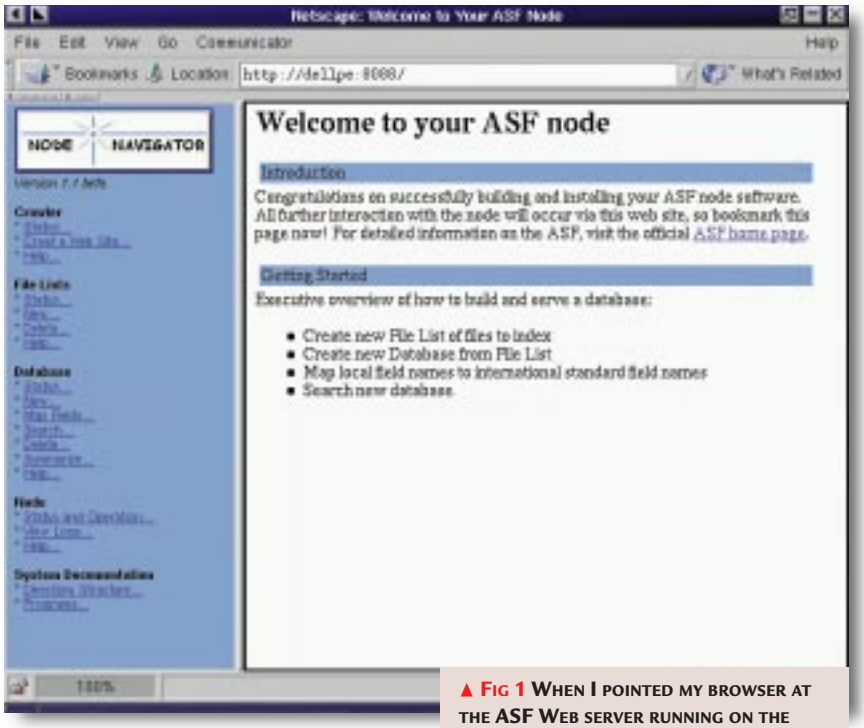
Casting the net

A web interface can be used for text search and retrieval — Chris Bidmead finds a free text database system for Unix.

Earlier this week the hard drive in one of my older servers took a dive, bringing down with it the Topic text retrieval database that is a key component of my writing activities. Thanks to the regular exercise I give my trusty HP SureStore DAT tape device nothing was lost but it set me thinking.

Verity's Topic is a product I've been using in one form or another for the best part of a decade. Designed around client-server architecture from the outset, it was one of the first text retrieval packages to migrate to web technology.

Irrespective of operating system or hardware you can query the text base from any machine on the network which is capable of running a web browser and get the results back in the form of web pages. So, you get complete freedom to choose your favourite workstation OS, independent of the software running Topic on the server. This was a key part of my migration to UNIX. Since then, the idea of using a web interface for text retrieval has become commonplace.



▲ FIG 1 WHEN I POINTED MY BROWSER AT THE ASF WEB SERVER RUNNING ON THE OTHER END OF MY NETWORK AND ENTERED MY USER NAME AND PASSWORD I CAME UP WITH THIS PROFESSIONAL-LOOKING SCREEN. THROUGH THIS, THE ADMINISTRATOR CONTROLS THE ENTIRE OPERATION OF THE DATABASE.

A number of free and commercial products offer search and retrieval through a browser, often using at the back end a standard web server connected to a set of small purpose-written programs through the CGI (common gateway interface) web server programming standard.

I'm running a Cobalt Qube 2700 Linux-based server on this system and the little blue box comes with an index and search engine called glimpse <glimpse.cs.arizona.edu> ready installed in conjunction with some CGI magic for rendering web pages. At the Arizona site you'll also find links to Harvest and WebGlimpse, which are similar systems based on the glimpse engine. A different approach is ht/Dig at www.htdig.org.

You can build it into a complete working system

The web search and retrieval mechanisms which come with the Qube are great on an intranet for storing HTML pages you know you'll need to refer to again, or want to publish to colleagues. But as far as I can make out they lack a key feature I've come to depend on with Topic and that is the ability to search on fields within a document.

For example, I find it very useful to be able to look up all relevant documents by a particular author over a particular range of dates. So Author and Date are fields which I establish in all my Topic documents, as well as Source and Title.

Topic runs on various flavours of UNIX but I have the Windows NT version dating back to the days when this column covered both these operating systems but since this is now resolutely the UNIX column I thought I'd scour the web to see what was on offer in the way of a free UNIX text database system.

The two faces of ASF

My quest brought me to something called ASF <asf.gils.net>. The initials stand for 'Advanced Search Facility' and I

A UK LINUX RESOURCE

I hadn't heard of Digital Networks UK until I stumbled on its ad banner on the slashdot.org website. Its own website at www.dnuk.com seems to be offering an excellent price on software. Official S.u.S.E. Linux 5.3, for example, is £20 (ex VAT and delivery) and hardware is available, too. If you're looking for a ready-made dual boot Windows/Linux system for fun and games, check out the company's PlayStar workstation, currently on offer for around £850 (inc VAT and delivery to the UK mainland). There's an interactive web page for you to juggle with the feature options and check out the resulting prices. The firm's Linux co-ordinator, Lee Chisnal, is on 0161 3398555.

◂Fig 2 An easy way of filling the database is to use the Web crawler that comes as part of the package. This can copy entire websites into your ASF system for indexing

was initially put off by the reams of information retrieval theory describing the system.

I waded through discussions about centroids and geospatial metadata until I twigged that ASF is at least two different things. On the one hand it's a comprehensive architecture for the implementation of a potentially vast collection of interlinked subject gateways — wow! On the other hand, it is a relatively small bundle of free software tools which have been assembled as a kind of 'proof of concept'. But this get-you-started kit also happens to be a truly practical software package [Figs 1&2]. And, yes, ASF does know how to handle fields, and is flexible about how it finds them inside documents.

The key component of the kit is Isearch, a set of tools for indexing and querying a collection of data files from the command line. Another component handles CGI interaction between Isearch and the web browser, Apache, which makes up the third key part of ASF.

The whole lot is ready tarballed as a 2.7Mb download. Now, I know from your emails that the fact that the distribution is source and you have to compile it yourself is going to strike many of you as bad news, but take my word for

LINUX FOR STARTERS

Reader John Horton sent me an interesting email but my reply was bounced back by the postmaster at his domain. I wrote back to the postmaster and was told that John had left the company, so I'm printing my reply here in the hope that he gets to read it, and because it covers some subjects which come up in much of the correspondence I receive.

John wrote: 'I've just installed Linux on my P200 PC at home, alongside Windows 95. I am fairly technically minded and used to DOS so I thought I would give Linux a whirl to see what all the hype was about. It would be a nice idea if you could start a section which includes all the basics. For instance, I see on the internet many pieces of software to download but have no idea how to install them.'

I think it's probably worth spending time getting to know the system you're running before you add more software. There are several quite different installation methods, depending on how the software is packaged. The most common ones are:

☛ **rpm** the RedHat Package Manager. A utility called rpm — or glint or xrpm if you're running under X — checks the dependencies (are the requisite libraries installed?), installs the software in the correct directories and runs any installation scripts that may be needed. These rpm packages are generally called <something>.rpm. See www.rpm.org.

☛ **Tarballs** are files and directories compressed into a single file. They may contain only source code *(see below)* or ready-to-run executables. Tarballs don't carry out dependency checks and need to be unpacked in the correct directories, or at least you need to know where to put the files manually. These files have a .tgz or .tar.gz extension.

☛ **Source code** may or may not come from a tarball and needs to be compiled for your system before you can use it. Often this is easier than it sounds, thanks to a config utility that reads your system and makes the required adjustments to the 'make file' (the config file which defines how the compilation runs). Small utilities designed to run across many UNIX platforms are often successfully distributed in this way.

John's email continued: 'Have you any ideas as to a good, ideally free web browser, word processor, etc? Also, I am aware of Gnome and Wine and would like to know more, especially about the latter. Any chance of a beginner's guide?'

☛ **Netscape** is available without charge, as is WordPerfect 8. You might like to check out the applications available at sal.kachinatech.com.

☛ **Gnome and Wine** are on my rather long list of things to write about. But in the meantime, why not cut out the middle man and head straight for www.gnome.org and www.winehq.com ?

it, the process could hardly be simpler. Provided you're running Linux or Solaris — the source can be ported easily to other UNIX platforms, but some twiddling is required — all you need to do once the tarball has been unwrapped is type 'make linux' or 'make solaris' and you're done.

All right, maybe it's not that simple as you'll need to make sure you have the appropriate development tools installed on your system. I had the GNU C compiler, gcc, already in place on the SuSE 5.3 system on my Dell PowerEdge but the string of errors I got in response to 'make linux' alerted me to the fact that I was missing the necessary C++ libraries.

I returned to the package selection menu in YaST, the SuSE system management and installation tool, and pulled in gpp and libgpp from the SuSE distribution CD-ROM. I don't know much about these but YaST describes them as 'The GNU C++ compiler and library' so they sounded right. Next time I ran 'make linux' there were a few warnings but no errors.

The top level 'make' which you trigger when you do this runs down the directory trees and fixes up the makefiles that control how each of the separate tools in the bundle are compiled.

Whoever is responsible for putting the bundle together has done an excellent job. Once you have the binaries you simply follow the instructions in the README file. This tells you first to run the bin/asf_setup script, which asks a few questions about user names, passwords and directory locations, and then to run bin/start_main_servers which launches Apache and the search engine.

Now the demo system is ready for use. When you access it from a browser it comes up looking like a fully-fledged professional piece of software, which it is. You can go on from there to build it into a complete working system — or even a 'subject gateway'!

IS IT LITESTEP OR AFTERSTEP?

MANY THANKS TO READER CARL ROBSON CARL@SKRAGGY.FREESERVE.CO.UK FOR SENDING ME WHAT LOOKS VERY LIKE THE AFTERSTEP DESKTOP I USE ON SEVERAL OF MY LINUX MACHINES. THE TWIST IS THAT THIS IS THE LITESTEP DESKTOP SHELL, RUNNING ON WINDOWS <LITESTEP.NET>. WELL, CARL, UNIX IT AIN'T — BUT IT'S A LITE STEP IN THE RIGHT DIRECTION...

■ The Open Source 1999 Conference

Earlier this year I attended the Open Source 1999 Conference at the Commonwealth Institute in London. This is a gathering of some 400 technical and managerial delegates hosted by a company called NetProject <www.netproject.com>.

NetProject is run by Eddie Bleasdale, a name which has a mellow resonance for anybody old enough to remember the early days of UNIX here in the UK. Bleasdale was the first European manufacturer of UNIX hardware back in the early eighties and in a lot of ways the Linux phenomenon is a rerun of that early excitement.

You will need to make sure you have the appropriate development tools

The keynote speaker of this year's conference was Eric S. Raymond. He is the Open Source guru whose considerable contribution to free software ranges from arcane technical documents such as *The Hitchhiker's Guide to X386/XFree86 Video Timing,* to the now classic Open Source position paper *The Cathedral and the Bazaar,* and along the way takes in practical programming efforts such as the ubiquitous fetchmail email collecting utility.

Undoubtedly, Raymond was the star of the conference with his exegesis of his latest paper *Homesteading the Noosphere*, which pursues a not always coherent analogy between the ownership of software and the Lockean theory of land tenure (you can read it yourself at www.tuxedo.org/~esr/writings/).

The noosphere is the sphere of human consciousness and mental activity in regard to its influence on the biosphere and in relation to evolution. Eric Raymond's noospheric aerobatics were perfectly complemented by two very down-to-earth lectures by Mike Bahahan of the IT consultancy and training company GBdirect <www.gbdirect.co.uk> who addressed the practical uses of Linux and free software in the context of business.

During a beer break at the end of the conference I also managed to catch up with Alan Cox, the kernel guru generally regarded as Linus Torvalds' representative on earth. You can see what Alan is doing at the moment by visiting the web pages at www.linux.org.uk/diary.

PCW CONTACTS

Chris Bidmead can be contacted via the PCW editorial office (address, p14) or email unix@pcw.co.uk