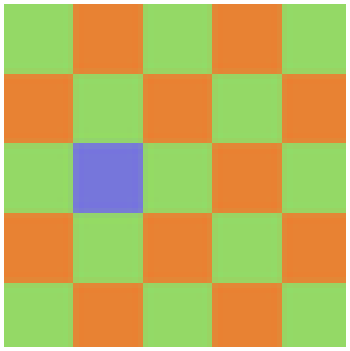
Kurz erklärt

Volltextsuche in Dateien unter GNU/Linux

Inklusive eingescannter PDF-Dateien

Günther Zander
29. Dezember 2020

Recoll



www.lug-hamburg.de

Finden ist besser als suchen

Mit der Zeit liegen zahlreiche unterschiedliche Dateien auf einem Speichermedium. Selbst bei einer sorgfältigen Organisation muss oft nach dem Gewünschten aufwendig gesucht werden. Also muss ein Programm her, dass auch für Anfänger geeignet ist und über eine grafische Oberfläche verfügt.

Das Programm *Recoll* berücksichtigt die Inhalte von Textdokumenten sowie Metadaten von Audio-, und Bilddateien genauso wie das Mailbox-Format von Thunderbird oder Evolution. Dafür erstellt dieses Programm einen eigenen Index über alle Dateien.

Eine Herausforderung ergibt sich bei den selbst eingescannte PDF-Dateien, da diese von der Qualität des Originals und dem Scan selbst abhängig ist. Diese werden beim scannen als Bilddateien abgespeichert und müssen für den Index mittels eines OCR-Programms erst einmal lesbar gemacht werden. Das dafür eingesetzte OCR-Programm *tesser* liefert bereits sehr gute Ergebnisse.

Wenn Sie es möchten, kann das Programm auch ihre epub-Dateien durchsuchbar gestalten. Falls dieses in Betracht kommt, sollten multiple Indexe verwendet werden. Auch ist es möglich den Index über ein Netzwerk zugänglich zu machen.

Ihre Konfiguration wird dabei im Verzeichnis $HOME/.recoll abgelegt. Dieses Verzeichnis kann jederzeit gelöscht werden. Nach einem erneuten Aufruf des Programms starten sie wieder mit der Grundkonfiguration und müssen den Index neu aufbauen.

Auf einem Laptop mit i5-CPU und einer SSD mit 113GB an Daten hat der Aufbau des ersten Index ca. 25 Minuten gedauert. Dabei hat der Index eine Größe von ca. 3GB. Die Aktualisierung des Indexes kann entweder manuell oder zeitgesteuert erfolgen und geht um einiges schneller. Die beschrieben Programme für Recoll Version 1.26.3 wurde unter Kubuntu 20.4 getestet.

In dem hier aufgeführtem Beispiel müssen sie die Aktualität der erzeugten Datenbank manuell sicherstellen. In dem Artikel „systemd als Timer einsetzen“ können sie dieses auch als wiederkehrende Aufgabe ihrem Computer überlassen.

Inhaltsverzeichnis

Installation der Programme 2
Konfiguration 3

Installation der Programme

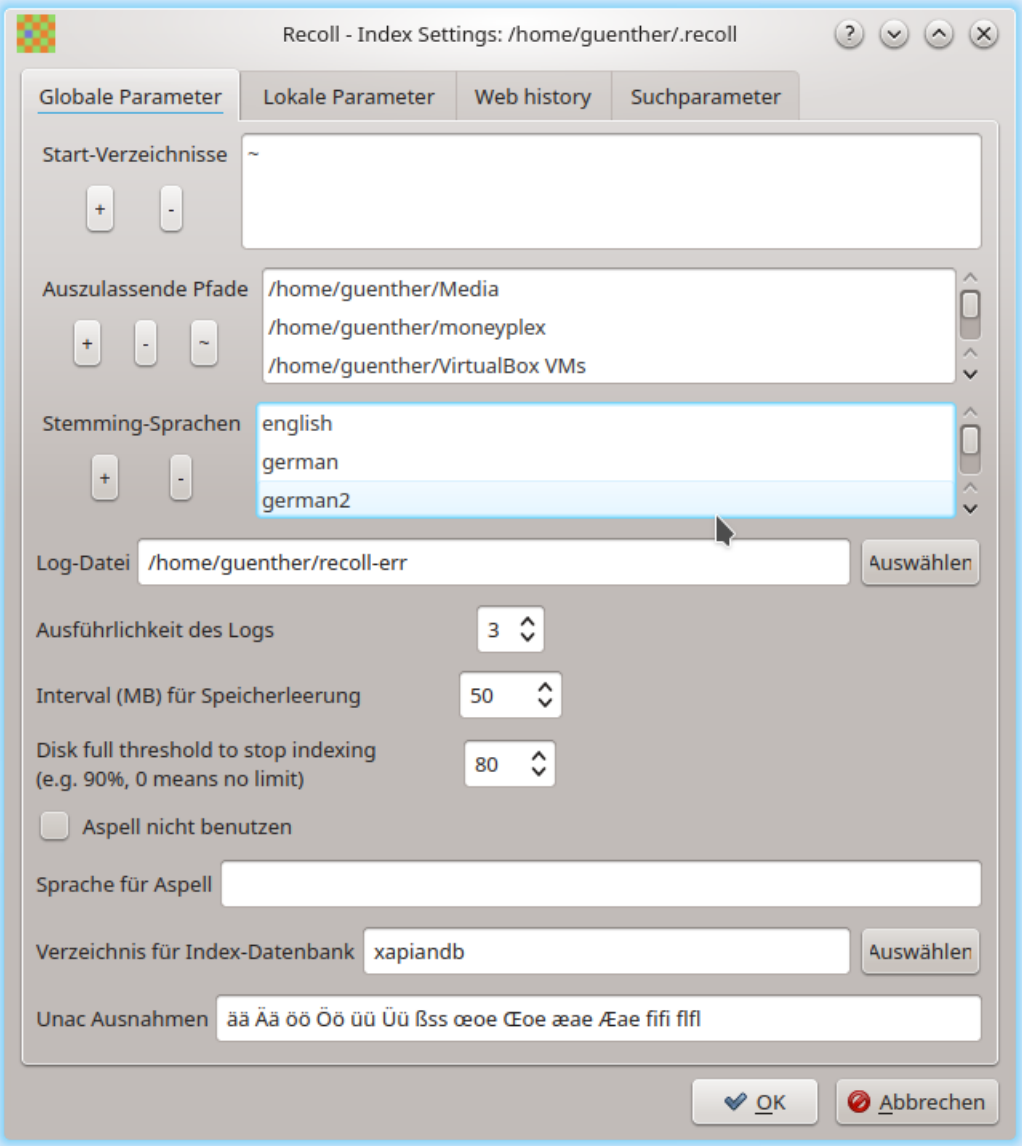
Um mit dem Programm vernünftig arbeiten zu können, müssen zuerst die dafür notwendigen Programm-Pakete installiert werden. Aus den Paketquellen lassen sich die Pakete *recoll, antiword, untex, python3-mutagen, catdvi, pff-tools, exif, unrtf, aspell, aspell-de, tesseract-ocr-deu, poppler-utils, evince* direkt installieren.

Das Packprogramm *lzma* muss über einen anderen Weg installiert werden. Zuerst muss mit dem Paketmanager das Paket *python3-pip* installiert werden, um im Anschluss mit root-Rechten in einem Terminal das Programm mit *"pip3 install pylzma"* als letzte Komponente zu installieren. Alternative können Sie auch das nachfolgende Script benutzen.

```
#! /bin/bash

# Install Recoll
# --------------
# Author:  Günther Zander, Hamburg 12/2020
# Licence: GPL
# File:    install-recoll.sh
# Version: 0.1
#
# Check root-Right
# ----------------
if [ "`id -gn`" != "root" ]; then
    echo; echo "Script bitte mit root-Rechten starten."; echo; exit 1
fi

# Programlist
# -----------

PRG=" recoll antiword untex python3-mutagen catdvi pff-tools"
PRG=$PRG" unrtf aspell aspell-de tesseract-ocr-deu poppler-utils"
PRG=$PRG" evince python3-pip exif"

PyPRG="pylzma"

# Execute
# -------

apt-get -y install $PRG
pip3 install $PyPRG

# pip3 install epub
```

Konfiguration

Bevor sie den ersten Index erzeugen, sollten Sie einige Änderungen vornehmen. Das Menü erreichen Sie über den Pfad *Einstellungen/Index-Einstellungen.*



Zu den Stemming-Sprachen müssen noch *german* und *german2* hinzugefügt werden um die Wörter später auch in Deutsch suchen zu können. Bei den *Auszulassende Pfade* sollten Sie überlegen welche in ihrem Umfeld sinnvoll sind. In dem Beispiel ist das der Link zum Verzeichnis /media sowie zu den Images von VirtualBox und zu den Daten der Bankingsoftware Moneyplex. Unter *Log-Datei* können Sie einfrei verfügbares Verzeichnis und Dateinamen für die LOG-Datei angeben.