grep everything
Clever Search in Different Data Formats

Dipl.-Inf. Frank Hofmann

9 July 2012

Overview

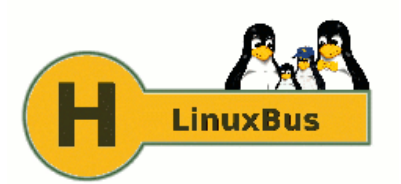
Frank Hofmann – Open Source Involvements and Projects

2000-2007

since 2006

since 2009

Regional LUG Meeting Berlin-Brandenburg since 2008



My Work

Linux, Layout & Satz

`http://www.efho.de`

- distribution of indoor and outdoor wireless devices
- pre-press preparation and print coordination

WIZARDS OF FOSS
Open Source Schulungen

`http://www.wizards-of-foss.de`

- open source training for experts
  co-founder and trainer

PARTNER VON: büro2.0
Open Source Bürogemeinschaft

`http://www.buero20.org`

- member of Büro 2.0 (Berlin)
  open source office community
  25 companies, 1300m^2, 60 members

grep Basics

- grep is based on
  g/re/p (**g**lobal / **r**egular **e**xpression / **p**rint)
- grep searches for according patterns in data streams

- patterns are Regular Expressions

- example:
  `grep Mikro[tT]ik invoice201100[1-5]`

  search for both patterns „Mikrotik“ and „MikroTik“ in all files named from `invoice2011001` to `invoice2011005`

zgrep and Friends

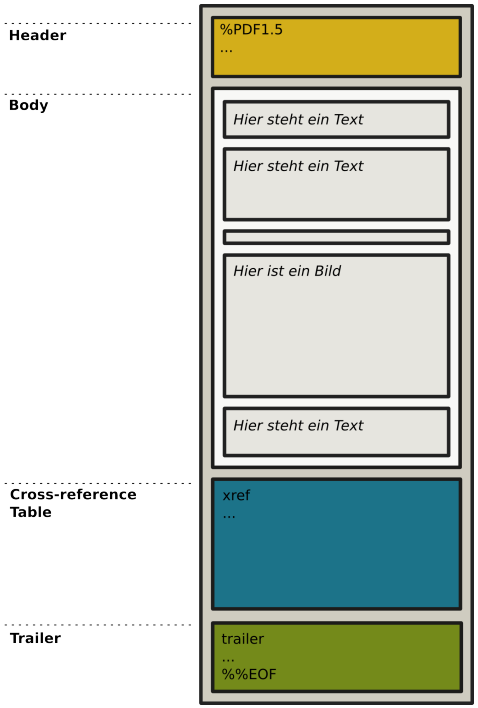
- `zcat file.gz | grep foo`

  can be shortened as

  `zgrep foo file.gz`

- is available for all grep variants: `zgrep`, `zfgrep`, `zegrep`
- is available several compression algorithms: `zgrep`, `bzgrep`, `xzgrep`
- ... and furthermore: `bzfgrep`, `bzegrep`, `xzfgrep`, `xzegrep`
- to make ist simple, use *zutils* instead: zgrep alternative automatically recognizes compression format, and works with gzip, bzip2, xz und lzip.

Combing through PDFs



- PDF is a data format with a fixed document structure
- `pdfgrep` package
  `http://pdfgrep.sourceforge.net/`
- example:
  `pdfgrep -i RouterBoard invoice*.pdf`

  search in files matching the file name pattern `invoice*.pdf` and the pattern „RouterBoard“ – no matter which spelling is used (option „-i“)

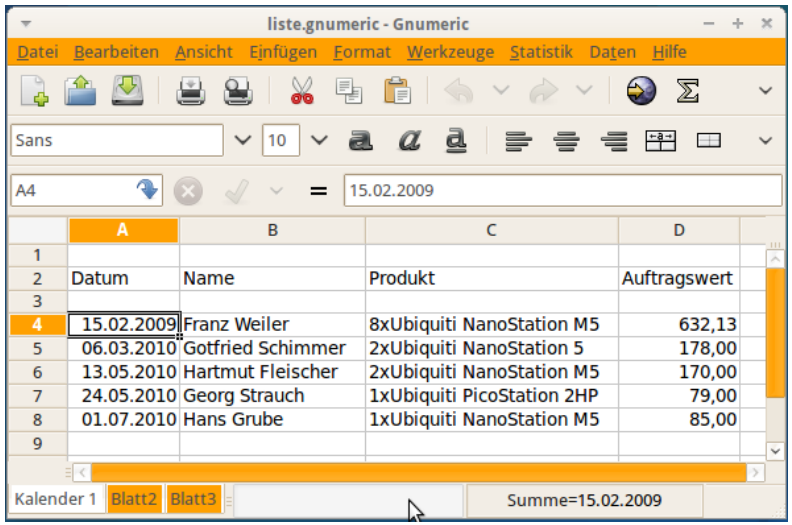
Searching in Audio Files

- `taggrepper` package
  `http://gitorious.org/taggrepper/pages/Home`
- search based on tags
  MP3: title, artist, album, year, genre, comment, title number, composer, original interpreter, copyright, and url
- taggrepper works with the formats MP3, Ogg Vorbis and FLAC

- example:
  `taggrepper --display-artist -a NDR *.mp3`

  select these MP3 files that have the author „NDR“; print both the filename und match

ssgrep



- ssgrep = spreadsheet grep
- Gnumeric file format is a gzip compressed XML
- ssgrep is part of the Gnumeric package

Example:

```
$ssgrep -Hn "Nano[Ss]tation" datei.gnumeric
datei.gnumeric:Blatt1!C4:8xUbiquiti NanoStation M5
$
```

search in the file `datei.gnumeric` according to the pattern „NanoStation“ (with S or s). „-Hn“ prints both the table cell, and the spreadsheet number

mboxgrep und grepmail

- searching the mail folder, in case of a match print the according mail
- support the mbox format; mboxgrep: also Maildir, MH, etc.
- output result is a valid mail folder as an mbox file
- mboxgrep can delete mails directly, that match (!)
- grepmail can limit the output: mail header or content, only
- grepmail can eliminate matches in signatures

xmlgrep

- xmlgrep: grep for search in xml data
- existing packages:
  Perl module XML::Twig, xmlstar, xml-coreutils, xmlclitools, XGrep, xml_grep2, sgrep

- locating an xml node in an xml tree: XPath
- example for Perl module XML::Twig, command xml_grep:

```
xml_grep --text_only "//collection/book/isbn" book.xml
```

  search for the subnode isbn in the file book.xml, and print the value of the node

Further Reading

- full grep overview – Axel Beckerts blog:
  `http://noone.org/blog/English/Computer/Shell/grep%20everything.futile`
- Frank Hofmann: PDF und PostScript durchsuchen, LinuxUser 2/2012
  `http://www.linux-community.de/Internal/Artikel/Print-Artikel/LinuxUser/2012/02/In-PDF-und-PS-Dateien-suchen`
- Axel Beckert, Frank Hofmann: Suche in komprimierten Dateien und Archiven, LinuxUser 4/2012
  `http://www.linux-community.de/Internal/Artikel/Print-Artikel/LinuxUser/2012/04/Suche-in-komprimierten-Dateien-und-Archiven`
- Axel Beckert, Frank Hofmann: grep in Anwendungsformaten (Teil 1), LinuxUser 6/2012
- Axel Beckert, Frank Hofmann: grep in Anwendungsformaten (Teil 2), LinuxUser 7/2012

Thanks for your attention!

Lassen Sie es setzen.

Contact:
Dipl.-Inf. Frank Hofmann
Hofmann EDV – Linux, Layout und Satz
c/o büro 2.0
Weigandufer 45 – 12059 Berlin
Email `frank.hofmann@efho.de`
web `http://www.efho.de`