foresight and diligent efforts, the task could easily become impossible. It's just a fact: there's a *lot* of stuff out there.

One tool became invaluable in my daily efforts: ange-ftp for GNU emacs. GNU emacs, if you aren't familiar with it, is an extremely flexible and powerful editor (it's most common on UNIX workstations, but versions exist for MS-DOS, Windows, OS/2, VMS, and a few other platforms). One of the editing modes of emacs, called *dired*, allows you to "edit" directories (a directory listing appears in a window on the screen). In dired mode, the editing keys let you rename, copy, delete, view, and edit files, among other things. Ange-ftp is an extension for emacs that lets you edit *remote* file systems via ftp in dired-mode. This lets me load the `/tex-archive/macros` directory from `ftp.shsu.edu` into an emacs buffer and view files simply by pointing to them and pressing "v". Ange-ftp handles all of the transactions with the ftp client in the background. Ange-ftp made gathering information from README files *much* easier.

Inspiration

What I really wanted wasn't an easier way to browse directories, no matter how grateful I was to have that, but a way of combining the `TeX-index` and other descriptions with a directory listing in some coherent way. A typical interaction with CTAN, in my experience, goes something like this: I need a *widget*, that's under the *something* directory. Oh! There are several things like that. This one looks interesting. Nope that's not it. How about this one. Yeah, that's better. Still, is this other one better? Nope. Ok, I'll try the second one.

I find this sort of interaction tedious via ftp.

As it happens, I was also beginning to explore the World Wide Web (WWW) at the same time, motivated, in part, by experimentation with LATEX2HTML and other tools that translate TEX documents into HTML for online documentation projects. Might this be the answer, I wondered ...? After several days of hacking, the first incarnation of CTAN-Web was born; the CTAN-Web home page is shown in Color Example 16.

What is the World Wide Web?

The WWW is a vast collection of network-accessible information. In an effort to make this information manageable, protocols have been developed for cross-referencing the Web and software written to browse documents in the Web. One of the most popular browsers is Mosaic, a browser from the NCSA.[1] WWW documents use hypertext to make traversing between documents transparent, allowing the user to follow a stream of ideas without regard to where the embodiment of the ideas exists in the Web.

Hypertext links allow you to build dynamic relationships between documents. For example, selecting a marked word or phrase in the current document can display more information about the topic, or a list of related topics.

Naturally, WWW documents can contain hypertext links to other WWW documents, but they can also contain links to documents available through other servers. For example, *Gopher* servers and anonymous ftp servers. Documents in the WWW are addressed by a "universal resource locator" (URL) that identifies the site from which they are available and the protocol that should be used to retrieve them. The general format of a URL is *protocol://site/pathname*. For example, the URL for the LaTeX help file that I maintain is:

```
http://jasper.ora.com/texhelp/LaTeX.html
```

In other words, it is available via the *http* protocol at `jasper.ora.com` in the file `/texhelp/LaTeX.html`.

Once retrieved, it is up to the browser to determine how they should be displayed. In addition to displaying HTML documents directly, many browsers can automatically spawn external viewers to view PostScript documents and image files in a variety of formats.

What is HTML?

WWW documents are plain ASCII files coded in HTML (Flynn 1994). HTML provides a convenient way to describe documents in terms of their structure (headings, paragraphs, lists, etc.). HTML is really a particular instance of an SGML document. SGML is the Standard Generalized Markup Language and it is defined by the ISO 8879 specification.

The relationship between SGML and HTML can be a little confusing. SGML provides a general mechanism for creating structured documents. HTML documents are SGML documents that conform to a single, fixed structure. (The HTML specification is available at `http://info.cern.ch/hypertext/WWW/MarkUp/MarkUp.html`.)

[1] The figures in this paper are of the X11 version of Mosiac.

A detailed exploration of structured documentation principles is beyond the scope of this article, however, a few words may help clarify the picture; users familiar with LaTeX are already familiar with structured documentation.

The key notion is that structures (characters, words, phrases, sentences, paragraphs, lists, chapters, etc.) in a document should be identified by *meaning* rather than appearance. For example, here is a sentence that you might find in an installation guide (this sentence is coded in TeX):

```
Use the {\bf cd} command to change to the
{\it /usr/tmp/install} directory.
```

The same sentence might be coded in a structured way like this:

```
Use the <command>cd</command> command to
change to the <directory>/usr/tmp/install
</directory> directory.
```

The advantage of the structured document is that it is possible to answer questions about the *content* of the document. For example, you might check to see if all of the commands that are mentioned in the installation guide are explained in an appendix. Since commands are explicitly identified, it is easy to make a list of all of them. In the unstructured case, it would be very difficult to identify all the commands accurately.

You can achieve structured documentation in TeX with macros, but you are never forbidden from using lower-level commands. The advantage of using a formal structured documentation system, like SGML, is that the document can be validated. You can be sure that the document obeys precisely the structure that you intended. The disadvantage of a formal system is that it must be translated into another form (or processed by a specialized application) before it can be printed, but that is becoming easier. In the case of HTML, many browsers already exist.

Since an HTML document is described in terms of its structure and not its appearance, most HTML documents can be effectively displayed by browsers in non-graphical environments. There is a browser for Emacs called W3 and a browser called Lynx for plain text presentation, for example.

What is CTAN-Web?

CTAN-Web is a collection of WWW documents that combines descriptions of many packages available from CTAN with pointers to each of the files in the archive. At present, the descriptions come from an early draft of my book, David Jones' `TeX-index`, and the `00Description` files in the archives. Over time, additional descriptions will be added. Figure 3 shows the top of the `/tex-archive/macros` directory.

The CTAN-Web also has the following features:

- Links are made directly to other online references in the Web. For example, the online help files provided in the `info/htmlhelp` directory are also available as WWW documents on the net. This fact is exploited in the descriptions of these files by creating a hypertext link directly to the online help.

  In addition, font samples can be displayed for several METAFONT fonts (viewing font samples requires a browser that understands GIF files).[2]
- The CTAN-Web documents are indexed. Users can perform online queries for material based upon any word that appears as a filename or in the online description of any file. Simple conditional searches can also be performed (for example, "x or y" or "x and y").

  A query for "verbatim and plain" finds 5 files and 9 directories.[3]
- Each instance of a file that appears in more than one place in the archive is identified. For example, any reference to the file `verbatim.sty` identifies all 7 instances of it in the archive.
- Want to know which files were modified within the last 12 days? Or between 1 Jan and 31 Jan of 1993? Information about the age of each file is maintained in a separate database, accessible via a script run by the CTAN-Web server. This allows you to perform online queries of the archive by age.
- A "permuted index" is constructed each time the Web is built. This allows you to quickly locate files by name.
- A list of files added or modified in the last 7 or 30 days is also constructed each time the Web is built.
- A tree (hierarchical) view of the archive is also available. The tree view provides a fast means of "walking" down into the lower levels of the archive.

Reaching CTAN-Web

You can reach the CTAN-Web pages by using the URL: `http://jasper.ora.com/ctan.html`

Behind the Scenes

For those who are curious, this section provides a brief description of how the CTAN-Web is constructed. The Web is now rebuilt on a daily basis using the most recent information from the `ftp.shsu.edu` server.

[2] Samples for all the METAFONT fonts will be generated shortly.

[3] In the Web built on 20 May 1994.

Handling the descriptions. In order to quickly locate descriptions for the various packages, I maintain the collection of descriptions in a directory structure that parallels the CTAN archives. Each description file is written in a mixture of TEX and HTML (a mixture is used so that it may one day be possible to produce a printed version of the Web). For example, the current description of `latex-help-html.zip` is shown in Figure 1.

Retrieving files from the archives. One of the first problems that had to be solved was how files would be retrieved from the archives. While it's easy to create a link to a file at an ftp site, in the case of CTAN-Web that isn't sufficient because CTAN exists at several sites. The link really needs to be made to the *closest* ftp site.

Although I suppose it is possible to identify the closest ftp site from the user's host id, that seemed impractical. The following compromise was selected instead: rather than linking files directly to an ftp site, they are linked to a script. The document server (`httpd`) provides a facility for making links that cause a program to be executed; the output produced by this program is then displayed as a WWW document. By passing the name of the file requested by the user as an argument to the script, it was possible to write a retrieval script that dynamically constructs a "retrieval document." The retrieval document contains links to the requested file at each of the CTAN hosts. It is then possible for the user to select the closest host. An example of the retrieval document created for `README.archive-features` is shown in Figure 2.

Selecting a link within the retrieval document causes the browser to actually retrieve the file via anonymous ftp from the selected site.

Documents in the Web. There are three kinds of documents in the CTAN-Web and within each document there are several kinds of links.

Directory documents. There is one directory document in the Web for each directory in the archive. Each directory document lists all of the files in the directory it represents along with their associated descriptions.

Directory names in each document are linked to the corresponding directory documents. File names are linked to filename documents (described below) or to the retrieval script, depending on whether the file occurs multiple times in the archive.

The directory document for the `tex-archive/macros` directory is shown in Figure 3.

Tree documents. There is one tree document in the Web for each directory in the archive that contains subdirectories. The tree document displays three levels of hierarchy starting at the directory it represents.

```
<!-- tex-archive/info/htmlhelp/latex-help-html.zip -->
An HTML version of the LaTeX help file created by George Greenwade.
This is the version provided online at <tt>jasper.ora.com</tt>.
It is also available in VMS format (formatted ASCII),
TeXinfo format, HTML format, and as a Microsoft Windows help file.
<!--ONLINE-->
<P>
The LaTeX help file is also
<A HREF="http://jasper.ora.com/texhelp/LaTeX.html">available online</a>.
<!--/ONLINE-->
```

Figure 1: The description of `latex-help-html.zip` in the Web sources.

Directory names in each document are linked to the corresponding tree document. If a directory in the tree does not have subdirectories, it is linked to its directory document instead.

The tree document for the `tex-archive/macros` directory is shown in Figure 4.

Filename documents. There is one filename document for each file that occurs in more than one place in the hierarchy. The filename document lists all of the instances of the filename.

Each instance of the filename in the document is a link to the directory document where that file resides in the archive.

The filename document for the `verbatim.sty` file is shown in Figure 5.

Building the Web document. Early versions of the Web document were constructed from the `FILES.byname` list from the server `ftp.shsu.edu`. Several *Perl* scripts manipulated the listing and constructed the Web document.

After a few weeks, it became clear that the `FILES.byname` listing was insufficient for constructing the Web document because the list contains no indication of symbolic links, for example. It is also poorly organized for my purposes (the necessity of making multiple passes was causing memory problems). George Greenwade kindly agreed to run a script on the archive that extracts more information and stores it in a form that can be translated into the CTAN-Web document in a single pass. (This information is provided in `/pub/fornorm.gz`, if you're interested).

Room for Improvement

I plan to improve CTAN-Web in a number of ways.

- One of the most important improvements is getting the Web off the node `jasper.ora.com` and distributed amongst all the CTAN hosts. The Internet connection from `jasper` to the outside world is actually quite slow and many users find that the performance is poor.
- Assign fixed URLs to each CTAN directory. At present, most of the URLs are assigned more-or-less sequentially when the Web is constructed. This means that the URL for the `tex-archive/macros/latex2e/contrib` directory, for example, changes over time. This prevents people from saving the URLs of frequently visited regions of CTAN-Web. The top level directories already have fixed names.
- Clean up the descriptions. Using an automatic tool to extract the descriptions from several sources in the archives was a fast way to get a large number of descriptions, but the process was not error free. A small, but significant, number of files in the Web have incorrect descriptions.
- Potentially add a report-generating function that can return an annotated list of the files that match a particular query.

Conclusion

I'm quite pleased with the CTAN-Web. There is room for improvement, but I already find it a faster and more flexible way to search the archives. If only I'd had it before I wrote the book. Ah well, there's always the next edition...

References

Flynn, Peter. "How to Write HTML Files." Hypertext electronic document available using the URL `http://www.ucc.ie/info/net/htmldoc.html`. 1994.

Walsh, Norman. *Making TEX Work.* O'Reilly & Associates, 1994.

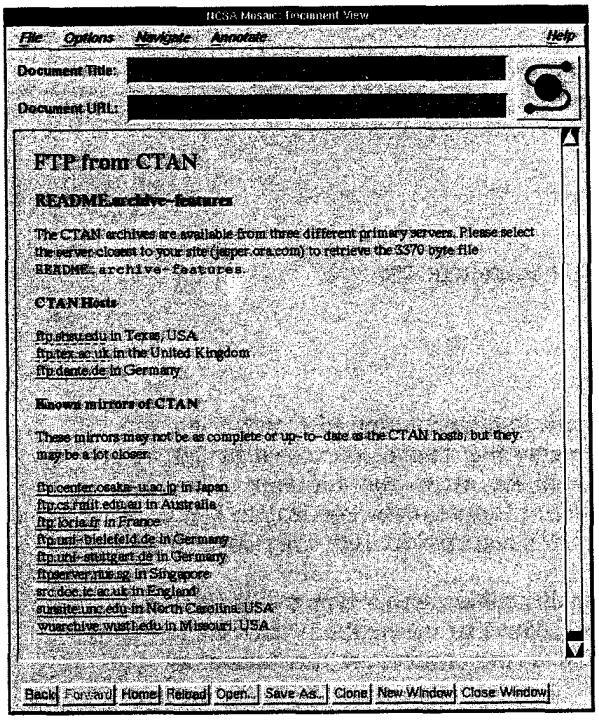

Figure 2: Example of a retrieval document.

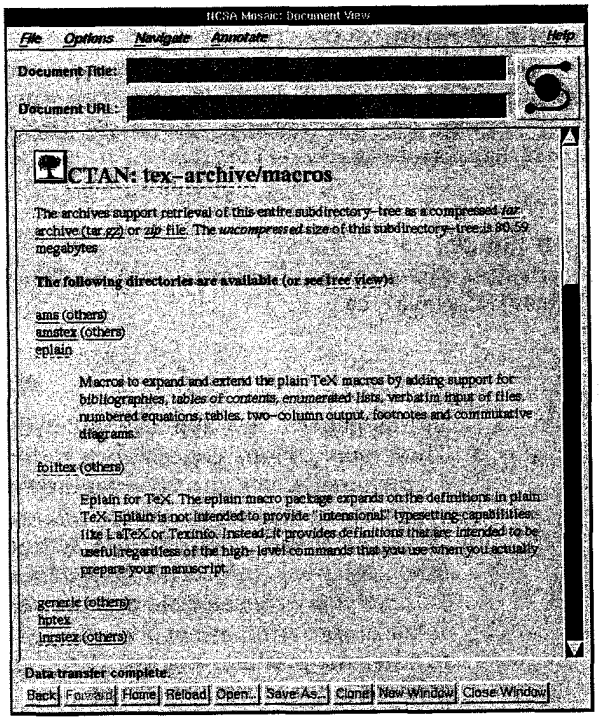

Figure 3: The `CTAN:/macros` directory.

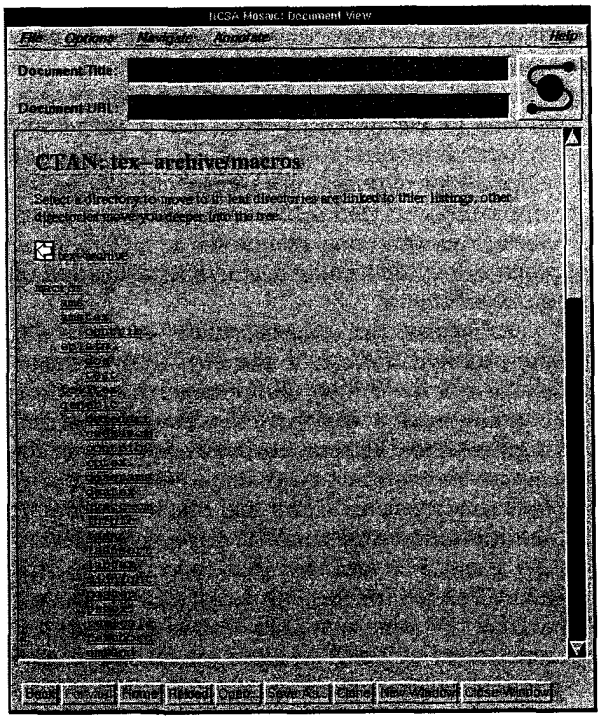

Figure 4: The `tex-archive/macros` tree document

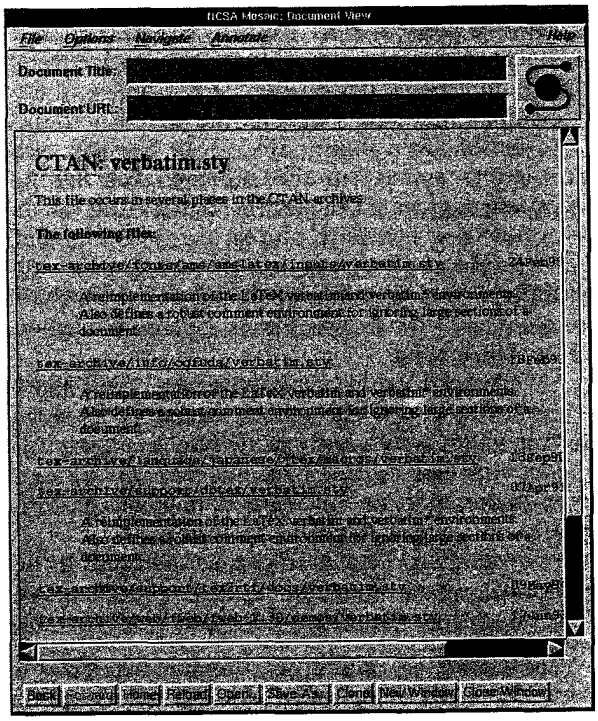

Figure 5: The `verbatim.sty` filename document