Improved PDF/A-1b

The idea behind an ‘improved’ PDF/A-1b is to enhance the conformance level by encompassing the advantages of Unicode. Unicode is obligatory only in PDF/A-1a, but that does not rule out the voluntary usage of Unicode in PDF/A-1b.

Conformance level in PDF/A

PDF/A-1 has two conformance levels that differ with regard to requirements and functions offered.

PDF/A-1b: Level B conformance (basic)

Level B conformance is the minimum requirement for PDF/A compliance. The focus here is on *reliable rendered visual appearance.*

PDF/A-1a: Level A conformance (accessible)

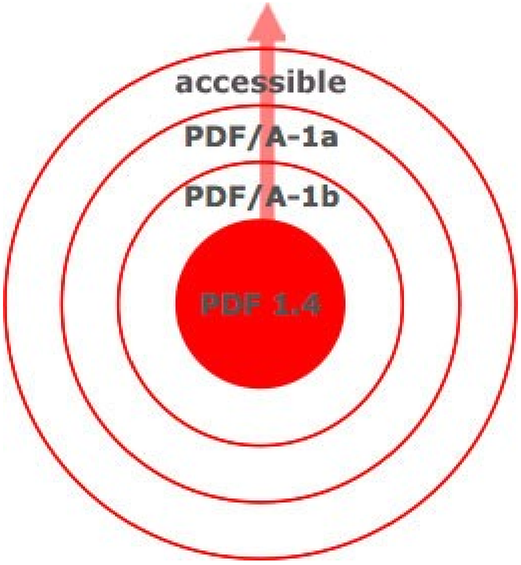
Level A conformance is the superset of PDF/A-1b. Over and above the features offered by level B, level A offers the following features that are important for providing accessible content:

+ Tagged PDF
+ Structure tree (hierarchy)
+ Language specification
+ Unicode mappings

Accessibility means that content (text, images, and graphics) is also accessible for visually impaired users via, for example, screen reader software. In addition, accessibility makes it easier to reuse content than with conventional PDF, thanks to functions such as text export.

Overview of accessibility levels

Accessibility levels can be schematically depicted as shown below. The top level is accessible PDF, which can be generated from the level below it – PDF/A-1a. Any PDF/A-1a-compliant file is also a valid PDF/A-1b file, and all variants are based on PDF 1.4 (the PDF version that was introduced with Acrobat 5).



Both PDF/A-1a and PDF/A-1b are based on PDF 1.4. Only PDF/A-1a files that meet additional criteria can be classified as 'accessible'.

Accessibility in law

Accessibility (primarily for Internet content, but increasingly for other documents, too) is regulated internationally by guidelines and laws.

In Germany, for example, the following decree governs accessibility: **BITV** (Barrierefreie Informationstechnik-Verordnung – http://www.gesetze-im-internet.de/bitv/).

In the US, accessibility has been grounded in law for several years already: **Section 508 Rehabilitation Act** (http://www.section508.gov/).

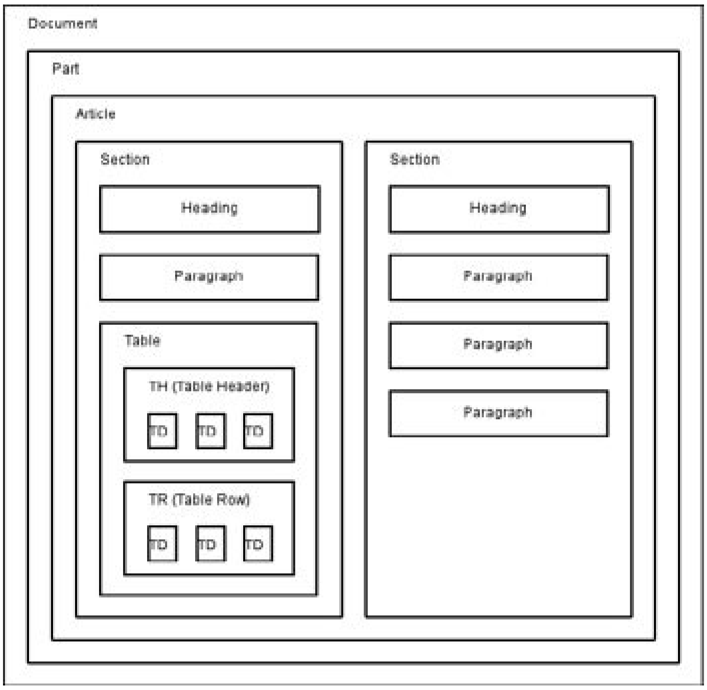
What do the terms 'tagged' and 'structured' PDF mean?

Tagged PDF:

Tagged PDF involves tagging text spans and giving them an ID. Alternative text can be given to graphics and images (alternative image captions as in HTML). Replacement text is prescribed for glyphs that do not represent letters, such as the telephone symbol or a logo character.

Structured PDF:

In the case of structured PDF, content is assigned to a structure tree. The structural elements are predetermined (Document, Header, Article, List, Table, and so on).



Structured PDF: An example layout

Additional layout information

This additional information refers to *artifacts* – elements that do not constitute part of the document content. This includes elements such as page numbers, headers and footers, footnotes, backgrounds, and crop marks.

Language specification

When ensuring that the content can be used unambiguously later on, the specification of the language is also important. This is governed by the *lang* attribute, with an entry such as ‘de-CH’ for Swiss German. The specification must be defined in the structure tree and in the *spans.*

Why is this not always possible?

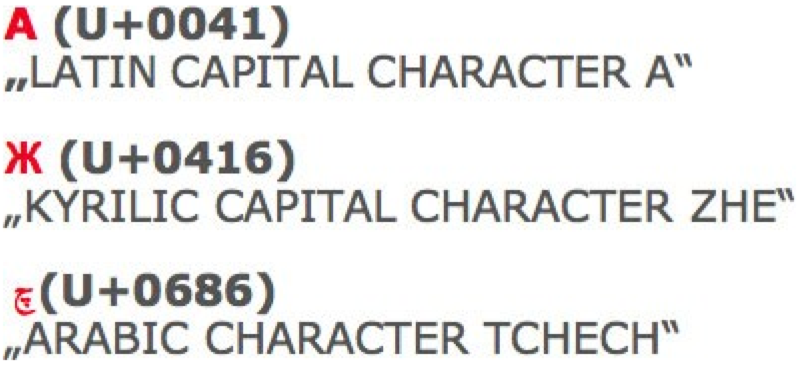
Structural information is only recognized by a formatter if additional information is available (for example, ‘this is a caption’). Converters of complete page data do not have this information, meaning that no assignment is possible. A later automatic interpretation is neither sensible nor permitted by the PDF/A-1a standard. This means that any later assignment must be carried out manually, which can involve a large amount of effort.

Unicode

PDF/A-1b files do not have to use Unicode, but they are permitted to do so. The use of Unicode brings advantages with regard to reusing PDF/A-1b files.

What is Unicode?

The term ‘Unicode’ has several meanings. Unicode is an organization (www.unicode.org), a system developed by a global team of experts, a large character table, a small database, and the description of many scripts.



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

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

چ(U+0686)
„ARABIC CHARACTER TCHECH“

Some Unicode characters from various languages.

Unicode character ranges

Unicode is subdivided into numerous blocks that help to arrange international and historical characters. The table below provides examples of these blocks and character ranges.

Unicode block	Character range	Number of characters
Basic Latin	U+0000 - U+007F	(128 characters)
Latin 1 Supplement	U+0080 - U+00FF	(128 characters)
…		
Currency Symbols	U+20A0 - U+20CF	(48 characters)
…		
Miscellaneous Symbols	U+2600 - U+26FF	(256 characters)
…		
CJK (Asian fonts)	U+4E00 - U+9FFF	(20,991 characters)

For a complete list of blocks, see the Unicode Website at http://www.unicode.org/Public/UNIDATA/Blocks.txt.

Unicode and PDF

Standard encodings (max. 256 characters) can be easily mapped to Unicode. This includes WinAnsi and MacRoman. Other than symbol fonts, all fonts must have a CMap entry (character set assignment) that references Unicode characters. This includes ligatures, for example:

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

Ligature	Combination	Individual characters
ff	U+005F	U+0066 U+0066
fi	U+0060	U+0066 U+0069
ffl	U+0061	U+0066 U+0066 U+006C

Why is Unicode so important?

Without Unicode, text cannot be read out loud by screen reader software. A precise and complete text search can only be ensured using Unicode. If Unicode is not used, content analysis is not possible.

This means that functions such as subsequent page break functionality, indexing, and address reading are not possible.

So what happens now with regard to an improved PDF/A-1b format?

PDF/A-1a cannot always be achieved. This is related to factors such as the source material not being suitable for conversion to PDF/A-1a or the effort involved in converting to level A being too high.

However, the PDF/A-1b format, which is easier to generate, can almost always contain Unicode. There are a few prerequisites for this. Sensible fonts must always be used when creating or converting the files. It may be necessary to provide Unicode information via additional mapping tables. Lastly, the output drivers must be set correctly.

A new compliance level would give users a better overview of their options. In addition to compliance levels A and B, PDF/A-2, which is currently being compiled, will also include PDF/A-2u (‘u’ for Unicode).

Jörg Palmer, Compart/aoe