Postscript- und PDF-Dokumente durchsuchen

Frank Hofmann

10. Juni 2012

Inhalt

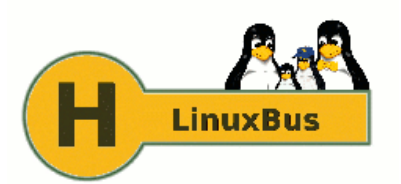
Frank Hofmann – OpenSource-Aktivitäten und Projekte

2000-2007

seit 2006

seit 2009

Regionales LUG-Treffen Berlin-Brandenburg seit 2008



Über Hofmann EDV – Linux, Layout und Satz

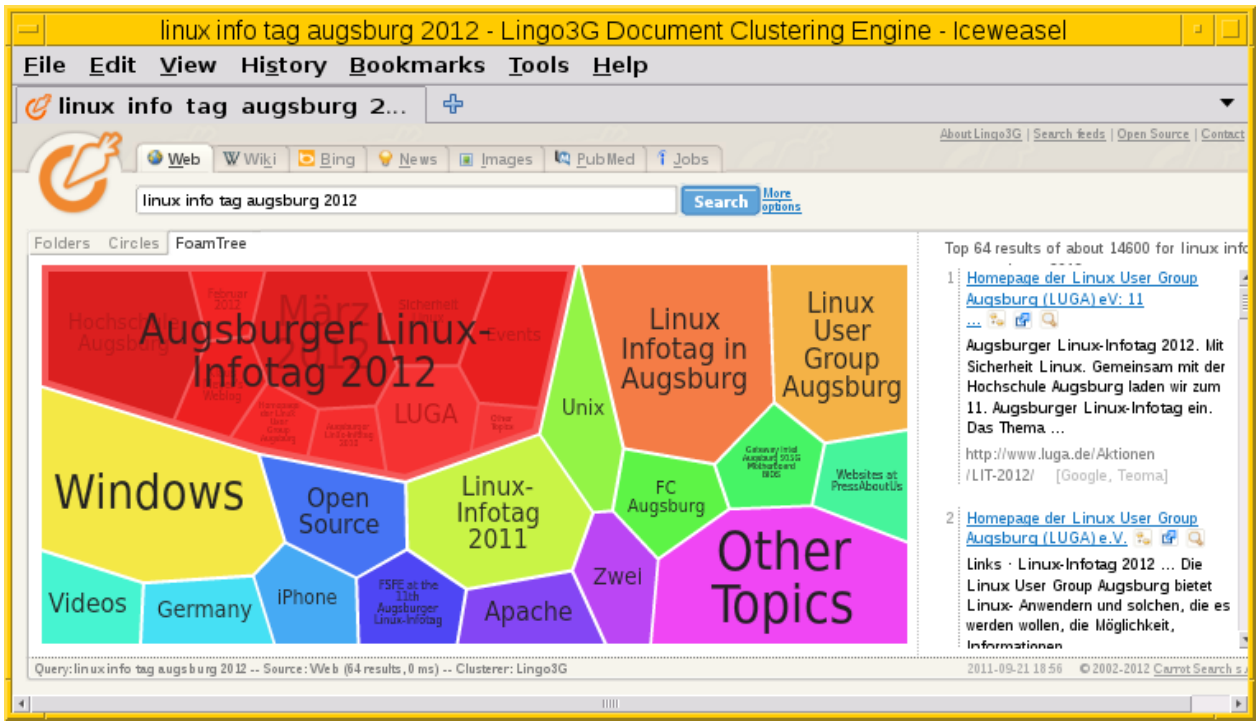
Linux, Layout & Satz

WIZARDS OF FOSS
Open Source Schulungen

- Layout und Satz, Druckvorstufe
- Administration und Service
  - Betreuung von Linux-Systemen (Debian)
  - Programmierung und Automatisierung auf der Basis von PHP und Python
  - Wireless Devices für den Innen- und Außeneinsatz
- Trainings für IT-Spezialisten

Suchen in Daten

- Werkzeug: Suchmaschine, Programm, Desktopsuche
- Darstellung der Suchergebnisse als Liste, Ringe oder Waben



Datenbasis

- Ziel: Dokumente und Daten wiederfinden und thematisch zuordnen
  - zweckdienliche Verzeichnishierarchie anlegen
  - sinnvolle Dateinamen vergeben
- Dokumente müssen durchsuchbar sein
  lesbar für uns und die Maschine (Programm)
- offene Dokumentation der Formate
- Text als Text im PDF einbinden, nicht als Bild ;-)
- Dokument klassifizieren
  Metainformationen setzen

Textseite – ohne Metainformationen

Sprüche

Vegetarische Gerichte schmecken besonders gut, wenn dazu ein kleines Schnitzel gereicht wird.

Sollten Sie einmal das Schnitzel nicht finden -- es liegt immer unter der Zitronenscheibe.

Milchreis schmeckt hervorragend, wenn man es vor dem Verzehr durch ein saftiges Steak ersetzt.

siehe: http://www.frag-mutti.de/

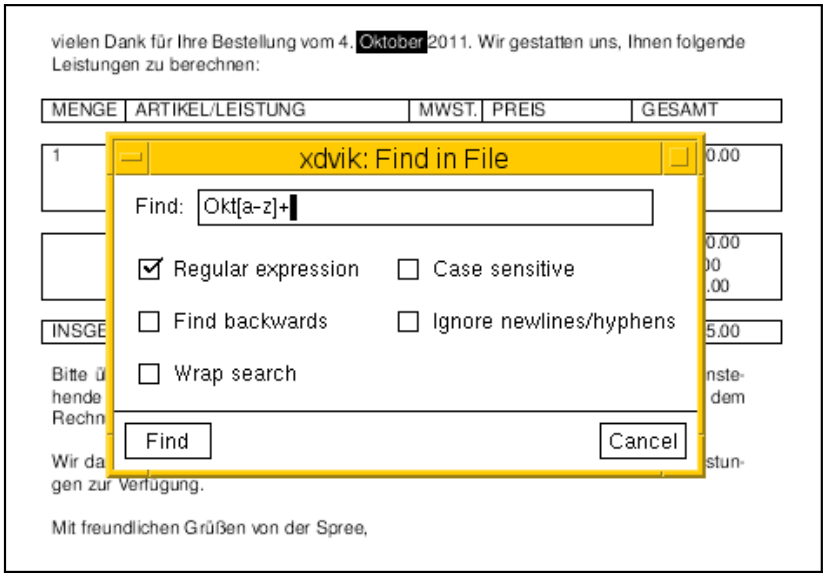
Metainformationen bei Docbook

Terminal - book.docbook + (~/projekte/li...pdf/wasserzeichen/docbook) - VIM

Datei Bearbeiten Anzeige Terminal Gehe zu Hilfe

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE book PUBLIC
  "-//OASIS//DTD DocBook XML V4.4//EN"
  "http://www.docbook.org/xml/4.4/docbookx.dtd">
  <book lang="de">
  <bookinfo>
        <title>Infrastruktur und Netzwerkkonfiguration</title>
        <subtitle>B&uuml;ronetzwerk</subtitle>
        <author>
                <firstname>Frank</firstname>
                <surname>Hofmann</surname>
        </author>
        <copyright>
                <year>2009</year>
                <holder>Hofmann EDV - Linux, Layout und Satz Berlin</holder>
        </copyright>
        <pubdate>2009</pubdate>
  </bookinfo>
  <preface><title>Einleitung</title>
-- EINFÜGEN --                                          17,15-22    Anfang
```

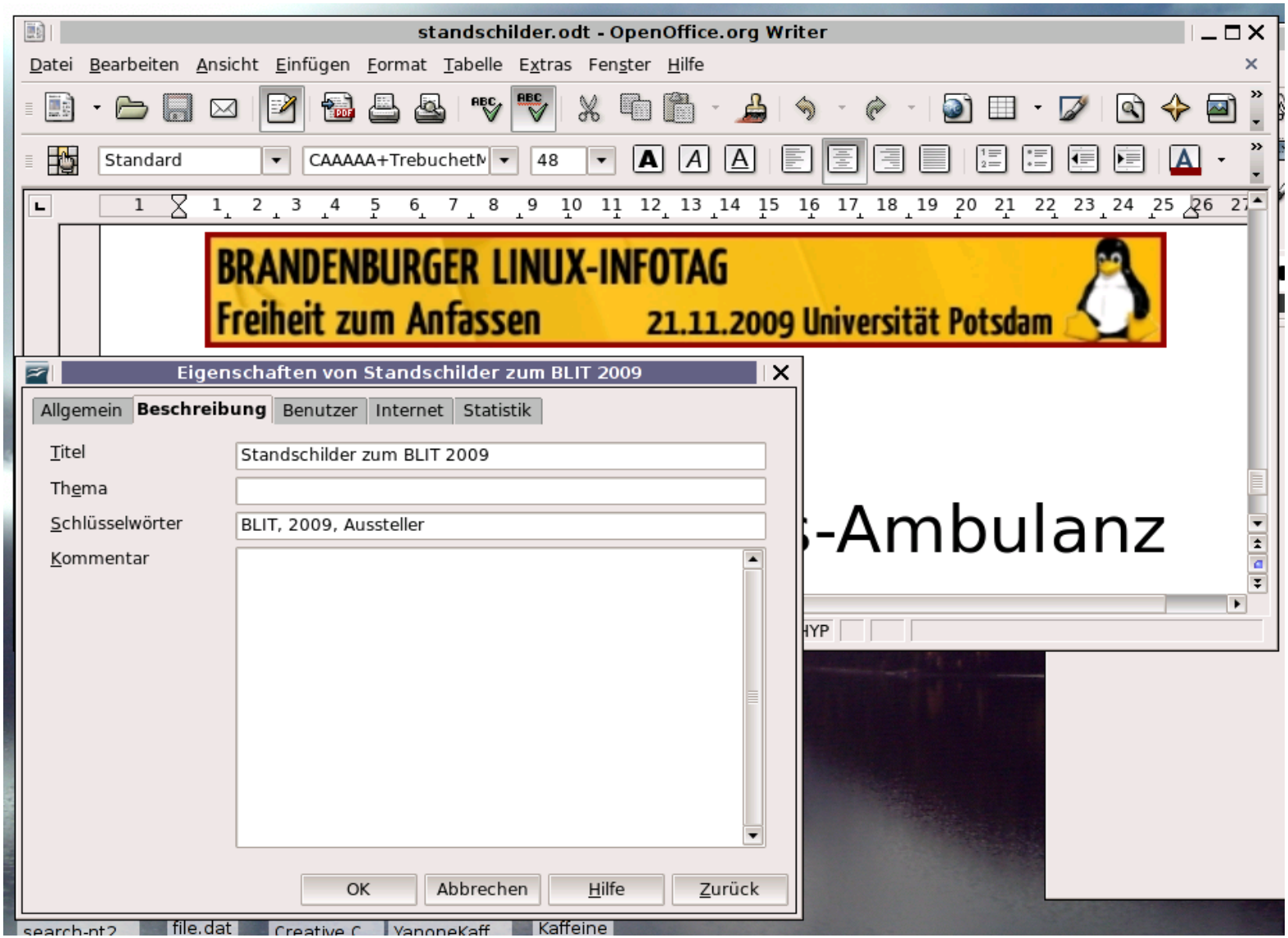
DVI-Dokumente



- DVI: geräteunabhängiges Dokumentenformat
- Suche in Xdvi und Okular
  - unabhängig von Groß- und Kleinschreibung
  - farbige Hervorhebung der Suchtreffer
  - Xdvi: unterstützt Reguläre Ausdrücke
- auf der Kommandozeile:

```
dvitype datei.dvi |
grep Muster
```

Metainformationen bei OpenOffice



Metainformationen im PDF

- LaTeX-Paket: hyperref

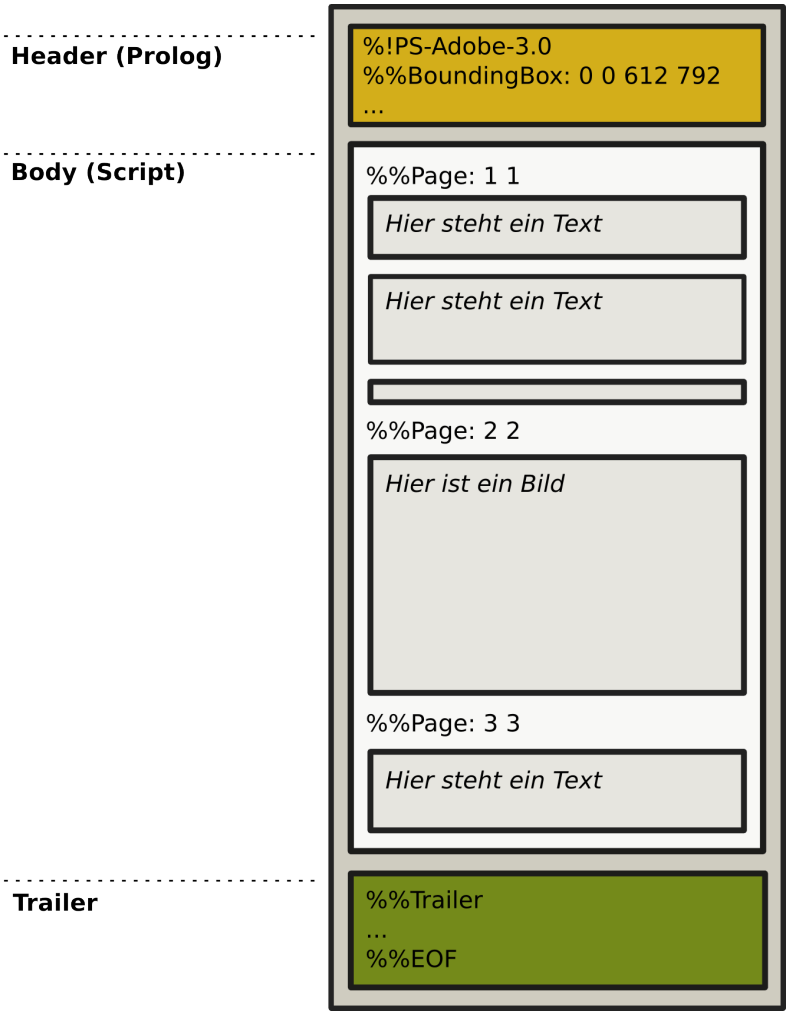
```
\usepackage{hyperref}
\hypersetup{
   pdfauthor={Frank Hofmann, Thomas Winde},
   pdftitle={Postscript- und PDF-Dokumente durchsuchen},
   pdfkeywords={PostScript; PDF; Suche},
   pdfsubject={Dokumentformate}
}
```

- Anzeigen der Metadaten mit UNIX-Kommando `pdfinfo`

  einfach: `pdfinfo` *`datei.pdf`*

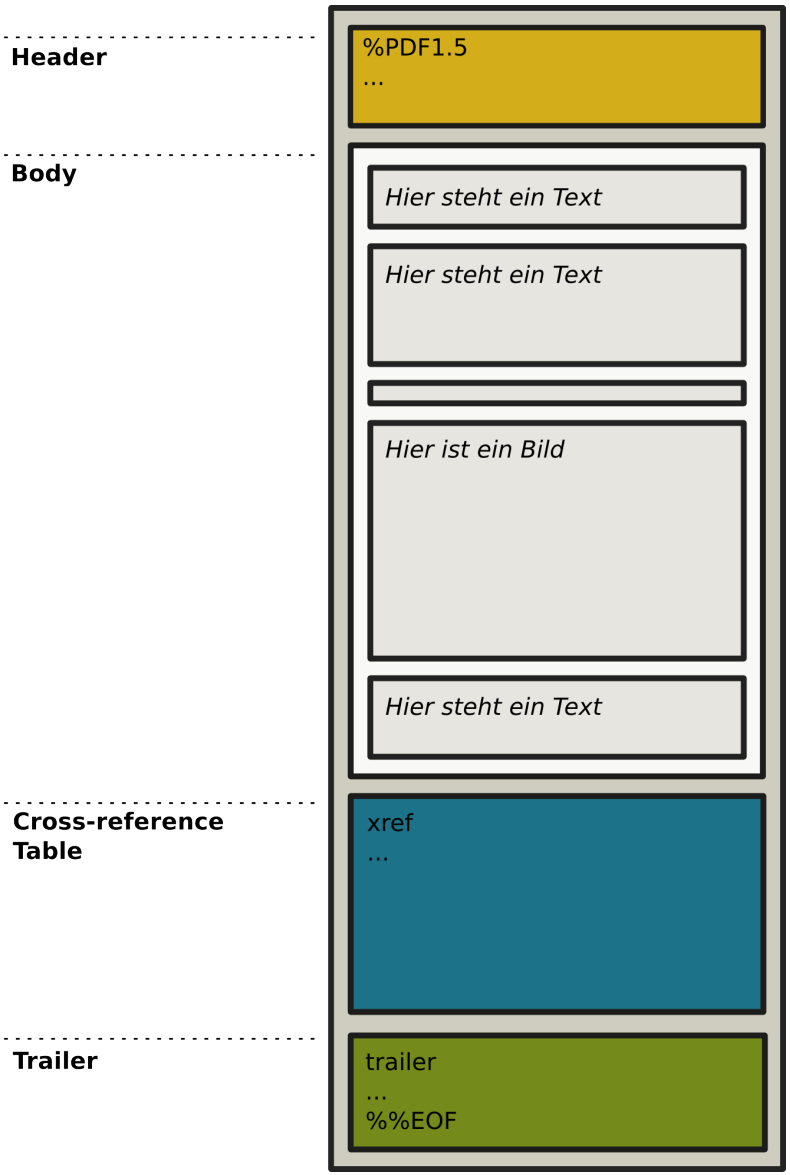
  ausführlich: `pdfinfo -box` *`datei.pdf`*

Das PostScript-Dokumentformat



- PostScript: stackbasierte Programmiersprache mit Variablendefinitionen, Prozeduren und Zuweisungen
- Prolog: Vorspann des Dokuments
- Script: Inhalt, Seitengröße, Grafik- und Textobjekte
- Trailer: Anhang

Das PDF-Dokumentformat



- PDF: Dokumentenformat von Adobe
- Header: Vorspann des Dokuments mit Metainformationen
- Body: Inhalt, Grafik- und Textobjekte mit Positionsangaben (genannt Object Stream)
- Cross-reference Table: Inhaltsverzeichnis der einzelnen Objekte
- Trailer: Anhang

PostScript-Dokumente

- Dokumentbetrachter: Ghostview, Kghostview, Evince und Okular bei unseren Tests funktionierte die Suchfunktion nicht

- auf der Kommandozeile:

  `pstotext` *`datei.ps`* `|` `grep` *`Muster`*
  `ps2ascii` *`datei.ps`* `|` `grep` *`Muster`*

  Nachteil: nur zuverlässig für Encoding ISO 8859-1 (Latin-1)

  Alternative über die Konvertierung nach PDF:
  `ps2pdf` *`datei.ps`*`;` `pdftotext` *`datei.pdf`* `|` `grep` *`Muster`*

  `ps2pdf` *`datei.ps`*`;` `pdfgrep` *`datei.pdf Muster`*

PDF-Betrachter

- Dokumentbetrachter:
  - Ghostview: keine Suche
  - Epdfview, Evince, Okular, Xpdf
    Suche über Button, „/“ (außer Xpdf) und Ctrl+F
  - Apvlv, Mupdf, Zathura
    Suche analog zu vi(m)-Tastenschema
- Suche im Dokument beginnt bereits während der Eingabe des Musters, Cursor springt zum nächsten Suchtreffer

Suche automatisieren – Variante 1

- `pdftotext` und `grep` mit einer Pipe in einer `for`-Schleife:

  `for datei in $(ls *.pdf); do pdftotext $datei - | grep --color` *`Muster`*`; done`

- pdftotext kommt problemlos mit den Encodings ISO 8859-1, 8859-15 und UTF-8 zurecht
- explizite Auswahl des Encodings über die Option `-enc` *`Encoding`*

Suche automatisieren – Variante 2

```
$pdfgrep -in -C 15 "RouterBoard RB.5" E*.pdf
E_20110027.pdf:2: Routerboard RB450G Level 5
E_20110027.pdf:2: Mikrotik Routerboard RB450G
E_20110066.pdf:2: MikroTik Routerboard RB450G (ID
E_20110111.pdf:2: MikroTik RouterBoard RB450G S/N
E_20110111.pdf:2: RouterBoard RB450G Wattac
E_20110124.pdf:2: MikroTik RouterBoard RB450G S/N
E_20110124.pdf:2: RouterBoard RB450G Wattac
E_20110182.pdf:1: 1 MikroTik RouterBoard RB250GS
E_20110185.pdf:2: 1 MikroTik RouterBoard RB250GS
$
```

- ... mit `pdfgrep`:

  `pdfgrep` *`Muster Dateiliste`*

- Option `-i`: unabhängig von Groß- und Kleinschreibung
- Option `-n`: Seite, auf der der Suchtreffer gefunden wurde
- Option `-C` *`Anzahl`*: Ausgabe max. *Anzahl* Zeichen

Vielen Dank

Danke für Eure Aufmerksamkeit :-)

Kontakt:
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 www.efho.de