Die Digitalisierung von Büchern

Peter Koppatz

1 Ein Erfahrungsbericht

Ich beschäftige mich seit 2012 mit dem Werk von Friedrich Rückert, einem Zeitgenossen Goethes, der wunderbare Gedichte geschrieben hat. Die Suche nach einer Gesamtausgabe endet oder beginnt eventuell mit folgenden Problemen:

- der Aufforderung zum Kauf eines Buches
- der Leser oder die Leserin muss die Schriften Fraktur lesen können
- für die verfügbaren Quellen steht kein copy & paste zur Verfügung
- es werden nur Fragmente sichtbar, die aus dem Gesamtwerk herausgelöst sind
- unterschiedliche Druckausgaben als Vorlage (Verlag, Erscheinungsjahr)

Ich möchte zum Linuxtag meinen bisherigen Projektstand vorstellen, der möglichst viele der oben genannten Probleme beseitigen soll. Jede(r), der selbst Interesse daran hat, an der Digitalisierung von Büchern aktiv mitzuwirken, wird ebenfalls auf seine Kosten kommen, da auch die verwendete Software Gegenstand des Vortrages ist. Dazu zählen:

- OCR-Technik von Google (Handhabung des Programmes „tresseract-ocr“)
- Konvertierung der Texte in das „REST“-Format
- Konvertierung in andere Formate wie z.B. HTML, Latex/PDF (mit Sphinx).
- Verwaltung der Inhalte im Versionskontrollsystem „Mercurial“

Die ersten 600 Seiten des Gesamtwerkes von Rückert sind bereits auf meiner Projektseite veröffentlicht. Wer sein Lieblingswerk aus alter Zeit ebenfalls digitalisieren will, ist auf der Website von **gesammelte-werke.org** herzlich willkommen. Und warum dieser Aufwand? Ganz einfach, weil es Spaß macht! Zum Schluss noch eine kleine Kostprobe:

Mein Liebchen hat das Herz sich abgeschlossen,

Den Schlüssel drauf geworfen in die See.

Dort hängt er tief, wo die Korallen sprossen,
Vergebens taucht nach ihm hinab mein Weh.