Effiziente und exakte Datensicherung

KI für Bild- und Büchererkennung

Martin Barghoorn APL-Berlin
März 2025

Datenwachstum und -veränderung

- Bereits 2017 hat jeder Mensch im Durchschnitt täglich über 600 MB Daten erzeugt, inzwischen sind es **mehr als ein Gigabyte pro Tag**.
- Das Datenzeitalter wird von neuen Technologien vorangetrieben, z.B. Beschleunigung durch 5 G und besonders KI

Datenmenge aktuell

- Das Whitepaper Data Age 2025 des Analystenhauses IDC, dessen Erstellung von Seagate unterstützt wurde, prognostiziert ein Anwachsen der weltweit generierten Datenmenge auf insgesamt **163 Zettabyte** (ZB) bis 2025.
- Die Anzahl der Bytes ist gleich **2 hoch 70, auch ausgedrückt als 1 Sextillion Byte**. Ein Zettabyte entspricht ungefähr eintausend Exabytes oder einer Milliarde Terabytes.

Wieviel Prozent der Daten sind oder werden zu Müll?

- Fotos, Videos, Dokumente:
  Weil Speicher so billig ist, löscht niemand mehr Daten. Der Berg an Datenmüll wird immer größer und frisst in einem Jahr mittlerweile so viel Strom wie halb Berlin.
- Auch Corona und Online-Office hat die Datenmenge stark erhöht

Qualifizierte Entsorgung von Daten

- A way between redundancy and loss
- Data or files are static or dynamic
- Most of files representing data are static but increasing fast (photos, sounds, videos)
- Principles of order: ordered data by name, type, size and date, außerdem exif
- Advantages and disadvantages of folders
- Programme suchen heute über alle Ordner hinweg

Standard Datensicherung mit APL

Dyalog APL

```
⎕sh'xcopy /s /d /y D:\DATABASE\ M:\DATABASE\ '
```

Was bedeuten diese Parameter?

/S- Zusätzlich zu den Dateien im Stammverzeichnis der Quelle können Sie mit diesem Parameter auch Verzeichnisse, Unterverzeichnisse und die darin enthaltenen Dateien kopieren.

`/D-` Parameter auch ohne Angabe eines bestimmten Datums ausführen, um nur Dateien in der Quelle zu kopieren, die neuer sind als die Dateien im Ziel.

/Y- Verwenden Sie diesen Parameter, um zu verhindern, dass der XCOPY-Befehl Sie auffordert, Quelldateien zu überschreiben, die bereits im Ziel vorhanden sind.

Differentielles Backup

Vor- und Nachteile von xcopy

Vorteil: hohe Geschwindigkeit

Nachteil: auf dem Quelllaufwerk aktuell mit Absicht gelöschte Dateien bleiben auf dem Zielllaufwerk erhalten (Entstehung von Datenmüll)

Welche Daten sollten regelmäßig gesichert werden

- Images, photos: *.jpg, *.tif, *.bmp
- Videos, films: *.mp4, *.mov
- Programs: *.apl, *.atf, *.dws
- Sounds: *.mp3, *.wav, *.wma
- Documents: *.txt, *.doc, *.docx, *.pdf

In der Regel sind nur bestimmte Verzeichnisse zu sichern

Exakte Datensicherung

In größeren Abständen sollte kontrolliert werden, ob sich Datenmüll abgesammelt hat.

Variante A: individuell z.B. mit APL2

Variante B: Standardsoftware

Regular and special fast APL-functions

- dyadic search functions, such as ⍳ Iota (index of) comparing data
- the unique items of a vector, the unique rows of a matrix and so forth
- APL2 external function RF (row find)
- Dyalog workspace fastfns: Code sequences as idioms implemented in C, function MatIota

APL2- indices ← matrix1 RF matrix2

- This external function returns the index of the first row in matrix1 of each row of matrix2. The function is equivalent to the expression:
- indices ← (⊂[1+⎕IO]matrix1)⍳⊂[1+⎕IO]matrix2
- The arguments must be rank 2 character, integer or nested arrays that have the same number of columns. Both arguments must be the same type. If the arguments are nested, their depth must be 2 or less, and all of their subitems must be character or integer arrays.
- This function may provide a performance improvement over the APL2 expression for some types of data.

First, looking for image data (internally) IBM APL2

```
3 11 ⎕NA 'RF'⍝ activate external function RF
Import der Daten vom Quelllaufwerk

⍴h1←⊃('dir /T/S ',' D:\bilder\Ansichtskarten\*.jpg ')
PIPE  ''
265 66 ⍝ 356 files at D directory
/T: date and time of creating the file
/S: scan also all subdirectories

    ⍴i←h1 RF h1  ⍝ Control of redundancy
```

Bearbeitung der importierten Daten

Datenträger in Laufwerk D: ist Daten

Volumeseriennummer: 6A66-42A9

Verzeichnis von D:\bilder\Ansichtskarten

```
13.04.2020  18:03           369.430 20200413_180304.jpg
14.04.2020  07:43           802.620 20200414_074359.jpg
14.04.2020  07:45            84.884 20200414_074520.jpg
14.04.2020  07:50           508.859 20200414_075016.jpg
03.07.2020  15:21            97.563 20200703_095215.jpg
03.07.2020  15:21           615.162 20200703_100258.jpg
04.12.2024  06:19         1.855.133 20201118_093015.jpg
```

Alle Zeilen, in denen an 1. Stelle ein ' ' steht entfernen

Removal of unnecessary lines

```
    h1←(~h1[;1]=' ')/[1]h1
ρh2←⊃('dir /T/S ',' H:\bilder\Ansichtskarten\*.jpg ' )
PIPE ''
Import der Daten vom Zielllaufwerk
Dateien nach Import und Bereinigung
  ρ⍤1 h1 h2
 251 66  243 66
 ρi←h2 RF h1                  251
alte←(i>1↑ρh2)/[1]h1
```

```
21.12.2023  21:30        488.208 bellinzona.jpg
25.11.2022  17:17        664.470 moench.jpg
31.01.2025  11:26        438.663 moench_kletter.jpg
12.03.2025  10:36        152.754 zittau_bergwerksschule.jpg
13.03.2025  11:02        124.799 zittau_panorama.jpg
10.03.2025  09:56        128.730 aushebung_militaer.jpg
07.03.2025  20:49        339.218 bergische_eisenfau_litho.jpg
10.03.2025  09:58        166.831 klingenthal.jpg
10.03.2025  10:02        115.886 neustadt_sachsen.jpg
10.03.2025  10:02        144.683 Phillippsdorf.jpg
```

Freie software für Windows

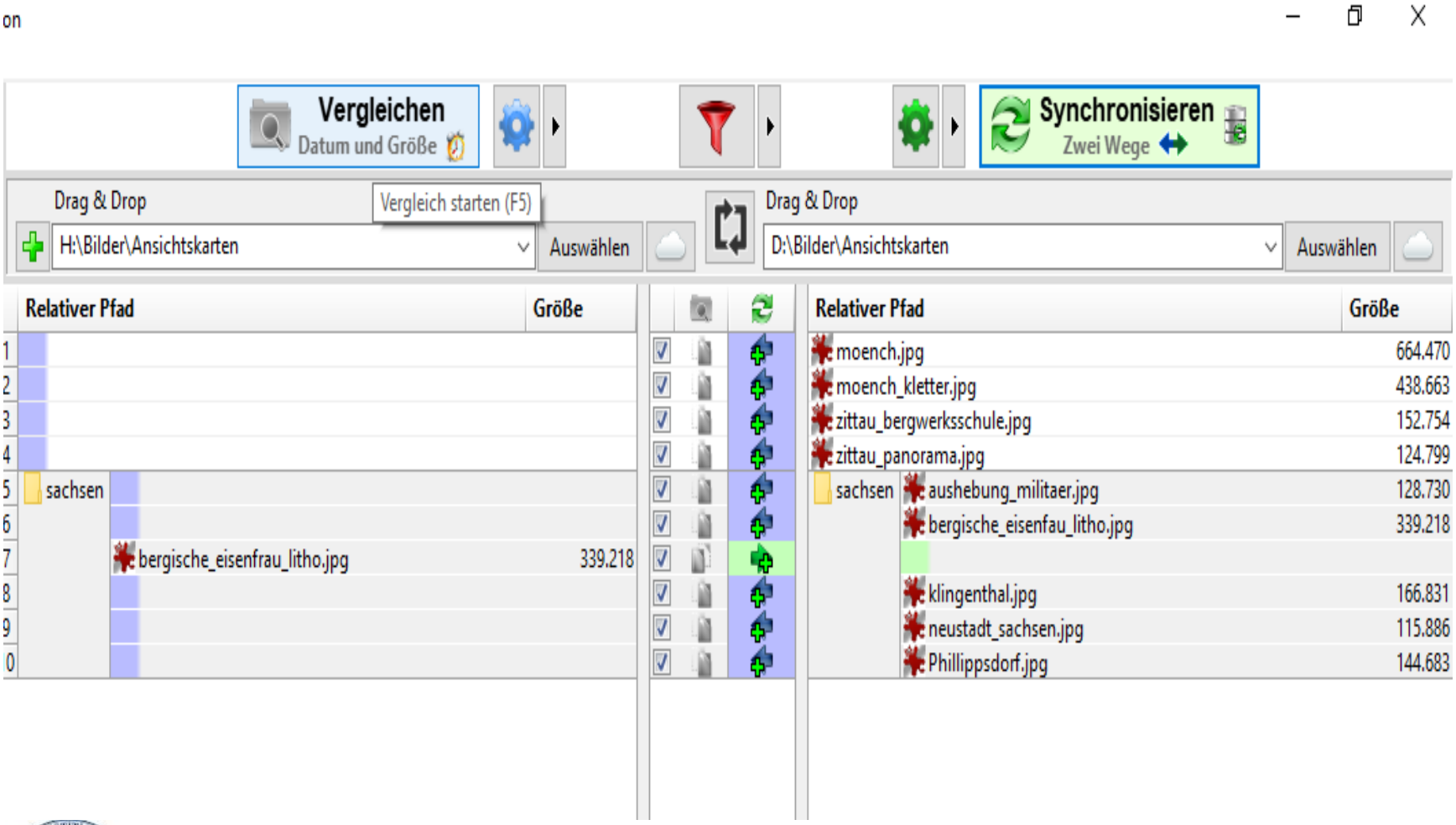
Automatische

oder

gezielte Synchronisation

FreeFileSync

FreeFileSync

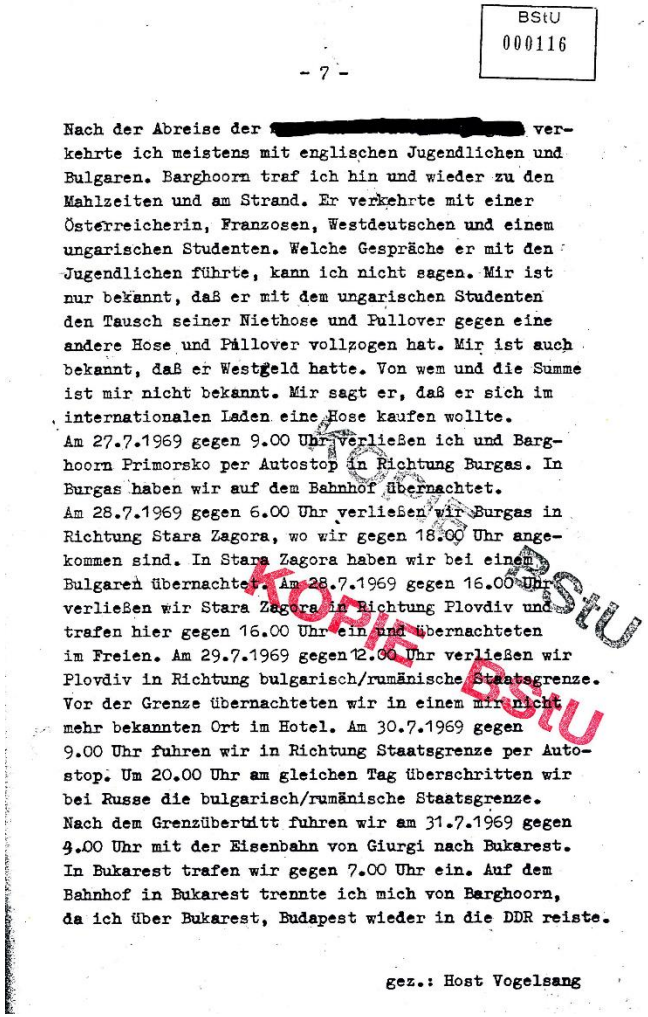


Martin Barghoorn APL-Berlin
März 2025

Teil 2, ganz frisch entdeckt

- Gezielter Einsatz von KI für
- Bild-,
- Text- und
- Bücherkennung
- Motivation, aus der Not geboren,
- Effizienz- und Umsatzsteigerung

App Samsung gallery findet die Bilder <u>oder</u> Textdateien nach Suchwörtern



Martin Barghoorn APL-Berlin
März 2025

Bücher als Informationsquelle

Historisch: Buch kommt auf *deutsch* von Buche, dem Baum.

In sehr alter Zeit wurde die Runen aus Stäbchen, die aus Buchenzweigen hergestellt wurden, geritzt. Das englische Wort write stammt aus dem altenglischen writan, was so viel wie "kratzen", "gravieren" oder "schreiben" bedeutet.

Auch Tafeln aus Buchenholz wurde beschrieben oder als Einbanddeckel verwendet.

Die Papyrusrollen, Vorläufer der Bücher, die in Körben, Krügen oder Töpfen aufbewahrt wurden, hatten außen eine Markierung zum bessern Auffinden der Rollen.

Buchrücken als Informationsquelle

- **Rücken-** engl. **Spine** (rückübersetzt Wirbelsäule)**,** franz. **dos du livre**
- Die schmale Seite eines Buches, an der der Buchblock an der Buchdecke befestigt ist. Bei Taschenbüchern ist der Rücken in der Regel fest mit dem Buchblock verklebt. Bei gebundenen Büchern mit festem Umschlag wird der Buchblock von innen am Bundsteg eingehängt.

Rückentitel

Name und Verfasser oder Herausgeber des Buches werden auf dem Buchrücken aufgeprägt oder aufgedruckt.

Häufig in Längsrichtung (Längstitel), damit die Bezeichnungen auch auf schmale Buchrücken passt.

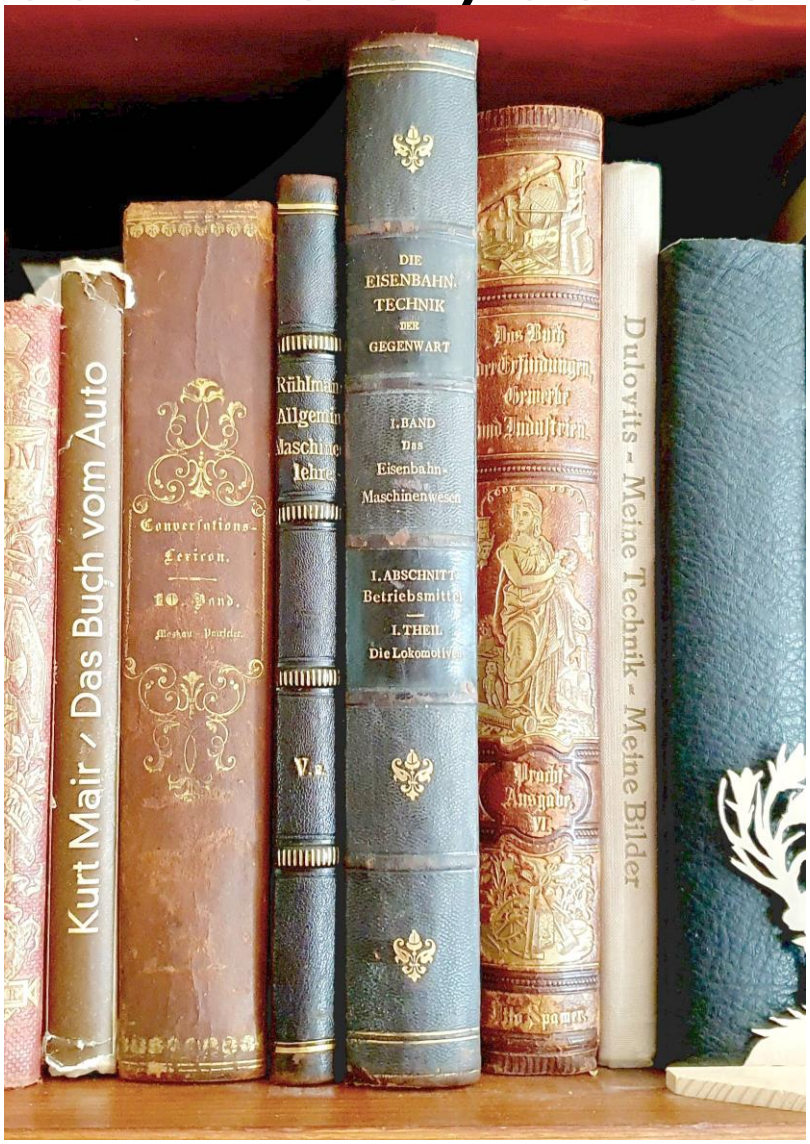
Ermöglicht die Erkennung eines Buches, wenn es im Regal steht oder liegt.

Richtung der Beschriftung

Bei aufgestelltem Buch kann die Schrift entweder senkrecht oder waagerecht, d. h. parallel zum Buchrückenfalz oder senkrecht dazu, ausgeführt sein.

Eine falzparallele Beschriftung ist in Deutschland, Frankreich und Italien in der Mehrzahl der Fälle von unten nach oben (also mit nach links geneigtem Kopf), in englischsprachigen Ländern von oben nach unten (mit nach rechts geneigtem Kopf) zu lesen.

Das neue Projekt, finden von Begriffen (Titel oder Autor) auf den Buchrücken



Martin Barghoorn APL-Berlin
März 2025

Erkennung von Schriften und Sprachen auf Buchrücken

- Alte Schriften? JA
- Beispiel Frakturschrift JA

- andere Sprachen, seltene Namen? Ja

- Dulovits - Meine Technik JA

Summary

Die Ki gallery gezielt und kreativ einsetzen kann die Arbeit äußerst erleichtern und den Umsatz signifkant erhöhen.

Thank you for your Attention!

Martin@Barghoorn.com