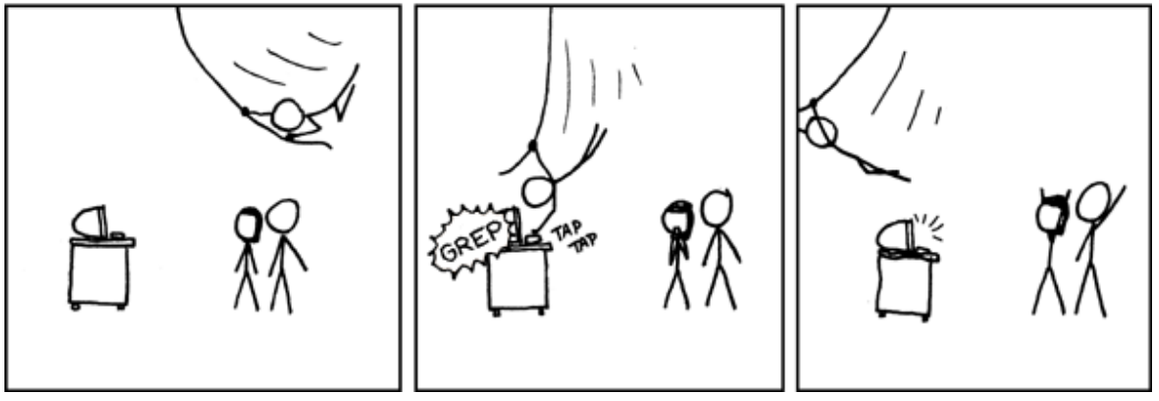
GREP Hardcore

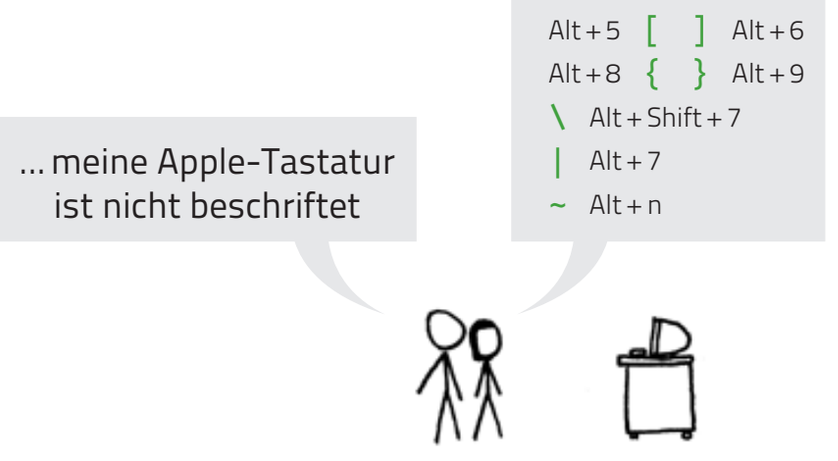
20. April 2012
Pubkon 2013

Kontakt: gregor.fellenz@publishingx.de
Folien: http://www.publishingx.de/dokumente



Schöner suchen und ersetzen

- InDesign bietet ab CS3 die Suche mit **Regulären Ausdrücken** bzw. **GREP**.
- **Was ist GREP?** Die Suche nach Mustern.
  und bei der **Ersetzung** die gefundenen Muster wiederverwenden.
- Seit InDesign CS4 **GREP-Stile.**
- Große Augen bei den Kollegen :-)



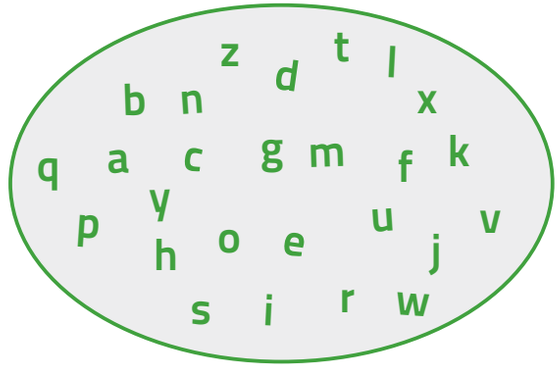
Das einfachste GREP Zeichen

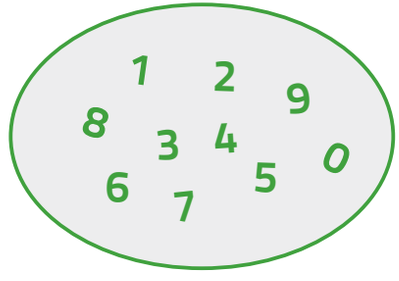
- Der Punkt `.` trifft alles Zeichen außer dem Absatzende.
- Der Modifier `(?s)` schaltet den Punkt um, so dass er wirklich alle Zeichen trifft.
- By the way: `(?m)` hat damit nix zu tun.

Finde einen ganzen Absatz!

InDesign zählt auch **Harte Zeilenumbrüche** als Absatzende.
Bzw. der Punkt trifft alles außer Absatzende und Harte Zeilenumbruch.

- `.+`
- oder `^.+$`
- oder `^(.|\n)+$`
- aber nicht `^[.\n]+$`





Zeichenklassen Extended

- **Eigene** Zeichenklassen bauen: `[\dabc,.]`
  Das Gegenteil mit `[^\s]`
- **Unicode Properties** werden mit `\p{}` bzw. `\P{}` gebildet.
  Brauchbar finde ich `\p{Zs}` – Alle Leerräume ohne Tabulator.
  Liste: *http://bit.ly/13n9nOw*
  - **Autsch:** Kombination in eigener Zeichenklasse geht nicht: `[\p{Zs}\t.]`
    Lösung: Als Variante kombinieren `(\p{Zs}|[\t.])`
- Wer braucht **Posix** Zeichenklassen? `[[:punct:]]`
  Kann übrigens auch nicht in einer Zeichenklasse kombiniert werden...

Positionen

- Positionen treffen KEIN Zeichen, sie nehmen keinen Raum ein.
- `^` und `$`
  Modifier `(?-m)`: `^` und `$` treffen nur am Anfang und Ende des Textabschnitts
- `\A` Anfang des gesamten Textabschnitts, Ende `\Z`
  In Tabellenzellen nicht ganz intuitiv!
  `\A\Z` = Leere Textrahmen
  `\r\Z` letztes, überflüssiges Absatzende-Zeichen entfernen.
- `\b` Wortgrenze - Übergang zwischen `\w` und `\W`
  Ggf. Probleme bei `Wort_Mit_Unterstrich`
- Wortanfang `\<` und Wortende `\>` verstehe ich nicht.

Fundstellen markieren

- Markierung von Bereichen im Muster mit `(GREP)`
- In der Ersetzung mit `$1`, `$2` ... aufrufen
- Klammer nicht mitzählen `(?:GREP)`

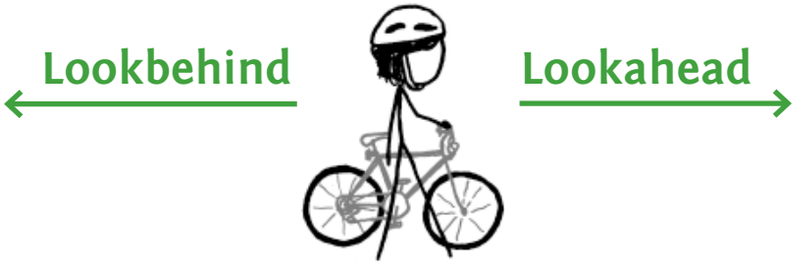
Doppelte Wörter finden

Fundstellen können innerhalb der Suche verwendet werden.

- Markierung der Fundstelle mit `()` – wie für die Ersetzung.
  Aufruf der Fundstellen mit `\1`, `\2` ...
  GREP für doppelte Worte `(\b[\u\l\-]+\b)\s+\1`

Doppelte Leerzeichen ersetz' ich einfach …

- **Erster Ansatz:** Suche mit der Zeichenklasse `\s` und dem Wiederholungszeichen `+`
- **Es zeigt sich:** `\s` trifft alle Leerräume (inkl. Umbruch)
- **Zunächst erstmal:** `+` verwenden.
- **Lösung:** `\p{Zs}` mit Tabulator `(\p{Zs}|\t)`
- Problem Indexmarken, XML-Tags …!
- Skriptansatz: *http://www.indesignblog.com/2010/11/whitespace-serachreplace-2/*



Ausschau halten

Bei der Verwendung von sogenannten **Look Around Assertions** wird geprüft, ob vor bzw. nach dem eigentlichen Ausdruck ein anderer Regulärer Ausdruck steht. Besonders wichtig für **GREP-Stile**

Lookbehind

- **Positiv** Brüche: Ziffern nach einem Bruchstrich: `(?<=/)\d+`
- **Negativ** Preise vor denen nicht EUR steht: `(?<!EUR)\s+\d+,\d\d`
- Achtung: Lookbehind **nur** mit fester Länge möglich (Widerholungen nicht einsetzbar)

Lookahead

- **Positiv** Ziffern denen die Einheit `cm` folgt: `\d+(?=\scm)`
- **Negativ** Geschäftsbericht, AG hinter dem Firmennamen prüfen: `Firma\s+(?!AG)`

Telefonnummern gliedern

Mit Look Around kein Problem

- **Suche** nach: `(?<=\d)(\d)(?=(\d\d)+\b)|(\d)(?=(\d\d){2,}\b)`
- **Ersetze** durch: `$0~<`
- Funktioniert leider nicht mit Zahlen, die am Ende eines Textabschnitts oder am Ende einer Tabellenzelle stehen.
  Workaround: Hilfszeichen ans Ende Stellen
  Suche nach: `(\d+)\Z`
  Ersetzen durch: `$1~|`

GREP der ein bestimmtes Wort nicht enthält

Anforderung: Alle Absätze, die das Wort `Treffer` **nicht** enthalten auswählen.

- Negierte Zeichenklasse klappt nicht `[^Treffer]`
- Lösung: `^((?!Treffer).)+$`
- Der GREP macht an jeder Position einen Lookahead ob dort NICHT „Treffer“ gefunden werden kann, wenn dies zutrifft sammelt der Punkt das nächste Zeichen ein. Durch die Klammerung wird dieser Test auf mehrere Zeichen ausgeweitet.
- **Bonus:** `eins` muss enthalten sein, `zwei` darf nicht enthalten sein: `^(?=.*eins)((?!zwei).)*$`
- Credits: *http://stackoverflow.com/questions/406230/regular-expression-to-match-string-not-containing-a-word*

Performance

- Generell nur bei langen Textabschnitten/Skripting relevant.
- Einfachste Regel `.*?` verbraucht viele Resourcen,
  Möglichst spezifischen GREP schreiben
  Wenn möglich durch eine Variante ersetzen, die den Begrenzer verneint: `"[^"]+"`
- Varianten sparsam einsetzen nach Möglichkeit optimieren:
  Statt `(abcd|abef)` besser `ab(cd|ef)`
- Nur benötigte Fundstellen markieren, ansonsten `(?:pattern)`

Tools und Skripte

Die Tool Sammlung von Peter Kahrel finden Sie unter
http://www.kahrel.plus.com/indesign/grep_matters.html

- Für komplexe Ausdrücke ist der **GREP Editor** zu empfehlen.
- Das Skript **Chain GREP queries** ermöglicht es, mehrere GREP-Abfragen zusammenzustellen und hintereinander ablaufen zu lassen.

Selber skripten

- Leseprobe *http://www.indd-skript.de/wp-content/uploads/2011/03/Kapitel_4-8__Suchen_und_Ersetzen_per_Skript.pdf*

Programme

- **Text Editor SublimeText**
  Mac/Windows *http://www.sublimetext.com/*
- **Dateien umbenennen**
  Mac/Windows *http://www.publicspace.net/windows/BetterFileRename/index.html*

Writing a regular expression is more than a skill – it's an art.

Jeffrey Friedl

Dank an

- **Peter Kahrel** Das E-Book/Buch für GREP in InDesign CS3/CS4
  http://shop.oreilly.com/product/9780596156015.do
  Skripte und Tools: http://www.kahrel.plus.com/indesign/grep_matters.html
- The Internet :-)
- http://xkcd.com/

In eigener Sache

InDesign automatisieren – Skripting, GREP & Co.

Das Buch zur InDesign Automation mit einer Skripting Einführung und vielen Praxistipps zu EPUB, XML und GREP.

Auf der Homepage zum Buch http://www.indd-skript.de gibt es Leseproben und alle Beispiele

- Klassisch auf Papier
  ISBN: 978-3-89864-734-2
  Preis: 34,90 Euro (D), 35,90 Euro (A)
- EPUB
  ISBN: 978-3-89864-882-0
  Preis: 27,90 EUR

Vielen Dank für Ihre Aufmerksamkeit!

Fragen und Anregungen?

Folien: http://www.publishingx.de/dokumente

E-Mail: gregor.fellenz@publishingx.de
Twitter: grefel
Blog: http://www.indesign.js