CTAN progress report

J Hefferon, ftpmaint@tug.ctan.org

2005-March-07

Abstract

This is an update on the work being done on CTAN, and is a follow-up to the prior article *CTAN Plans*.[1] We sketch some challenges, outline our goals, and describe the work to date.

CTAN could be considered, at present, a great success. It is the authoritative collection of TEX-related materials, with about eight gigabytes of community contributions. It has many visitors, who seem to be leaving with what they want — for instance, every day the tug.ctan.org site alone sees seven thousand web visitors and has ten gigabytes of FTP downloads.

But despite this, CTAN must change. It was developed in early 1990's as an FTP archive with an expected audience of sysadmins and TEXperts. Now, with the rise of personal systems, the average TEX maintainer is a struggling user who relies on their distribution and who works in isolation (i.e., has no easy access to an expert, either in computer systems or in TEX).

In recent years, DANTE has generously sponsored meetings among CTAN maintainers and others to discuss how to move forward in response to this new world.

What's wrong?

We will sketch the most pressing issues facing CTAN, but for more discussion see the *CTAN Plans* article.

Issue number one is that our visitors often have trouble finding things. Some reasons are that the archive is big, that many packages could fit into it in more than one place, and that some places on the archive hold so many packages that browsing through them all is impossible. That is, CTAN's growth has become sprawl.

One reason for this is our second issue, that administration is labor-intensive. Each of the maintainers is a volunteer. Yet each spends hours a day on the archive and energy for organizational efforts beyond routine installations can be hard to find. (Note, however, that a sometimes-proposed alternative to an administered archive where individual package authors decide what files to include and where their package should go could easily make the archive even more poorly organized than it is now.)

The third issue for CTAN is that we should offload more traffic to our mirrors. Users are poorly served by being forced to contact the core locations. The chief roadblock with mirrors has been that users often want to get a .zip bundle of an entire directory and most mirrors do not provide the bundling functionality.

Another issue is our desire to keep histories of at least some packages. This would allow users with a document that works only with an old version of a package to regenerate their output.

The fifth issue that has been a theme in our discussions is that we must do at least some customized stuff. For instance, a Google search of CTAN for "margins" yields many hits that are not helpful to inexperienced visitors, such as .dtx files. We need directed searching.

The final issue — perhaps the most important — is that we need to interface with distributions. If a new package is uploaded, or a new version of an existing package, we should be able to right away offer it to people who rely on a distribution. This means packaging it for the distributions. Everyone involved regards this need as critical.

What has been done?

The change that has been the most important, even though it was behind the scenes, is that Graham Williams has moved his wonderful *Catalogue* of CTAN packages to a CVS tree. This database now consists of about 1500 XML files. He has also expanded the data model, the DTD for the XML files, to allow more kinds of metadata, such as documentation links

[1] See http://www.tug.org/TUGboat/Articles/tb24_2/tb77heff.pdf.

and keywords. Maintaining these files is by-hand work, and so has increased the administrative load, but moving to CVS has cleared the flow of information in.

Another change is that we now ask for documentation from all contributors. We request it in PDF format, using Type 1 fonts. These have two advantages: in contrast with `.dvi` or `.ps` files they are readable to almost all of our visitors and so are better suited to be the target of search engine hits, and they showcase TEX's typographic excellence.

We have also merged the directory `/macros/latex/contrib/supported` with its sibling `unsupported` to make the single directory `/macros/latex/contrib`, as the distinction between the two proved to be not useful.

These changes have something in common. All American schoolchildren know the joke, "Where does an 800 pound gorilla sit? Anywhere that he wants." and so in the US the phrase "800 pound gorilla" has come to mean something so big that it is hard to move. In some ways, CTAN is such a beast. All three changes were clearly the right thing to do, but all three were more involved than a person would think possible. For instance, a Google web search for '`/macros/latex/contrib/supported`' shows that even now, quite some time after the directory change was made, the Internet still contains many links to the old location.

What is being done now?

These are our major goals.

- We are moving to further integrating package metadata into the archive. An example is that we plan to drop the `nonfree` tree and maintain the license distinctions in metadata.
- For this, we are enriching the metadata. For instance, it will include more extensive package descriptions, keywords, lists of related packages, and categorizations of packages by functionality.
- We are engineering our processes to keep the metadata information accurate as our holdings change over time. In particular, we will give uploaders the ability to edit the metadata using web forms at the time of upload. (Actually, it can be edited anytime but we expect that most information will come in as the packages come in.)
- With the information accessible to us in those files, we will give our visitors ways to leverage it such as a database-backed web site, and extensive search facilities including full-text searches of the descriptions and of the package documentation and keyword searches.
- We will offer visitors better integration with our mirrors.

Our present targets are: Rainer's Schöpf's initiative to store the package `.zip` archives right in the file system (instead of having them generated on the fly) and the resulting upgrades of install script will greatly help us refer visitors to mirrors, Graham Williams's and Robin Fairbairns's work on enriching the *Catalogue* information is at the core of being able to build an information-rich site, and Jim Hefferon's work on a web interface that is further described below.

The outstanding major goal that remains stalled is to work out and implement a standard for interfacing with distributions.

Example: ctanWeb

Development of the web component of the project is proceeding steadily. The code is in a Subversion archive, and there is now a beta site development testing. This section describes some features that will likely be present in the end.

One of the main goals of the updated site is to be welcoming to TEX beginners. For instance, the present top page has, in its second sentence, a link to an overview of TEX and friends, and another link to a short description of the steps needed to get started on a Unix system, on Windows, or on OS X.

Visitors can browse the archive's directories in a way similar to that offered by `tug.ctan.org` today. However there have been some improvements, such as the incorporation of package descriptions into the browsing. (Also, these web pages are now static for faster response and decreased server load.)

Visitors can browse the packages by functionality. The exact description tree is still under development, but perhaps a person looking for information on how to set page headers under LATEX would navigate their way through `Top > LaTeX > Page layout > Headers and footers` to get to a page listing a small number of packages, one of which is `fancyhdr`.

The updated site includes a provision to keep a history of certain packages as part of the package installation script.

Uploading of packages by developers is metadata-driven. For instance, a package author who gives us an update will get a screen to edit the metadata — description, version number, related packages, keywords, etc. This puts maintenance of the metadata in the hands of those best qualified to do it.

All metadata will continue to require approval by a CTAN administrator. However much of the approval process has been moved to the web so we can have people administering parts of CTAN remotely. In this way, we hope to increase the number of people involved in CTAN without requiring them to be system administrators.

Search facilities are considerably more advanced than on the present site. There is a search of package descriptions. There is also a search of all documentation files, including `README` files and `.pdf` files. This is better targeted than, say, a general `htDig` result; for instance, results show the package caption along with a link to the documentation so a person can better tell which responses are incidental hits and which are really of interest. (And, behind the screen, all of the searching relies on a database instead of `tug.ctan.org`'s present kludge.)

Conclusion

We've made progress on CTAN, and many of the elements of the final form have appeared (but much remains to do). We hope that the changes will help our visitors.

Acknowledgments

The CTAN team would like to gratefully acknowledge the support of the TEX community in general, and especially that of the user groups TUG, UK-TUG, and DANTE. In particular, the author would like to thank DANTE for support to attend recent conferences, including EuroTEX 2005. Without that support the work described here would not have been done.