PDF/A Competence Center

Das bessere PDF/A-1b

Die Idee hinter einem „besseren" PDF/A-1b ist, dass dieser Conformance Level um die Vorteile von Unicode aufgestockt wird. Unicode ist in PDF/A-1a Pflicht, gegen die freiwillige Verwendung in PDF/A-1b spricht nichts.

Conformance Level in PDF/A

Zu PDF/A-1 sind zwei Conformance Level erschienen, die sich in den Anforderungen und in den gebotenen Funktionen unterscheiden.

PDF/A-1b: Level B Conformance (Basic)

Der Level B stellt die Mindestanforderung an PDF/A dar. Hier liegt der Schwerpunkt auf der Wiedergabetreue (*rendered visual appearance*).

PDF/A-1a: Level A Conformance (Accessible)

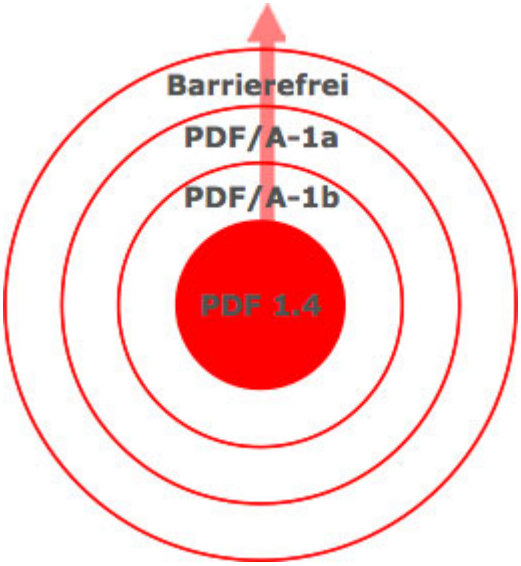
Level A Conformance bedeutet die Obermenge von PDF/A-1b. Zusätzlich zu Level B bietet Level A die folgenden Merkmale, welche für die Accessibility/Zugänglichkeit von Inhalten wichtig sind:

+ Tagged PDF
+ Strukturbaum (Hierarchie)
+ Sprach-Angabe
+ Unicode Mappings

Accessibility bedeutet, dass Inhalte (Text, aber auch Bilder und Grafiken) barrierefrei auch für sehbehinderte Rezipienten zugänglich ist, zum Beispiel über Vorlese-Software. Außerdem ist über Accessibility auch die Wiederverwertung von Inhalten viel besser möglich, als bei konventionellem PDF, etwa über Textexport.

Die Ebenen der „Accessibility" im Überblick

Die Stufen der Zugänglichkeit lassen sich wie folgt schematisch darstellen: Die höchste Ebene ist barrierefreies PDF, das sich gut aus der darunter liegenden Stufe PDF/A-1a herstellen lässt. Jedes PDF/A-1a ist automatisch auch ein gültiges PDF/A-1b und alle basieren auf PDF 1.4 (die PDF-Version, die mit Acrobat 5 eingeführt wurde.)



Im Kern basieren PDF/A-1a und -1b PDF auf PDF 1.4. Nur PDF/A-1a-Dateien, die weiteren Maßstäben gerecht werden, sind auch barrierefrei.

Barrierefrei im Sinne des Gesetzgebers:
International ist die Barrierefreiheit (vor allem für Internet-Angebote; in zunehmendem Maße aber auch für andere Dokumente) in Richtlinien oder Gesetzen geregelt.

In Deutschland regelt eine Verordnung die Barrierefreiheit:
BITV (Barrierefreie Informationstechnik-Verordnung – http://www.gesetze-im-internet.de/bitv/).

In den USA ist Barrierefreiheit bereits seit Jahren im Gesetz verankert:
Section 508 Rehabilitation Act (http://www.section508.gov/).

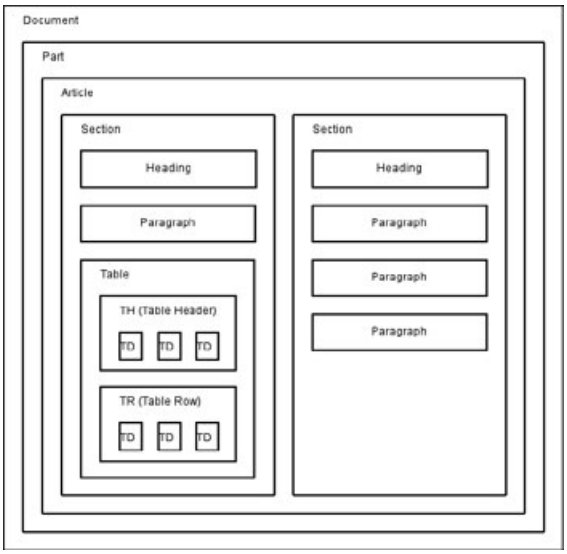
Was bedeutet tagged und structured PDF?

Tagged PDF:

Bei *tagged PDF* (PDF mit Markierungen) sind *Text-Spans* gekennzeichnet und mit ID versehen. Für Grafiken und Bilder ist *alternate text* vorgesehen (alternative Bildbezeichnung wie in HTML). *Replacement text* ist vorgeschrieben für Glyphen, die keine Buchstaben darstellen, zum Beispiel das Telefon-Symbol, oder ein Logo-Zeichen.

Structured PDF:

Die logischen Inhalte werden einem Strukturbaum zugewiesen. Die Strukturelemente sind vorgegeben (etwa *Document, Header, Article, List, Table* und weiteres mehr).



Structured PDF: Ein Beispielaufbau.

Zusätzliche Layout-Informationen

Diese zusätzlichen Informationen zum Layout betreffen *Artifacts*, die kein inhaltlicher Bestandteil sind. Dazu zählen unter anderem Elemente wie Seitenzähler, Kopfzeile und Fußzeile, die Fußnotenverwaltung, Hintergründe und Schnittmarken.

Angabe der Sprache

Wichtig ist auch die Angabe der Sprache, um die spätere, exakte Nutzung der Inhalte zu ermöglichen. Dies wird über das Attribut *lang* geregelt, als Beispiel „de-CH". Die Angabe muss sowohl im Strukturbaum als auch in den Abschnitten (*Spans*) hinterlegt sein.

Und warum geht das nicht immer?

Die Strukturinformation kennt nur ein Formatierer mit zusätzlichen Anweisungen (etwa: „dies ist eine Bildunterschrift"). Konvertierern von fertigen Seitendaten fehlen diese Information, so dass eine Zuordnung nicht möglich ist. Ein nachträgliches automatisches Interpretieren ist nicht sinnvoll und gemäß PDF/A-1a-Norm nicht erlaubt. Die nachträgliche Zuordnung muss also gegebenenfalls „händisch" durchgeführt werden, was recht aufwändig sein kann.

Unicode

PDF/A-1b muss nicht – aber kann – Unicode verwenden. Mit Unicode verfügt das PDF/A-1b über Vorteile für die weitere Nutzung.

Was ist Unicode?

Die Bezeichnung „Unicode" hat mehrere Bedeutungen. Unicode ist eine Organisation (www.unicode.org); ein System, das von einem weltweiten Expertenteam entwickelt wird; eine große Zeichentabelle; eine kleine Datenbank sowie die Beschreibung vieler Skripte.

A (U+0041)
„LATIN CAPITAL CHARACTER A"

Ж (U+0416)
„KYRILIC CAPITAL CHARACTER ZHE"

ڃ(U+0686)
„ARABIC CHARACTER TCHECH"

Einige Unicode-Zeichen aus verschiedenen Sprachen

Unicode-Bereiche

Unicode ist in zahlreiche Blöcke aufgegliedert, mit deren Hilfe die internationalen und historischen Zeichen systematisiert werden. Ein Blick auf einen Ausschnitt dieser Systematik stellt sich wie folgt dar:

Unicode-Block	Bereich	Anzahl der Zeichen
Basic Latin	U+0000 - U+007F	(128 Zeichen)
Latin 1 Supplement	U+0080 - U+00FF	(128 Zeichen)
…		
Currency Symbols	U+20A0 - U+20CF	(48 Zeichen)
…		
Miscellaneous Symbols	U+2600 - U+26FF	(256 Zeichen)
…		
CJK (asiat. Schriften)	U+4E00 - U+9FFF	(20991 Zeichen)

Eine komplette Liste der Blöcke ist auf der Unicode Website (http://www.unicode.org/Public/UNIDATA/Blocks.txt) zu finden.

Unicode und PDF

Standard-Encodings (max. 256 Zeichen) können einfach nach Unicode gemapped werden. Dazu zählen etwa WinAnsiEncoding und MacRomanEncoding. Eine Ausnahme stellen hier die Symbol Fonts dar.

Alle anderen Fonts müssen einen CMap-Eintrag (Zeichensatzzuordnung) haben, der auf Unicode-Zeichen verweist. Dazu zählen zum Beispiel Ligaturen:

<005F> <0060> <0061> [<00660066> <00660069> <00660066006C>]

Ligatur	Kombination	Einzelzeichen
ff	U+005F	U+0066 U+0066
fi	U+0060	U+0066 U+0069
ffl	U+0061	U+0066 U+0066 U+006C

Warum ist Unicode so wichtig?

Ohne Unicode gibt es kein Vorlesen von Text. Eine Textsuche kann nur über Unicode exakt und komplett ablaufen. Ohne Unicode ist auch keine inhaltliche Erschließung möglich.

Das heißt: Es ist unter anderem keine nachträgliche Seitentrenn-Logik, keine nachträgliche Indexierung und keine nachträgliche Adresslesung realisierbar.

Was ist nun das bessere PDF/A-1b?

PDF/A-1a kann nicht immer erfüllt werden. Das hängt unter anderem damit zusammen, dass das Ausgangsmaterial nicht für die Wandlung nach PDF/A-1a geeignet ist, oder dass der Aufwand für eine Konvertierung gemäß Level A zu hoch ist.

Aber auch das einfacher zu erstellende PDF/A-1b kann fast immer auch Unicode enthalten. Dafür sind einige Voraussetzungen zu bedenken. Bei der Erstellung oder Konvertierung müssen immer vernünftige Fonts verwendet werden. Es müssen gegebenenfalls über zusätzliche Mapping-Tabellen Unicode-Informationen beigestellt werden. Schließlich sollte auch der Ausgabe-Treiber richtig eingestellt sein.

Ein neuer Compliance Level würde den Anwendern mehr Übersicht über die Möglichkeiten geben. Das in Arbeit befindliche PDF/A-2 wird außer den beiden Compliance Level A und B auch die Variante PDF/A-2u („u" für Unicode) bieten.

Jörg Palmer, Compart/aoe