Dateiformate und Umwandlungsstrategien

Gebräuchliche Dateiformate für Texte im World Wide Web

Generell gilt: Verweise auf Texte, z.B. in digitalen Bibliotheken, lassen auch ohne explizite Kennzeichnung des jeweiligen Dokumenttyps vorab eine sofortige Bestimmung des zu Grunde liegenden Dateiformats zu. Gleiten Sie nämlich mit dem Mauszeiger über den betreffenden Link, erscheint in der Statusleiste am unteren Bildschirmrand des Browsers neben der Adresse (URL), unter der der Text erreichbar ist, auch eine Extensionsangabe zum Dateityp (manchmal kann diese freilich etwas versteckt eingeschachtelt sein). Wollen Sie nun PDF- oder Word-Dateien (etc.) nicht an Ort und Stelle öffnen, sondern zunächst einmal herunterladen, stehen Ihnen auf diese Weise die nötigen Informationen, die bei großen Dateien eine längere Online-Ladezeit des betreffenden Dokuments vermeiden helfen, zur Verfügung.

Die wichtigsten Formate, in denen Texte im Web präsentiert werden, seien hier – unter Nennung der jeweiligen Erweiterung – noch einmal kurz beschrieben:

- ***.html** (bzw. ***.htm**): das gebräuchliche Format für Webseiten; schreibt man das Kürzel aus (*Hypertext Markup Language*), werden bereits wesentliche Erscheinungsformen des Internets greifbar: allgemeine Verweisungsstruktur und Fundierung in einer Auszeichnungssprache (beim Herunterladen von Webseiten kann übrigens eine Wahl zwischen verschiedenen Speichermethoden getroffen werden: Webseite komplett, Webseite nur HTML oder Textformat)
- ***.xml**: die kennzeichnende Endung für Dokumente, die auf der *Extensible Markup Language* bzw. deren verschiedenen Dokumentgrammatiken beruhen; im Browser zugänglich durch die Verwendung spezieller Stylesheets (bzw. durch die Transformation in andere Dokumenttypen)
- ***.pdf**: das Portable-Document-Format, das von der Firma *Adobe* auf der Grundlage der Drucker- und Seitenbeschreibungssprache *Postscript* entwickelt wurde und demzufolge die seiten- und elementgenaue Darstellung von Büchern und Dokumenten aller Art (inklusive Grafiken, Formeln, Sonderschriften, etc.) ermöglicht; anzumerken ist allerdings, dass es sich bei PDF einerseits um ein proprietäres Format handelt (mit dem auch Zugriffsrechte der Nutzer/innen eingeschränkt werden können), und dass andererseits ein weiter gehendes Markup der Texte, wie es moderne wissenschaftliche Webeditionen grundsätzlich erfordern, nicht möglich ist
- ***.djvu**: relativ neues Format, in dem Texte und Bücher aller Art als durchsuchbare Images zur Verfügung gestellt werden können; im Gegensatz zu PDF nicht proprietär gebunden
- ***.txt**: gleichsam das roheste aller Textformate, die Speicherung als reiner Text, wodurch indes eine beliebige Weiterverwendung des Textes in anderen Formatzusammenhängen gesichert ist
- ***.rtf**: das sog. »Rich Text Format«, das ein Öffnen des Textes durch jede beliebige Textverarbeitung ermöglicht
- ***.doc**: das Dateiformat von *Microsoft Word*
- ***.odt**: die Textvariante des mittlerweile ISO-zertifizierten Open-Document-Formats, ab Version 2 das Dateiformat von *OpenOffice.org Writer* (XML-basiert, mit einer spe-

ziellen, Speicherplatz sparenden Kompression; die Vorgängerversion dieses Formats – kenntlich an der Endung ***.sxw** – ist übrigens auch noch nicht gänzlich entschwunden)

Neben diesen bekannten Dateitypen seien pauschal noch die diversen Ebook-Formate erwähnt, in denen Texte zum Download angeboten werden: das *Palm*-Format (***.pdb**), das *MS-Reader*-Format (***.lit**), das *Rocket-eBook*-Format (***.rb**), und mehrere andere. Will man die so kodierten Texte nicht nur unterwegs auf den entsprechenden elektronischen Ausgabegeräten lesen, muss natürlich ein spezieller Reader auf dem PC installiert sein.

Möglichkeiten der Umwandlung von Dateiformaten

Immer wieder einmal steht man vor der Notwendigkeit, Texte aus einem Format in ein anderes zu transformieren (zur Durchführung textstatistischer Untersuchungen mit verschiedenen Textanalyse-Tools, zum Import in spezialisierte Anwendungen beispielsweise im Information-Retrieval-Bereich, zur einheitlichen Präsentation in einem speziellen Publikationsrahmen, zur Sicherung von Textinhalten aus einem nicht mehr unterstützten Dateiformat heraus oder auch nur zur Erleichterung der Druckfunktion).

Nachstehend sollen daher einige der zur Verfügung stehenden Möglichkeiten zur Umwandlung von Dateiformaten bzw. Dateitypen kurz aufgeführt werden:

- *Überführung von PDF- in Text-Dateien:* Es sind grundsätzlich zwei Vorgehensweisen möglich: **Erstens**: Sie wählen im Menüpunkt *Datei* die Funktion *Als Text speichern*, den Rest erledigt der *Adobe Reader* selbst (was aber, je nach Voreinstellungen und Größe der betreffenden Datei, einige Zeit dauern kann). **Zweitens**: Sie markieren den kompletten Text der PDF-Datei (Achtung: in der PDF-Datei muss der Anzeigemodus *Fortlaufende Seiten* gewählt sein, sonst kann es passieren, dass nur eine einzelne Seite »ankommt«), kopieren diesen in die Zwischenablage und fügen ihn in einen geöffneten Texteditor ein. Wird der so entstandene Neu-Text nun mit der passenden Zeichenkodierung (UTF-8 oder auch ANSI) abgespeichert, ist eine Ausgangsbasis für die weitere Verarbeitung des Textes in Anwendungen, die das TXT-Format erfordern, hergestellt. Grafische Elemente oder das genaue Seitenlayout der Vorlage gehen bei beiden Vorgehensweisen allerdings verloren. (Statt in einen Editor lässt sich der markierte und kopierte PDF-Text natürlich auch in eine beliebige Textverarbeitung einfügen und dort im eigenen (oder einem anderen) Dokumentformat bzw. im Rich Text Format abspeichern.)
- *Umwandlung eines Textdokuments ins PDF-Format:* Verwendet man als Textverarbeitung *OpenOffice.org Writer* (oder *StarOffice Writer*), ist die Umwandlung denkbar einfach, da *OpenOffice* über eine eigene Exportfunktion nach PDF verfügt. Bedenkt man zudem, dass *OpenOffice* die verschiedensten Dateitypen öffnen kann, so ist über diese Anwendung eine generelle Möglichkeit des Transfers aus Fremdformaten ins PDF-Format gegeben. Der gängigste Weg der Umwandlung läuft allerdings über die Installation eines speziellen Tools, das den PDF-Export bewerkstelligt (z.B. *PDF Creator*, auf der Kurs-CD enthalten). Zur Umwandlung eines geöffneten Textes rufen Sie in der Textverarbeitung die Druckfunktion auf; dort findet sich auf der Liste der Drucker auch das PDF-Tool wieder, das nun statt des Druckers gewählt werden muss. Nach Betätigen des Druck-Befehls wird das Textdokument dann in eine PDF-Datei umgewandelt.

- *Abspeichern von Texten im HTML-Format:* Textverarbeitungen bieten den Nutzerinn-en prinzipiell verschiedene Speicheroptionen, so auch die Möglichkeit, den geöffneten Text im Webseiten-Format zu speichern. Allerdings wird dabei meist auch viel überflüssiger Code produziert, der den Stilkriterien wohlgeformter HTML-Files geradezu Hohn spricht. Insbesondere ist dies bei *Word* der Fall, *OpenOffice.org* liefert hier zwar deutlich bessere Ergebnisse, ist aber auch nicht gerade perfekt. Dave Raggett's Tool *HTMLTidy* (das auch in etlichen Webeditoren wie z.B. *HTML-Kit* oder *tsWebEditor* eingebaut ist) kann hier Abhilfe schaffen; von diesem freien Tool sind inzwischen auch Versionen mit grafischer Benutzeroberfläche herunterladbar (erwähnenswert v.a. *HTML Trim* und *TidyUI*). Übrigens können markierte und kopierte Texte immer auch in HTML-Editoren eingefügt und dort über die erhalten gebliebenen mehr oder minder rudimentären Formatierungsmerkmale hinaus weiter bearbeitet werden.
- *Erzeugen von PDF-Dateien aus HTML-Files:* Mit Hilfe eines Tools wie *HTMLDOC* (auf der Kurs-CD enthalten) lassen sich aus herunter geladenen oder selbst edierten HTML-Dateien navigierbare PDF-Dateien (unter Erhalt der ursprünglichen Link-Struktur) generieren. Erstreckt sich ein HTML-Text über mehrere (Unter-)Seiten, stellt auch dies kein Hindernis dar. Hat man mit *HTMLDOC* aus HTML-Files nun ein PDF-Dokument erzeugt, lassen sich übrigens einige der Probleme, die das Drucken von Webseiten herauf beschwören kann, relativ wirkungsvoll umgehen.

Generelle Informationen zu Dateikennzeichnungen

http://www.wotsit.org/