Zeitschriften Digitalisieren

PC-Treff-BB, Peter Rudolph

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 Seiten x 6 Ausgaben/Jahr x ca. 15 Jahrgänge = über 6000 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)

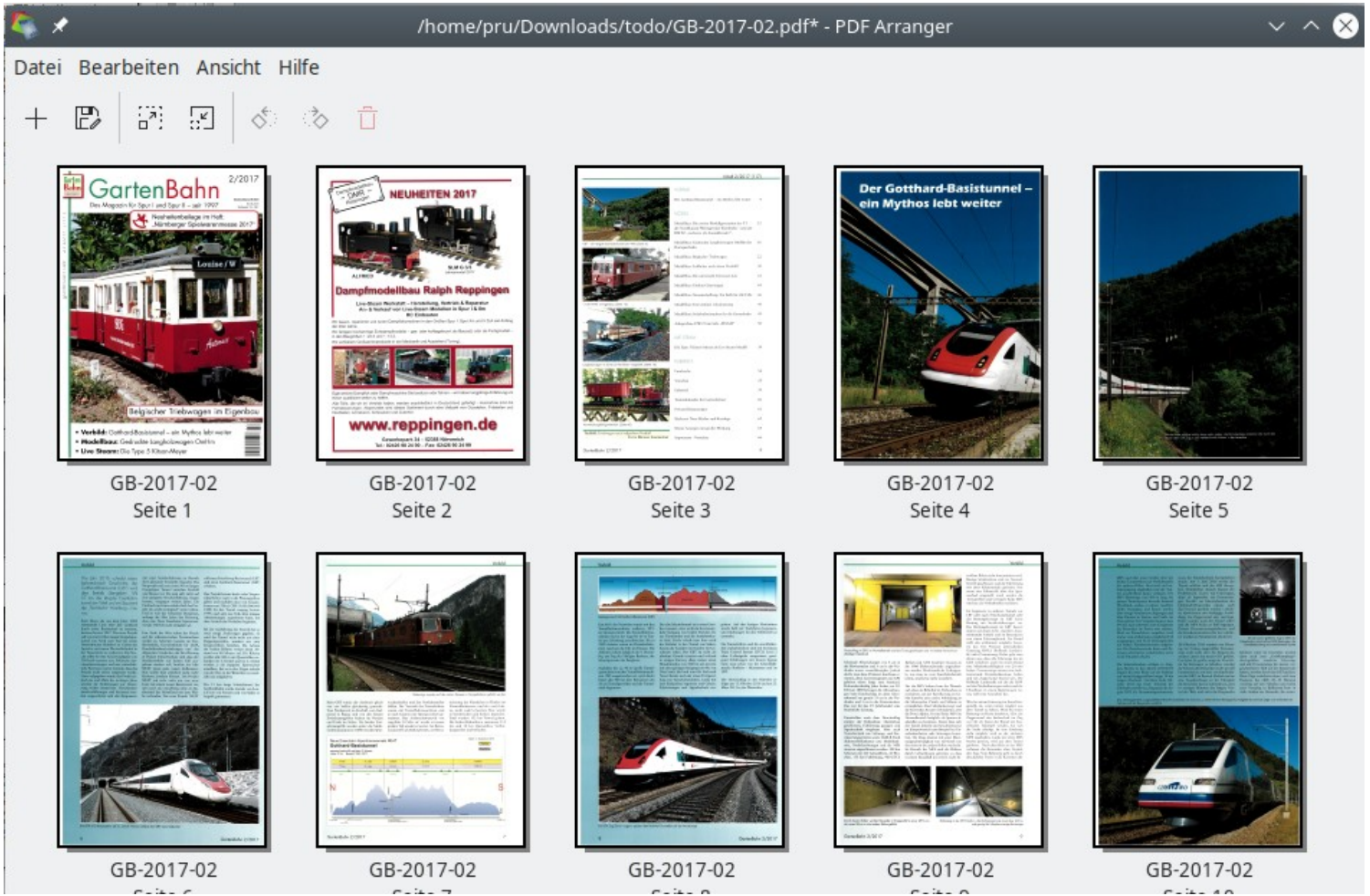
- Deckblatt ist zu dick:
  - Bei vorderem Deckblatt muss dem Einzug oft etwas geholfen werden
  - Hinteres Deckblatt wird entfernt (enthält nur Werbung)
- Scanner: Druck/Scan/FAX-Kombi Lexmark CX510de
  - "Scan to USB"
  - 300dpi, Farbe
  - beidseitig
  - Stapeleinzug

Scannen (2)

PDF zusammenbasteln

- Bei Einzugsproblemen müssen mehrere PDFs aneinandergehängt 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>

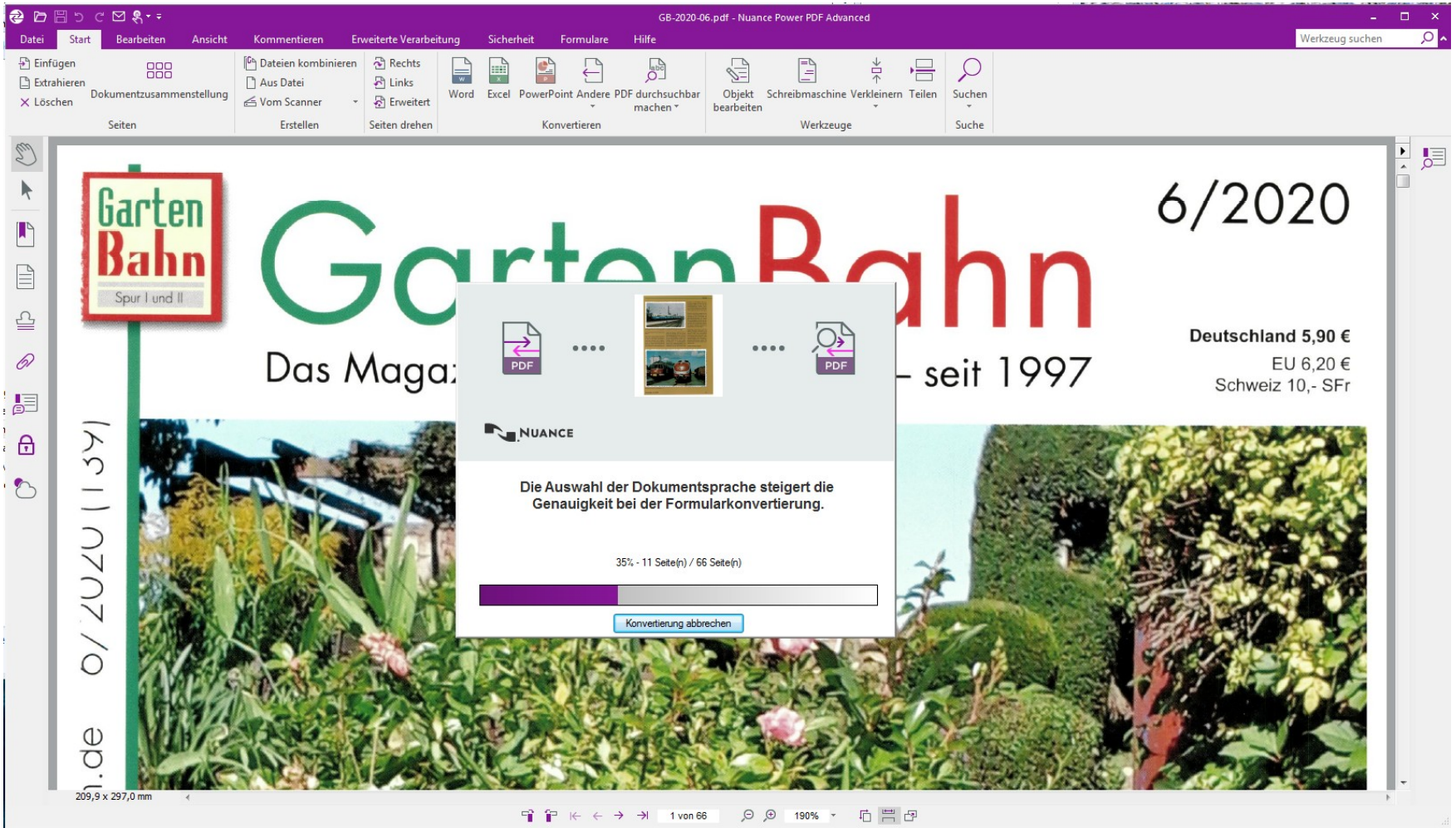
PDF Arranger



OCR

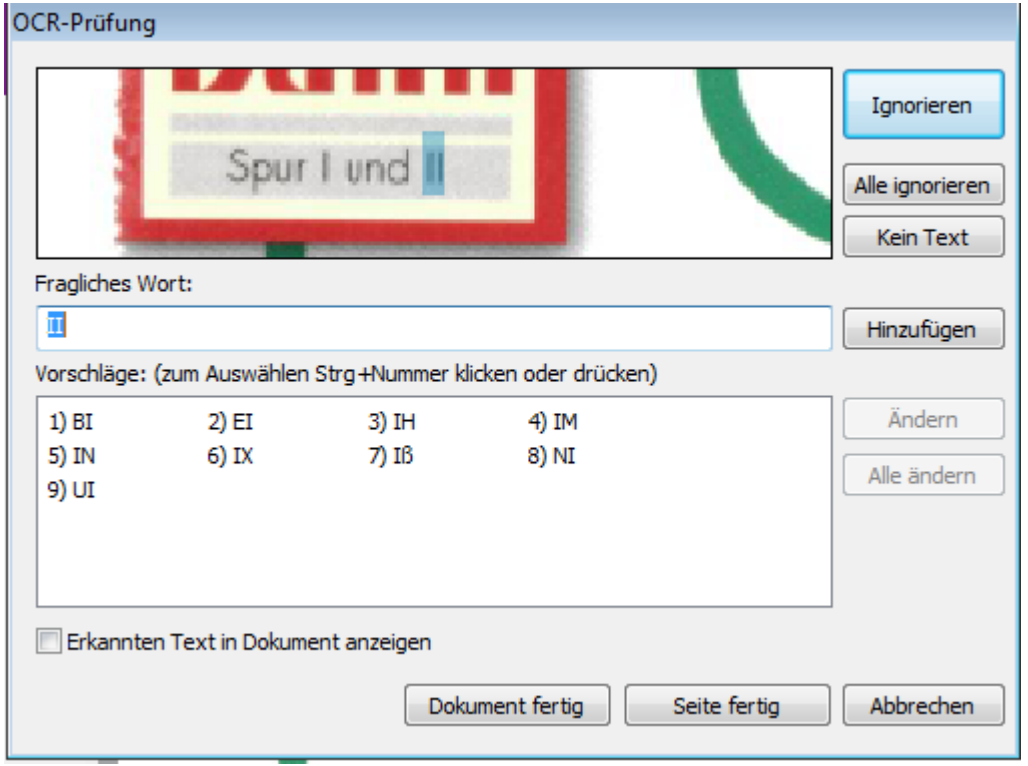
- Linux: ocrMyPDF
  - Gutes Scanergebnis, aber schlechte Positionierung
- Nuance Power PDF (nur Windows)
  - Test bei Ingo mit Version Standard 1.1
  - Gekauft: Version Advanced 2.1 für 17,99 EUR
- Vorgehen
  - PDF öffnen
  - Button "PDF durchsuchbar machen"
  - Datei speichern

Screenshot OCR



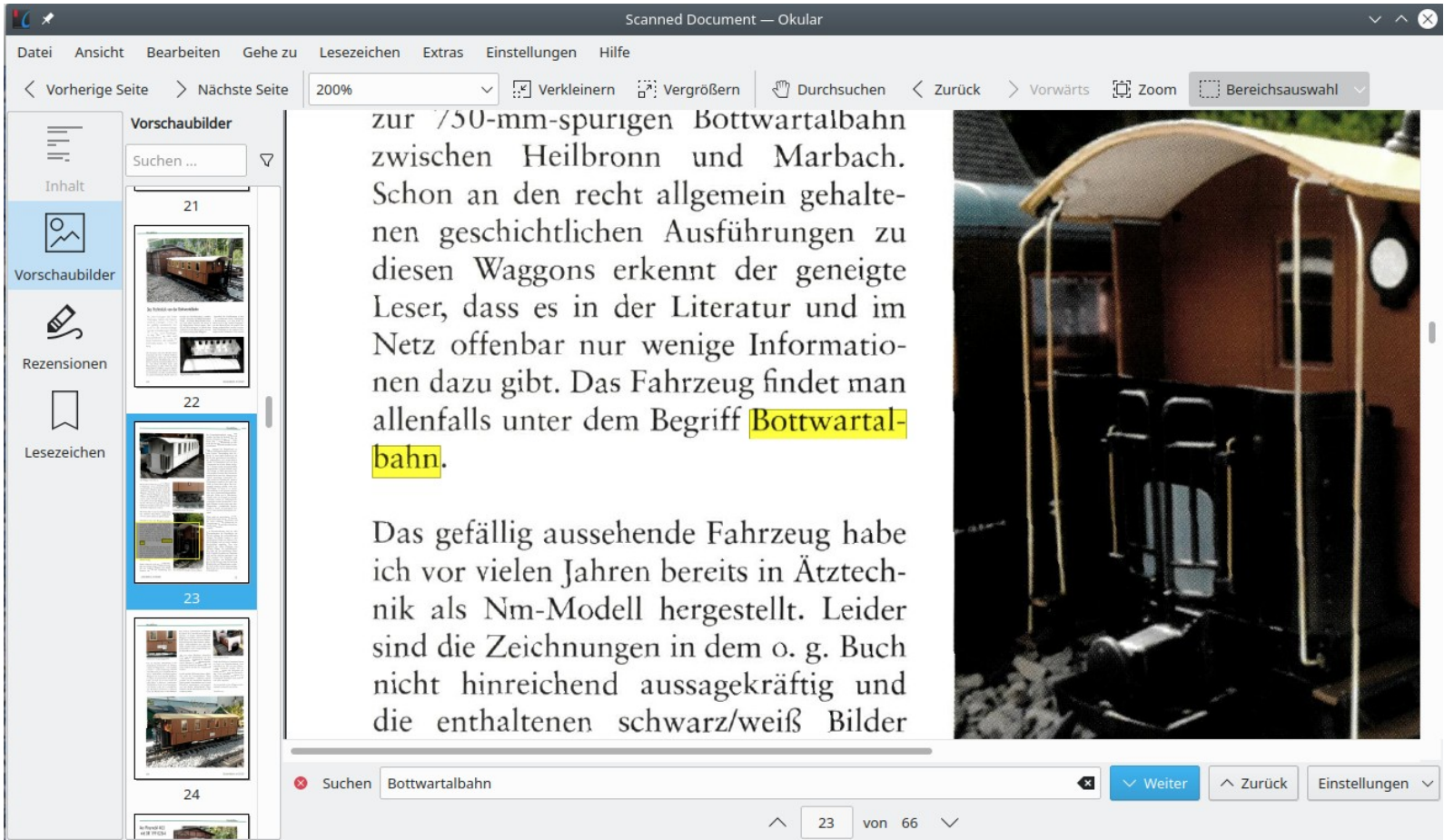
OCR Abschluss

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



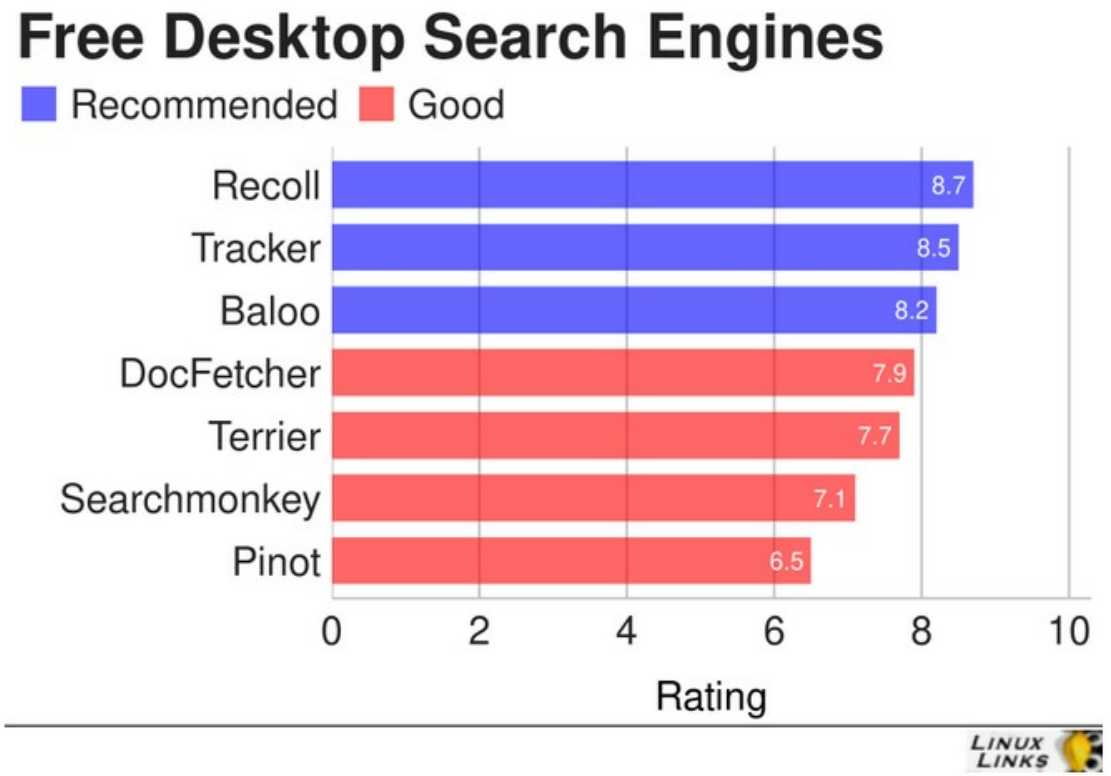
Ergebnis im PDF-Viewer "Okular"

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

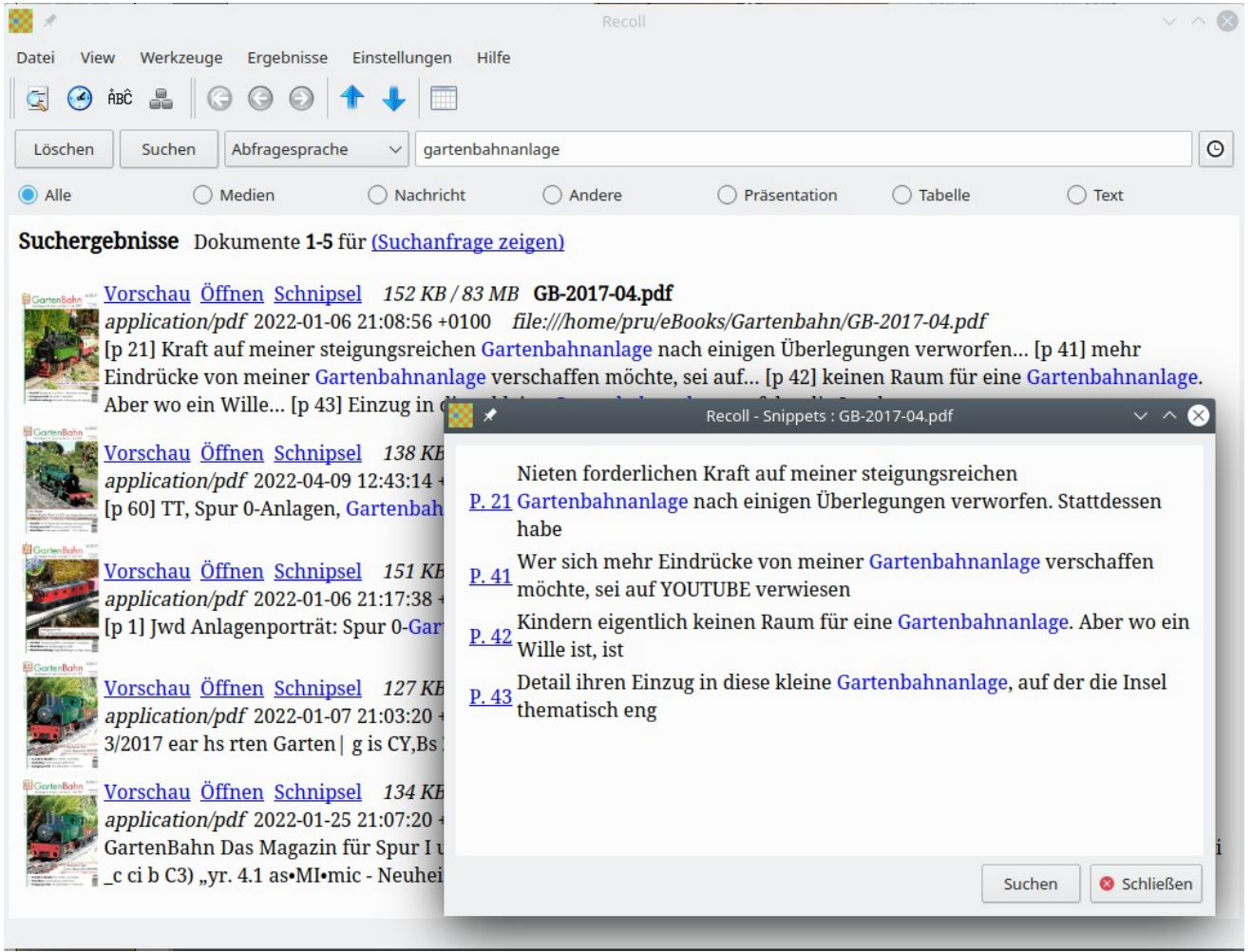


Index / Suche

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

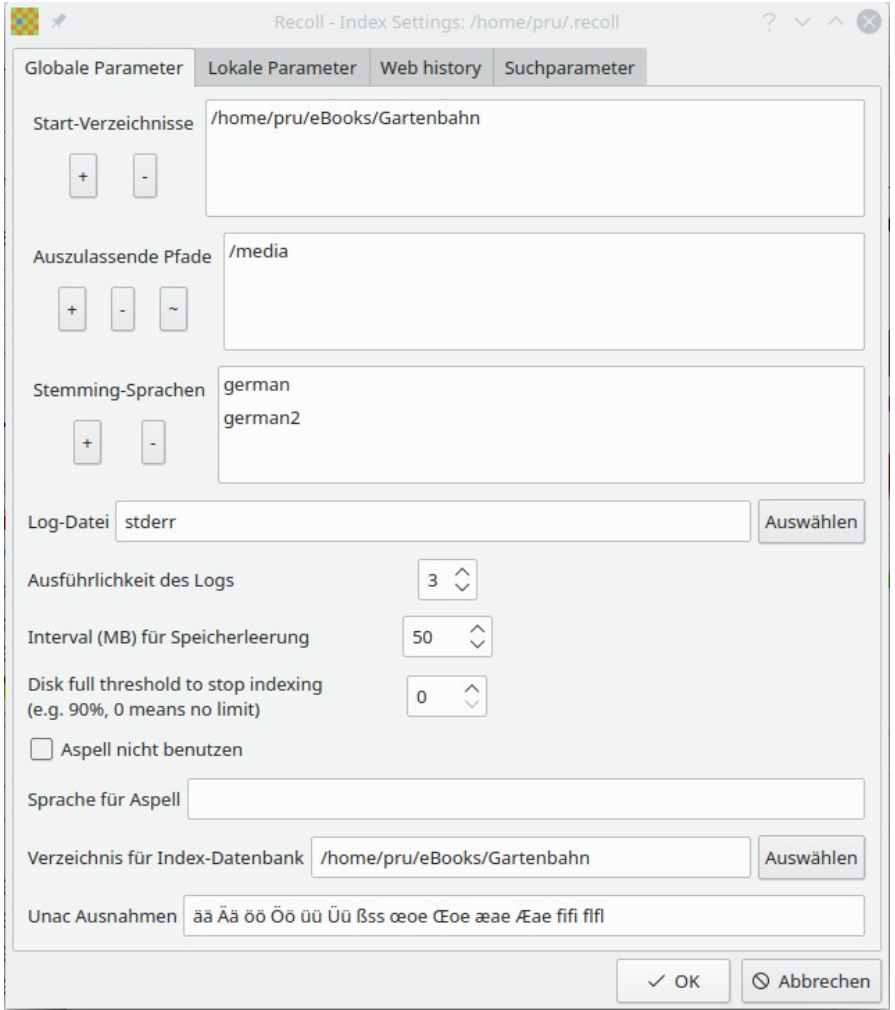


Recoll



Recoll: Einstellungen

- Beschränken auf den Ordner mit den Zeitschriften-PDFs



Recoll: Einstellungen PDF-Reader

- Parameter für Okular: