

A categorised search of the CTAN

Peter Flynn

Abstract

The search engines accessible through the search page of the Comprehensive \TeX Archive Network (CTAN) allow you to search in three places: *a)* the CTAN directory structure and its filenames; *b)* Google; or *c)* the Graham Williams catalogue. While each has its advantages, they have a tendency to provide too much information. A new interface to *(a)* is being tested, which only shows direct matches, and categorises the output into different types of file.

1 Seek and ye shall find¹

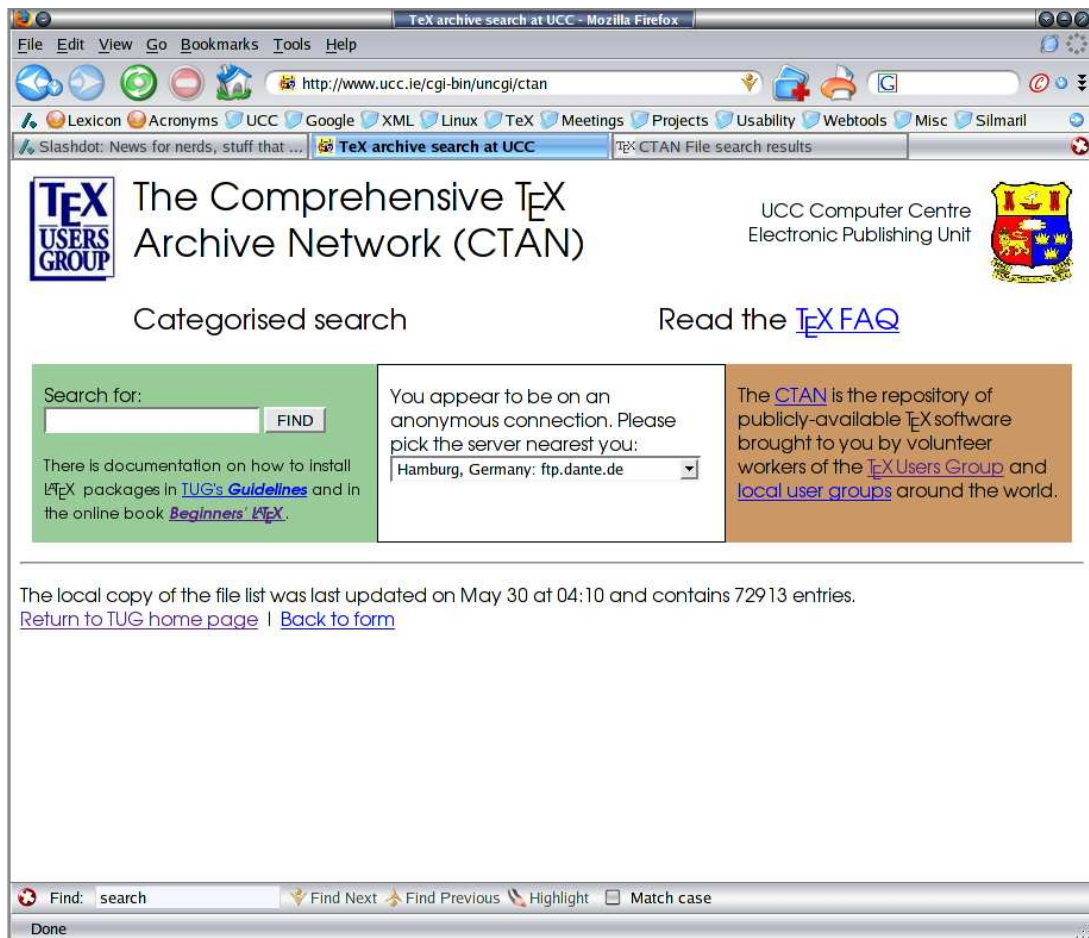
Readers of the `comp.text.tex` Usenet newsgroup and other online sources of information about \TeX and friends will have noticed the increase in recent years in comments like “I searched Google for it,” [“it” being the subject of the query], “but I didn’t find anything useful.”

While Google is unquestionably the primary network search resource used today, it is unavoidably indiscriminating between different senses of the words you search for. Expecting it to provide usable information about a system like \LaTeX is asking a little much.

The directory search at CTAN is ideal if you want to see the entire contents of each directory which contains a file which matched your search term, but this often returns huge directories containing many unrelated files, and requires considerable foreknowledge of file types to make use of.

¹Matt. 7:7

Figure 1: The initial search web page at <http://www.ucc.ie/cgi-bin/ctan>



The catalogue search is equally good for finding packages whose name or description contains your search terms, but this too can return far more choice than the user wants.

2 Enhanced directory searching

There has been a CTAN search engine of a sort at <http://www.ucc.ie/cgi-bin/ctan> for many years, but all it did was show files with matching names. It too was not discriminating, and made no concessions to the needs of the user.

A local user requested a way to see what different types of files there were, and this has resulted in a major revision of the search mechanism, and in the way the results are presented. It is still experimental, so suggestions for improvement are welcomed.

2.1 The interface

The initial HTML page at <http://www.ucc.ie/cgi-bin/uncgi/ctan> is shown in Figure 1. Note that the nearest CTAN mirror is selected automatically. This may not always work, however — users on connections without assigned hostnames (often dial-up or broadband), for example, cannot usually be identified accurately. In either case the user has the option to pick a nearer server.² On the left is the search box and links to installation guides; on the right is a “credits” panel.

Typing a word and clicking “FIND” performs a simple search for that string in the `FILES.byname` list as before. The search term can contain multiple words, and the default is to require all of them to be matched (as if they were separated by AND). They can be separated by OR if the terms are optional. The application of simple Boolean division is echoed in the redisplay of the search term (see Figure 3).

2.2 The output

Results are sorted into four categories (for the moment: more could be added — see Figure 2):

²In an earlier version of this site, the user’s preference was recorded against their second-level domain for future reference and re-use by others from the same domain, but this proved to be unreliable and has been dropped for the moment.

Figure 2: A simple search

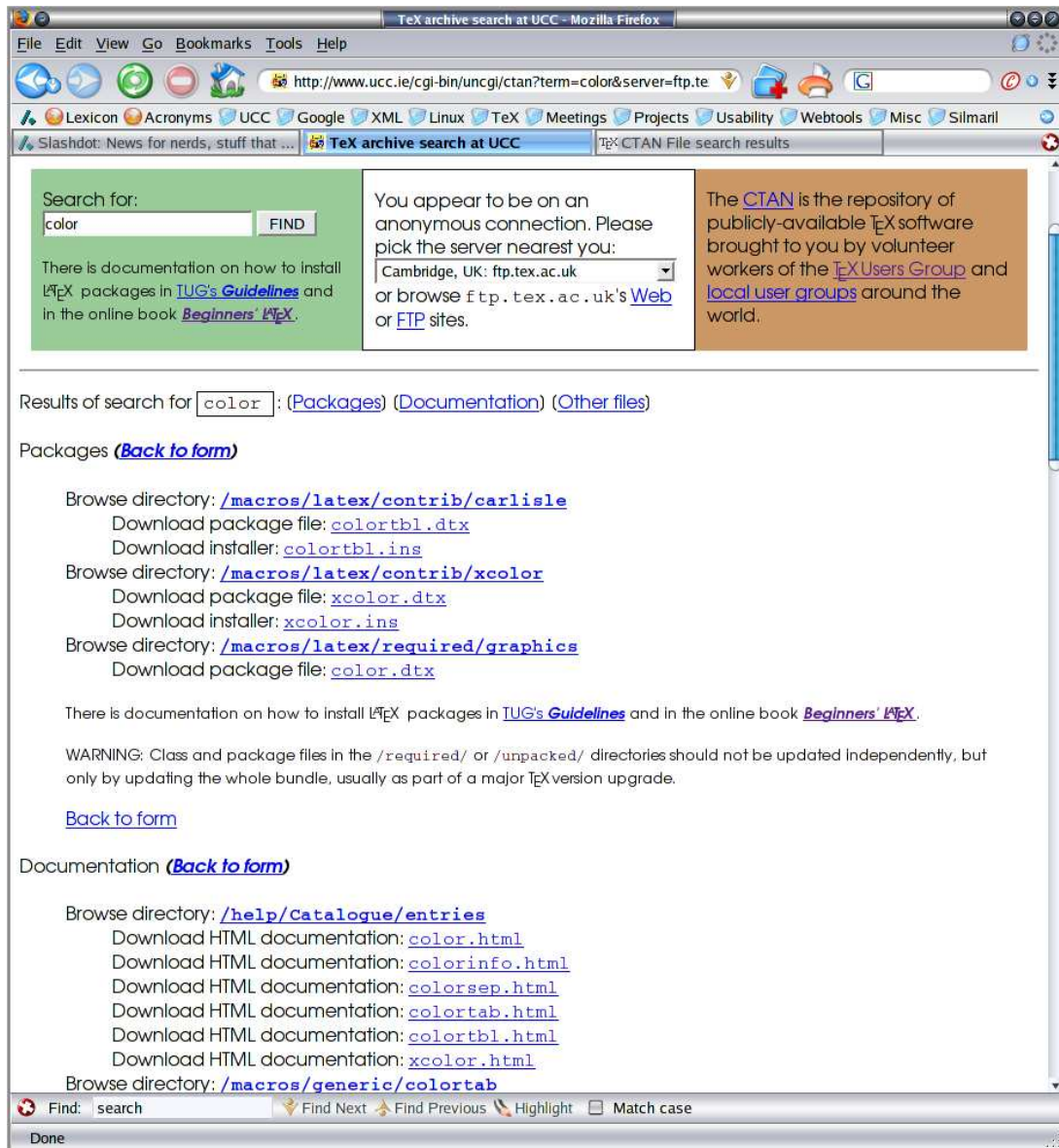


Figure 3: A more complex search

The screenshot shows a Mozilla Firefox browser window with the address bar displaying `http://www.ucc.ie/cgi-bin/uncgi/ctan?term=mf+logo&server=ftp`. The page title is "TeX archive search at UCC - Mozilla Firefox". The browser's menu bar includes "File", "Edit", "View", "Go", "Bookmarks", "Tools", and "Help". The address bar contains several icons and a search bar. Below the address bar, there are several tabs, including "Slashdot: News for nerds, stuff that ..." and "TeX archive search at UCC". The main content area of the browser shows the CTAN website. The website header includes the "TeX USERS GROUP" logo, the text "The Comprehensive TeX Archive Network (CTAN)", and the "UCC Computer Centre Electronic Publishing Unit" logo. Below the header, there are two main sections: "Categorised search" and "Read the TeX FAQ". The "Categorised search" section contains a search box with the text "mf logo" and a "FIND" button. Below the search box, there is a message: "There is documentation on how to install TeX packages in TUG's Guidelines and in the online book Beginners' L^AT_EX." The "Read the TeX FAQ" section contains a message: "You appear to be on an anonymous connection. Please pick the server nearest you: Cambridge, UK: ftp.tex.ac.uk or browse ftp.tex.ac.uk's Web or FTP sites." Below the search box, there is a message: "The CTAN is the repository of publicly-available TeX software brought to you by volunteer workers of the TeX Users Group and local user groups around the world." The search results section shows "Results of search for mf and logo" with links for "(Packages)", "(Fonts & Graphics)", "(Documentation)", and "(Other files)". Under "Packages", there are links for "Browse directory: /macros/latex/contrib/mflogo", "Download package file: mflogo.dtx", and "Download installer: mflogo.ins". There is also a link for "Back to form". Under "Fonts and Graphics", there are links for "Browse directory: /fonts/mflogo", "Download: CATALOGUE", "Download: Makefile", "Download: README", "Browse directory: /fonts/mflogo/mf", "Download METAFONT file: logos18.mf", "Browse directory: /fonts/mflogo/ps-typel/hoekwater", and "Download: README". The browser's status bar at the bottom shows "Find: search" and "Done".

1. Document Classes: class files (.cls) and class options;
2. Packages: style files (.sty), DOCT_EX (.dtx) and installer files (.ins);
3. Documentation: HTML files, DOCT_EX (.dtx) (again), PDF, PostScript, and text files;
4. Fonts: METAFONT and PostScript font files (metrics, virtual fonts, font definitions, etc.)

The directory is a link to the chosen server for browsing; each matched file is typed and the explanation displayed alongside the filename. Not all file types are yet covered.

The intention is to provide the user — especially the newcomer — with enough information to let them decide which files they need without having to undertake extensive research, and to help ensure that (for example) they remember to download *both* .dtx *and* .ins files for a package.

3 To Do

Much remains to be done. US and British spellings need to be conflated, so that a search for “colour” and “color” are made equivalent unless enclosed in quotes. User-typed Regular Expressions would be nice, but more work is needed on the Boolean filter first to add a NOT operator. The remaining file types need to be added.

Adding a test search of the catalogue is high on the agenda, but it needs some linguistic expertise to add filters capable of ranking the hits by likely relevance, which is out of my field entirely.

As I said, suggestions are welcomed!