Herausforderungen in der digitalen Lanzeitarchiverung im Hinblick auf das PDF Format

Leistungsnachweis 4

von

Andrés Baumeler

Telefon: 076 443 04 71, E-Mail: andres@baumeler.dev

Alte Riedikerstrasse 5c, 8610 Uster

Betreuerin

Eliane Blumer

Universität Zürich, Philosophische Fakultät /
Zentralbibliothek Zürich

CAS Datenmanagement und Informationstechnologien

Inhaltsverzeichnis

1 Einleitung **2**

2 Grundlagen **3**

2.1 PDF Format und PDF Standard 3

2.2 PDF für die Langzeitarchivierung 4

3 Herausforderungen **6**

3.1 Sicherstellung der Authentizität 6

3.2 Erschliessung und Zugänglichkeit 7

3.3 Kurzlebigkeit von Technologien 8

4 Schluss **9**

4.1 Einschätzung 9

4.2 Ausblick 10

Abkürzungen **11**

Literaturverzeichnis **12**

Kapitel 1

Einleitung

Die rasante digitale Transformation hat zu einem exponentiellen Anstieg der digitalen Dokumente geführt, die in verschiedenen Bereichen wie Verwaltung, Wissenschaft, Bildung und Wirtschaft genutzt werden. Angesichts der Notwendigkeit einer langfristigen Aufbewahrung und Archivierung dieser Dokumente ist das Portable Document Format (PDF) zu einem der bevorzugten Formate geworden. PDF bietet eine plattformunabhängige und konsistente Darstellung von Dokumenten auf verschiedenen Systemen und Geräten.

In Kapitel 2 wird das PDF-Format und der PDF-Standard vorgestellt, sowie einen Überblick über die Anwendung von PDF in der digitalen Langzeitarchivierung gegeben.

Im Kapitel 3 wird auf eine Auswahl von Herausforderungen eingegangen welche sich durch die Verwendung des PDF-Formates in der Langzeitarchivierung ergeben.

Im Kapitel 4 findet eine Einschätzung der vorgestellten Herausforderungen statt. Weiter wird versucht abzuschätzen wie sich diese Herausforderungen in näherer Zukunft verändern werden.

Kapitel 2

Grundlagen

2.1 PDF Format und PDF Standard

Das PDF Format wurde von Adobe Systems entwickelt und im Jahr 1993 vorgestellt. In den folgenden Jahren wurde das Format immer bekannter und zeigte Potenzial für die digitale Langzeitarchivierung. Im Jahr 2002 wurde eine Arbeitsgruppe innerhalb der International Organization for Standardization (ISO) gegründet, um ein standard Format für die digitale Langzeitarchivierung zu entwickeln. Vertreter einer Vielzahl von US-amerikanischen Verbänden und Bundesbehörden, darunter AIIM (Association for Information and Image Management), NPES (Association for Suppliers of Printing, Publishing and Converting Technologies) und NARA (National Archives and Records Administration), trafen sich mit Experten aus dem Bibliothekswesen (Harvard University Libraries, Library of Congress), dem Justizsystem (Administrative Office of the United States Courts) und der Industrie (einschließlich Adobe Systems und Kodak). Am 1. Oktober 2005 wurde der PDF/A-Standard unter der Bezeichnung ISO 19005-1:2005 (PDF/A-1) veröffentlicht. Es war das weltweit erste standardisierte Dateiformat für die digitale Langzeitarchivierung. Seitdem sind drei weitere Teile des Standards erschienen: PDF/A-2 (Open-Type Schriftarten und digitale Signaturen) und PDF/A-3 (einbetten von original Dateien im PDF). 2020 erschien mit PDF/A-4 (Einbetten von 3D Objekten) der jüngste Teil. [1]

Der PDF/A Standard ist als mehrteilige Serie angelegt. Das bedeutet, nachfolgende Versionen verdrängen vorhergehende Versionen nicht. Der Standard PDF/A-1 ist weiterhin gültig auch wenn mittlerweile PDF/A-2 und PDF/A-3 erschienen sind. Für die Teile PDF/A-1 bis PDF/A-3 gibt es jeweils drei Konformitätsstufen. Diese Stufen werden mit A, B, oder U beschrieben. Ein PDF ist konform auf Stufe A wenn alle Anforderungen des PDF/A Standards erfüllt sind. Das beinhaltet unteranderem auch, dass der Text innerhalb des Dokuments in der natürlichen Lesereihenfolge angeordnet sein muss. Stufe B sagt aus, dass ein Dokument eindeutig reproduziert werden kann. Ein Dokument auf Stufe B muss aber im Gegensatz zu Stufe A nicht 100% Textextraktion und Durchsuchbarkeit bieten. Dokumente auf Stufe U garantieren, dass sämtlicher enthaltener Text zu standart Unicode Character Codes gemappt werden kann. [1], [2]

2.2 PDF für die Langzeitarchivierung

Für die digitale Langzeitarchivierung ist es wichtig, dass der Inhalt von Dokumenten jederzeit und auf jeder Plattform gleich aussieht. Neben der Reproduzierbarkeit müssen Dokumente auch in Zukunft, unter Umständen auf einer heute noch nicht existierenden Plattform geöffnet werden können. Da sich nur schwer abschätzen lässt, wie sich die heute etablierten Betriebssysteme entwickeln, kann nicht vorausgessetzt werden, dass die heute verwendeten Programme auch in Zukunft noch zur Verfügung stehen oder auf den in Zukunft etablierten Betriebsystemen lauffähig sind.

Der PDF/A Standard und das darin beschriebene Dateiformat bieten für diese Herausforderungen eine Lösung. Der PDF/A Standard baut auf dem PDF Standard. Um die Reproduzierbarkeit sicherzustellen, stellt der PDF/A Standard zusätzlich zum PDF Standard weitere Anforderungen an ein Dokument. So muss im PDF/A Standard etwa sichergestelt sein, dass Ressourcen welche für die Darstellung des Inhalts, wie etwa Schriftarten, in die Datei eingebetet sein müssen. Der PDF/A Standard setzt zudem voraus, dass Inhalte nicht verschlüsselt sein dürfen und dass gewisse Java Script Funktionen nicht verwendet werden dürfen. Weiter wird vom PDF/A Standard auch

verlangt, dass Metadaten im Extensible Metadata Platform (XMP) Format eingebettet werden. Damit das PDF Format auch in Zukunft noch interpretiert werden kann, ist es wichtig, dass die Spezifikation für das PDF Format offen zugänglich ist. Die ISO stellt sicher, dass die Standards öffentlich Verfügbar sind. Durch diese Zugänglichkeit können Programme zum Erstellen und Betrachten von PDF Dokumenten auf für zukünftige Platformen und Betriebssysteme entwicklet oder angepasst werden. So ist sichergestellt, dass PDF Dokumente auch noch betrachtet werden können, wenn ein Hersteller die Entwicklung eines Programs einstellt.[5], [3]

Kapitel 3

Herausforderungen

In einer Studie der Schweizer Nationalbibliothek wurden Schweizer Gedächtnisinstitutionen zu Herausforderungen der digitalen Langzeitarchivierung befragt. Dabei wurden verschiedene Bereiche genannt bei welchen das PDF Format Unterstützung bieten kann. Die Teilnehmenden Institutionen nannten unter anderem Herausforderungen bei der Sicherstellung der Authentitiztät, der Erschliessung und Zugänglichkeit sowie bei der kurzlebigkeit von Technologien. Nachfolgend wird auf diese drei Bereiche eingegangen und erläutert wie der PDF Standard bei diesen Herausforderungen helfen kann. [4]

3.1 Sicherstellung der Authentizität

Der PDF Standard bietet mittels digitalen Signaturen eine Möglichkeit die Authentizität eines Dokuments zu prüfen. Eine digitale Signatur erlaubt es aber noch nicht die Authentizität eines Dokuments sicherzustellen. Dazu werden Prozesse innerhalb der erstellenden Organisation benötigt welche sicherstellen, dass die digitalen Signaturen korrekt erstellt und interpretiert werden. Der PDF Standard wirkt hier unterstützend und bietet Möglichkeiten welche von anderen Prozessen und Software innerhalb der Institution verwendet werden können.

Das Problem der Sicherstellung der Authentizität wird noch verschärft, wenn Dokumente archiviert werden sollen, welche nicht innerhalb der eigenen

Organisation erstellt wurden. Für die archivierung solcher Fremddokumente stellt die digitale Signatur ebenfalls nur ein Bauteil dar. Es werden weiter Prozesse und Schnittstellen benötigt um die Authentizität sicherzustellen.

3.2 Erschliessung und Zugänglichkeit

Eine zentrale Herausforderung besteht in der Gewährleistung der langfristigen Lesbarkeit und Interpretierbarkeit von PDF-Dokumenten. Durch die kontinuierliche Weiterentwicklung des PDF-Formats und die Einführung neuer Versionen besteht das Risiko, dass ältere Versionen möglicherweise nicht mehr von zukünftiger Software und Hardware unterstützt werden. Diese Herausforderung kann nur teilweise durch das PDF Format gelöst werden. Durch die ISO wird sichergestellt, dass der Standard langfristig weiterentwickelt wird und offen verfügbar ist. Denoch liegt es hier in der Verantwortug der Institution geeignete Prozesse einzuführen um die verwendeten Formate und Software regelmässig auf ihre Eignung zu Prüfen und gegebenfalls zu aktualisieren. Nur so kann sichergestellt werden, das archivierte Inhalte auf den gewünschten Platformen und Schnittstellen zugänglich bleiben.

Bei der Erschliessung bietet das PDF Format Unterstützung indem es die Erfassung von Metadaten direkt im Dokument im XMP Format erlaubt. Dadurch kann eine langfristige Erschliessung der Inhalte durch das PDF Format unterstützt werden. Die Erfassung der Metadaten in einem File reicht aber nicht aus um eine wirklich Langfristige Erschliessung sicherzustellen - dazu werden geeignete Prozesse und Software welche diese Prozesse unterstützen in den Institutionen vorausgesetzt welche das PDF Format für die digitale Langzeitarchivierung verwenden.

Der PDF/A-3 Standart erlaubt es jedes Fileformat in ein PDF Dokument einzubetten. Diese Funktion kann etwa verwendet werden um die original Datei aus welcher ein PDF generiert wurde in ein PDF einzubetten. Der PDF Standard macht aber keine Angaben darüber ob die eingebeteten Dateien für die digitale Langzeitarchivierung geeignet sein müssen. Durch ein fehlendes Verständnis des PDF/A-3 Standards kann es vorkommen, dass eine Institution fälschlicherweise davon ausgeht, dass Dokumente für die langfristige

Archivierung geeignet sind, nur weil diese im PDF/A-3 Format vorliegen. Dadurch kann es einerseits bei der Erschliessung zu Problemen kommen, da die eingebetteten Dokumente nicht sauber erschlossen werden, sonder nur das Umfassende PDF Dokument. Andererseits kann es durch das Einbetten von Dokumenten auch zu Problemen bei der Zugänglichkeit kommen wenn in Zukunft eingebetete Dateien nicht mehr interpretiert werden können.

3.3 Kurzlebigkeit von Technologien

Grundsätzlich ist das PDF Format eine Technologie wie andere Dateiformate auch in läuft deshalb auch Gefahr in Vergessenheit zu geraten. Was das PDF Format aber von anderen Formaten unterscheidet ist, dass hier die ISO dahintersteht und sicherstellt, dass die Spezifikation des Formats auch in Zukunft noch öffentlich verfügbar ist und weiterentwickelt wird.

Institutionen können aber trotz Einsatzt des PDF Formats von kurzlebigen Technologien betroffen sein, etwa wenn für die Erstellung von PDF Dateien spezifische Softwar vorausgesetzt wird oder wenn die verwaltung von PDF Dateien in einem Archiv- oder Dokumentenmanagementsystem erfolgt.

Hier kann der Einsatz von PDF Formaten, insbesonders PDF/A insfoern abhilfe schaffen, dass eine allfällige Migration eines Archiv- oder Dokumententenmanagemensystems vereinfacht wird, da PDF Dokumente von einer vielzahl verschiedener Programme gelesen und bearbeitet werden können. Wenn konsequent auf PDF/A gesetzt wird sollte es auch keine externen Abhängigkeiten geben welche eine Migration erschweren können.

Der PDF Standard und das PDF Format sind aber auch für diese Herausforderung keine direkte Lösung. Auch hier wirkt der PDF Stahdard nur unterstützend und setzt etablierte Prozesse innerhalb der Institutionen voraus, um eine digitale Langzeitarchivierung zu gewährleisten.

Kapitel 4

Schluss

4.1 Einschätzung

Der PDF/A Standard und das darin beschriebene PDF Dateiformat sind eine geeignete Lösung für die digitale Langzeitarchivierung von Inhalten welche sich auf geeigntete Art und Weise als Dokument reproduzieren lassen, auch wenn es organisatorische oder technische Herausforderungen bei der Verwendung des PDF Formats entstehen können. Mit der ISO steht eine Organisation hinter dem Format welche in der Lage ist die notwendige Langfristigkeit und Offenheit des Standards sicherzustellen welche für die Verwendung in der Archivierung vorausgesetzt wird. Der PDF/A Standard ist aber nur ein Bauteil einer Lösung zur digitalen Langzeitarchivierung. Neben der Wahl der Dateiformate gehören auch eine sorgfältige Planung und eine laufende Überprüfung der getroffenen Annahmen zu einer seriösen digitalen Langzeit Archivierung.

Meiner Einschätzung nach ist das PDF/A Format eine solide Entscheidung für die digitale Langzeitarchivierung von Inhalten welche in Dokumentenform repräsentiert werden können. Mindestens genau so wichtig wie die Wahl des Dateiformats, sind aber auch das Etablieren von geigneten Prozessen innerhalb der Institution um die Langzeitarchvierung sicherzustellen.

Weiter ist es unerlässlich, die verschiedenen PDF Standards genau zu verstehen und nicht automatisch davon auszugehen, das alleine durch eine

Konvertierung zu PDF/A-1b eine langfristige Lesbarkeit und Reproduzierbarkeit gewährleistet ist.

4.2 Ausblick

Die digitale Langzeitarchivierung muss sich stetig anpassen, da immer neue Inhalte mit neuen Formaten aufbewahrt werden müssen. Das bedeutet auch, dass der PDF Standard weiterentwicklet werden muss um zukünftigen Entwicklungen gerecht zu werden. In der nahen Zukunft werden immer mehr und verschiedenartige Daten produziert werden welche als PDF aufbewahrt werden sollen. Hierbei wird der PDF Standard weitherhin einen wertvollen Beitrag leisten können. Durch die Veröffentlichung von weiteren Standards wie PDF/A-4, PDF/X-4 und PDF/VT wurden in der näheren Vergangenheit bereits neue PDF Standards veröffentlicht um sich ändernden Anforderungen gerecht zu werden. Durch Forschritte in der Forschung wird es wohl dazu kommen, dass der Standard um weitere Kompressionsalgorithmen und Bildformate ergänzt wird um mit den steigenden Dateigrössen Schritt zu halten.

Abkürzungen

PDF	Portable Document Format
PDF/A	Portable Document Format for Archiving
ISO	International Organization for Standardization
XMP	Extensible Metadata Platform

Literaturverzeichnis

[1] 06 2023. URL https://en.wikipedia.org/wiki/History_of_PDF.

[2] 06 2023. URL https://www.pdf-tools.com/de/pdf-knowledge/all-about-pdf-a-long-term-archiving/.

[3] Caroline R. Arms and Carl Fleischhauer. Digital formats: Factors for sustainability, functionality, and quality. *Archiving Conference*, 2005.

[4] Daniel Burda, Angelina Dungga Winterleitner, and Beat Estermann. Digitale langzeitarchivierung in der schweiz. Technical report, Berner Fachhochschule, 09 2017.

[5] Association for Digital Document Standards e.V. Pdf/a in a nutshell 2.0. info@pdfa.org, 2013.