# Dateiformate und ihre Tücken ☺

tekom Herbsttagung 2011

Angelika Zerfaß

# Was wünschen wir uns?

- Saubere Dateien, die nach den Regeln des Textverarbeitungsprogramms erstellt wurden
- Dateilieferanten, die uns professionell mit den richtigen Dateiformaten beliefern und nicht alles nach Word kopieren ☺
- Genaue Anweisungen wo übersetzbarer Text ist und wie mit dem Rest verfahren werden soll
- Dateien in einer verarbeitbaren Größe

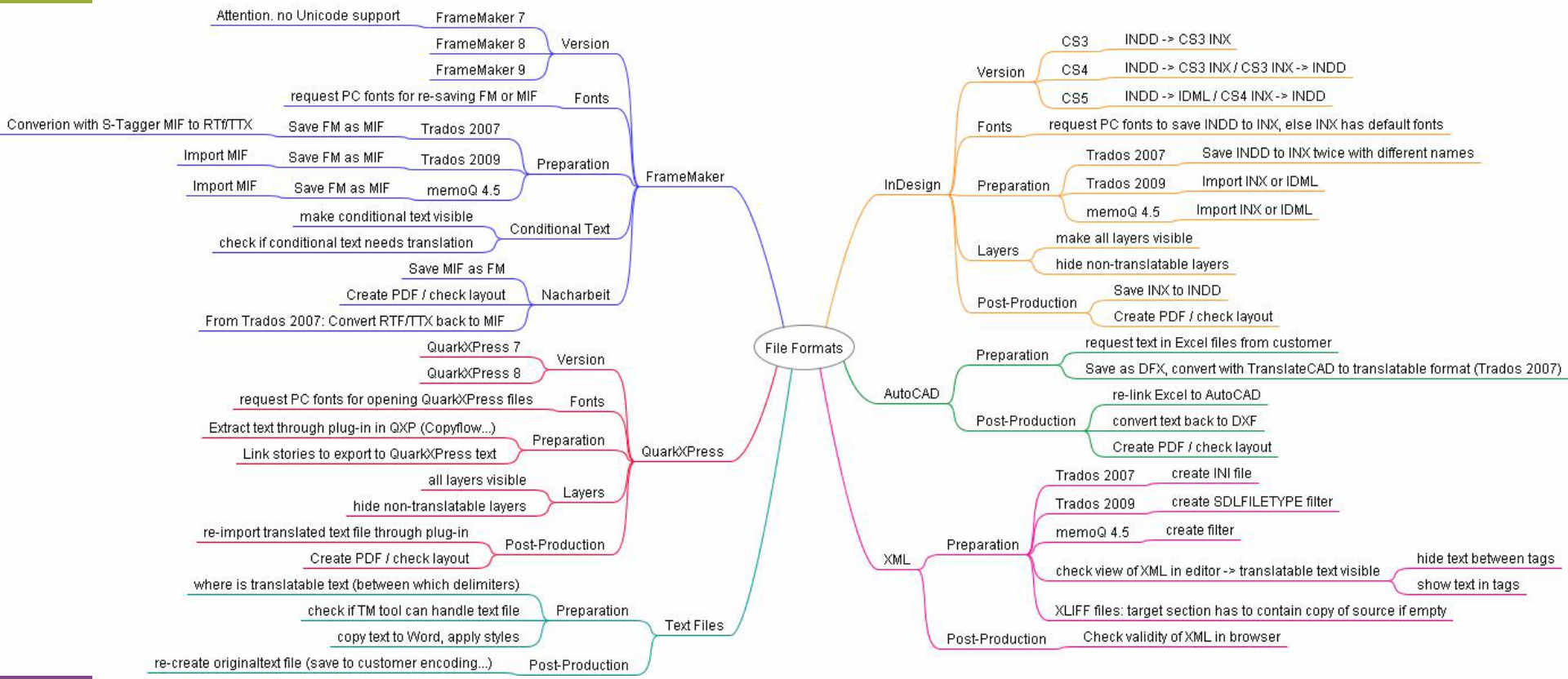
# Was bekommen wir?

- Ausgangsdateien, deren „Layout“ über Zeilenschaltungen, Tabulatoren und Leerzeichen erstellt wurde
- (Word) Dateien, die schon seit Jahren kopiert und umgespeichert werden und so viele interne Fehler haben, dass sie irgendwann einfach nicht mehr funktionieren
- Dokumente in denen dem Spieltrieb beim Formatieren (von Leerzeichen…) freien Lauf gelassen wurde.

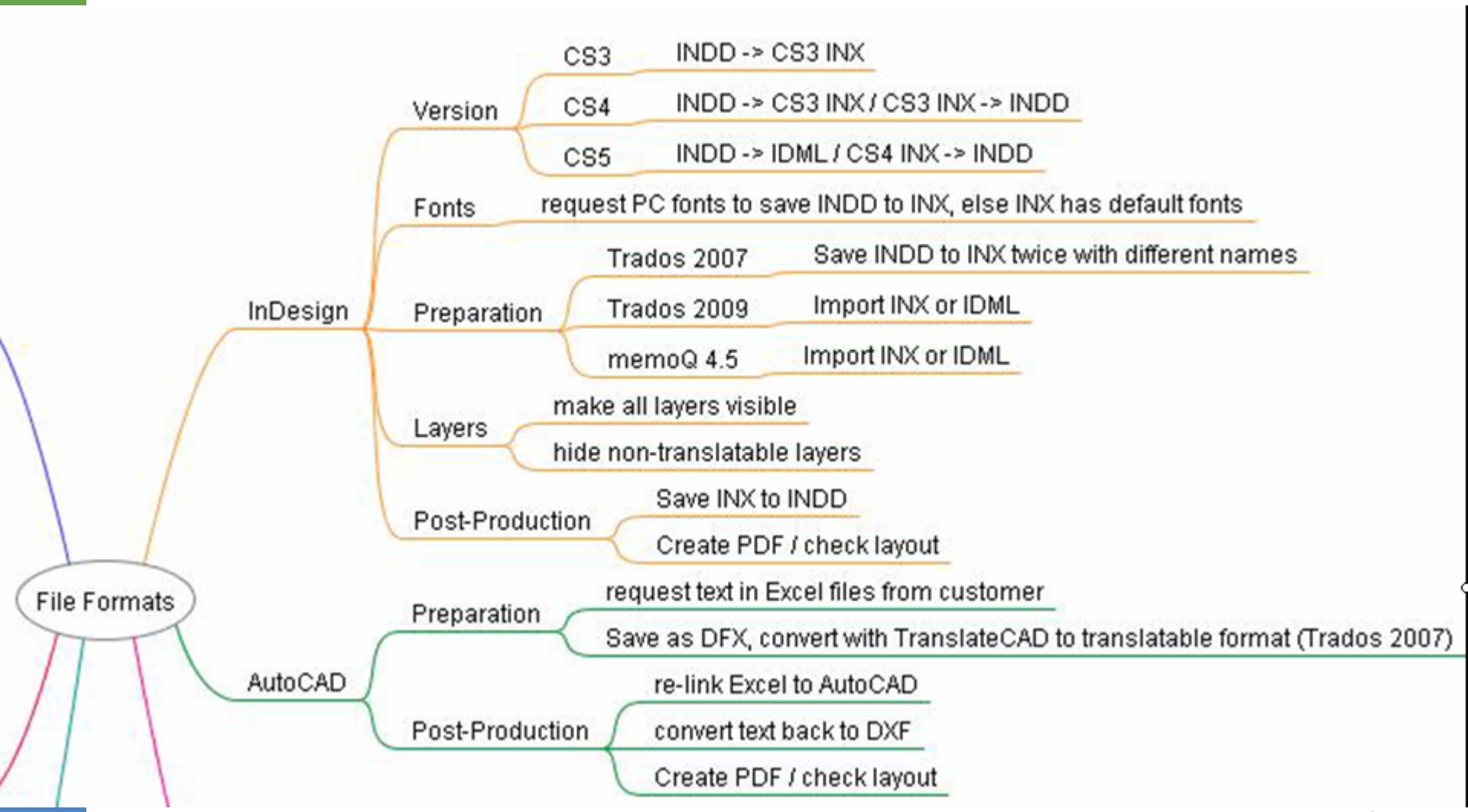
# Grundlegendes

- Was sollte man prüfen, bevor eine Datei in die Übersetzung geht oder für die Ermittlung Statistiken für Wortzahlen und Match-Raten genutzt wird?
- Welche Version eines Dateiformates wird in welchem Übersetzungsprogramm unterstützt und gibt es Workarounds falls ein Übersetzer eingesetzt werden muss, der das geeignete Tool nicht hat?
- Mit welchen Formaten können die Übersetzungsprogramme überhaupt arbeiten, mit welchen nicht.

# Mindmap Dateiformate



# Mindmap Dateiformate



# Was gilt für alle Formate?

- Ungünstige Segmentierung des Textes durch überflüssige Zeilenschaltungen
- Texte auf Grafiken
- Eingebettete oder gruppierte Objekte (Office)
- Kritische Dateigröße (DOC->RTF, InDesign, XML)
- Falsche Segmentierung durch unbekannte Abkürzungen
- Existiert außer dem offensichtlich zu übersetzenden Text noch weiterer Text, z.B. innerhalb von Tags (XML)

# Office Dateien

DOC/X – PPT/X – XLS/X

# Beispieldatei.doc

- Indexeintrag – muss im Übersetzungsprogramm unterstützt und in der Analyse gezählt werden
- Referenz auf Seitenzahl ist kein automatisches Feld, muss ggf. manuell angepasst werden
- Text in Tabelle mit harter Zeilenschaltung umgebrochen
- Abbdlg. ist eine unbekannte Abkürzung
- Überschrift ist mit einer harten Zeilenschaltung getrennt
- Es ist noch ein Kommentar in der Datei
- Hinweistext, wo ein Screenshot erscheinen sollte – nicht übersetzbar
- Grafiknamen in Sätzen sind speziell geschrieben, sollten geschützt werden und in den Statistiken nicht als übersetzbar gezählt werden
- Ein eingebettetes Objekt kann von vielen Übersetzungsprogrammen nicht ausgelesen werden

# Beispieldatei.xls

- Tabellenblattnamen sind schon im Deutschen zu lang (2. Blatt)
- Text enthält HTML Codes (nicht alle Übersetzungsprogramme können das herausfiltern)
- Zellen auf Blatt 2 haben Namen, die aber nicht im Übersetzungseditor auftauchen
- Blatt 3 ist geschützt (Passwort: Info), Texte erscheinen nicht im Übersetzungseditor

# Beispieldatei.ppt

- Folie 1 enthält Notizen: übersetzbar oder nicht?
- Folie 2 enthält Grafik: Platzierung nach Übersetzung anpassen
- Folie 3, 4 und 5 enthalten eingebettete Objekte
- Folie 6 enthält Text in Grafiken: nicht übersetzbar, da Grafik neu gestaltet werden müsste

# DTP Dateien

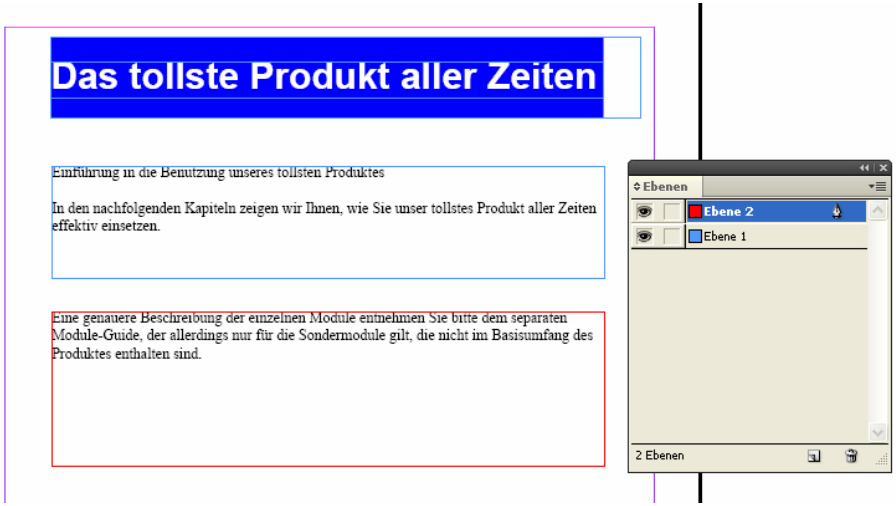
InDesign – FrameMaker - QuarkXPress

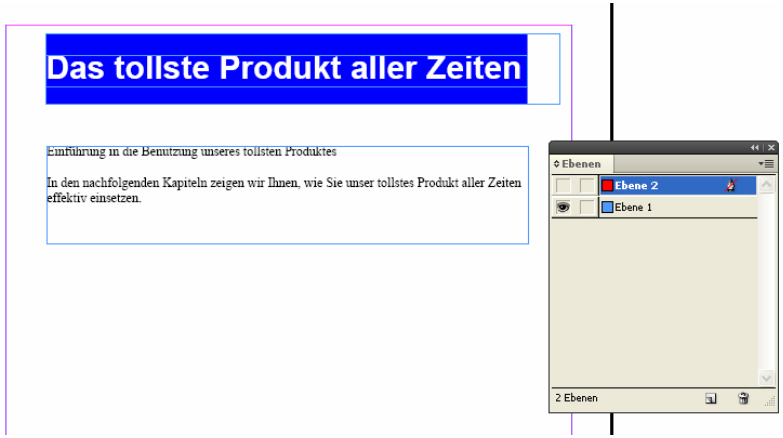
# Beispieldatei.indd/inx/idml

- Welche Formate aus InDesign können Übersetzungsprogramme lesen?
  - INX aus CS2, CS3, CS4
  - IDML aus CS4, CS5
- Um selbst INDD nach IDML oder INX zu speichern, braucht man die Kundenschriftarten, sonst ersetzt InDesign die Schriftarten mit Standardschriftarten
- InDesign ist sehr empfindlich, was Tag-Fehler beim Übersetzen angeht. Es müssen alle Tags aus dem Ausgangstext behalten werden!

# Beispieldatei.inx/idml

- Überschrift ist weiß: Prüfen ob Text im Übersetzungseditor sichtbar ist
- PDF zeigt mehr Text als INX/IDML: eine Ebene wurde ausgeblendet



# Beispieldatei.indd/inx/idml

- Da InDesign INX nicht immer sauber gespeichert wird, nach dem Öffnen im Übersetzungstool ein Segment öffnen, kopieren, schließen und die Datei exportieren, um zu sehen, ob sich die Zieldatei mit dieser INX Datei problemlos wieder herstellen lässt
- Wenn nicht, INX ein zweites Mal exportieren oder von INDD zweimal exportieren – Achtung: Kundenschriftarten nötig!

# Beispieldatei.indd/inx/idml

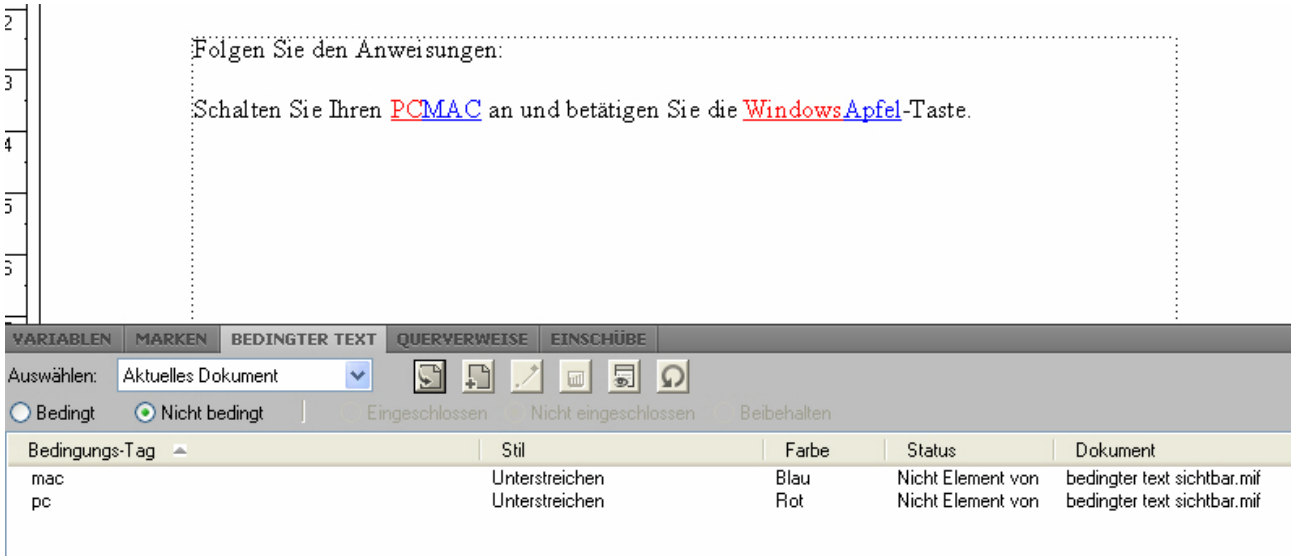
- InDesign CS3 und CS4 produzieren das INX Format
- InDesign CS4 und CS5 produzieren das IDML Format
- Erst ab Trados 2009, memoQ 5… wird IDML unterstützt
- IDML ist ein gezipptes Format. Die eigentlichen Texte befinden sich im Ordner Stories – jede Textbox ergibt eine eigene XML Datei

# Beispieldatei.fm/mif

- Welche Formate aus FrameMaker können Übersetzungsprogramme lesen?
  - MIF 7 und 8 (Trados 2007)
  - MIF 8 und 9 (Trados 2009)
  - Mif 7 (bedingt), 8 und 9 (memoQ 5)
- Um von FM nach MIF zu speichern müssen die Kundenschriftarten installiert sein, da sonst FrameMaker diese mit Standardsschriftarten ersetzt.

# Beispieldatei.fm/mif

- Bedingter Text:
- PDF zeigt Text für PC, MIF zeigt Text für MAC
- FM zeigt beide mit bedingtem Text
- Die meisten Übersetzungsprogramme lesen nur den eingeblendeten bedingten Text aus



# Kundenbeispiel

- Bedingter Text:
- Texte die mit 45/50 UND 40/45 belegt waren, tauchten nicht in der Übersetzung auf, da nur eine der beiden Formatvorlagen sichtbar

Texte mit Bedingung 45/50 erscheinen


Texte mit Bedingung 40/45 erscheinen nicht

# ML Formate

HTML – XML - XLIFF

# Beispieldatei.html

- HTML Dateien können Fehler enthalten, die im Browser nicht auffallen, aber im Übersetzungseditor schon (z.B. vergessene spitze Klammern in Tags)
- HTML kann fixe Eingabefelder enthalten -> Länge der Übersetzung beachten
- HTML kann pop-ups enthalten, Textinhalte befinden sich IN einem Tag und müssen unter Umständen freigegeben werden
- Manche Texte (z.B. mit Scripten erstellte) werden nicht im Übersetzungseditor angezeigt

# XML

- Für XML muss eine Filterdatei (Einstellungsdatei, INI, SDLFILETYPE) erstellt werden, damit der Editor Tag von Text unterscheiden kann
- **Text zwischen Tags** kann ein- oder ausgeblendet werden
- **Text innerhalb von Tags** kann ein- oder ausgeblendet werden
- Text kann anhand einer **Bedingung** ein- oder ausgeblendet werden
- XML kann Elemente in HTML enthalten, die gesondert markiert werden müssen (Snippet Marker Trados 2007, Cascading Filters memoQ5)

# Sonstiges

AutoCAD – PDF – TXT – PHP…

# AutoCAD Zeichnungen

- AutoCAD Dateien (DWG-Format) können im Austauschformat DXF gespeichert werden
- DXF kann ins TXT Format konvertiert werden um dann in die Übersetzung zu gehen.
- Rückkonvertierung TXT -> DXF -> DWG
- Teilweise ist es einfacher, eine PDF zu erstellen, die Übersetzung per Kommentar einzufügen und in der DWG direkt nachzutragen
- AutoCAD erlaubt es auch, Text in Excel Tabellen zu hinterlegen, die für die Übersetzung genutzt werden können

# PDF

- PDF ist eigentlich kein bearbeitbares Format
- Die Übersetzungsprogramme behelfen sich mit PDF Konvertern
  - Trados 2009 -> Solid Converter, produziert Word Dateien
  - memoQ -> Extraktion des Textes ohne Layout, produziert TXT Dateien
- Aus einem Scan entstandene PDFs sind KEIN Text, sondern Bilder, die kein Tool übersetzen kann. Hier muss zunächst mit einem OCR (optical character recognition) Programm aus dem Bild ein Text (Word) gemacht werden

# PHP

- PHP Dateien sind Textdateien, bei denen der bearbeitbare Text meist zwischen Anführungszeichen steht.
- Je nach Übersetzungsprogramm ist es möglich, den Text zwischen Anführungszeichen zu extrahieren