Dateinamen sind Schall & Rauch

Von Seegras zur CoSin 2018

seegras@discordia.ch
gpg: FBF8 2114 3988 3FFC 9840 0472 0A09 6559 09AF A732

Kapitel das Erste

Worin erklärt wird weshalb Dateinamen ephemeral sind und%20was man mit den 187539 numerierten Dateien in /lost+found anfXngt, ebenfalls was man mit Dateien von Freunden anf&#65533;ngt welche ein propriet&#65533;res Betriebssystem benutzen.

Dateisysteme

- Dateisysteme können ebenfalls unterschiedliche Zeichensätze benutzen.
- Resource Forks, Extended Attributes & ADS werden nicht mit-transferiert.
- Es passieren Fehler. fsck kann «verlorene» Dateien mit Nummern als Dateinamen nach /lost+found verlinken.

Server, Services & Users

- Dateinamen werden umkodiert. Wegen unterschiedlichen Auffassungen über den Zeichensatz (CIFS, NFS, Web, FTP, Filesharing) hin zu notwendigen encodings (HTTP).
- Der Benutzer ändert Dateinamen absichtlich um dumme Devices (VFAT!) oder Shells (Wegen Leerzeichen, Ampersand etc.) zu akkomodieren.

Lösung: Metadaten

Das Problem lässt sich ganz einfach lösen:

- Metadaten zum Inhalt (Titel, Autor, Publikationsdatum, etc.) gehören in die Datei.
- Aus den Metadaten lassen sich dann automatisch Dateinamen und -Hierarchien im lokalen Dateisystem generieren.

Kapitel das Zweite

Welches erklärt was man benötigt um Metadaten in Dateien zu speichern, anhand von Erkenntnissen die wir von Bibliothekaren und Archivaren gelernt haben.

Vier Dinge

- Ein Dateiformat welches das Speichern von Metadaten unterstützt
- Einen Katalog
- Ein Vokabular
- Software die damit interagieren kann

Dateiformate

Es gibt zu fast jeder Art Daten die man in grossen Mengen speichern möchte ein Dateiformat mit Metadaten-Unterstützung:

- PNG, JPG (EXIF),
- PDF (XMP), DJVU, EPUB
- Matroska
- MP3 (IDv3), AAC, FLAC

Der Katalog

Metadaten sind nutzlos ohne Katalog. Der Katalog definiert:

- Feldnamen (Autor, Titel, Publikationsdatum)
- Möglicher Feldinhalt (Text, Bild)
- Zeichensatz (UTF8)

Das Vokabular

- Ein Vokabular ist eine nicht-auschliessliche Wertemenge eines Feldes, welche definierte Werte oder Wertformate enthält.
- Beispiel: «various artists» deklariert für IDv3 im «album artist»-Feld dass auf dem Album Werke mehrerer Künstler sind.
- Notebene kann es sein dass exakt eine Schreibweise korrekt ist, in genau einer Sprache.

Fortgeschrittenes Vokabular

- Das Vokabular definiert auch den Umgang mit Mittel-Initialen/Namen, Artikel (der/die/das) oder wie mehrere Namen getrennt werden (z.b. mit &)
- Suchreihenfolge: Das ist ein Problem für Software und soll niemals in Metadaten einberechnet werden («Titel, Der»)
- Es definiert was Werte für eine Kategorie sind (textbook, novel) und welche für ein Genre (crime, fantasy)
- Und es definiert dass das Publikationsdatum das Datum der Erstpublikation des Werkes ist; nicht das der Datei.

Kapitel das Dritte

Von der Identifikation von Dateien anhand historischer Fakten, arbiträren legalen Merkmalen, sonstig involvierten Entitäten und fantasievollen Beschreibungen.

Eindeutigkeit

- Filme identifiziert man Anhand von Titel, Regisseur, Erscheinungsjahr und Ausgabe («extended edition», «directors cut»>)
- Bücher anhand von Titel, Autor, Erscheinungsjahr, Verlag. Weil ISBN gibt es erst seit 1969, ISSN (für Magazine) seit 1975.

Eindeutig Mehrdeutig

Audio. Wir haben ein Problem. Weniger mit Hörbüchern, aber mit Musik.

- Komponist, Interpreter, Publikationsdaten der Musik, der Aufname, der Abmischung; Opus- und Katalognummern, Titel, Tonart, Albumtitel, Track-Nummer, Ausgabe…
- Und es gibt AcoustID.

Kapitel das Vierte

Wie wir mit Mühe, RegEx und Webscraping Metadaten in Dateien einfügen

Theorie und Praxis

- Calibre (+Extract ISBN), picard, MediaElch können das theoretisch. Meist falsch.
- Praxis: Datei zuerst umbenennen, oder aufgrund vom Dateinamen einige Metatags automatisch vor-befüllen.
- Für Audiodateien kann kid3 Tags aus Dateinamen extrahieren.

Alles Falshc

- Viele Dateitypen kommen ohne jegliche Metadaten daher. PDF, MKV, DJVU
- Einige andere haben meist: MP3, FLAC, AAC, EPUB
- Nur ist alles falsch: «Titel, der», «Nachname, Vorname», «Titel: Autor; Autor: Titel» «Autor: Release-Group», «Autor: Verlag», etc..

RegExit

- Mittels perl & etwas RegEx kann man die Fehler reparieren.
- epub-rename: 75 regexen, hauptsächlich misformatierte/verdrehte Titel/Autor Tags
- exif-meta: 664 regexen, hauptsächlich bogus Titel und Autoren-Tags in PDF Dateien («Windows User» und «Unknown»).

Kapitel das Fünfte

Worin wir hierarchische Datenbanken, genannt «Dateisysteme» anhand der Metadaten aus den Dateien automatisch befüllen; auf dass wir mit primitiven Werkzeugen wie «ls» und «locate» Erkenntnis gewinnen können.

Do-it-yourself

- Audiodateien: kid3 oder picard machens automatisch, man kann ein Template spezifizieren.
- exif-rename (via exiftool, d.h. funktioniert mit extrem vielem), epub-rename (via epub-meta und ebook-meta) beide selbstgemacht.
- MediaInfo extrahiert beliebige Metadaten und erlaubt via Template beliebige Ausgaben davon: `mediainfo --inform=file:///home/user/rename.csv file.mkv`

rename.csv

```
General;mkdir "%Original% (%Released_Date%)"\nmv
%CompleteName% "%Original% (%Released_Date%)"
```

- File_Begin;
- File_End;
- Page_Begin;
- Page_Middle;
- Page_End;
- General_Begin;
- General_End;

Summarum

- Mit allen relevanten Metadaten in der Datei selber sind spezifische Dateinamen irrelevant
- Stattdessen ist es nun möglich Dateien so zu benennen und zu sortieren wie es der Benutzer wünscht; mit oder ohne Leerzeichen, Umlaute, Satzzeichen. Beliebig, alles automatisch generiert.

Ende

Was wo zu finden ist.

- https://seegras.discordia.ch/Programs/ (epub-meta, epub-rename, exif-rename, exif-meta, exif-info, titlemkv, nfo2xml, thek2xml, mkvattachcover, Avinfo.csv, djvu-meta. Also: bicapitalize, respace, respacefilter, lowercase)
- https://www.kvibes.de/mediaelch/ (MediaElch; movieDB scraper)
- https://picard.musicbrainz.org/ (picard; automatischer AcoustID audio tagger)
- https://kid3.sourceforge.io/ (kid3; manueller audio tagger kann automatisch umbenennen)
- https://mkvtoolnix.download/ (mkvtoolnix; konvertiert beliebige container nach matroska)
- https://calibre-ebook.com/ (ebook-convert; teil von calibre)
- https://www.mobileread.com/forums/showthread.php?t=126727 (extract ISBN)