# Recoll/Xapian - effiziente Dokumenten- / Desktopsuche

Michael Schwipps

26. Januar 2016

# Übersicht

- Recoll ist ein Google für zu hause
- Motivation / Warum?
- Features
- Erweiterungsmöglichkeiten und Grenzen

# Motivation, Eigenschaften klassischer Unix-Tools

- Unix-Tool: Xgrep, ctags, cscope
- zuverlässig, stabil, sicher, schnell – aber
- nur textbasierte Dateiformate, keine Binärformate
- RegEx, boolesche Verknüpfung über etwas Shell-Magie
- sehr gute Vim-Integration

# Recoll

- Unterstützt direkt die üblichen Dokumenten- und Containerformate
- Container „multilevel“, z.B. tgz in Email
- boolesche Ausdrücke in der üblichen Suchmaschinen-Notation, Wildcard*
- die Textextraktion erfolgt mit Linux-Standardtools(z.B. pdftotext)
- leicht erweiterbar
- fuzzy-Suche, mehrsprachiges Stemming während der Suche (z.B. deutsch und englisch), Stammformbildung: fliegen, fliege, flog
- Priorisierung/Ranking
- aspell-basiertes „Meinten Sie“-Feature / Anti-Tippfehlervorschlag

# Recoll 2, Xapian

- gibt's fertig als CLI-Tool und X-Programm
- Integration in Web-Tools (z.B. MediaWiki, redmine) häufig leicht durch fertig Plugins möglich
- Datenbank-Indizierung via Sprachintegration python, php
- Indizierungstrigger erfolgt über expliziten Aufruf (z.B. cronjob) oder via FAM/inotify

- Xapian ist das Speicher-Backend (Search Engine Library)
- in C++ geschrieben und mit noch mehr Sprachbindung für Perl, Python, PHP, Java, Tcl, C#, Ruby, Lua, Erlang and Node.js
- Indizierung via Recoll und Suche via Xapian möglich

# Erweiterbarkeit bei der Textextraktion

- Beispiel: vermailte Pdf-Dateien aus einem Kopierer indizieren
- Besonderheit dabei: der Textinhalt steht nicht unmittelbar in der Datei
- OCR/Tesseract-Integration (Idee/Bug/Issue von mir)
- Anhänge kann kein (mir) bekannte MUA durchsuchen, der IMAP-Standard kann das imho auch nicht
- Container mbox / Email auf

# Integration in mutt

- Voraussetzung: Speicherung der Email in Maildir
- Indizierung erfolgt wie üblich
- Suche über ein Shellskript das eine Maildir mit gesymlinkten Sucherergebnissen generiert

# Querverweise auf vergleichbare Tools

- Solr (fett, nur Web-Interface, Erweiterbarkeit nur mit java?), hat dafür noch weitere fuzzy-Methoden (z.B. Levenshtein) und ist clusterfähig, skaliert besser

# Fragen und Quellen

- Fragen?
- `http://www.lesbonscomptes.com/recoll`
- `http://xapian.org`