Resources

CTAN plans

Robin Fairbairns, Jim Hefferon,
Rainer Schöpf, Joachim Schrod,
Graham Williams, Reinhard Zierke
ctan@dante.de

Abstract

The readers of *TUGboat* likely know the Comprehensive TEX Archive Network as a great pile of TEX stuff. That is, it is full of TEX materials and it is great, but it is also perhaps a pile — a bit of a mess.

We will sketch some plans for improving CTAN. As part of that, we will outline its architecture, history, and some issues.

1 Preamble

Taking it from the top: CTAN is an Internet archive of material related to TEX that is available for public download. We now hold five gigabytes of material. Each day about ten thousand visitors download a large number of files, and others upload some more.

We try hard to be definitive, to live up to our "Comprehensive" name. We hold everything from LATEX macro packages up to entire distributions such as MikTEX and teTEX.

2 Present

CTAN is not a single site, but instead is a set of sites. Three of these actively manage the material, for instance installing new or updated packages.

- `dante.ctan.org` in Germany is sponsored by the German TEX group Dante, and is maintained by Rainer Schöpf and Rienhard Zierke.
- `cam.ctan.org` is sponsored by UK-TUG and is maintained by Robin Fairbairns, in England.
- `tug.ctan.org` in the USA is sponsored by TUG, and maintained by Jim Hefferon.

To ensure that we have the same policies, we rely on an active mailing list. To ensure that we hold the same material, we rely on a number of custom scripts.

In addition to the core sites, about seventy-five sites around the world help out by offering a mirror — every day they sync up with a core site and then make their copy also publicly available. This gives users more options and eases network traffic. We encourage people to use the mirrors.[1]

[1] See `http://www.dante.de/mirmon/` and also `ftp://tug.ctan.org/tex-archive/README.mirrors`.

3 Past

Before CTAN there were a number of sites with TEX materials available for download but there was no authoritative collection. At a podium discussion that Joachim Schrod organized at the 1991 EuroTEX conference, the idea arose to bring together the separate collections. (Joachim was involved because he ran one of the largest FTP servers in Germany at this time, and had heavily modified the basic tool `mirror.pl` for this purpose.)

CTAN was built in 1992, by Rainer Schöpf and Joachim Schrod in Germany, Sebastian Rahtz in the UK, and George Greenwade in the US (George came up with the name). The site structure was put together at the start of 1992 — Sebastian did the main work — and synchronized at the start of 1993. TUG provided a framework, a Technical Working Group, for this task's organization. CTAN was officially announced at the EuroTEX conference in Aston, 1993.

When CTAN was founded, the main way to access files over the network was FTP. So the system was built with an expectation that visitors would get materials that way (and perhaps also with an expectation that visitors are experienced users). In 1999 a try at a more extensive web interface was put on the TUG site, but it is weak and was not adopted by the other two core sites.

4 Problems

Nobody likes complainers, but to describe our plans we must describe the issues that they address. There are problems with the collection itself, and problems with the administration of that collection.

One problem with the collection is that it is big. Its structure has been outgrown and needs updating. Most people in the TEX community have had the experience of being unable to find the solution to a problem, only to later discover that a solution was in fact on CTAN. That is, we have found that as we have grown, the information available to archive users to help locate materials has not grown fast enough to allow them to find what they need. This has been eased by the metadata[2] assembled by Graham Williams into his *Catalogue*.[3] But nonetheless, we need to be more information-rich.

A longstanding request about the collection[4] has been for CTAN to keep histories of packages, so that users can compile documents that rely on old versions. (Now, when a package author sends an update, we overwrite the old material.)

Another problem in the didn't-think-we'd-get-big (or-old) category involves mirrors. Often, the best way for a user to get a package from CTAN is to get its entire directory at once, so that they don't miss some files. To that end, the core sites support on-the-fly creation of `.zip` and `.tar.gz` file bundles.[5] The web front end at `http://www.ctan.org` uses this capability, and something like it must be a part of any future interface. However, it doesn't work with mirrors. In order to send users who want a bundle to a mirror, the system needs to know which mirrors correctly do on-the-fly-ing. So we wrote a script to check. When we ran it we found that not one mirror actually made both `.zip` and `.tar.gz` bundles without error. Consequently, the great majority of downloads come from the three core sites.

The flip side of people getting things from us is our getting things from the community. We are concerned by a trend whereby some package authors do not upload their work, but instead leave it on a personal web server. This is bad because it brings us back to the pre-CTAN days of materials that are scattered and that may disappear; that is, CTAN could end up not-comprehensive. It is also bad because, even if we know that the author's work exists, we have trouble gathering this material since the web protocol HTTP makes it hard for us to fetch things into our holdings.[6] Obviously we must work with the world as it is, but this is a problem.

The final collection issue that we will mention is that early developers, including Knuth, expected that most users would be fairly sophisticated: they would have developed basic typographic knowledge, and would need a minimum of computer and development support (e.g., they would write their own macros). This has not proved to be so. We perceive instead that the majority of TEX users want a distribution that comes with LATEX, etc., already set up. If they need to get something else, then they would like the distribution to have a module that can interface with CTAN and set it up for them. So a key goal is that we must — in conjunction with distributions — produce a system that meets those expectations.

We next describe some issues with the administration of the archive. Users do not see these directly, but they have an effect on what users do see.

[2] Data about data, that is, information about the packages.

[3] `http://www.ctan.org/tex-archive/info/Catalogue`

[4] Notably by Nelson Beebe.

[5] For example, visiting the url `ftp://ftp.ctan.org/tex-archive/macros/latex/contrib/shadethm.zip` will get the entire `shadethm` directory as a zip archive. From a command line FTP client, `get /tex-archive/macros/latex/contrib/shadethm.zip` will do the same.

[6] On the scale at which we work, this is not as simple as just using a program like `wget`.

The first is that we are a shoestring operation. The machines have been granted by user groups, but the critical network connections are donated by each maintainers's institution. The maintainers are far apart — some mailing list members have never met any other member — which slows progress and adds chances for miscommunication. All of the maintainers are volunteers and satisfy CTAN's time demands in the face of other things that must come first.

The time demands on maintainers are relevant because they have slowed our development. In particular, we must promptly handle materials that are uploaded. To help a reader get a sense of it — and for the satisfaction of bellyaching — consider a new package upload. The machine's maintainer gets it from the upload area and unpacks it. He checks the license, and decides where the package will go. He checks that there is a `README` file, and that there is documentation in PDF format that uses Type 1 fonts. Often, these checks involve corresponding with the author or with the CTAN mailing list, introducing a delay of a day or more. He then uses our custom install script to copy the material into the public area, and to trigger the mirroring process. He notifies the CTAN announcement list (and thus `comp.text.tex`). Finally, he edits the package *Catalogue* metadata and puts it into the CVS tree. In all, it averages perhaps a half hour per package.

More people in the administration might help. However, in addition to the necessary expertise with TEX, systems administration, and with the layout of CTAN, our work takes place in the context of an increasingly complex computing world. To name one example, in recent years licenses have become a big issue. We need people, but we also need a way to bring them in so that they can learn gradually.

5 Plans

Suppose that a colleague gives you a paper that requires a package not in your TEX setup. At present, you would visit CTAN, find the package, and then install it. Imagine if, instead, your TEX distribution got the package, installed it, and proceeded with running the paper, without your having to know anything about it. Technology that would make this kind of negotiation between the user's computer and CTAN reasonable is called "web services".[7] Our most important goal is to develop — in coordination with existing distributions — a capable spectrum of web services for CTAN to offer.

One step toward that, and toward accomplishing present goals also, is to better organize our holdings. For instance, we have already combined the subdirectories `supported` and `other` of `/macros/latex/contrib`, and we plan also to meld the `/info` and `/help` directories. A bigger job is to break all of our holdings into packages, and have each package in its own directory (no more `misc`). This is the natural way to answer a web services query, "what is the latest version of the file `f` in the package `p`?"

In support of web services, and also to help visitors get more information out we must get more information into the *Catalogue*. We must both (1) expand the information of the kind that is in there already, and (2) also expand the kinds of information that can go in there.

Part of (1) is an effort to provide an easy way to edit this metadata on the web, for instance, when an author uploads or updates a package. As a bonus, this may provide a way to bring people in to help CTAN. A person could make a reasonable contribution by editing metadata and checking it into the CVS tree, without having to do system administration of a CTAN site.

An example of (2) is that we need to retain keywords, so that users can search for a package in this way (this search could happen on a CTAN web page, or from a user's desktop through a web service). For some time we've been discussing the underlying model for the metadata and in support of this, the *Catalogue* recently moved to a CVS tree.[8]

All that information should be in a database. This fits into our plans in many ways because, while CTAN grew up as an FTP archive, the web has changed everything and we need to fix our web system to be database-backed. It should provide an interface that is uniform across all three core sites.

That interface could allow users to find packages in alternate ways (this was first suggested by a comment made by Sebastian). At present, users can look through the FTP directories, can search the list of all files, can do a crude text search of the *Catalogue*, or can do a web search of our holdings using a standard search engine, and we've mentioned above that we'd like to add a keyword search. But, we'd also like to add a search of packages by functionality: a user trying to work with page headers might click through a branch of choices like `Top > LaTeX > Page layout > Headers and footers`.

One of the things that a modern site should have is the ability to search documentation. At present, many packages do not have documentation,

[7] A rough definition is: a web server will respond to queries beyond just requests for pages.

[8] `http://texcatalogue.sarovar.org`

or have it in a format that is not suitable as a search result (e.g., if the result of a search is a link to a `.dtx` file then clicking on it is unlikely to be helpful). We have begun enforcing that package contributers provide documentation just in PDF format, which is the only format that combines widespread accessability and typographic excellence.

Two problems listed above are the question of keeping package histories, and the question of mirrors providing `.tar.gz` and `.zip` bundles. We believe that we can solve these together, saving each version of a package as a bundle — then we have a bundle available, and mirrors need not create them.

We need to convince authors to upload their materials. We have in the past urged authors to do so,[9] but here also a volunteer, who can find materials and politely persuade authors, would help.

Finally, we are constantly thinking about the maintainer's work flow. There has been some wild talk about an administration GUI, but the problem is that there are so many exceptions and special cases that we often cannot see how to do it any other way than by hand.

6 Prognosis

We plan to make CTAN more of a "Comprehensible" TEX Archive Network. These plans have been under discussion and in development for two to three years.

Dante has helped out greatly by sponsoring key people to come to meetings in the last two years, at Bremen and at Darmstadt, for in-person discussions. We must say that even beyond the grace of the invitations, the Dante people were kindness itself: in particular, Volker and Klaus moved the entire process forward greatly.

At present, the *Catalogue* format has been adjusted to allow development (in addition, moving it to the CVS tree allows contributions in parallel), structures for the databases are in place, and we have beta code for the web editing of metadata and other parts of the new web system. Now, we must test that system. Also, we must supply the static data: the web page content, the keywords, the by-function categories, etc.. Finally, we need more data about packages for the database.

Briefly: progress is maddeningly slow, but there *is* progress.

[9] If you have something that others would find useful, please consider sharing it!