Linux WWW-HOWTO

by Peter Dreuw, pdreuw@wing.gun.de v0.7.6, 6 October 1996

This document contains information about setting up WWW services under Linux (both server and client) and how to maintain them. It tries not to be a in detail manual but an overview and a good pointer to further information.

Contents

1 Introduction

Many people are steping into Linux 'cause they are looking for a really good *internet capable* platform. Others use Linux for the fun installing a free OS on their system. Some of those want to get in touch with the internet, of course. Furthermore, there are institutes, universities and other mostly not-for-profit organisations which want to or need to set up internet sites on small expenses. This is, where the WWW HowTo comes in. This document tries to explain how to set up clients and servers for the (in my mind) largest online part of the net - *The World Wide Web*.

1.1 Copyright

1.2 Disclaimer

This document is meant as an introduction into WWW techniques used or usable on Linux. I an not an WWW nor security expert ! I AM NOT RESPONSIBLE FOR ANY DAMAGES INCURRED DUE TO ACTIONS TAKEN BASED ON THE INFORMATION INCLUDED IN THIS DOCUMENT.

1.3 Feedback

Any feedback is really welcome. Just mail to pdreuw@wing.gun.de.

1.4 New versions of this Document

New versions of this document can be retrieved via anonymous FTP from sunsite.unc.edu under /pub/Linux/docs/HOWTO and almost any friendly Linux ftp mirror site.

Furthermore, you can download it via `<http://ourworld.compuserve.com/homepages/dreuw/lxwwwh2.tgz>` as gzipped tar archive containing a sgml, text, latex and ps version. The html version is directly available under `<http://ourworld.compuserve.com/homepages/dreuw/lxwwwh2.htm>`

2 Setting up WWW client software

The following chapter is dedicated to the web users. Some hacks and tricks setting up current versions of common web browsers. Please feel free to contact me, if your favorite web browser is not mentioned here.

(As this is a really early version of the WWW-HOWTO, most of them are likely not to be listed...)

Personally, I prefer the Emacs - W3 browser and Lynx as they have some speed advantages and there is no need to retrieve the complete graphics through my slow speeded dial up line ;)

2.1 Overview

Lynx is the smallest Web browser I know and use - but ist has many special features, so don't skip this chapter.

Emacs - well there is nothing to say about the Emacs W3 browser, its just Emacs, like the Emacs news reader, the Emacs mail reader etc. pp.

Netscape Navigator is the only browser mentioned here, which is capable of this new funny things like *JavaScript* and these nice <APPLET> tag feature needed tu run *Java*. Please report if there is any other web browser which can do the one or other. I'd really like to know.

There are rumors, that Microsoft is going to port the Internet Explorer to varios Unix platforms - maybe including Linux. If you DO know something more reliable, please drop me a mail.

2.2 Lynx

The smallest (?, hm, something around 650 K executable) and maybe fastest Web browser available. It does not eat up much bandwidth nor system resources as it only deals with text displays like any console, terminal or xterm. You don't need any *X Window system* nor additional megabytes of system memory running this little browser.

Furthermore, the source code is available, too.

2.2.1 Where to get

The latest version is 2.5 and can be retrieved from `<http://www.wfbr.edu/dir/lynx>` or from almost any friendly Linux ftp server like ftp://sunsite.unc.edu under /pub/Linux/system/Network/info-systems/www/ or mirror site.

Or, take a look at the *Lynx enhanced pages* `<http://www.nyu.edu/pages/wsn/subir/lynx.html>` for information on using Lynx.

2.2.2 How to install

Just retrieve the archive, unpack it, *read* the `README` and *follow* the steps told in the `INSTALLATION` file.

If you don't want a source distribution, you'd maybe retrieve a binary distribution for the Linux on Intel based systems available on sunsite.

Lynx compiles and runs on my system without any problems on both Linux 1.2.13 and 2.0.x.

2.2.3 Special features

Well, there are. For a complete description, just read the manuals and doc files that come with Lynx.

To get a nice glimpse, just type in

```
lynx --help
```

and be impressed.

In my humble opinion, the most special feature of Lynx against all other web browsers is the capability for batch mode retrival. One can write a shell script which retrieves a document, file or anything like that via *http*, *ftp*, *gopher*, *WAIS*, *NNTP* or *file://* - url's and save it to disk. Furthermore, one can fill in data into HTML forms in batch mode by simply redirecting the standard input and using the *-post_data* option.

2.3 Emacs-W3

There is one sad thing about the Emacs W3 browser ;) If you got GNU Emacs or XEmacs running, you probably got the W3 browser running to. Not much work in this HOWTO. If you feel, that there should be more information about this, please let me know.

The Emacs W3 mode is a nearly fully featured web browser system written in the Emacs Lisp system. It mostly deals with text, but can display graphics, too - at least - if you run the emacs under the X Window system.

The most recent GNU emacs package is available under `<ftp://prep.mit.ai.edu>`, the most recent XEmacs could be retrieved from `<ftp://ftp.xemacs.org>`.

2.4 Netscape Navigator Gold 3.0

Yeah, you made it. The Queen of WWW browsers. Something almost like Emacs is in the world of text editors. Netscape Navigator can do nearly everything (except cooking coffee... but maybe *java* will do...). But on the other hand, the most memory hungry and resource eating pice of *web browser*, *news* reader, *mail* reader (pop3), mail & news editor I've ever seen.

My latest version of the Netscape Navigator Gold (export version) is from 28-Aug-1996 and (c) 1995, 1996 Netscape Communications Corp.

(As I live in Europe, I can only get the export version...)

2.4.1 Where to get

The first place to get the Netscape Navigator for Linux as binary distribution is on `<ftp://ftp.netscape.com>`. The second - as these servers are heavily loaded - may be any friendly netscape mirror site. You might as well ask archie about this. Maybe, you'll be happy and find it on a cd rom - this will save some bandwidth as the archive is quite large (2.5 MB).

2.4.2 Unpacking & Installing

Unpack the archive und *read* the `README` file ! There is really nothing strange about this, you know.

2.4.3 Java applets with the navigator

There are some reports telling that there are problems running java applets with the Netscape Navigator Gold 3.0 even if java is activated in the otions dialog. The archive known to me contained a file *java_30* which must be renamed to *java_30.zip*. After this, any java applet should work fine within the netscape environment.

If you continue to have problems using java applets like Netscape Navigator hangs or just terminates after downloading a java applet, take a look at your *libc version*. Just do a

```
ldconfig -v | less
```

(maybe, you have to be root doing so...) and watch out for an entry

```
libc.so.5 => libc.so.5.xx.yy
```

where your libc version is 5.xx.yy. If your libc isn't 5.2.18, this may be the problem. There are many reports for Linux 1.2.13 systems, that they should upgrade to libc 5.2.18 when the need to run Netscape Navigator in general. Additionally, it may be a good idea to downgrade your libc from 5.3.xx to the 5.2.18 if you run Netscape Navigator and a Linux 2.0.x kernel. (In fact, the libc 5.3.xx series is for beta testing purposes, so you should know what you're doing.) Some of the 5.3.xx series break Netscape Navigator and the Java classes code.

For more information on *Java on Linux* or *Java programming*, please read the JAVA-HOWTO or visit `<http://www.sun.com>`.

3 Setting up WWW server systems

This section contains information on different http server software packages and additional server side tools like script languages for CGI programs etc.

For a technical description on the http mechanism, take a look at the RFC documents menitoned in the chapter "For further reading" of this HOWTO.

3.1 cern httpd

As the cern original httpd server is reported to have some ugly bugs, to be quite slow and resource hungry, it is not described in this HOWTO by now. If you volunteer to admit some facts or chapters, please send them to me, I'll add them to this doc.

3.2 apache

-*To be written* - *sorry* Features, Overview, Advantages

3.2.1 Where to get

3.2.2 Installing

3.2.3 Configuring

3.2.4 Special Features

Apache httpd has got some special features in the actual version.

Host multicasting BlaBla??? how to setup ....

Module system how to include other modules ... where to get infos about module programming ...

3.3 CGI scripts systems

- *to be written* - *sorry* - CGI (common gateway interface)

3.3.1 How does CGI work in principle ?

- *to be written* - *sorry* - calling structure, http structure, program parameter format (slightly touched), things to keep in mind

3.3.2 Perl

- *to be written* - *sorry* - something easy in perl (sample script)

3.3.3 PHP/FI

- *to be written* - *sorry* - something easy in PHP/FI (sample script)

3.3.4 W3-mSQL

- *to be written* - *sorry* - something even more easy (sample script) hint about setting up !!!

3.3.5 some useful scripts

- *to be written* - *sorry* -FaxInbound to nice Table including php/fi script and shell script

4 Maintaining a WWW site or some Web Pages

If you have to maintain a web site or if you maintain at least a web page, you have to think about your offer to the network and you have to spend some thoughts about approaching the reader / user of your web pages.

4.1 The mainstream: HTML technical

Well, I'm not gonna tell you, how HTML is encoded an how you have to design your pages. I'll just give you some pointers where you can find more advanced information.

You should take a look at `<http://www.w3.org/>` for the latest HTML language specification.

Take a look at the list at the end of this article, you'll find more hints, where to read on.

4.2 Some thoughts about bandwidth

Many users connect to the internet via slow speed modem lines. A speed range from 14,400 bps to 28,800 bps is state-of-art for "private sites". In europe, there are ISDN systems growing, but a speed of 64,000 bps isn't that more fast in comparison to - let's keep it simple - 10,000,000 bps ethernet. And 10 Mbps ethernet isn't really a high speed LAN connection nowadays.

As you realize that many users don't have this fast access to the net, you should keep in mind to put up the relation between information and bytes. Optimize it at 1:1 - if you can. You *may* use graphics in your web pages following the multi media trend, but always remember the goals of your page and of the graphic you're going to put in. If most of your users are connected via a small modem line and the graphic severes only for estethic reasons or some eye-catching effects, you'd better bann it from your pages, or -at least- rerender it to the smallest possible file size and use best compression. Your users will like it.

Always remember, nobody really likes an eye-catcher, that comes up about 3-5 minutes *after* the text message.

4.3 Some thoughts about server load

On a web server, there is normally at least one server task running. If this task reads a request from a http client, it duplicates itself (on Linux it's called *fork*ing) and the new copy serves the request, while the original keeps listening for new requests. After finishing the request, the copy terminates. (In fact, some servers - like the apache - always keep up a default of five ready waiting server copies for requests parallel to the master incarnation for speed reasons.)

Some web browsers like the Netscape Navigator series do many requests parallel on the same server, which increases the server load spend on the same user. These browsers e.g. retrieve the HTML page and parse them while retrieving and issue new requests for other information like the embedded graphics, applet files, sound files or any other additional mime-encoded data. In opposition, 'simple' browsers request and retrieve one file after another, which keeps the server per user load relation as low as possible.

Many users prefer browsers that use the multi request technique like the Netscape Navigator, because they bring up a more complete overview on the requested page before the single request browser does.

This is in my opinion because many page designers do stick on embedding the information into the graphics, denying the text-only browsers.

So, we - as server maintainers - got the problem, that most of the users cast multiple requests on out server within the same page retrival. We can limit this by limiting the server software not to serve more requests than "x" from the same requesting system at the same time. But how to get this "x" ? It's not easy to calculate and a lot of personal expirience on your site is necessary to depict it. But I'll give you some hints. We have to take our connection bandwidth into account, our server memory size, some feeling about our servers cpu/disk performance and ... well, that's enough for the first glimpse. You should take a look at the memory usage a single server task has. Then think, how many of them could kept in memory at all. Think, how many per cents of your web pages could remain in your servers disk cache. Optimize the count of web server tasks against the disk cache size and you're really near to your personal "x". Furthermore, you can put in other jobs the server got. E.g. if your system also serves for ftp, you might limit the maximum possible connections to keep up some minimum room for the ftp server task. If your web server also does some database services, you'd better keep up some cpu cycles and also shrink your "x". Play somewhat around with these values and test them. And (!) read the following chapter about CGI scripting, which also takes server performance and - depending on the CGI jobs - amount of memory.

4.4 CGI vs. Applet / Client side script

- *to be written* - *sorry* - overview ond advantage/disadvantage and hints when to use which.

4.5 Style ideas

Uh, a really difficult theme for beeing on a short sentence. I don't try to mix up your genious design ideas. Nor I'm gonna put you into my personal design strategies. I'd just like to add one or two statements to the above ideas on server load and bandwidth.

Numerous research on human behavior on user interfaces and on-screen presentation have brought out interesting results. There are some simple facts one should keep in mind designing WWW pages.

- Keep text in short blocks. This HOWTO is ugly to read on screen, but nice to read in paper print. (Try it yourself!) Human beeings often have difficulties to read lengthy text printed on screen. They loose their point in the sentence; their concentration suffers.
- Don't mix up graphics and text blocks. This is a good-looking but ugly-to-read feature. You can spread Headlines, eyecatchers but, please, don't mix up block text with graphics. Behaviorists found out, that human are much more attracted by graphis on screen than by text. People find it more easy to realize a graphic on screen than on paper, in opposition to text which is more easy to "see and decode" on paper than screen.

Did you know this ? If you'd like to get more information on that, search for GUI style guides and ergonomy research results done by many universities and software companies (including MS).

4.6 HTML editors under Linux

Hm, there are some. In fact, there are reported to be many. But as I already shot my shoot, I didn't test them all. But I am really curiosly looking forward to read the reports you're gonna mail.

4.6.1 vi, vim

vi and vim are perfectly usable for writing HTML code... (don't flame me on that) because HTML code only uses ASCII text chars. I don't want to give stuff for another editor war. Those who know vi/vim and use it daily can use it for HTML code either. You can make vi/vim help you developing HTML code by doing some macros for vi/vim. But as this is no VI-HOWTO, I'll leave this fact alone here. Just take it, that it is possible to use vi/vim for HTML editing (at least for some short changes). If you already know how to program vi/vim, you'll certainly know how to abstract for HTML either. If you don't do so, well, don't care.

4.6.2 emacs & XEmacs

- *to be written* - *sorry* -

4.6.3 asWedit

- *to be written* - *sorry* -

4.6.4 other pointers

Ah, there was some reference for a package named phoenix, based on tkWWW, but I was not able to get them running on my system. I think, it was a problem with my tcl/tk versions but you'll never know. I didn't spend much time around with them, so, maybe they'll run on your system both. Just go'n ask archie. Maybe, you can drop me a mail, if you are sucessful.

If you miss your faivorite HTML editor here, just write a mail to me. Maybe, I'll add some pointers to web pages about HTML editors for Linux to. Just send me some nice URL's.

4.7 Graphics

Thoughts, Ideas, Hints ? Well, you may read the *comp.graphics* newsgroup. And, you can visit `<http://www.w3.org/pub/WWW/Graphics/>`.

4.7.1 Format gif

GIF (Graphics Interchange Format) was introduced 1987 by Compuserve, Inc. an revised 1989. It uses a LZ algorythm, which underlies U.S. copyright or patent law. So there might exist some legal problems using this graphics format in the internet - despite the fact that nearly anybody does.

Gif is a good format for small pictures with simple structured graphics like computer graphics or banners.

Gif has some advantages as it is one of the (if not the) widest spread graphic formats in online systems:

- offers a good compression
- compresses without information loss
- has a interlace capability, i.e. pictures could be viewed in full size (with less resolution) before they're retrieved completely.
- can hold more than one picture within one file
- can hold a small animation in one file
- nearly any graphical web browser supports gif
- can hold a transparent color
- fast decompression system

The disadvantages are:

- only 256 color pictures possible
- license and copyright problems (?)
- not ideal file size

4.7.2 Format jpeg

The Joint Graphic Experts Group (JPEG) did the design for the jpeg/jpg/jiff graphic format. This format is based on a discrete cosinus transformation (DCT) and a Huffmann encode compression. JPEG works with an significant information loss, which can make your pictures somewhath less colorous or less sharp. Typical compression factor is 1:5 ranging to 1:50. (Above 1:10 anybody is able to see the artefacts risen through the compression/decompression cyle.)

JPEG is a good format for photographies, large graphics and really complex pictures.

The advantages are:

- strong compression, small files and therefor fast download...
- any graphical browser knows about jpeg

The disadvantages are:

- slow compression/decompression
- possible information loss

4.7.3 Format png

Portable Network Graphics (PNG) - the new format on the net. PNG is favorised by the W3 consortium. For some more special information visit `<http://www.w3.org/pub/WWW/TR/WD-png.html>` and `<http://www.w3.org/pub/WWW/Graphics/PNG/Overview.html>`. Here you'll find a technical specification, some programmers information etc. PNG is a ideal format replacing GIF. The PNG homepage is on `<http://quest.jpl.nasa.gov/PNG/>`. For the users, PNG will have some advantages and some disadvantages. Here they are:

For the advantages:

- can replace the license loaded GIF - PNG has no license problems
- 256 palette system as well as grayscale and true color capability including a transparency element
- complex interlace mode where not only different lines are sequenced but a two dimensional serialize system retrieves the picture resulting the user to realise the picture content more early.
- fast decompression algorythm is possible
- public available description - license free
- public available sample code - license free
- extensible design

For the disadvantages:

- not widely spread (Netscape does not support it by now, some plugins do)
- not so strong compressing pictures
- no final specification ready, in working draft state.

PNG is currently supported on Linux through the following programs: *ImageMagick* (Version ¿=3.7), *GhostScript 4.0*, *Gimp*, *PovRay 3.0*, the *netpbm* package. For *xv* 3.10a there exists an inofficial patch.

4.7.4 Converters

- *to be written* - *sorry* - netpbm, xv, ghostscript, gimp, ImageMagick, CorelDraw auf Wine :-)))

4.8 Specials

There are now many specials beyond the HTML'n'Image range. There are Applets written in Java and JavaScript pages and many things beyond.

4.8.1 Java

There is nothing to add about Java in general, just read the java section in the Netscape Navigator chapter of this HOWTO and the overview on Java Applett vs. CGI script in this HOWTO. Then, you can also read the really good and compact Linux JAVA HOWTO. For programming Java, please refer really good books on that.

4.8.2 ActiveX

ActiveX is at the time of writing still a Microsoft child. Microsoft claimed, that they would release it to the public domain or at least to release it to a ActiveX consortium.

ActiveX has nothing to do with the X Window system nor with XFree.

It is derived from the Microsoft and IBM OLE system. After releasing the specs, there should be a Unix port. But, we have to wait till then. Nothing for Linux, yet.

5 FAQ

There aren't any frequent asked questions - yet...

6 For further reading

- RFC1866 written by T. Berners-Lee and D. Connolly, "Hypertext Markup Language - 2.0", 11/03/1995
- RFC1867 writtenm by E. Nebel and L. Masinter, "Form-based File Upload in HTML", 11/07/1995
- RFC1942 written by D. Raggett, "HTML Tables", 05/15/1996
- RFC1945 by T. Berners-Lee, R. Fielding, H. Nielsen, "Hypertext Transfer Protocol – HTTP/1.0", 05/17/1996.
- RFC1630 by T. Berners-Lee, "Universal Resource Identifiers in WWW: A Unifying Syntax for the Expression of Names and Addresses of Objects on the Network as used in the World-Wide Web", 06/09/1994
- RFC1959 by T. Howes, M. Smith, "An LDAP URL Format", 06/19/1996

7 Thanks

Special thanks to Greg Hankins gregh@cc.gatech.edu for encuraging me to write this work and the fun I had doing it.

I'd also like to thank Chris Hendricks, Fido: 2:2433/443@fidonet.org for his engagement in Linux and my personal race to keep at least one nose ahead :-)