07 - PDF in der Langzeitarchivierung

mit PDF/A Exkurs

Was ist PDF?

Das Dateiformat PDF ist wie ein digitaler Ausdruck. Ein PDF ist nie das Ursprungsformat, sondern wird aus einem bearbeitbaren Format erstellt. Es ist dem Anwender relativ leicht möglich mit einem herkömmlichen Reader Markierungen oder Kommentierungen vorzunehmen, ganz ähnlich wie es auch mit einem Papierausdruck ist. Tiefergehende Änderungen in den Inhalt einer PDF Datei, wie z.B. Seiten hinzuzufügen, zu entfernen, oder Textinhalt abzuändern, sind möglich, benötigen jedoch entsprechende Tools.

Ziel der PDF-Datei ist, dass der Inhalt sich unabhängig vom verwendeten Rechner und Betriebssystem für alle Anwender gleich darstellt. Daher wird das PDF in der Regel für bereits abgeschlossene Inhalte verwendet, die nicht mehr weiter bearbeitet werden. In der Langzeitarchivierung kommt dieses Format für textbasierte Inhalte sehr oft vor.

Was ist PDF/A?

Ein PDF/A ist ein PDF, für das eine Reihe von Restriktionen gilt.
So muss bei einem PDF/A-1 beispielsweise folgendes zwingend beachtet werden:

Nicht erlaubt:

- Einbettung von Audio- oder Video-Dateien
- LZW-Kompression
- Verschlüsselung
- Transparenz
- JavaScript
- Datei grösser als 2GB

Obligatorisch:

- Vollständigkeit der Daten. Dies beinhaltet die Einbettung aller im Dokument verwendeter Schriften und Bilder, und einiger vorgegebenen Metadaten.
- Visuelle Reproduzierbarkeit (z. B. bzgl. Schriftarten und Farbdarstellung)

PDF/A-1 basiert auf dem PDF 1.4-Standard.
Somit gehen alle ab nach PDF 1.4 eingeführten Funktionen wie die Verwendung mehrerer Ebenen verloren.

Im Gegensatz dazu basiert **PDF/A-2** auf dem PDF Standard der Version 1.7.
Daher sind hier folgende Funktionen erlaubt:

- LZW-Kompression (das Patent ist abgelaufen und entsprechend seit PDF/A-2 erlaubt)
- JPEG2000-Kompression
- Ebenen
- Transparenz
- Einbettung von Open Type Fonts

Außerdem bietet PDF/A-2 die Möglichkeit andere PDF/A-Dateien in die PDF/A-2 Datei einzubetten, so dass mehrere zusammengehörige PDF/A-Dateien gemeinsam archiviert werden können.

PDF/A-2 ersetzt oder verdrängt PDF/A-1nicht. Bereits erstellte PDF/A-1-konforme Dokumente bleiben für die Langzeitarchivierung weiterhin gültig und müssen nicht konvertiert werden.

Darüber hinaus ist oft von **Konformitätsebenen** a, b oder auch u die Rede:

a Das „a" steht hier für „accessible". Einem PDF/A-1a oder auch PDF/A-2a sind strukturelle Informationen hinzugefügt, z. B. ob es sich bei der Textstelle um eine Überschrift, um das Inhaltsverzeichnis oder einen Paragraphen handelt. Es kann auch deskriptive Informationen enthalten, beispielsweise weiterführende Metadaten zu einem enthaltenen Foto. Zudem muss jedes im PDF vorhandene Zeichen einem Unicode Zeichen zugeordnet werden können.
Mit diesen Erweiterungen ist die Barrierefreiheit gegeben, indem das PDF/A von einem Screenreader vorgelesen werden kann.

b Das „b" steht für „basic", dies bedeutet, dass das PDF/A-1b oder PDF/A-2b nur die visuell Reproduzierbarkeit gewährleistet.

u Das „u" steht für „unicode". Es wurde erst mit der Version 2 eingefügt und ist ein Kompromiss zwischen „a" und „b". Jedes Zeichen muss in Unicode abbildbar sein, Strukturinformationen und deskriptive Metadaten sind aber nicht notwendig.

PDF/A-3 basiert ebenfalls auf PDF 1.7 und wurde gleichzeitig mit PDF/A-2 entwickelt. Der einzige Unterschied zwischen diesen beiden Formaten ist, dass in ein PDF/A-3 jede Art von Datei eingebettet werden kann, völlig unabhängig vom Format. Rein technisch kann auch ein Virus oder eine Exe-Datei in ein PDF/A-3 eingebettet werden. Die Verwendung von PDF/A-3 ist entsprechend für die digitale Langzeitarchivierung ungeeignet.

Yvonne Tunnat
Projektmanagement Langzeitarchivierung
ZBW – Deutsche Zentralbibliothek für Wirtschaftswissenschaften
Leibniz-Informationszentrum Wirtschaft, Düsternbrooker Weg 120, D-24105 Kiel
Tel: +49 431. 88 14-610
y.friese@zbw.eu

Claire Röthlisberger-Jourdan
KOST - Koordinationsstelle für die dauerhafte Archivierung elektronischer Unterlagen
c/o Schweizerisches Bundesarchiv, Archivstrasse 24, CH-3003 Bern
claire.roethlisberger@kost.admin.ch

Yvonne Tunnat und Claire Röthlisberger-Jourdan sind Mitglieder der nestor AG Formaterkennung

Weitere Kurzartikel aus der Reihe „nestor Thema" finden Sie auf www.langzeitarchivierung.de -
der Webseite von **nestor – Kompetenznetzwerk Langzeitarchivierung**.