

Zeitschriften Digitalisieren

Motivation

- Viele Jahre Gartenbahn-Zeitschrift
- Inhalt:
 - jede Menge Tips, Produktberichte, Bastelanleitungen
 - Anlagenvorstellungen liefern Ideen
 - Großteil des Inhalts ist zeitlos
- Warum Digitalisierung?
 - Papier braucht Platz
 - Artikelsuche ist mangels Index sehr aufwändig

Idee

- Lösung:
 - Zeitschriften einscannen als PDF
 - Texterkennung um PDF durchsuchbar zu machen
 - Tool um Suche zu ermöglichen
- Probleme:
 - Seitenweises Einscannen dauert zu lange
 - $68 \text{ Seiten} \times 6 \text{ Ausgaben/Jahr} \times \text{ca. } 15 \text{ Jahrgänge} = \text{über } 6000 \text{ Seiten}$
 - OCR-Software muss ohne manuellen Eingriff gute Erkennung haben
 - Erkannter Text sollte möglichst passgenau über dem Hintergrund liegen
 - Index-Erstellung muss automatisch erfolgen

Vorbereitung (1)

- Zeitschrift-Rücken absägen (Kreissäge)
 - Holz oben und unten reduziert Fetzen
 - Stapel eher dicker als dünner
 - Fest anpressen und eher langsam Durchschieben
 - Angekokeltes Papier macht nichts

Vorbereitung (2)



Vorbereitung (3)



Vorbereitung (4)

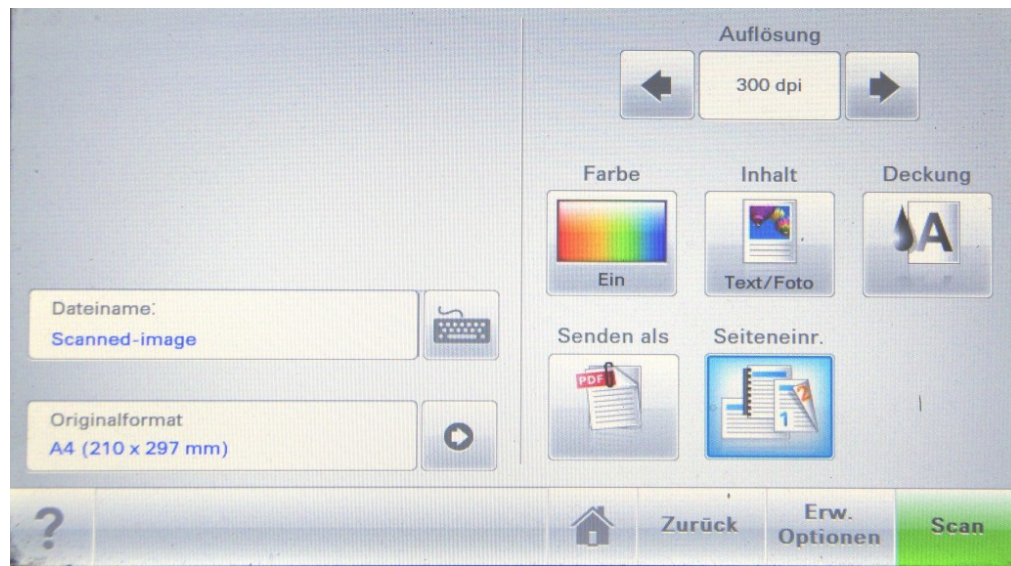


Vorbereitung (5)

- Schnittkante glatt machen
 - Abstehende Fetzen können verhakt sein
- Ordentlich Aufschlagen
 - vom Staub befreien
- Von der Schnittkante her einzeln aufblättern
 - Löst Verhakungen
 - Fetzen ggf. abreißen
- Transportrollen (Gummiwalzen) des Scanners regelmäßig mit Spiritus reinigen
 - spätestens wenn es Einzugprobleme gibt

Scannen (1)

- Deckblatt ist zu dick:
 - Bei vorderem Deckblatt muss dem Einzug oft etwas geholfen werden
 - Hinteres Deckblatt wird entfernt (enthält nur Werbung)
- Scanner: Druck/Scan/FAX-Kombi Lexmark CX510de
 - "Scan to USB"
 - 300dpi, Farbe
 - beidseitig
 - Stapelinzug



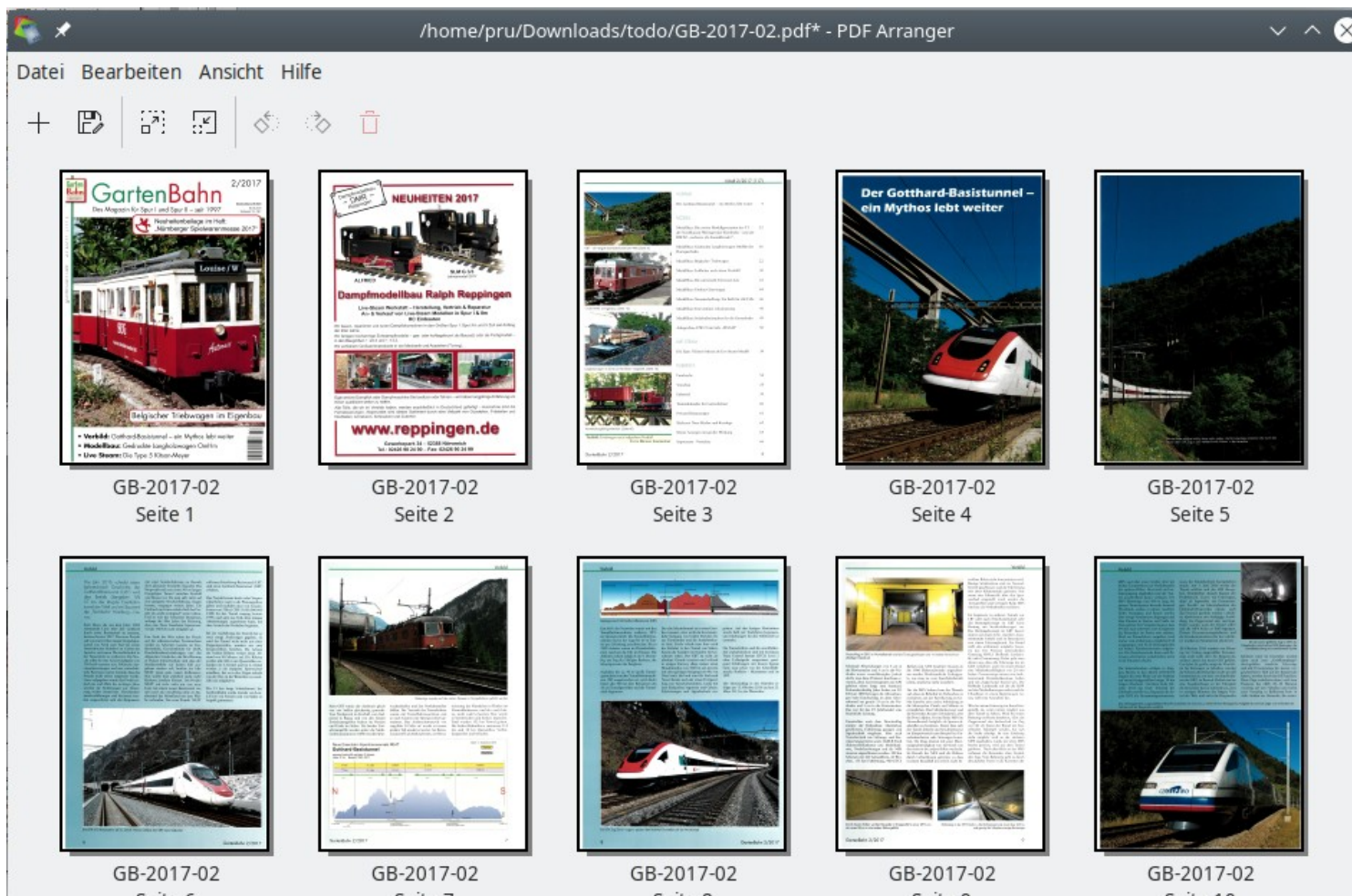
Scannen (2)



PDF zusammenbasteln

- Bei Einzugsproblemen müssen mehrere PDFs aneinandergeschoben werden
- Manchmal werden 2 Seiten auf einmal durchgezogen, dann muss eine Seite eingefügt werden
- Tools, siehe <https://wiki.ubuntuusers.de/PDF/>
 - PDFsam
 - PDF Arranger
 - Kommandozeile:
 - `pdfunite <datei1> <datei2> ... <ausgabedatei>`

PDF Arranger



OCR

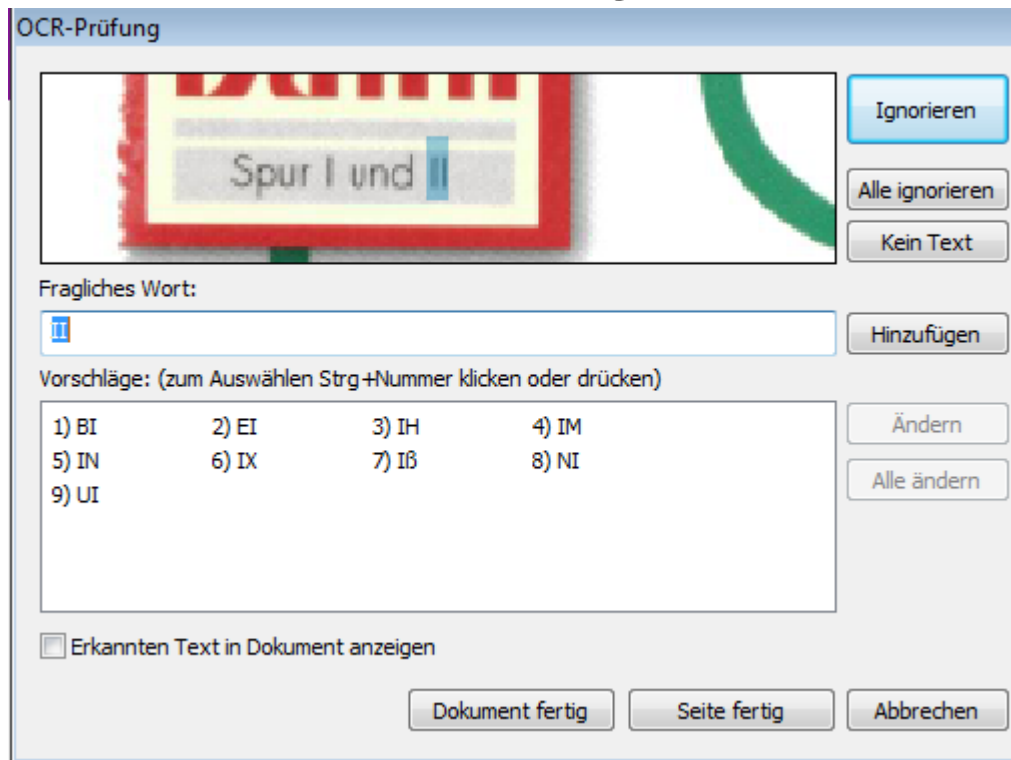
- Linux: ocrMyPDF
 - Gutes Scanergebnis, aber schlechte Positionierung
- Nuance Power PDF (nur Windows)
 - Test bei Ingo mit Version Standard 1.1
 - Gekauft: Version Advanced 2.1 für 17,99 EUR
- Vorgehen
 - PDF öffnen
 - Button "PDF durchsuchbar machen"
 - Datei speichern

Screenshot OCR

The screenshot shows the Nuance Power PDF Advanced interface. The main document is a magazine page titled "Garten Bahn" with the subtitle "Das Magazin für den Gartenliebhaber seit 1997". The date "6/2020" is visible in the top right corner. The page features a large photograph of a garden with various plants and flowers. A conversion dialog box is open in the center, displaying the Nuance logo and the text: "Die Auswahl der Dokumentsprache steigert die Genauigkeit bei der Formularkonvertierung." Below this text is a progress bar showing "35% - 11 Seite(n) / 66 Seite(n)" and a button labeled "Konvertierung abbrechen". The software's menu bar includes options like "Datei", "Start", "Bearbeiten", "Ansicht", "Kommentieren", "Erweiterte Verarbeitung", "Sicherheit", "Formulare", and "Hilfe". The toolbar contains icons for various functions such as "Einfügen", "Extrahieren", "Löschen", "Dateien kombinieren", "Aus Datei", "Vom Scanner", "Rechts", "Links", "Erweitert", "Word", "Excel", "PowerPoint", "Andere", "PDF durchsuchbar machen", "Objekt bearbeiten", "Schreibmaschine", "Verkleinern", "Teilen", "Suchen", and "Suche".

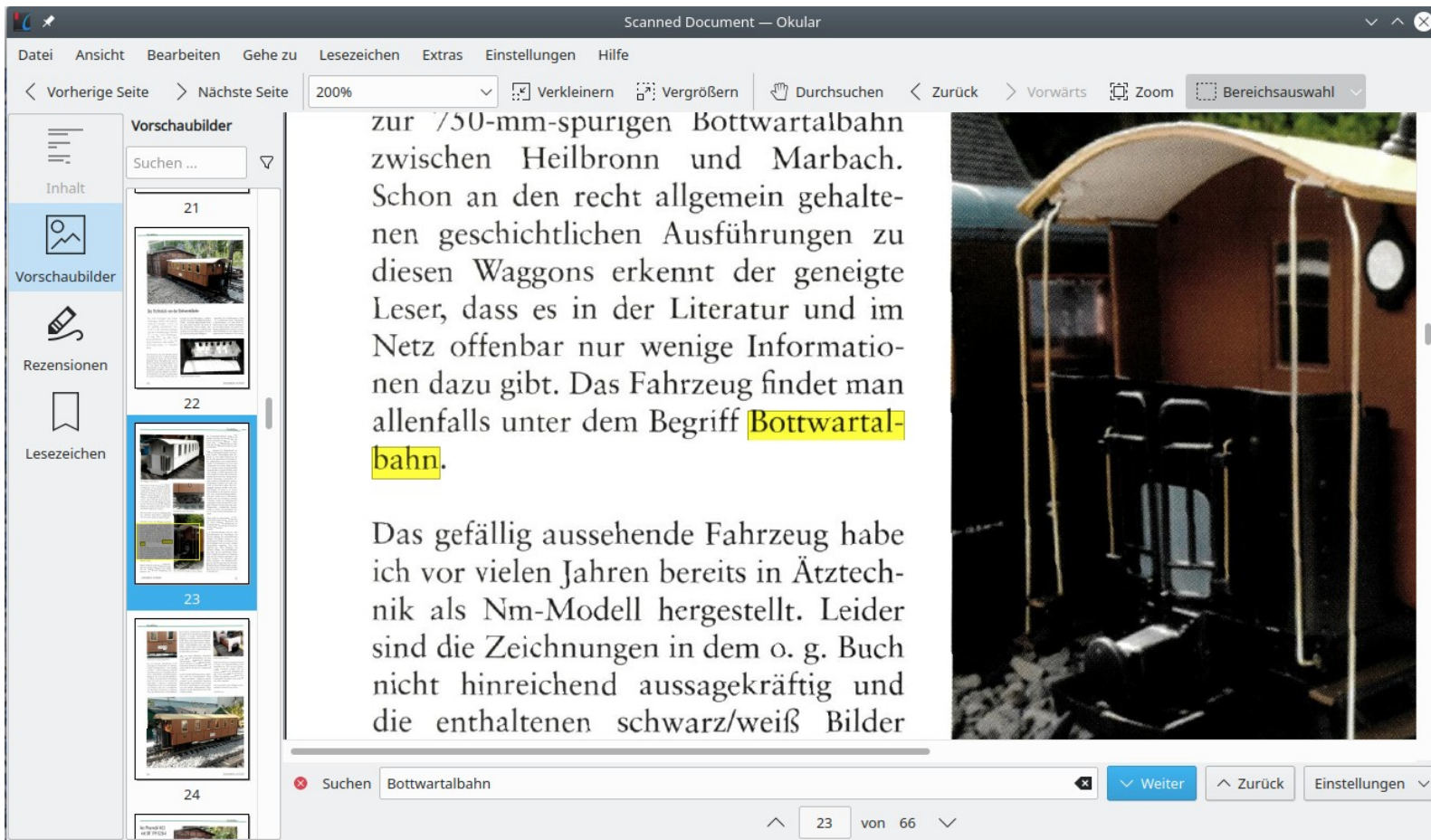
OCR Abschluss

- Zum Schluss kann noch Rechtschreibprüfung laufen
 - ich verzichte darauf so genau brauche ich es nicht



Ergebnis im PDF-Viewer "Okular"

- Der Adobe Akrobat-Reader ist übrigens zu blöd das getrennte Wort zu finden:



The screenshot shows the Okular PDF viewer interface. The main window displays a scanned document page with the following text:

zur 750-mm-spurigen Bottwartalbahn zwischen Heilbronn und Marbach. Schon an den recht allgemein gehaltenen geschichtlichen Ausführungen zu diesen Waggons erkennt der geneigte Leser, dass es in der Literatur und im Netz offenbar nur wenige Informationen dazu gibt. Das Fahrzeug findet man allenfalls unter dem Begriff **Bottwartalbahn**.

Das gefällig aussehende Fahrzeug habe ich vor vielen Jahren bereits in Ätztechnik als Nm-Modell hergestellt. Leider sind die Zeichnungen in dem o. g. Buch nicht hinreichend aussagekräftig und die enthaltenen schwarz/weiß Bilder

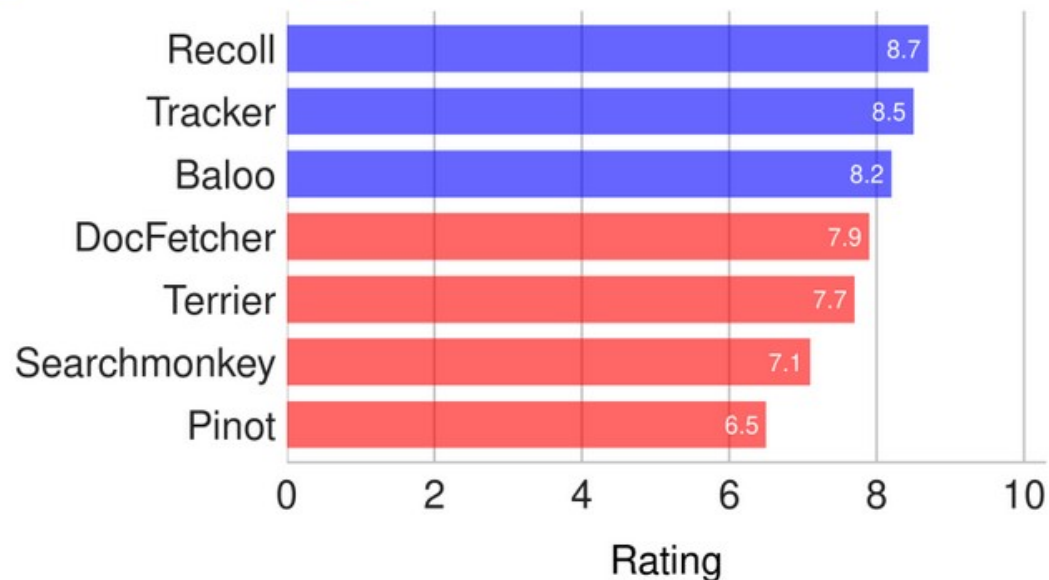
To the right of the text is a photograph of a model train car, which is a narrow-gauge passenger car with a yellow roof and black body.

The search bar at the bottom of the viewer shows the search term "Bottwartalbahn" and the results are displayed on page 23 of 66.

- Vergleich Desktop Search Engines
 - <https://www.linuxlinks.com/desktopsearchengines/>
- Meine Wahl fiel auf Recoll

Free Desktop Search Engines

■ Recommended ■ Good



Recoll

The screenshot shows the Recoll application window with a search for 'gartenbahnanlage'. The search results list several PDF documents. A preview window is open for the first result, 'GB-2017-04.pdf', showing a snippet of text from page 21.

Suchergebnisse Dokumente 1-5 für [Suchanfrage zeigen](#)

Vorschau Öffnen Schnipsel 152 KB / 83 MB **GB-2017-04.pdf**
application/pdf 2022-01-06 21:08:56 +0100 *file:///home/prue/eBooks/Gartenbahn/GB-2017-04.pdf*
 [p 21] Kraft auf meiner steigungsreichen [Gartenbahnanlage](#) nach einigen Überlegungen verworfen... [p 41] mehr Eindrücke von meiner [Gartenbahnanlage](#) verschaffen möchte, sei auf... [p 42] keinen Raum für eine [Gartenbahnanlage](#). Aber wo ein Wille... [p 43] Einzug in

Vorschau Öffnen Schnipsel 138 KB
application/pdf 2022-04-09 12:43:14
 [p 60] TT, Spur 0-Anlagen, [Gartenbah](#)

Vorschau Öffnen Schnipsel 151 KB
application/pdf 2022-01-06 21:17:38
 [p 1] Jwd Anlagenporträt: Spur 0-Gar

Vorschau Öffnen Schnipsel 127 KB
application/pdf 2022-01-07 21:03:20
 3/2017 ear hs rten Garten | g is CY,Bs

Vorschau Öffnen Schnipsel 134 KB
application/pdf 2022-01-25 21:07:20
 GartenBahn Das Magazin für Spur I u
 _c ci b C3) „yr. 4.1 as•MI•mic - Neuhei

Recoll - Snippets : GB-2017-04.pdf

Nieten forderlichen Kraft auf meiner steigungsreichen
P. 21 [Gartenbahnanlage](#) nach einigen Überlegungen verworfen. Stattdessen habe
P. 41 Wer sich mehr Eindrücke von meiner [Gartenbahnanlage](#) verschaffen möchte, sei auf YOUTUBE verwiesen
P. 42 Kindern eigentlich keinen Raum für eine [Gartenbahnanlage](#). Aber wo ein Wille ist, ist
P. 43 Detail ihren Einzug in diese kleine [Gartenbahnanlage](#), auf der die Insel thematisch eng

Suchen Schließen

Recoll: Einstellungen

- Beschränken auf den Ordner mit den Zeitschriften-PDFs

Recoll - Index Settings: /home/pru/recoll

Globale Parameter Lokale Parameter Web history Suchparameter

Start-Verzeichnisse /home/pru/eBooks/Gartenbahn

Auszulassende Pfade /media

Stemming-Sprachen german
german2

Log-Datei stderr Auswählen

Ausführlichkeit des Logs 3

Interval (MB) für Speicherleerung 50

Disk full threshold to stop indexing (e.g. 90%, 0 means no limit) 0

Aspell nicht benutzen

Sprache für Aspell

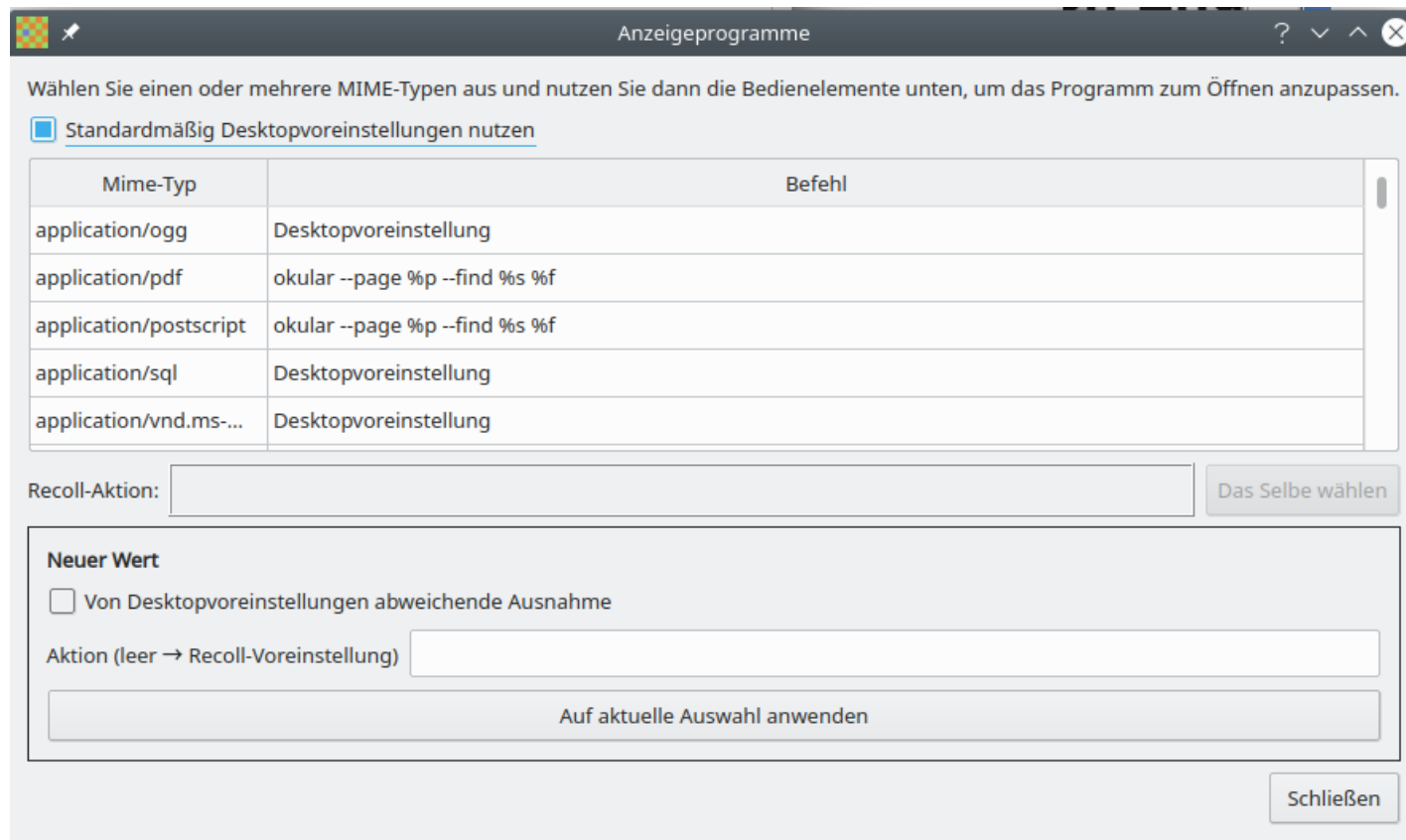
Verzeichnis für Index-Datenbank /home/pru/eBooks/Gartenbahn Auswählen

Unac Ausnahmen ää Ää öö Öö üü Üü ßss œoe Œoe æae Æae fifi fifl

✓ OK ⊗ Abbrechen

Recoll: Einstellungen PDF-Reader

- Parameter für Okular:



Wählen Sie einen oder mehrere MIME-Typen aus und nutzen Sie dann die Bedienelemente unten, um das Programm zum Öffnen anzupassen.

Standardmäßig Desktopvoreinstellungen nutzen

Mime-Typ	Befehl
application/ogg	Desktopvoreinstellung
application/pdf	okular --page %p --find %s %f
application/postscript	okular --page %p --find %s %f
application/sql	Desktopvoreinstellung
application/vnd.ms-...	Desktopvoreinstellung

Recoll-Aktion: Das Selbe wählen

Neuer Wert

Von Desktopvoreinstellungen abweichende Ausnahme

Aktion (leer → Recoll-Voreinstellung)

Auf aktuelle Auswahl anwenden

Schließen