Dokumentqualität bei Batch-Konvertierung nach PDF/A

Roland E. Suri
ETH Zürich, Fachstelle Digitaler Datenerhalt, ETH-Bibliothek, roland.suri@library.ethz.ch, www.library.ethz.ch/Digitaler-Datenerhalt

1 Einführung

Änderungen in den Eigenschaften von Dateiformaten erschweren die langfristige Lesbarkeit digitaler Dokumente. Es ist deshalb unklar, ob digitale Inhalte langfristig erhalten werden können. Die Erhaltung digitaler Dokumente über eine Periode von 20 Jahren benötigt jedenfalls meist erhebliche Ressourcen (siehe Suri und El-Saad, in Vorbereitung). Digitale Archive empfehlen deshalb bestimmte Dateiformate, die sich gut zur Langzeitarchivierung eignen, wie das PDF/A-Format. Da nur wenige Kunden PDF/A-Dateien einreichen, könnten digitale Archive eingereichte Dateien selbst in das PDF/A-Format konvertieren.[1,2]

Das ETH Data Archive evaluierte deshalb drei Software-Tools für die Batch-Konvertierung gängiger Dateiformate nach PDF/A-1b:

- LuraDocument PDF Compressor
- Adobe Acrobat XI Pro
- Document Converter von 3-Heights

Das Testset bestand aus total 80 Dateien der Dateiarten JPEG, MS PowerPoint, PDF, PNG, MS Word, MS Excel, MSG und "Webseite".[3]

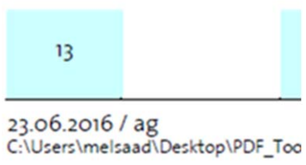
2 Beispiele für Konvertierungsfehler

MS Word

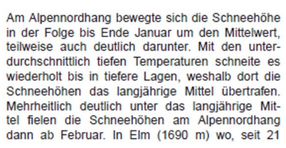
PDF/A-1b von 3-Heights

MS Word

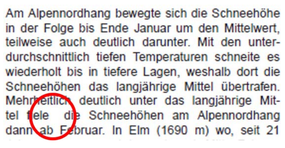
PDF/A-1b (alle drei Tools)



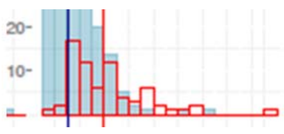
PDF 1.6



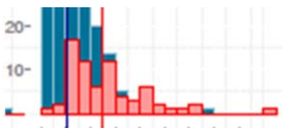
PDF/A-1b v. Acrobat



PDF 1.5



PDF/A-1b v. 3-Heights



3 Resultate

Die Programme stoppten manchmal die Stapelverarbeitung der Testdateien, sodass manuelle Eingriffe erforderlich waren (3-4 Stopps pro Software für die 80 Testdateien).

Die Konvertierungstools produzierten auch für unterstützte Inputformate manchmal keine Ausgabedatei: Drei (Adobe Pro) bis zu sieben (LuraTech und 3-Heights) PDF/A-1b-Dateien fehlten.

Von den produzierten PDF/A-1b Dateien wurden (je nach Validator) zwischen 5% und 20% der Ausgabedateien als ungültig beurteilt.

Die Qualität der konvertierten Dateien wurde untersucht, indem das visuelle Erscheinungsbild des Eingabedokuments mit dem des erzeugten PDF/A-1b-Dokuments auf einem Computerbildschirm verglichen wurde. Sorgfältige Sichtprüfung ergab, dass die Umwandlung in PDF/A-1b den Informationsgehalt in 11% der konvertierten Dateien beeinträchtigte. Es traten folgende Reproduzierbarkeitsfehler auf

- Verlorene Links
- Verlust anderer Dokumentinhalte (unlesbare Zeichen, fehlender Text, fehlender Dokumentteil)
- Aktualisierte Felder (Zeitstempel und Konvertierungsordner)
- Fehler in Vektorgrafiken
- Schreibfehler

4 Schlussfolgerungen

Batch-Konvertierung heterogener Dateien nach PDF/A-1b verursacht komplexe Probleme, deren Lösung eine aufwendige Bearbeitung einzelner Files benötigen würde. Auch bei erheblichem Zeitaufwand sind gewisse Qualitätsverluste unvermeidlich. Die Einhaltung der PDF/A Standards würde zudem die Mithilfe von PDF Experten erfordern. Archive sehen sich deshalb häufig gezwungen, auch schwach-standardisierte Dateiformate zu akzeptieren, und damit Lösungen zur Langzeiterhaltung auf später zu verschieben.

5 Referenzen

1. **Aussonderung digitaler Unterlagen und deren Archivierung im Bundesarchiv. Ein Leitfaden, Bundesarchiv 2010, Version 1.2, Zugriff am 20. Februar 2018, erhältlich bei: https://www.bundesarchiv.de/imperia/md/content/abteilungen/abtb/bbea/behoerdenleitfaden-v1.2-2010-08-27-internet.pdf**
2. **KOST (2016) "PDF/A: Produktereview batchtauglicher PDF/A-Konverter", Koordinationsstelle für die dauerhafte Archivierung elektronischer Unterlagen (KOST), Dateiname „konvert2pdfa_2016_v1.0_DE.pdf", erhältlich bei: http://kost-ceco.ch/cms/index.php?pdfa_konverter_de (Zugriff am 14. Juni 2016).**
3. **Suri R. E. (2017) "Lost in Migration: Document Quality for Batch Conversion to PDF/A", ETH Zürich, ETH Research Collection, (Datensammlung), erhältlich bei: http://doi.org/10.3929/ethz-b-000200243**