Entwicklung einer webbasierter Applikation zur Bearbeitung von PDF Dateien

Bachelorarbeit zur Erlangung des akademischen Grades
Bachelor of Science
im Studiengang Technische Informatik
an der Fakultät für Informations-, Medien- und Elektrotechnik
der Technischen Hochschule Köln

vorgelegt von:	Janina Schroeder
Matrikel-Nr.:	11132206
Adresse:	Laurentiusweg 10 50321 Brühl janina_jessika_jelena.schroeder@smail.th-koeln.de
eingereicht bei:	Prof. Dr. Chunrong Yuan
Zweitgutachter:	Prof. Dr. René Wörzberger

Köln, 04.03.2024

Bachelorarbeit

Titel: Entwicklung einer webbasierter Applikation zur Bearbeitung von PDF Dateien

Gutachter:

- Prof. Dr. Chunrong Yuan

- Prof. Dr. René Wörzberger

Zusammenfassung: Für die Bachelorarbeit habe ich eine Open Source offline Webseite zur Bearbeitung von PDF Dateien im Firefox Browser programmiert. Seit Adobe den PDF Standard entwickelt hat, tauchten zahlreiche meist kostenpflichtige PDF Anwendungen, um PDF Dateien zu bearbeiten auf dem Markt auf. Ich habe den Markt an PDF Programmen analysiert und diese mit meiner Webapplikation verglichen. Daraufhin beleuchte ich den aktuellen Stand der Technik des PDF Standards. Im späteren Verlauf erkläre ich die Implementierung meiner Webapp und meine Erfahrungen mit anderen Browsern, sowie auf MacOS, Linux, Android und iOS. Die Javascript Libraries PDF.js und PDF-LIB sind das tragende Fundament meiner PDF Webapp. Die PDF Webapp vereint alle Funktionalitäten, die man für gängige PDF Bearbeitung benötigt. Man kann PDFs lesen, splitten, mergen, erstellen, sowie mit Texten, Bildern, Geometrie und Zeichnungen versehen. Am Ende diskutiere ich, was man hätte besser machen können, welche Funktionalitäten fehlen und welche Features in Zukunft noch geplant sind.

Stichtwörter: PDF Bearbeitung, Adobe, Javascript, Vue JS 3, auf PDF zeichnen, Splitten, Mergen, PDF.js, PDF-LIB

Datum: 04. März 2024

Inhaltsverzeichnis

Tabellenverzeichnis V

Abbildungsverzeichnis VI

Abkürzungsverzeichnis VII

1 Grundlagen 2
1.1 PDF Vorstellung . . . 2
1.2 Wichtigste Features . . . 3
1.2.1 What You See Is What You Get (WYSIWYG) . . . 3
1.2.2 Fonts . . . 3
1.2.3 Bilder . . . 4
1.2.4 3D-Daten . . . 4
1.2.5 Kommentare . . . 4
1.2.6 Verweise . . . 5
1.2.7 Formulare . . . 5
1.2.8 Inkrementelles Update . . . 6
1.2.9 Kompression . . . 6
1.2.10 Ebenen . . . 7
1.2.11 Portfolio . . . 7
1.2.12 JavaScript . . . 7
1.3 PDF Dateiformate . . . 8
1.3.1 PAdES . . . 8
1.3.2 PDF/X . . . 9
1.3.3 PDF/A . . . 11
1.3.4 PDF/E . . . 12
1.3.5 PDF/H . . . 13
1.3.6 PDF/VT . . . 13
1.3.7 PDF/UA . . . 14
1.3.8 PDF/R . . . 14
1.3.9 Durchsuchbares PDF . . . 14
1.4 PDF Dateiversionen . . . 15
1.4.1 PDF 1.0 . . . 15
1.4.2 PDF 1.1 . . . 15

1.4.3 PDF 1.2 . . . 16
1.4.4 PDF 1.3 . . . 16
1.4.5 PDF 1.5 . . . 16
1.4.6 PDF 1.4 . . . 16
1.4.7 PDF 1.6 . . . 17
1.4.8 PDF 1.7 . . . 17
1.4.9 PDF 2.0 . . . 17
1.5 PDF Implementierung . . . 17
1.5.1 PostScript . . . 18
1.5.2 Adobe Imaging Model . . . 19
1.5.3 Dateiformataufbau . . . 20
1.5.4 Implementierung von Fonts . . . 22
1.5.5 Implementierung von Transparenzen . . . 23
1.6 PDF Sicherheitsaspekte . . . 23
1.6.1 Digitale Unterschrift . . . 23

2 PDF Programme auf dem Markt 25
2.1 Die Firma Adobe Systems Incorporated . . . 25
2.2 Aktueller Stand von Forschung und Technik . . . 25
2.3 Freie PDF Programme und Onlinedienste . . . 25
2.3.1 PDFCreator . . . 25
2.3.2 LibreOffice . . . 26
2.3.3 OpenOffice . . . 26
2.3.4 ghostscript . . . 26
2.4 Kostenpflichtige PDF Programme und Onlinedienste . . . 26
2.4.1 Adobe Acrobat . . . 26
2.4.2 Adobe Acrobat Pro . . . 26
2.4.3 Onlinetools von Acrobat . . . 27
2.4.4 Acrobat Distiller . . . 27
2.4.5 Microsoft Word . . . 27
2.5 PDF zu Word Konvertierung . . . 27
2.6 PDF zu Latex Konvertierung . . . 28
2.7 Relevanz von PDF in verschiedenen Marktbranchen . . . 28
2.8 Rolle von PDF in der Druckvorstufe und Designbranche . . . 29
2.8.1 Farbdarstellung . . . 30
2.8.2 Preflight . . . 30
2.8.3 Fontformate . . . 30

3 Open Source PDF Web App 31
3.1 Problemstellung und Anforderungen . . . 31
3.2 Konzept und Methodik . . . 31

3.3 Funktionalität der PDF Web App 31
3.4 Bedienung der PDF Web App 31
3.5 Implementierung der PDF Web App 31
3.6 Testdurchführung der PDF Web App 31
3.6.1 Funktionale User Tests 31
3.6.2 Stress Tests 31

4 Diskussion und Kritik 32

Literatur 34

Anhang 38

Tabellenverzeichnis

Abbildungsverzeichnis

Abkürzungsverzeichnis

AES Advanced Encryption Standard. 17

BITV Barrierefreie-Informationstechnik-Verordnung. 14

CAD Computer-Aided Design. 4, 12

CAdES CMS Advanced Electronic Signatures. 8

CEPS Cisco Enterprise Print System. 30

CID Character Identifier Font. 4, 16, 30

ETSI European Telecommunications Standards Institute. 8

GDI Graphics Device Interface. 18

ICC International Color Consortium. 9, 10, 13, 16, 29

ISO International Organization for Standardization. 2, 6, 8–17

MIME Multipurpose Internet Mail Extension. 13, 17

OCR Optical Character Recognition. 14, 26

OPI Open Prepress Interface. 16, 30

PAdES PDF Advanced Electronic Signatures. 8, 28

PCS Profile Connection Space. 29, 30

PDF Portable Document Format. 25

PDL Page Description Language. 17–19

PPD PostScript Printer Description. 18, 26

ppi Pixels per inch. 20

RIP Raster Image Processor. 17–19

RLE Run Length Encoding. 6, 7

WCAG Web Content Accessibility Guidelines. 14

WYSIWYG What You See Is What You Get. 2, 3, II

XAdES XML Advanced Electronic Signatures. 8

XFA XML Forms Architecture. 6, 16, 17

XML Extensible Markup Language. 6, 17, 21

XMP Extensible Metadata Platform. 11, 21

ZSA Zeitstempel-Anbieter. 24

Einleitung

Motivation

Aufbau der Arbeit

1 Grundlagen

1.1 PDF Vorstellung

Das PDF Dateiformat steht für Plattformunabhängigkeit, Hardwareunabhängigkeit, Konsistenz in Formatierung und Layout und soll ein möglichst originalgetreues Druckergebnis liefern. Der Leser soll ein PDF Dokument immer nach dem Prinzip WYSIWYG (What You See Is What You Get) in der Form betrachten und ausdrucken können wie vom Ersteller des Dokuments festgelegt.

PDF wurde 1993 von der Firma Adobe Systems Incorporated veröffentlicht und ging aus dem 1991 von Adobe-Mitbegründer John Warnock gestarteten „Project Camelot"hervor. Ziel dieses Projektes war, ein Dateiformat für elektronische Dokumente zu kreieren, sodass diese Anwendungsprogramm, Betriebssystem und Hardware unabhängig originalgetreu wiedergegeben werden können. **scheeberger**, [1] Anfangs war der Adobe Reader kostenpflichtig und PDF war für einen langen Zeitraum ein proprietäres Dateiformat, welches offengelegt im PDF Reference Manual von Adobe dokumentiert ist. Die Spezifikation von PDF ist seit 1993 kostenlos einsehbar. [2] Die International Organization for Standardization (ISO) übernahm PDF 2007 in den Standardisierungsprozess und seit der Veröffentlichung von PDF Version 1.7 am 1. Juli 2008 gilt PDF als Offener Standard als ISO 32000-1:2008. [1], [2] Vorher war PDF ein proprietäres Dateiformat von Adobe. Der Begriff Offener Standard bezeichnet einen Standard, der für alle Teilhaber am Markt besonders leicht zugänglich, weiterentwickelbar und einsetzbar ist. Das bedeutet, dass der Standard von einer gemeinnützigen Organisation eingeführt, veröffentlicht, weiter bearbeitet wird und gleichmäßige Einflussnahme aller interessierten Parteien ermöglicht. [3] Im gleichen Jahr publizierte Adobe eine Public Patent Licence zum ISO Standard 23000-1, also PDF Version 1.7, die royalty-free Rechte für Adobes gesamte Patentsammlung einräumt, um PDF Implementierungen zu programmieren, verkaufen und verbreiten. [2] Royalty-free bedeutet hierbei, dass Computerherstellerfirmen pro verkauftes Endgerät keine Lizenzgebühr (royalties) bezahlen müssen, sowie keine fixe Jahrespauschale. [4] Heute wird PDF seit 2006 von der PDF Association weiterentwickelt. [1]

1.2 Wichtigste Features

Die in den Unterkapiteln genannten Operationen auf dem PDF-Dateiformat beziehen sich hauptsächlich auf Adobe Acrobat-Werkzeuge. PDFs können Texte, Tabellen, Bilder, Pfade, Links, Buttons, Formulare, Audio-, Videoelemente und Funktionen enthalten. Rich Media PDFs können interaktiven Inhalt enthalten, der eingebettet oder verlinkt werden kann. Sie enthalten Bilder, Audio, Video oder Buttons, z.B. als digitalen Katalog. [2]
In PDFs werden alle Informationen als nummerierte Objekte gespeichert. Objekte können zu Gruppen kombiniert werden. Der aktuelle Farbmodus im Dokument kann in andere Farbmodi konvertiert werden. Fonts und Bilder sollten grundsätzlich immer eingebettet werden.
Um die Navigation innerhalb eines PDF Dokuments zu erleichtern kann man anklickbare Inhaltsverzeichnisse und miniaturisierte Seitenvorschauen verwenden. Optional ist eine Gliederung mit hierarchischer Baumstruktur in Form von Lesezeichen möglich, mit der der Betrachter leichter durch das Dokument geführt werden kann.
PDF-Dateien enthalten grundsätzlich Metadaten. Bei Metadaten oder Metainformationen handelt es sich um strukturierte Daten, die sich auf Merkmale anderer Daten beziehen. Beispiele für Metadaten sind Name, Titel der Datei, Autor, Stichwörter zum Inhalt, das Datum der Speicherung. PDF-Dateien können Dateianhänge enthalten, die geöffnet und im lokalen Dateisystem abgespeichert werden können. [2]

1.2.1 WYSIWYG

Ein PDF-Dokument hat ein festes Layout und eine feste Anzahl von Seiten. Unabhängig von der Software mit der das Dokument angezeigt wird oder mit welcher Hardware es ausgedruckt wird bleiben alle Elemente auf den Seiten immer exakt an derselben Position. Alle Layout- und Formatierungsangaben stammen aus der Erstellungsanwendung. Bei der Konvertierung von Dokumenten mit variablem Layout zu PDF, wie z.B. .txt-Dateien oder HTML muss der Inhalt auf die vorhandenen Seiten und den verfügbaren Platz verteilt werden.

1.2.2 Fonts

Jedes Textzeichen ist ein abstraktes Symbol und ein Schriftzeichen beruht auf eine graphische Darstellung. Eine Schriftart ist in PDF als Objekt enthalten. Die Schriftart als Objekt kann mit Werkzeugen in Acrobat bearbeitet werden. Der Text muss ausgewählt werden und es können folgende Operationen angewendet werden: Farbveränderung in

RGB, Transparenzen, Verschiebung, Löschen, Skalierung, Verzerrung, Spiegelung, Drehung, Beschneidung und Ersetzung. In Acrobat Pro kann der gesamte Text pro Seite in Pfade konvertiert werden. PDF unterstützt Type-1 Fontformate, Multiple-Master-Fonts, TrueType-Fontformate, OpenType-Fontformate, Dfonts, Character Identifier Font (CID)-codierte Fonts und Composite-Fonts. Falls die Schriftart nicht im Dokument eingebettet wurde, wird die Schriftart aus der Ursprungsdatei möglicherweise durch eine Ersatzschrift des Benutzersystems im PDF-Programm substituiert. [5]

1.2.3 Bilder

Generell sollte für das Bearbeiten von Bildern ein externes Bildbearbeitungsprogramm verwendet werden, z.B. Adobe Photoshop oder das kostenlose Gimp. Dafür kann für die Bearbeitung von Photoshop das Bild aus Acrobat Pro extrahiert werden aus dem PDF und später wieder ersetzt werden. Vektorgrafiken als Pfadobjekte und Bilder Pixelobjekte können nach Auswahl verschoben, gelöscht, skaliert, verzerrt, gespiegelt, gedreht, die Deckkraft verändert, beschnitten oder ersetzt werden. [5] Bilder können in Acrobat neu berechnet werden, d.h. ihre Auflösung wird neu berechnet. Niedrig aufgelöste Bilder behalten ihre Auflösung. Ein guter Neuberechnungsalgorithmus heißt bikubische Neuberechnung. Bei Schwarzweißbildern kann eine Neuberechnung zu unschönen Artefakten führen. [6] Generell führt eine Neuberechnung der Auflösung in Bildbearbeitungsprogrammen zu besseren Ergebnissen als in Acrobat. Etwaige Pixelbearbeitungen wie Tonwertkorrekturen oder das Schärfen von Bildern kann ausschließlich in Bildbearbeitungsprogrammen vorgenommen werden.

1.2.4 3D-Daten

PDFs mit 3D-Inhalten bestehen aus dem U3D-Flächenmodell oder dem BREP/Flächenmodell PRC. Sie werden vorwiegend bei der Visualisierung von Computer-Aided Design (CAD)-Daten verwendet. Beide Formate können im Adobe Reader angezeigt, animiert, geschnitten und gemessen werden. Viele Drittanbieter PDF-Reader und die PDF-Viewer im Browser können eingebettete 3D-Daten meist nicht darstellen. Einige CAD-Programme ermöglichen einen 3D-PDF-Export oder Import. [1]

1.2.5 Kommentare

Ein Kommentarobjekt, das mit ein oder mehreren Dokumentenseiten verlinkt ist, besteht aus 2 technisch separaten Bausteinen. Zum einen werden Kommentare durch ein grafisches Element auf den zugehörigen Seiten symbolisiert, zum anderen wird der Kommentarinhalt in einem rechteckigen Kommentarbereich dargestellt. Ein Anwender

kann die Darstellung des Kommentarobjekts je nach Geschmack modifizieren. Unüblicherweise kann ein Kommentar sogar als Video-Kommentar abgespielt werden. Die wichtigsten Kommentartypen sind Notizzettel, Textmarkierung, Stempel, Wasserzeichen, Textboxen, Formen, Freihand-Markierung, Audio, Video und 3D-Illustrationen. Kommentare können optional mit ausgedruckt werden. [7]

1.2.6 Verweise

Technisch gesehen sind Verweise oder Hyperlinks spezialisierte Kommentare ohne Symboldarstellung. Auf der Seite wird ein Ausschnitt zur Platzierung des Verweises gewählt, der über einem Inhaltselement (Text oder Bild) liegt. Der Verweis zeigt auf eine Seite oder Seitenbereich im geöffneten Dokument, eine andere PDF-Datei, eine E-Mailadresse oder URL. Man kann sogar Zielobjekte mit einem im gesamten Dokument eindeutigen Namen einstellen. [7]

1.2.7 Formulare

In PDFs kann man Formularfelder erstellen vom Typ Textfeld, Kontrollkästchen, Auswahlknopf, Kombinationsfeld, Auswahlliste, Schaltfläche, Barcode- oder Unterschriftsfeld. Ein Formularfeld ist ein Objekt zum befüllen und speichern von Felddaten. Die unterschiedlichen Formularfeldtypen weisen verschiedene Eigenschaften in Bezug auf Interaktivität und Gestaltung auf und jedes Feld hat einen nur einmal vorkommenden Namen im gesamten Dokument. Mit dem eindeutigen Namen können Namensgruppen realisiert werden. Durch eine hierarchische Struktur mittels Teilnamen die mit einem Punkt voneinander getrennt sind mit dem äußersten Gruppennamen zuerst können Felddaten noch besser und logischer beschrieben und strukturiert werden. Jedes Feldobjekt geht Hand in Hand mit einem Widget, welches ein spezielles Kommentarobjekt zur Steuerung darstellt. Diese Widgets stehen für Werte oder Zustände der Felder und sind dafür verantwortlich, dass man Formulare im PDF-Dokument mit dem Computer, Tablet oder Smartphone ausfüllen kann. Außerdem ist es möglich unsichtbare Feldobjekte, die ohne das Widget platziert werden können, zu erstellen, um das PDF-Programm anzusprechen. Häufiger werden mehrere Widgets verknüpft mit einem Feldobjekt verwendet. [7] In der Praxis werden Formulare in einem Grafik- oder Layoutprogramm gestaltet und als PDF exportiert. Um elektronisch ausfüllbare Formulare zu verwenden müssen zusätzlich in Acrobat Formularfelder auf die entsprechenden Stellen platziert werden. Falls ein Listenfeld verwendet wird, sollte man eine Schrift für die Listeneinträge im PDF einbetten. Formulare können einen druckbaren und nicht druckbaren Teil enthalten. In der Druckvorstufe müssen vor dem Druck alle Formularfelder eliminiert werden, damit alle Schriften eingebettet werden können.

[5] Es gibt 2 verschiedene Möglichkeiten von PDF Formularen: AcroForms (Acrobat Forms) oder Adobes proprietäre XML Forms Architecture (XFA) forms, welche mit Version 2.0 von der ISO als veraltet markiert wurden. XFAs Haupterweiterungen zu Extensible Markup Language (XML) sind rechnergestützte, aktive Tags und sein Datenformat ist kompatibel mit anderen Systemen, Anwendungen und Technologiestandards. [8] AcroForms unterstützen das abschicken (submit), zurücksetzen und importieren von Daten. Die submit-Aktion transferiert die Namen und Werte eines ausgewählten interaktiven Formularfelds zu einer vordefinierten URL.

1.2.8 Inkrementelles Update

Die ursprüngliche Version einer PDF-Datei bleibt erhalten, während das inkrementelle Update die Änderungen im Dokument enthält. Professionelle PDF-Programme können wie eine Versionsverwaltung jede geänderte Version des Dokuments laden. Bei einfacheren PDF-Programmen wird lediglich die letzte Version geladen. Bei Verwendung von inkrementellen Updates kann man digital unterschriebene Dokumente ändern ohne dass die Unterschrift ungültig wird, da die Dokumentversion mit der digitalen Unterschrift ein andere Version ist als die nachträgliche Änderungen. Dabei muss die digitale Unterschrift als inkrementelles Update gespeichert werden, sonst würde sie verfallen bei nachträglicher Dokumentenänderung unabhängig von der Änderungsart. Folglich sollten mehrfach signierte Dokumente ebenfalls mit der Option inkrementelles Update gespeichert werden. [7] Alle Änderungen in einem inkrementellem Update werden am Ende der Datei normalerweise nach dem Katalog angehängt. Hingegen werden gelöschte Seiten aus dem Katalog, nicht aber aus der PDF-Datei entfernt. Folglich steigt der Speicherbedarf einer PDF-Datei pro inkrementellem Update. [5]

1.2.9 Kompression

PDF-Dateien sind komprimiert und haben üblicherweise einen Bruchteil der Größe des Ursprungsformats oder von Bilddateien. Dies wird durch Vermeidung von Redundanzen und Erhöhung der Entropie (Zeichendichte) und Weglassen von Informationen bewerkstelligt. Kompressionsalgorithmen sind nicht auf bestimmte Dateiformate beschränkt. [5] In PDF können die folgenden Kompressionsalgorithmen für Bilder verwendet werden: IP, Run Length Encoding (RLE), JPEG, JPEG2000, CCITT und JBIG2. Eine hohe Bildqualität im PDF bedeutet eine größere Datei. Faktoren, die die Bildqualität beeinflussen, sind Breite x Höhe des Bildes, Farbtiefe, Farbraum und die Kompressionsmethode. [7] Im Allgemeinen gibt es verlustlose und verlustbehaftete Kompression. Die Kompressionsalgorithmen RLE, die genauso effiziente LZW, Flate-Komprimierung, ZIP und CCITT gehören zur verlustfreien Kompression. Zur

verlustbehafteten Kompression gehören JPEG, JBIG2 und JPEG2000. Außerdem ist es möglich eine Datenreduktion durch Neuberechnung zu erzielen. Hierbei wird das verlustbehaftete Downsampling verwendet und führt häufig zu nicht befriedigenden Ergebnissen. Es gibt die eher im Ergebnis mangelhafte Kurzberechnung, durchschnittliche und bikubische Neuberechnung. Neuberechnung in Adobe Photoshop führen zu besseren Ergebnissen als in Acrobat Distiller. [5] PDF-Dateien können außerdem zur Weboptimierung serialisiert werden, sodass Teile des PDFs während des Ladevorgangs dargestellt werden. Liegen unkomprimierte Elemente im Dokument vor, werden diese beim Speichern durch die Flate-Komprimierung, die auch den ZIP-Algorithmus verwendet, komprimiert.

1.2.10 Ebenen

Ebenen werde auch als Gruppen mit optional sichtbarem Inhalt bezeichnet und stellen quasi mehrere Inhaltsschichten auf einer einzelnen PDF-Seite, wobei jede alle Seiten im Dokument beliebig viele Ebenen enthalten kann. Jede Ebene kann PDF-Inhalt sozusagen gruppieren wie eine Seitenschicht und Bearbeitung von Inhalten auf einer Ebene wirkt sich nur auf diese Ebene aus. Man kann Inhalte auch mehreren Ebenen zuordnen oder keiner Ebene. Ebenen können ein- und ausgeblendet werden, ihre Reihenfolge verändert werden, gesperrt werden, zusammengeführt werden, aus anderen PDF-Dateien importiert werden und für unterstützende Dateiformate von Adobeprogrammen, z.B. Photoshop, Illustrator oder InDesign, exportiert werden. Zusätzlich kann man eine Ebenennavigation aufbauen mit Hilfe von Links und Lesezeichen, um Ebenensichbarkeit zu steuern. [9]

1.2.11 Portfolio

Ein Portfolio bezeichnet eine Datei bestehend aus anderen Dateien, die kein Hauptdokument enthält, sondern lediglich eine Pseudo-Seite. Diese Pseudo-Seite wird von Portfolio inkompatiblen PDF-Programmen angezeigt. Außerdem können andere PDF-Dateien und andere Dateiformate im PDF-Hauptdokument eingebettet werden. [7]

1.2.12 JavaScript

In PDF kann man Ereignisse Aktionen zuordnen, d.h. bei Eintreffen eines Ereignisses wird automatisch eine Aktion ausgeführt. Ein Ereignis ist eine bestimmte Statusänderung von Objekten oder eine interaktives Anwenderereignis. Dabei kann man als Aktion JavaScript-Code aufrufen, dessen Aktion mit Lesezeichen, Verweisen, Seiten und Dokumentereignisse verknüpft ist. Auf Formularfeldern kann ebenfalls JavaScript

angewandt werden. [7] Diese JavaScript-Erweiterung für Acrobat ist eine proprietäre Technologie von Adobe. Viele andere nicht-Adobe PDF-Programme bieten keine Unterstützung für JavaScript. [2]

1.3 PDF Dateiformate

PDF hat zahlreiche Dateiformate, von denen die meisten standardisiert wurden, hervorgebracht. Jedes Dateiformat ist einem individuellen Anwendungsbereich zugeordnet und adressiert spezifische Industriebranchen: PAdES für elektronische Signaturen, PDF/X für den professionellen Druck, PDF/A für die Archivierung, PDF/E für den Ingenieurbereich, PDF/H für das Gesundheitswesen, PDF/VT für den Druck mit variablen Daten, PDF/UA für Barrierefreiheit, PDF/R für gescannte Dokumente und Durchsuchbare PDFs für Stichwortsuche. Im Folgenden stelle ich jedes Format vor und beschreibe seine speziellen Merkmale.

1.3.1 PAdES

PDF Advanced Electronic Signatures (PAdES) ergänzt den Funktionsumfang um Werkzeuge mit denen man elektronische Signaturen erzeugen, anpassen und prüfen kann. Folglich soll dieses Dateiformat die Integrität, Authentizität, Verbindlichkeit und Rechtssicherheit von digital signierten PDF-Dokumenten herstellen. Es wurde vom European Telecommunications Standards Institute (ETSI) veröffentlicht und 1999 in PDF 1.3 eingeführt und basiert auf der ISO 32000-1 Spezifikation. Nachfolgend wurde dessen Konzept weiterentwickelt. PAdES erweitert PDF um kryptographische Techniken und ermöglicht sichtbare und unsichtbare Signaturen. Elektronisch signierte Dokumente können über lange Zeit ihre Gültigkeit behalten, selbst wenn die teilhabenden kryptografischen Algorithmen kompromittiert wurden. PAdES implementiert verschiedene Signaturformate, wie CMS Advanced Electronic Signatures (CAdES) und XML Advanced Electronic Signatures (XAdES), unterstützt Zeitstempel und die Validierung des Zertifikatwiderrufsstatuses. Bei der Validierung des Zertifikatwiderrufstatuses wird die Gültigkeit der Signatur untermauert, obwohl das Zertifikat des Unterzeichners widerrufen wurde. Zertifikatbasierte Signaturen sollen die Identität des Unterzeichners und die Unabänderlichkeit des Dokuments sichern. Eine zertifizierte PDF-Datei ermöglicht die Umsetzung bestimmter Nutzungsrechte, wie eingeschränkte Bearbeitung, Ausfüllen von Formularen oder gesperrtes Drucken. Eine elektronische Unterschrift kann mit dem Programm Adobe Acrobat Sign erstellt werden. [10]

1.3.2 PDF/X

Speziell für den simpleren Datenaustausch in der Druckvorstufe und der professionellen Druckindustrie wurde PDF/X (Exchange) als ISO 15930:2001 entwickelt. Dieser erste 2001 entwickelte Dateiformatstandard beschreibt die speziellen Eigenschaften von Druckvorlagen und vereinfacht die Datenübermittlung von der Design-Agentur und Druckvorstufe bis zum finalen Druck. Besonderen Wert wurde darauf gelegt, dass in offenen Dateiformaten aus Layoutprogrammen keine Informationen über Farbe und Schrift verloren gehen und einer Verfälschung im Druckergebnis vorgebeugt werden kann. [11] Die Entwicklung von PDF/X zielt auf eine Verminderung von Druckfehlern und Mehraufwand in der Druckerei. In der Umsetzung bedeutet das, dass Elemente, die sich nicht sinnvoll drucken lassen, z.B. Video und Audio, nicht berücksichtigt werden. Beschnitt, Farbangaben und verwendete Schriften sind u.a. für den Druck notwendig und sollten verwendet werden. Qualitätsanforderungen, die sich auf bestimmte Druckverfahren beziehen, sind nicht implementiert, sondern werden abstrakter definiert. Besondere Qualitätsanforderungen liegen vor allem im Zeitungsdruck, Akzidenzdruck oder Bilderdruck vor. Des weiteren werden schwarze Schrift oder Linien im Drucker durch 3 oder 4 Farben zusammengesetzt und fehlende Schriften werden häufig durch den Font Courier kompensiert. Im Druck sollten keine verlustlosen Kompressionsalgorithmen für Bilder verwendet werden wie JPEG, da Artefakte auftreten können. Ebenso gibt es keine automatischen Einstellungen für passende Auflösungen von Vollton-, Halbton- oder Strichbildern. Vielmehr geht es darum, Grundvoraussetzungen für den Druck sicherzustellen, wie z.B. ob der richtige Farbraum gesetzt wurde oder korrekte Einstellungen für Überdrucken und Überfüllen vorliegen. PDF/X-kompatibel bezeichnet die Eigenschaft von Dokumenten, dass sie ohne vorherige Prüfung von der Druckerei direkt verwendet werden können. [12]

Es gibt verschiedene Varianten von PDF/X, die jeweils einen verbesserten Farbspielraum ermöglichen.

PDF/X-1a

In der a-Version sind lediglich CMYK und Sonderfarben möglich. Farben können nicht auf Grundlage von International Color Consortium (ICC)-Profilen definiert werden. Transparenzen, [12] Ebenen, Verschlüsselung, JavaScript, LZW-Kompression, Formularfunktionen und interaktive Elemente sind nicht implementiert. Dieser Standard wurde von ISO 15930-1:2001 auf ISO 15930-4:2003 überarbeitet. Lediglich in der überarbeiteten Version, die die Version von 2001 ersetzt, werden auch Sonderfarben unterstützt. [13]

PDF/X-2

Die 2. Variante ist als ISO 15930-6:2003 erschienen und garantiert dominante Voraussetzungen zum farbigen Qualitätsdruck wie Farbmanagement, CMYK- und Sonderfarbdaten in beliebiger Kombination. [13]

PDF/X-3

Zusätzlich erweitert Version 3 als ISO 15930-3:2002 [13] um die Farbräume RGB und LAB, sowie ICC-Profile. Möglicherweise wird in der Druckvorstufe der im Dokument eingestellte Farbraum in CMYK umgewandelt. Es findet eine automatische Tranzparenz- und Ebenenreduzierung statt. [12]

PDF/X-4

Transparenzen, Ebenen und Graustufen können in dieser PDF/X-Variante als ISO 15930-7:2008 [13] verwendet werden, wodurch sie für das Bedrucken von Textilien besonders gut geeignet ist. [12]

PDF/X-5

PDF/X-5 wird wurde als ISO 15930-8:2010 als vorletzter Standard des PDF/X-Formats verabschiedet und inkludiert externe Elemente und Multichannel-ICC-Profile. [13]

PDF/X-6

Der letzte PDF/X-Standard als Version 6 wurde in ISO 15930-9:2020 offenbart und basiert auf dem PDF 2.0 Standard. In dieser Version sind neben maßgeblichen Neuerungen für die heutigen Print-Anforderungen Lockerungen im Vergleich zu vorherigen PDF/X-Standards eingeführt worden. Die wichtigsten Neuerungen sind Parameter für Tiefenkompensierung, separate Ausgabebedingungen, DPart Metadaten, Informationen zu Sonderfarben mit CxF/X-4 und Mixing Hints. Zu den Lockerungen zählen die Möglichkeit von Notizen und grafischen Anmerkungen, strukturelle Aktionen, Formularfelder und digitale Signaturen. Außerdem besteht dieser Standard aus 2 Konformitätsstufen: PDF/X-6p zur Referenzierung von ICC-Profilen und PDF/X-6n für Multicolor-Profile. [13]

1.3.3 PDF/A

Das PDF/A Dateiformat (Archivable) wurde zur gesetzeskonformen Langzeitarchivierung von digitalen Dokumenten entwickelt und solche Dokumente sind zunächst schreibgeschützt. Der ISO-Standard definiert die Konformität der Form von Elementen wie Schriften oder Layout für eine Langzeitarchivierung. Dadurch ist die Lesbarkeit der Dokumente über lange Zeiträume gesichert und die Bedingungen einer revisionssicheren Archivierung gewährleistet. [14] Revisionssichere Archivierung bedeutet, dass gespeicherte Daten vor nachträglichen Modifikationen, Fälschung oder Manipulation geschützt sind. [15] Der Fokus in diesem Dateiformat liegt auf langfristige und einfache Speicherung der PDF-Dateien. Folglich ist die Einbettung von Audio und Video nicht implementiert, aktive Komponenten wie Links, sowie externe Ressourcen, wie Grafiken und Schriftarten werden nicht unterstützt, sondern müssen direkt eingebettet werden. Ebenso können Dokumente nicht verschlüsselt werden. Die Einbettung von Metadaten als Extensible Metadata Platform (XMP) wird unterstützt, was die Identifizierung und Suche von Dokumenten erleichtert. Es gibt einige Nachteile von PDF/A. Nicht alle Dokumente können problemlos in dieses Dateiformat umgewandelt werden, wie beispielsweise Dokumente mit Audio, Video oder JavaScript. Nach der Konvertierung zu PDF/A kann es zu Fehlern in der visuellen Darstellung kommen und die Dateigröße kann enorm werden, da alle Elemente direkt eingebettet werden müssen. [14]

PDF/A-1

Seit der ersten Version von PDF/A wurde es in die ISO-Norm übernommen worden als PDF/A-1 in ISO 19005-1:2005. [13] Die Originalversion stellt sicher, dass alle externen Quellen wie Schriften oder Bilder eingebettet sind, unterstützt digitale Signaturen und Hyperlinks. PDF/A-1 ist abwärtskompatibel. Es gibt 2 Qualitätsebenen von PDF/A-1: PDF/A-1b (Basic) und PDF/A-1a (Accessible). Die Basic Variante legt Wert darauf, dass Dokumente eindeutig visuell reproduzierbar sind und Accessible ist zusätzlich für Barrierefreiheit optimiert. Bei Accessible können Text und inhaltliche Struktur von einem Screenreader vorgelesen werden. [14] Des weiteren werden Tagged PDFs, Sprach-Angabe und Unicode Mappings unterstützt. [13]

PDF/A-2

Im Jahr 2011 wurde die PDF/A-2 Version als ISO 19005-2:2011 auf den Markt gebracht. Sie ermöglicht die Kompression von Grafikformaten mit JPEG-2000, Transparenzen, PDF-Ebenen, Portfolios, Object Level XMP Metadaten, Kommentartypen und Annotationen und digitale Signaturen. [13] PDF/A-1-Dateien können in PDF/A-2-Dateien

eingebunden werden. Es gibt 3 Varianten von PDF/A-2: PDF/A-2b (Basic), PDF/A-2u (Unicode-Textsemantik) und PDF/A-2a (Accessible). Basic gewährleistet das unveränderte Erscheinungsbild eines Dokuments und definiert die Mindestanforderungen. Die Unicode-Version ergänzt um Unicode-Unterstützung und Indexierung. Accessible setzt alle Anforderungen der ISO-Norm 19005-2 um. [14]

PDF/A-3

Ein Jahr später wurde PDF/A-3 im Standard ISO-19005-3:2012 veröffentlicht. Er basiert auf PDF 1.7 und ermöglicht die Einbettung dynamischer, zur Laufzeit interpretierbare Komponenten und Dateiformate. Gleichfalls definiert PDF/A-3 die Konformitätsebenen 3b, 3u und 3a. Die u-Variante bietet eine Vereinfachung in der Durchsuchbarkeit von Texten und das Kopieren von Unicode-Text. [13]

PDF/A-4

Viel später im Jahr 2020 wurde PDF/A-4 als ISO 19005-4:2020 herausgebracht. Dieser Standard basiert auf der PDF 2.0 Dateiversion. Sie spezifiziert 2 neue Konformitätsebenen PDF/A-4f für nicht-PDF/A konforme Dateianhänge und PDF/A-4e für Einbindung von 3D-Inhalten in den Formaten U3D oder PRC für den Engineering-Bereich. [13]

1.3.4 PDF/E

PDF/E (Engineering) gilt als international standardisiertes Austauschformat als Norm ISO 24517 für technische Dokumente und wird im Maschinenbau, in der Fertigung und im Baugewerbe für Fertigungspläne, Konstruktionszeichnungen oder technische Dokumentationen verwendet. Das PDF/E Dateiformat von 2008 ist speziell für das Ingenieurwesen entworfen und kann interaktive 3D-Elemente und Animationen darstellen. Im einzelnen können CAD-Dateien im 3D- und 2D-Format eingebettet werden. Die 3D-Elemente können im Dokument ausgeklappt oder gedreht werden. Metadaten und interaktive Funktionen wie Lesezeichen, Formulare oder Hyperlinks werden unterstützt. [11]

1.3.5 PDF/H

Das PDF/H (Healthcare) Dateiformat soll im Gesundheitswesen Patientendaten erfassen, austauschen und archivieren. Hierbei wird besonders Wert auf die Anforderungen des Datenschutzes in gesundheitsspezifischen Ämtern, Institutionen und Arztpraxen gelegt. PDF/H wurde 2008 kreiert, jedoch wurde es nicht in den Normierungsprozess der ISO eingebunden. [13] Es handelt sich eher um eine Best-Practice für die Umstellung von Papiergesundheitsakten auf PDF als e-Akte. Die Digitalisierung soll gesundheitsbezogene Daten strukturieren, verwalten und so präsentieren, dass Forscher*innen und Beschäftigte im Gesundheitswesen effizienter auf sie zugreifen können.

1.3.6 PDF/VT

Basierend auf PDF/X wurde PDF/VT als spezielles Austauschformat im variablen Datendruck (Variable Data) und Transaktionsdruck (Transactional Printing) im Jahr 2010 auf den Markt gebracht. [16] Wiederkehrende Elemente wie Texte, Grafiken oder Bilder sollen effizienter verarbeitet und an den Drucker übertragen werden können. [11] Variabler Datendruck bezeichnet ein digitales Druckverfahren bei dem einzelne Parameter von Printprodukten individuell variiert werden können, wobei das Grundlayout beständig bleibt. Folglich können große Mengen von Printprodukten mit Personalisierung hergestellt werden, z.B. Werbebriefe mit konstanten grafischen Elementen wie Firmenlogos und individuellen Namen der Kund*innen. Dadurch können Firmen ihr Corporate Identity-Layout behalten und ihre Kund*innen persönlicher ansprechen. Der Begriff Transaktionsdruck definiert das Drucken von Transaktionsdokumentationen wie Rechnungen, Mahnungen, Lieferscheine oder Quittungen im Waren- und Dienstleistungssektor. Herausstechend ist, dass PDF/VT große Mengen an variablen Daten in einer einzigen PDF-Datei speichern kann, wobei es immer einen Satz von statischen und variablen Daten gibt. Diese Vorgehensweise spart Zeit, Kosten und reduziert Fehler. Vorteilhafterweise werden ICC-Profile unterstützt. [16] Die erste Version als ISO 16612-1:2005 konnte sich auf dem Markt nicht durchsetzen. [13]

PDF/VT-2

Die zweite Version als ISO 16612-2:2010 implementiert die Verwendung externer grafischer Inhalte und das Streamen von mehrteiligen Multipurpose Internet Mail Extension (MIME)-Paketen in der Version PDF/VT-2s. Zum Lesen solcher Dokumente wird ein PDF/X-4- bzw. PDF/X-5- oder PDF/VT-konformer PDF-Reader benötigt. [13]

PDF/VT-3

Spezialisiert auf die Integration von variablen Daten (DPart-Metadaten) und den Transaktionsdruck ist die dritte PDF/VT Variante als ISO 16612-3:2020. [13]

1.3.7 PDF/UA

Das PDF/UA (Universal Accessibility) Dateiformat dient der Erstellung barrierefreier Dokumente. Die PDF/UA-Kennzeichnung stellt eine Klassifizierung für barrierefreie Dokumente dar und orientiert sich an den Anforderungen der Web Content Accessibility Guidelines (WCAG) 2.0 des World Wide Web Consortiums. Als Rechtliche Grundlage dient die Barrierefreie-Informationstechnik-Verordnung (BITV) 2.0. Um die Anforderungen der PDF/UA-Kennzeichnung zu erfüllen, müssen Dokumente bestimmte technische und inhaltliche Vorgaben erfüllen. Auf Basis des Matterhorn-Protokolls, welches aus 31 Prüfpunkten und 136 Konformitätskriterien besteht, müssen Aufbau von Texten, Bildern, Listen, Tabellen und Formularefeldern festgelegte Einstellungen haben. Diese Konformitätskriterien können auf der einen Seite nur von einer Software geprüft werden und auf der anderen Seite nur von Menschen. Menschen mit Einschränkungen sollen das Dokument optimal nutzen. Zur Erleichterung des Verständnisses sollten Überschriften, alternative Texte für Bilder, Beschreibungen für Tabellen, Tags und eine klare Lesereihenfolge verwendet werden. Mittels der Tags können Screenreader Inhalt und Struktur des Dokuments erfassen. [17] Es gibt die PDF/UA-1 Version aus ISO 14289-1:2012 und die überarbeitete Version PDF/UA-1 aus ISO 14289-1:2014, die erst 2020 als gültig erklärt wurde. [13]

1.3.8 PDF/R

Speziell für die Speicherung, Transport und Austausch von gescannten Dokumenten gibt es das ISO 23504-1:2020 standardisierte Format PDF/R-1. Es bietet die grundsätzlichen Funktionalitäten von TIFF, bitonalen, Graustufen- und Echtfarbbilder. [13]

1.3.9 Durchsuchbares PDF

Searchable PDFs können mit Suchfunktionalitäten eines PDF-Readers durchsucht werden. Es kann gezielt nach Zahlen oder Stichwörtern durchsucht und Inhalte können zur Bearbeitung in anderen Programmen kopiert werden. Man erkennt durchsuchbare PDFs daran, dass man den Text markieren kann. Diese PDF-Art wird üblicherweise

durch die Optical Character Recognition (OCR) Technologie erstellt. Bei OCR handelt es sich um optische Zeichenerkennung, die Textzeichen und Dokumentstruktur analysiert. Auf diese Weise können gescannte Dokumente oder Pixelbilder als PDF abgespeichert und in ein Durchsuchbares PDF in Adobe Acrobat umgewandelt werden. Während der Umwandlung wird dem Dokument eine zusätzliche unsichtbare Textebene, die unter der Bildebene liegt und durchsuchbar ist, auf der Seite hinzugefügt. In Acrobat ist es außerdem möglich mit der einfachen Suche innerhalb einer Datei nach Suchbegriffen zu suchen, mit der erweiterten Suche oder der Suchen-Werkzeugleiste mehrere PDF-Dokumente zu durchsuchen und speziell in der erweiterten Suche u.a. Objektdaten und Bildern zu lokalisieren. Textbasierte PDF-Dateien können grundsätzlich durchsucht werden und auch in andere Dateiformate wie Microsoft Word, Excel oder PowerPoint umgewandelt werden. Durchsuchbare PDFs ermöglichen Barrierefreiheit. Sie können von Bildschirmleseprogrammen für Sehbehinderte vorgelesen oder vergrößert werden. [18]

1.4 PDF Dateiversionen

Gestartet mit Version 1.0 war PDF lediglich ein proprietäres Dateiformat von Adobe. Die Freigabe von PDF als offenes und kostenlose Dateiformat führte letztendlich erst zu seiner weltweiten Verbreitung und Anerkennung. Erst im Jahr 2005 entwickelte sich Version 1.4 zu einem internationalen ISO Standard. Die letzte Dateiversion 2.0 von 2017 ist schon eine Weile her und es hat sich zeitlich nur das PDF/R-Dateiformat später entwickelt als PDF 2.0.

1.4.1 PDF 1.0

PDF 1.0 wurde 1992/1993 entwickelt und ist wurde nicht normiert. 1992 wurde die Spezifikation als Buch verkauft und 1993 das der Spezifikation entsprechende digitale Format entwickelt, welches ausschließlich den RGB Farbraum darstellen konnte. Medien, die einen anderen Farbraum besitzen wurden in RGB konvertiert. In der Druckindustrie ist jedoch der CMYK-Farbraum von Bedeutung. Folglich ist PDF 1.0 nicht für den Printbereich geeignet. Damals war Adoba Acrobat 1.0 das einzige Programm mit dem man diese Dateiversion bearbeiten konnte. [13]

1.4.2 PDF 1.1

Genauso ist das 1994 kreierte PDF 1.1 keine Norm und implementiert weiterhin nur den RGB Farbraum, jedoch geräteunabhängig. Zusätzlich benötigte man ein Update

von Adobe Acrobat auf Version 2.0. Erstmals sind in diesem Format das Einbetten von externen Links, mehrseitige Artikel und Threads, Passwortverschlüsselung und Notizen bzw. Anmerkungen erschienen. [13]

1.4.3 PDF 1.2

Das 1996 erschienene PDF 1.2 wurde ebenfalls nicht standardisiert, jedoch ermöglichte es erstmals den druckbaren CMYK-Farbraum und Sonderfarben zu verwenden. Des weiteren wurden interaktive Formularfunktionen, Unicode, Multimedia Kompatibilität, Unterstützung der Open Prepress Interface (OPI) 1.3 Spezifikationen und eine Druckrasterfunktion implementiert. [13] In PDF 1.2 wurden erstmalig AcroForms vorgestellt.

1.4.4 PDF 1.3

1999 wurde PDF 1.3 mit fehlender Normierung auf den Markt gebracht und trug seinen Teil 2001 und 2002 bei zur Standardisierung des ISO PDF/X Standards. Es ist kompatibel mit PostScript 3 und bietet die Neuerungen der 2-Byte CID Schrifttypen, OPI 2.0 Unterstützung, Farbraumerweiterung für Sonderfarben durch ICC-Profile und den DeviceN Farbraum, weiche Schatten und Farbübergänge (Smooth Shading), digitale Signaturen, RC4-Verschlüsselung (40 Bit in Acrobat 4 und 56 Bit in Acrobat 4.05) und JavaScript. [13]

1.4.5 PDF 1.5

Im Jahr 2003 kam PDF 1.5 auf den Markt und hat sich nicht zur Norm entwickelt. In dieser Version wurden erstmals Ebenen implementiert. Des weiteren wurden gesteigerte Kompressionstechniken einschließlich Objekt-Streams und JPEG 2000-Kompression, sowie eine verbesserte XRef-Tabelle und XRef-Streams implementiert. 12 weitere Seitenübergänge für Präsentationen, verbesserte Unterstützung für Tagged PDF und die Adobe proprietäre Technologie XFA wurden außerdem hinzugefügt. [13]

1.4.6 PDF 1.4

Der erste PDF ISO-Standard ISO 16612-1:2005 wurde zeitlich nach Version 1.5 verabschiedet. In diesem Format sind Transparenzen, JavaScript 1.5, bessere Integration von Datenbanken, Titel, Textblockdefinition, JBIG2-Komprimierung und 128-Bit-RC4-Verschlüsselung erstmalig eingeführt worden. [13]

1.4.7 PDF 1.6

In diesem ISO 15930-8:2008 Standard sind erstmals folgende Technologien in diesem Format eingeführt worden: NChannel, welches eine Erweiterung von NDevice mit Sonderfarben ist, JPEG 2000-Kompression, Advanced Encryption Standard (AES) Verschlüsselung, direkte Einbettung von OpenType-Schriften, Containerisierung, 3D-Daten (U3D) und XML Formulare. [13]

1.4.8 PDF 1.7

Veröffentlichung am 1. Juli 2008 ist PDF in Version 1.7 als ISO 32000-1:2008 als Offener Standard definiert worden. Es wurden komplexer 3D-Objekte, Kontrolle über 3D-Animationen und Einbettung von Standard-Druckeinstellungen wie Papierauswahl, Anzahl der Kopien und Skalierung hinzugefügt. [13]

1.4.9 PDF 2.0

XFA ist in PDF 2.0 als ISO 32000-2:2017 vom ISO-Gremium als veraltet markiert worden. Mehr Einstellungsmöglichkeiten und Funktionen sind diesem Format beigefügt worden: Erweiterte Definitionen der Halbtöne der Rasterung, z.B. für Flex- oder Tiefdruck, konsistente Transparenz, erweiterte Tagged-PDF Funktion für Barrierefreiheit, Definition von Sonderfarben über Spektralfarben, alternierende Reihenfolge der zu druckenden Farben, Steuerung der Schwarzpunktkompensation, AES-256-Bit-Verschlüsselung und Einbettung von 3D-Messungen oder Querschnittsdaten. [13]

1.5 PDF Implementierung

PDF ist eine vektorbasierte Page Description Language (PDL) (Seitenbeschreibungssprache) und basiert auf dem PostScript-Format. Der MIME-Type von PDF heißt application/pdf. Eine PDL beschreibt den Seitenaufbau, wie die Seite in einem Ausgabeprogramm bzw. Ausgabegerät, z.B. einem Drucker, aussehen soll. PDLs können Seiten mit Vektoren beschreiben. Vektorielle Seitenbeschreibung bedeutet, dass das Format beliebig skalierbar ist ohne Qualitätseinbußen, jedoch eingebettete Pixelgrafiken erhalten durchaus mittels genügend Skalierung Qualitätsverluste. Deren Ausgabeformat ist normalerweise nicht zur weiteren Bearbeitung vorgesehen. An den Drucker wird durch die PDL ein Datenstrom der zu druckenden Aufgabe erzeugt und an den Drucker gesendet. Der Raster Image Processor (RIP) eines Druckers wandelt die Bildschirmausgabe in die gerasterte Druckerausgabe um. Viele APIs der Hardwareabstraktionsschicht

im Computer wie Graphics Device Interface (GDI) oder OpenGL können in PDL ausgeben. Speichert ein Satzprogramm den Seitenbeschreibungscode eines Dokuments in einer Datei, müssen Drucker die PDL nicht selbst verarbeiten. Eine PostScript Printer Description (PPD) Druckerbeschreibungsdatei definiert Fonts, Papiergröße, Auflösung und andere Standardeigenschaften für einen bestimmten PostScript-Drucker. [19] Im Common Unix Printing System, der Standard-Druckersteuerung von Linux hat der PostScript und der PDF-Interpreter ghostscript die Aufgabe eines RIP, d.h. er ist für die Umwandlung in die gerasterte Druckausgabe auf dem Drucker zuständig. Zudem stellen PDLs eine Schnittstelle zum Quellcode eines Dokuments bzw. zu Programmen, die Quellcode verwalten oder das Dokument formatieren können, dar. Die PDL PDF erweitert die Funktionalität von PostScript um anklickbare Links (Hypertextfunktionalität), die die Navigation im Dokument erleichtern oder um URLs, die sich automatisch im Browser öffnen. [20]

1.5.1 PostScript

Sowohl PostScript als auch PDF haben zum Ziel die Seiten eines Dokuments vollständig für die Ausgabe in der Druckvorstufe zu beschreiben. Die abwärtskompatible, stackorientierte, Turning-vollständige Hochsprache PostScript PDL wurde in den 1980er Jahren von Adobe erfunden. [21], [22] Hinzu wurden weiter PostScript Technologien entwickelt, die aus der Programmiersprache PostScript, Grafik-, Textformatierungsanwendungen, Treibern und Abbildungssystemen bestehen. PostScript hat sich als Industriestandard etabliert. Die letzte Version ist PostScript 3 von 1997. Seine primäre Anwendung gemäß des Adobe Imaging Models findet sich in der Beschreibung von Textdarstellung, graphische Formen und Bildern auf gedruckten oder auf dem Bildschirm angezeigten Seiten. Dabei ist die Beschreibung des Dokuments geräteunabhängig und eine PostScript-Datei ist sequentiell organisiert. PostScript unterstützt unter anderem beliebige geometrische Formen, Zeichenoperationen in Graustufen, RGB, CMYK und CIE (Yxy-Farbraum) und vorinstallierte oder benutzerdefinierte Fonts und Digitalbilder jeglicher Auflösung je nach Farbmodell und ein allgemeines Koordinatensystem. In PostScript wird eine Seite, die ein Koordinatensystem umspannt, als Grafik betrachtet, die verschiedene Grafikelemente enthalten kann. Dabei werden die Textzeichen eines Fonts, gemäß des Adobe Imaging Models, als graphische Formen betrachtet auf denen Grafikoperationen möglich sind. Das Koordinatensystem unterstützt alle linearen Transformationen, die auf alle Seitenelemente angewandt werden können. Die Seitenbeschreibung in PostScript kann auf jedem Gerät, was einen PostScript Interpreter implementiert, gerendert werden. In diesem Prozess wird die high-level PostScript-Beschreibung in low-level Rasterdatenformate für das jeweilige Gerät übersetzt. PostScript Programme können erstellt, übertragen und als ASCII Quellcode interpretiert werden. [21] Jede PostScript-Datei muss durch einen RIP

interpretiert werden. Der PostScript-Interpreter rechnet die Benutzerkoordinaten in Gerätepixel um, wobei auch die technischen Eckdaten des jeweiligen Geräts mitberücksichtigt werden. Theoretisch kann derselbe PostScript-Code auf verschiedenen Endgeräten mit unterschiedlichen Auflösungen mehr oder weniger identische Ausgaben erreichen. Den Interpreter gab es früher als Hardware-RIP, der allerdings nicht mehr zum Einsatz kommt. Heute gibt es lediglich Software-RIPs, die von einem Betriebssystem kontrolliert werden und hardwareunabhängig arbeiten. Fast alle RIP-Hersteller orientieren sich am Standard und somit sind PostScript-Fehler in der Druckvorstufe minimiert worden. [5]

1.5.2 Adobe Imaging Model

PDF und die PostScript Programmiersprache haben das Adobe Imaging Model als Gemeinsamkeit. Es kann nahtlos zwischen PDF und PostScript konvertiert werden und beide erzielen das gleiche Ausgabeergebnis beim Druck. Dennoch fehlt PDF das general-purpose Framework der PostScript Programmiersprache. Stattdessen stellt ein PDF Dokument eine statische Datenstruktur optimiert für den random-access dar und enthält zusätzlich Seitennavigationsinformationen für interaktives Lesen. Das high-level Imaging Model beschreibt die Elemente, die auf der Seite dargestellt werden, also Text, Geometrie oder Bilder, als abstrakte graphische Elemente aus Vektorobjekten und Bézierkurven, anstatt als Pixeldefinitionen. Dadurch wird das Imaging Model zu einem geräteunabhängigem Modell und kann hochwertige Ausgaben auf vielen verschiedenen Druckern und Bildschirmen liefern. Die PDL beschreibt dieses Imaging Model. Eine Anwendung generiert zuerst die geräteunabhängige Beschreibung des gewünschten Ausgabegeräts in der PDL. Daraufhin interpretiert eine Firmware oder Software eines spezifischen Ausgabegeräts für Rasterausgaben die Beschreibung und rendert sie im Ausgabegerät. Hierbei hat die PDL die Rolle eines Austauschstandards für die Übertragung und Speicherung von druckbarem oder auf Displays darstellbaren Dokumenten. [21] Später wurde das Imaging Model für die Unterstützung von Transparenzen erweitert. Diese Funktionalität wurde speziell für PDF implementiert und wird nicht von PostScript unterstützt. Bei PostScript überschreibt das zuletzt gezeichnete Objekt alle darunterliegenden Objekte im Hintergrund. Das Imaging Model beschreibt Pfad-Objekte aus verbundenen Punkten, Linien und Kurven. Text-Objekte bilden eine eigene Datenstruktur (Fonts), die als Glyphen aus Pfad-Objekten bestehen. Bild-Objekte sind aus einzelnen Pixelwerten in einer rechteckigen Fläche aufgebaut und enthalten eine eindeutige Position im Rechteck und einen Farbwert. Die Flexibilität des Adobe Imaging Models zeichnet sich durch seine flexible Ausgabefähigkeit auf jeglichen Rastergeräten und hochauflösenden Displays. Die Größe des Pixels wird durch die Ausgabeauflösung des Rastergeräts bestimmt, die bei Monitoren zwischen

75 und 110 Pixels per inch (ppi) und bei Tintentstrahl- bzw. Laserdruckern zwischen 300 und 1400 ppi liegt. [5]

1.5.3 Dateiformataufbau

PDF ist ein reines objektbasiertes Dateiformat und PDF-Dateien enthalten Dokumentdaten in binärer Form. PDF-Dateien bestehen aus Sequenzen von 8-Bit-Binär bzw. 7-Bit-ASCII. [5] Ein Dokument entspricht immer einer Datei. Das Einbetten von binären Dateien in beliebigen Formaten oder anderer PDF-Dateien ist möglich. Die Struktur besteht im Wesentlichen aus 4 Komponenten. Zunächst spezifiziert der Header die Version der PDF-Spezifikation (Signature) und den Charset Identifier. [23] Der Body enthält die Daten der Objekte, aus denen das Dokument besteht und die Cross-Reference Table (Xref) deckt die Informationen über die Position der indirekten Objekte in der Datei ab. Die Xref-Sektion wird auch als Katalog bezeichnet und speichert genauer gesagt die Byte-Positionen der Objekte im Body. Ein Verweis auf ein Objekt im Katalog kann es für andere Seiten wiederverwenden. Zuletzt definiert der Trailer die Startposition der Cross-Reference Table als Pointer startxref und von speziellen Objekten im Body. [24] Außerdem enthält der Trailer einen Size Entry und die Markierung %%EOF für das Dateiende. [23] Die Objekte im Body sind in einer komplizierten hierarchischen Struktur, dem Dokument, verknüpft. Zur Dateigrößenoptimierung werden komplexe Verbindungen zwischen den Daten hergestellt und die Daten eines mehrfach vorkommenden Objektes müssen nur einmal gespeichert werden. [7] Text in AcroForms wird als Stream gespeichert. In Streams kann alles gespeichert werden und sie werden nicht interpretiert. Die Objekte in Streams können Referenzen auf andere Objekte vor allem Seiten enthalten. [24] Objekte können so gruppiert werden, wodurch eine bessere Komprimierung erreicht wird, vor allem bei Gruppierung von Linien. [5] Der Trailer enthält eine Referenz zum root Element im Body [24] und Metadaten. PDF-Dokumente können Dictionary Objekte enthalten, was ein Paar von Objekten darstellt, genannt Entries. Aktionen werden beispielsweise als Entries gespeichert. [25] Die PDF-Dokumentstruktur ist auf einen schnellen, wahlfreien Zugriff (random-access) auf beliebigen Seiten optimiert. Im Unterschied dazu sind PostScript-Dateien seriell organisiert. In Xref sind alle Informationen für den random-access eingetragen. Neben Objekteinträgen können auch Cross-Reference-Streams hinterlegt werden. Xref ist der einzige Teil in PDF mit einem konstanten Format und kann aus mehreren Cross-Reference-Sections und Subsections für das inkrementelle Update bestehen jeweils mit Objekteinträgen. Ein Objekteintrag ist wie folgt aufgebaut: Die ersten 10 Bytes für die Byteposition (Offset), mit Leerzeichen als 1 Byte getrennt die folgenden 5 Bytes für die eindeutige Generation Number und zuletzt eine ebenfalls durch ein Leerzeichen getrennte Markierung mit f oder n. f steht für free entry, d.h. gelöschtes Objekt, und n für in use entry. Das erste Objekt in Xref hat eine Generation

Number von 0 und wird nicht verwendet. [24] Jedes Objekt wird im Body mit obj und endobject eingekapselt und jeder Stream mit stream und endstream. [5]

PDF Viewer prozessieren PDF-Dateien im Prinzip vom Ende bis zum Anfang, d.h. vom Trailer zum Body. [24] Beim Parsen wird zunächst die Signature überprüft, dann wird die Position von %%EOF und startxref gesucht, was die Position der Xref angibt. Die Xref stellt die Offsets jedes Objektes zur Verfügung und der Trailer zeigt auf den /Root Entry des Root-Objekts. Nachfolgend werden alle Objekte geparst und überprüft, ob /Root den /Pages Entry, /Pages ein Seiten-Array, jede /Page eine Größe der /MediaBox hat, /Contents als Stream-Objekt vorliegt und /Resources das /Font dictionary definiert. Zuletzt wird die Seite gerendert durch BeginText, Auswahl des Fonts, Bewegung des Cursors, Anzeigen des Strings und EndText. [23] Referenzen werden nicht in der parse time ausgewertet, sondern nur nach Verwendung. Strings können als <Length> <string> oder <string> <terminating symbol> definiert werden. [26] Metadaten wurden bis PDF 1.7 durch den XMP Standard codiert und als XML formatierte Daten in PDF-Dateien abgelegt. [7] XML ist eine Sprache zur Markierung von Inhalten mit Hilfe von Tags, um die Struktur zu beschreiben und Elemente zu identifizieren. [5] Unicode wird in den Metadaten unterstützt. Seitenobjekte und die meisten PDF-Strukturen sind gerichtete azyklische Graphen. [26] Beim inkrementellen Update wird am Ende der Original-Trailers ein Body Update, eine neue Cross-Reference-Section und ein Updated Trailer der Datei hinzugefügt. In der neuen Version der Cross-Reference-Section werden alle Objekte aufgeführt, die gelöscht, geändert oder ersetzt wurden. Der Updated Trailer umfasst alle Änderungen bezüglich des Original Trailers. [5]

Eine PDF-Datei besitzt 3 verschiedene Layers, womit nicht die Optional Content Layers gemeint sind. Zunächst enthält die Content-Layer alle druckbaren Objekte, sprich Grafiken, Bilder und Texte. Elemente dieses Layers können ausschließlich mit Acrobat Pro bearbeitet werden. Oberhalb der Content-Layer liegt die Enhancement Layer. In ihr sind Lesezeichen, Hyperlinks, Thumbnails, digitale Signaturen, Annotationen, Formularfelder und alle Multimedia-Elemente wie Video und Audio abgelegt. Zuletzt in der unsichtbaren Information Layer liegen alle Basisinformationen zu Schriftdaten, Formularfeldinhalten, Verschlüsselungsinformationen, Querverweistabellen und PDF-spezifische Informationen. [5]

Grundsätzlich besteht eine PDF-Datei aus 5 Seitenrahmen als Boxen für jede Seite. Diese Boxen werden weder gedruckt, noch standardmäßig angezeigt. Die äußerste und größte Box ist die MediaBox. Sie entspricht der physischen Größe des Mediums und dem Papierformat. Sie muss immer in einer PDF-Datei vorhanden sein und enthält auch alle Objekte, die über den Rand der Seitengröße hinausragen, wobei diese über die MediaBox gehen können. Innerhalb der MediaBox liegt als nächste Box die CropBox. Sie entsteht durch das Beschneiden der Seite und definiert den Ausschnitt zur Anzeige

in Acrobat. Zusätzlich wird sie meist zum Platzieren von PDF-Dokumenten in anderen Programmen und zum Ausdrucken aus Acrobat verwendet. Beim Erstellen der PDF-Datei hat die CropBox die Größe der MediaBox und ist immer vorhanden. Darunter liegt die BleedBox. Sie definiert die Beschnittzugabe, die in der Praxis meist auf 3 mm gesetzt wird. Beschnittzugabe wird in der Druckvorstufe verwendet, damit keine weißen Blitzer am beschnittenen Druckbogen zu sehen sind. Die optionale BleedBox sollte kleiner als die MediaBox sein. Druckmarken wie Passkreuze, Schnittmarken oder Farbbalken sollten immer außerhalb der BleedBox liegen. Die TrimBox steht für die finale Größe des gedruckten und zugeschnittenen Dokuments. Ein zu druckendes Dokument benötigt zwingend die im PDF-Dateiformat optionale TrimBox, deren Größe kleiner oder gleich der BleedBox und MediaBox sein sollte. Der Standardwert für die TrimBox ist die Größe der CropBox. Ganz innen im Boxmodell von PDF liegt die ArtBox. Sie stellt einen Rahmen um alle druckbaren Objekte dar und legt somit den Inhalt fest. Meist sind ArtBox und TrimBox von der Größe her identisch. Im Bezug auf PDF/X-Dateien darf nur entweder die ArtBox oder die TrimBox vorhanden sein. Da die ArtBox optional ist, dient sie vor allem in Ausschussprogrammen als Default-Box, falls keine TrimBox angelegt wurde.

1.5.4 Implementierung von Fonts

Die Beschreibung von Glypen ist bei eingebettete Schriften als Datenstrom im Eintrag FontFile registriert. Falls die Schrift nicht eingebettet wurde fehlt dieser Eintrag. Ein optionales Unicode-Mapping ToUnicode ist von Nöten, damit die Glyphen auch über Unicode verarbeitet werden kann. Ist dieses Mapping nicht vorhanden, so kann keine Textsuche und das Kopieren von Text stattfinden. Fehler im Mapping oder Modifikation von Schriften können zu falsche Ausgabebuchstaben, mangelnde Wiederverwendung und fehlerhafte Textkonvertierung führen. Jede Glyphe im Dokument wird über einen Character-Code prozessiert. Daraufhin erfolgt eine Zuordnung des Character Codes zum hinterlegten Encoding (Mapping). Zuletzt wird die Glyphe im aktuellen Font über die Glyphen-ID zum Zeichen der Glyphe aufgerufen. Folglich erzielt das Mapping des Codes und der Aufruf der Glyphe die benötigte Konturbeschreibung. Schriftsubstitution findet immer dann statt, wenn der Character-Code nicht mit der Encoding-Tabelle übereinstimmt. Häufig fehlen bestimmte Glypen im Font. Falls eine Outline-Beschreibung des Fonts zum Erstellungszeitpunkt nicht verfügbar ist, wird die Einbettung des Fonts verhindert. Dies kommt vor allem dann vor, wenn ein Font ein Schutzflag besitzt. Weitere Probleme bei der Schrifteingbettung sind u.a. Laufweitenfehler in Schriften, Fehler in der Buchstabenbeschreibung oder beim Cachen von Fonts. Zwecks der Schriftsubstitution müssen folgende allgemeine Informationen zu einem Font in der PDF-Datei gespeichert sein: Name der Schrift, Typ, Subtyp,

Schriftstärke, Zeichenbreite, Laufweite, maximale Ausprägung der FontBox, Dickteninformationen, Positionsangaben über Versal- und x-Höhe und Winkel für Italic (kursiv). Diese Informationen sind selbst bei nicht eingebetteten Schriften vorhanden. Für jeden verwendeten Font wird ein Font Descriptor in der Datei hinterlegt. [5]

1.5.5 Implementierung von Transparenzen

Wird eine PostScript oder PDF-Datei erstellt, werden die Transparenzen vom Flattener reduziert (verflacht). Um den gewohnten visuellen Effekt der Transparenzen beizubehalten gibt es unterschiedliche Verfahren bei der Reduzierung auf Vektor- und Pixelebene.

1.6 PDF Sicherheitsaspekte

Etwa 40 % der Unternehmen setzen PDFs für geschützte Inhalte ein. In den letzten 2 Jahren ist die Nutzung der elektronischen Signaturfunktion in PDFs um mehr als 150 % gestiegen. [27]

In den Sicherheitseinstellungen eines PDF-Dokuments können Dokumentensicherheit und Zugriffsregeln justiert werden. PDF unterstützt Verschlüsselung und die Vergabe von 2 Passworttypen. Eventuell kann beim Öffnen einer Datei ein Passwort gefordert werden oder das Kopieren von Teilinhalten, jeglichem Inhalt, Ausfüllen von Formularfeldern, Dokumentveränderungen (z.B. Struktur, Inhalt, Kommentare) oder das Ausdrucken kann vom Ersteller des Dokuments gesperrt worden sein.

1.6.1 Digitale Unterschrift

Digitale Unterschriften sollen die Identität des Unterzeichners des Dokuments authentifizieren und dass der Inhalt nach der digitalen Unterschrift nicht geändert wurde. Der Verfasser kann sein PDF-Dokument mit einem digitalen Zertifikat signieren. Das Zertifikat bescheinigt die Echtheit der Unterschrift und der Herkunft und wird von einem Zertifizierungsanbieter ausgestellt. Zusätzlich können Zertifikate ablaufen oder entzogen werden und müssen gültig sein. Dabei sollte ein vertrauenswürdiger Zertifizierungsanbieter gewählt werden. Digitale Signaturen werden durch einen Hash basierend auf das erstellte PDF-Dokument berechnet und geben der PDF-Datei einen eindeutigen Fingerabdruck. Dieser Hash wird im Dokument gespeichtert und wird überprüft, wenn die Unterschrift validiert werden soll, indem er neu berechnet wird. Unterscheiden sich beide Hashs voneinander wurde die PDF Datei verändert. Jede

Unterschrift kann mit einem Zeitstempel versehen werden. Ein vertrauenswürdiger Zeitstempel-Anbieter (ZSA) belegt den Zeitpunkt, wann diese Unterschrift geleistet wurde.
Eine PDF-Datei ermöglicht mehrere digitale Unterschriften, jedoch muss jede neue Unterschrift in einem inkrementellen Update geleistet werden. Jede Unterschrift muss mit einem Unterschriftsfeld im Dokument verbunden sein. Optional kann das Unterschriftsfeld mit einem Widget gekoppelt sein. Dann wird die Unterschrift graphisch dargestellte. Unterschriften ohne Widgets sind versteckte Unterschriften. [7] Eine digitale Signatur ist eine spezielle Art von elektronischer Signatur, die kryptographische Techniken implementiert. Diese Techniken beinhalten asymmetrische Schlüssel zur Verschlüsselung. Der Unterzeichner verwendet einen privaten Schlüssel, um seine digitale Signatur zu erstellen, welcher an das Dokument angehängt wird. Zur Validierung der digitalen Signatur wird der öffentliche Schlüssel verwendet. Im Gegensatz dazu kann eine elektronische Signatur verschiedene Formen annehmen, beispielsweise eine gescannte handschriftliche Unterschrift, ein getippter Name, eine biometrische Signatur oder eine digitale Signatur. Elektronische Signaturen sind der Oberbegriff und digitale Signaturen sind eine Teilmenge davon. Elektronische Signaturen bieten variierende Sicherheitsgrade. [10]

2 PDF Programme auf dem Markt

Die Popularität von Portable Document Format (PDF) Dateien ist seit 2008 rasant angestiegen in der globalen Informationsübertragung. Täglich werden weltweit 2,5 Milliarden PDF Dokumente erzeugt. Seine Beliebtheit verdankt PDF vor allem an der plattformübergreifenden Kompatibilität (Desktop-Computer, Tablets und Smartphones), denn PDF Dokumente ist auf mehr als 1,5 Milliarden Geräten ohne zusätzliche Software lesbar. Über 80% der geschäftlichen Dokumente werden als PDF Datei weitergegeben. [27] 90 % der Büroangestellten wollen auf das PDF Dateiformat nicht mehr verzichten. Drei Viertel aller archivierten Dokumente sind PDF Dokumente. [28] Bis 2025 werden über 3 Millarden Dollar jährlich für PDF Editoren ausgegeben werden. [28] Im Jahr 2015 gab es 1,6 Milliarde PDF-Dokumente im Web und im Jahr 2019 wurden PDF-Dateien bei ca. 99 % Firmen und Regierungsinstitutionen weltweit verwendet. **ccc-break-pdfs**

2.1 Die Firma Adobe Systems Incorporated

1982 wurde die Firma Adobe Systems Incorporated von John Warnock und Chick Geschke gegründet. Adobe Illustrator war das erste auf PostScript basierende Grafikprogramm, welches 1988 auf den Markt gebracht wurde. [5]

2.2 Aktueller Stand von Forschung und Technik

2.3 Freie PDF Programme und Onlinedienste

PDF Dateien lassen sich in vielen Programmen einfach über den Druckdialog erstellen. Apple hat das Lesen von PDF Dokumenten in seiner Apples Vorschau integriert. Viele Webbrowser stellen PDF Viewer bereit, so Google Chrome seit 2010. [1]

2.3.1 PDFCreator

PDF Dokumente und Dateien erzeugen

2.3.2 LibreOffice

PDF Dokumente und Dateien erzeugen

2.3.3 OpenOffice

PDF Dokumente und Dateien erzeugen

2.3.4 ghostscript

2.4 Kostenpflichtige PDF Programme und Onlinedienste

2.4.1 Adobe Acrobat

Die nicht-Pro Version von Acrobat kann prüfen, ob es sich bei dem geöffneten PDF-Dokument um ein PDF/A-Dokument handelt und auf dessen Konformität prüfen. Zusätzlich kann man sich die Kompatibilität mit anderen PDF-Dateiformaten PDF/X, PDF/E, PDF/VT und PDF/UA anzeigen lassen. [14] Acrobat kann über JavaScript ferngesteuert werden. Dazu muss man die Berechtigung zur Ausführung von JavaScript erteilen. [5]

2.4.2 Adobe Acrobat Pro

PDF-Inhalte können bearbeitet und Dokumente mit digitalen Signaturen unterzeichnet werden. [18] Eine PDF-Datei kann in eine PDF/A-Datei inklusive seiner Varianten, PDF/X, PDF/UA, PDF/VT oder PDF/E konvertiert werden. Außerdem kann die Kompatibilität mit diesen Formaten überprüft werden in Preflight-Profilen. [14] Die Barrierefreiheit kann automatisch validiert werden oder ein neues Dokument kann direkt barrierefrei erstellt werden. Adobe Acrobat Pro kann andere Dokumentenformate wie HTML, DOC, DOCX, TXT und RTF in PDF konvertieren, PDF in andere Dateiformate wie Microsoft Word exportieren oder Dokumente unterschreiben. [29] Mit dem Werkzeug Scan & OCR kann man Pixelbilder als PDF und gescannte PDF-Dokumente in ein durchsuchbares PDF umwandeln. [18] Installiert man Acrobat Pro in Windows, steht dem Anwender ein Adobe PDF-Drucker mit entsprechender PPD-Datei zur Verfügung. [5]

2.4.3 Onlinetools von Acrobat

Produktseite:
`https://www.adobe.com/de/acrobat/online.html`
`https://www.adobe.com/de/acrobat/online/convert-pdf.html` Mit den Adobe Acrobat Onlinetools kann man über den Browser verschiedene Dateitypen in PDF umwandeln, unter anderem PDF in JPEG oder andere Bildformate, PDF Dateien bearbeiten und Kompression anwenden. Die Onlinetools können außerdem PDF in Word umwandeln. [18] Der Adobe Acrobat PDF-Converter der Onlinetools kann DOCX, DOC, XLSX, XLS, PPTX, PPT, TXT, RTF, JPEG, PNG, TIFF, BMP, sowie Adobe eigene AI-, INDD- und PSD-Dateien in PDF konvertieren. [29] Die kostenlose Version des PDF-Converters kann nur begrenzt oft genutzt werden.

2.4.4 Acrobat Distiller

In Acrobat Distiller als Software-Interpreter können PostScript-Dateien in PDF konvertiert werden und umgekehrt. Bei der Erstellung einer PDF-Datei schneidet Distiller an der MediaBox-Grenze herausragende Elemente ab. [5]

2.4.5 Microsoft Word

Aus einem Word-Dokument lässt sich in Microsofts Word eine PDF-Datei, inklusive im PDF/A-Dateiformat, erstellen. [14]

2.5 PDF zu Word Konvertierung

In PDF ist keine automatische Anpassung des Seiteninhalt-Layouts, wie z.B. in Microsoft Word, möglich. Daher kann ein PDF-Dokument nicht sinnvoll in das Word-Format umgewandelt werden ohne möglicherweise das ursprüngliche PDF-Layout zu beeinflussen und zu ändern, sowie die maximalen Bearbeitungsmöglichkeiten von Word ausschöpfen zu können.

2.6 PDF zu Latex Konvertierung

2.7 Relevanz von PDF in verschiedenen Marktbranchen

Durchsuchbare PDFs werden in Verträgen, Rechnungen und Geschäftsunterlagen verwendet, damit Mitarbeiter*innen Informationen gezielter suchen und Daten abteilungsübergreifend effizienter verwaltet werden können. In Forschungsarbeiten und wissenschaftlichen Artikeln werden durchsuchbare PDFs bei der Überprüfung von Quellen oder dem Extrahieren von Zitaten hauptsächlich verwendet. Behörden, Bibliotheken und Unternehmen digitalisieren Dokumente zur Archivierung und wandeln sie in ein durchsuchbares PDF um, was den langzeitigen Bestand der Dokumente sichert. [18]

Das PDF/A-Dateiformat wird in Bibliotheken und Archiven zur digitalen Archivierung von Büchern, Zeitschriften und historischen Dokumenten verwendet. Auch im Behördenzweig und Verwaltungssektor hat PDF/A für die Aufbewahrung von Verwaltungsakten und rechtlichen Dokumenten seine Existenzberechtigung. Im Gesundheitswegen wird es außerdem zur Speicherung von Patientenakten und medizinischen Unterlagen verwendet. Hingegen im Finanzwesen werden mit ihm Geschäftsunterlagen und Finanzdokumente verwahrt. Unternehmen und Organisationen können mit PDF/A gesetzliche und Compliance-Vorschriften einhalten. [14]

PDF/VT-Dateien werden im Direktmarketing verwendet. Personalisierte Werbematerialien erhöhen die Wahrscheinlichkeit einer positiven Reaktion bei den Kundi*nnen auf die Werbebotschaft und verbessert die Bindung von Unternehmen und Kund*innen. Der Transaktionsdruck findet bei Finanzdienstleistungen, Versicherungen und E-Commerce besonderen Anklang. Beliebte Transaktionsdokumente sind Rechnungen, Kontoauszüge, Versicherungspolicen oder Bestellbestätigungen. [16]

PDF-Dokumente mit der PDF/UA-Kennzeichnung stärken den Ruf und die Reputation eines Unternehmens oder einer Organisation durch Engagement für Inklusion. [17]

Digitale Signaturen werden bei digitalen Freigabe-, Abnahme-, Genehmigungs- und Vertragsprozesse verwendet. PAdES wird in Rechtssystemen, Finanzwesen und Regierungssektor eingesetzt. [10]

2.8 Rolle von PDF in der Druckvorstufe und Designbranche

Vor allem in der grafischen Industrie wird PDF gerne verwendet, weil es eine plattformübergreifende Visualisierung bietet auf allen Betriebssystemen. Schriften können bei

Einbettung exakt wiedergegeben werden, unabhängig ob es sich um eine Windows oder MacOS Schrift handelt. Im Vergleich zu PostScript-Dateien erzielen die kompaktere Codierung von Seiteninhalten, dem einmaligen Speichern von identischen Bildern und die Verwendung von Kompressionsalgorithmen eine maßgeblich kleinere Dateigröße bei PDF. Korrekturänderungen in PDF-Dateien sind kurz vor dem Druck noch möglich und PDF entwickelte sich zunehmend zum Containerformat für alle grafischen Elemente. Die Produktion von Druckerzeugnissen wird somit wesentlich flexibler und sicherer. Downsampling und Kompression beschleunigt den Transport von der Agentur zum Dienstleistungsbüro enorm. In der Ausgabe ist die effektive Auflösung maßgeblich. Effektive Auflösung ist die Bildauflösung, die aus der Anzahl der Bildpunkte und der Fläche resultieren, auf der das Bild platziert wurde. Downsampling beeinflusst diese effektive Auflösung. Starke Artefakte fallen im Offsetdruck weniger auf als im Digitaldruck.
Für die Betrachtung von Druckvorstufen-PDF-Dateien sollte immer Acrobat Pro bzw. Adobe Reader verwendet werden, da viele Drittanbieter-PDF-Viewer druckvorstufenrelevante Informationen nicht fehlerlos darstellen können. [5]

Erstellt man eine PostScript-Datei muss man zwischen composite oder separierte Ausgabe wählen. Bei einer separierten Ausgabe gibt es für jeden Farbauszug bei CMYK inklusive aller Sonderfarben einen eigenen Farbauszug. Inhalte einer Composite-PDF-Datei können einfach verändert werden. Farbsimulationen und die Nutzung der Überdruckvorschau sind in einem separierten PDF-Dokument nicht mehr möglich. Separierte PDF-Dateien sind nicht medienneutral, da eine Farbverrechnung bei der Separation erfolgt. Überfüllung bleibt im composite PDF nicht erhalten, jedoch Überdrucken und Aussparen schon. In modernen Workflows hat man sich vom separierten Workflow abgewendet. [5]

Seit PDF 1.3 werden ICC-Profile unterstützt, die die Farbeigenschaften, Helligkeit, Weißpunkt, Gammakurve und Farbumfang eines bestimmten Monitors eines spezifischen Geräts beschreiben, sprich ein ICC-Profil beschreibt, wie Farben von diesem Gerät dargestellt werden können. Außerdem wird die Transformation zwischen dem Gerät und dem Profilverbindungsraum Profile Connection Space (PCS) definiert. Dabei gibt es die Variante Eingabeprofile für Kameras und Scanner und Ausgabeprofile für Monitore und Drucker. Zweck des ICC-Profils ist möglichst Farbübereinstimmungen zwischen verschiedenen Geräten zu erzielen. [30]
Beim PCS handelt es sich um ein neutrales Farbmodell im ICC-Colormanagement, welches den Quellfarbraum mit dem Zielfarbraum verbindet und somit geräteunabhängig ist. Der PCS kann entweder der LAB oder XYZ Farbraum sein. [31]
Der DeviceN-Farbraum, der seit PDF 1.3 verwendet werden kann, wird auch in PostScript 3 unterstützt und erlaubt die willkürliche Kombinationen von Farbkanälen beim Composite-Druck. Dokumente mit Schmuckfarben müssen auf einem Gerät mit physikalisch getrennten Kanälen für jede verwendete Schmuckfarbe ausgegeben

werden. Folglich kann kein CMYK- oder RGB-Gerät Dokumente mit Schmuckfarben farblich korrekt darstellen. Davon sind fast alle Farbdruckersysteme betroffen, sowie die von Adobe Acrobat erzeugte Bildschirmdarstellung von PDF Dokumenten mit Schmuckfarben. Ohne den DeviceN Farbraum können Bilder mit Kombinationen von z.B. CMYK und 2 Schmuckfarben oder Schwarz und eine Schmuckfarbe nicht im Composite-PostScript und Composite-PDF wiedergegeben werden, sondern höchstens mit CMYK als Näherung. [32] OPI ist ein Workflow Protokoll, welches in der elektronischen Druckvorstufe verwendet werden kann, um Desktop Publishing Systeme und high-end Cisco Enterprise Print System (CEPS) zu verknüpfen und optimiert die Übertragung von hochauflösenden Dateien in Netzwerken. [33]
Seit PDF 1.3 werden CID Schrifttypen unterstützt. CID ist ein Synonym für das PostScript Type o Format, das eine Adressierung von mehr als 256 Zeichen ermöglicht und für Fonts mit einer großen Zeichenanzahl verwendet wurde. [34]

2.8.1 Farbdarstellung

Der RGB Farbraum eignet sich lediglich für die Bildschirmdarstellung und beschreibt die für den Menschen 16,7 Mio. sichtbaren Farben mit Hilfe von additiver Farbmischung.

2.8.2 Preflight

2.8.3 Fontformate

Da allgemeine Schriftinformationen immer eingebettet sind und die Zeilenlängen im Prinzip immer stimmen, können Druckvorstufenbetriebe zumindest immer erkennen, welche Schrift bzw. Schriftschnitt der Ersteller der PDF-Datei ursprünglich vorgesehen hatte, falls die Schrift nicht eingebettet wurde. Composite-Fonts sind Basisschriften mit hierarchischem System. Die oberste hierarchische Ebene stellt den root font dar alle folgenden Fonts sind descendant fonts. Sie ermöglichen die Einführung von Type-1-Schriften im asiatischen Markt. [5]

3 Open Source PDF Web App

3.1 Problemstellung und Anforderungen

3.2 Konzept und Methodik

3.3 Funktionalität der PDF Web App

3.4 Bedienung der PDF Web App

3.5 Implementierung der PDF Web App

3.6 Testdurchführung der PDF Web App

3.6.1 Funktionale User Tests

3.6.2 Stress Tests

4 Diskussion und Kritik

Fazit und Ausblick

Literatur

[1] Wikipedia. „Portable Document Format.“ (2023), Adresse: https://de.wikipedia.org/wiki/Portable_Document_Format (besucht am 19. 12. 2023).

[2] Wikipedia. „PDF.“ (2023), Adresse: https://en.wikipedia.org/wiki/PDF (besucht am 23. 12. 2023).

[3] Wikipedia. „Offener Standard.“ (2023), Adresse: https://de.wikipedia.org/wiki/Offener_Standard (besucht am 20. 12. 2023).

[4] Wikipedia. „Royalty-free.“ (2023), Adresse: https://en.wikipedia.org/wiki/Royalty-free (besucht am 23. 12. 2023).

[5] H. P. schneeberger, *PDF in der Druckvorstufe das umfassende Handbuch, PDF-Dateien erstellen, prüfen, korrigieren und ausgeben; PDF/X-1a bis PDF/X-5 sicher im Griff; Preflighting, Automatisierung, Standards u.v.m.* (Galileo Design), de. Bonn: Galileo Press, 2014, 910 S., Für Beruf und Ausbildung, ISBN: 978-3-642-38552-0.

[6] D. S. Peter Bühler Patrick Schlaich, *PDF, Grundlagen - Print-PDF - Interaktives PDF*, de. Berlin: Springer-Verlag GmbH Deutschland, 2018, 97 S., ISBN: 978-3-662-54615-4. DOI: 0.1007/978-3-662-54615-4.

[7] Soft Xpansion GmbH & Co. KG. „PDF: Grundlagen eines Dateiformats.“ (2013), Adresse: https://soft-xpansion.com/files/cc/PDF-Grundlagen.pdf (besucht am 21. 12. 2023).

[8] Wikipedia. „XFA.“ (2023), Adresse: https://en.wikipedia.org/wiki/XFA (besucht am 21. 12. 2023).

[9] Adobe Systems Incorporated. „PDF-Ebenen.“ (2023), Adresse: https://helpx.adobe.com/de/acrobat/using/pdf-layers.html (besucht am 23. 12. 2023).

[10] Adobe Systems Incorporated. „PAdES: Elektronische Signaturen in PDF-Dokumenten. Für was steht die Abkürzung PAdES, welche Vorteile hat das Format und wie kannst du selbst ein Dokument digital signieren. Wir zeigen es dir.“ (o. D.), Adresse: https://www.adobe.com/de/acrobat/resources/document-files/pdf-types/pades.html (besucht am 26. 12. 2023).

[11] Adobe Systems Incorporated. „PDF/E: das Dateiformat für Engineering und technische Kommunikation. Erfahre, was PDF/E-Dateien auszeichnet, in welchen Bereichen sie Anwendung finden und wie du selbst eine PDF/E-Datei erstellst.“ (o. D.), Adresse: `https://www.adobe.com/de/acrobat/resources/document-files/pdf-types/pdf-e.html` (besucht am 26. 12. 2023).

[12] Adobe Systems Incorporated. „PDF/X-Dateien: effizienter drucken. Wir erklären dir, was man unter PDF/X-Dateien versteht, wie sie genau funktionieren und was der Unterschied zu originären PDFs ist.“ (o. D.), Adresse: `https://www.adobe.com/de/acrobat/resources/document-files/pdf-types/pdf-x.html` (besucht am 26. 12. 2023).

[13] PROJECT CONSULT. „PDF Standards.“ (o. D.), Adresse: `https://www.project-consult.de/themen/pdf-standards/` (besucht am 20. 12. 2023).

[14] Adobe Systems Incorporated. „PDF/A: Wie unterscheidet es sich von PDF? Erfahre, was PDF/A-Dateien auszeichnet, was der Unterschied zu einem PDF ist und wie du selbst eine PDF/A-Datei erstellen kannst.“ (o. D.), Adresse: `https://www.adobe.com/de/acrobat/resources/document-files/pdf-types/pdf-a.html` (besucht am 25. 12. 2023).

[15] Adobe Systems Incorporated. „Revisionssichere Archivierung mit einem Dokumentenmanagementsystem (DMS). Lerne, wie du Dateien revisionssicher mit einem Dokumentenmanagementsystem archivierst.“ (o. D.), Adresse: `https://www.adobe.com/de/acrobat/resources/audit-proof-archiving.html` (besucht am 25. 12. 2023).

[16] Adobe Systems Incorporated. „PDF/X-Dateien: effizienter drucken. Wir erklären dir, was der ISO-Standard PDF/VT bedeutet, wofür du den Dateityp brauchst und wie du selbst ein PDF/VT erstellen kannst.“ (o. D.), Adresse: `https://www.adobe.com/de/acrobat/resources/document-files/pdf-types/pdf-vt.html` (besucht am 26. 12. 2023).

[17] Adobe Systems Incorporated. „PDF/UA: So erstellst du barrierefreie PDFs.PDF/UA: So erstellst du barrierefreie PDFs. Wir erklären dir, was eine PDF/UA-Kennzeichnung ist, wofür du sie brauchst und wie du selbst barrierefreie PDF-Dokumente erstellen kannst.“ (o. D.), Adresse: `https://www.adobe.com/de/acrobat/resources/document-files/pdf-types/pdf-ua.html` (besucht am 26. 12. 2023).

[18] Adobe Systems Incorporated. „So erstellst du ein durchsuchbares PDF. Lerne, was ein Searchable PDF ist, welche PDF-Arten es gibt und wie du selbst eine durchsuchbare PDF-Datei erstellst.“ (o. D.), Adresse: `https://www.adobe.com/de/acrobat/resources/document-files.html` (besucht am 25. 12. 2023).

[19] TechTarget. „PPD file (Postscript Printer Description file).“ (o. D.), Adresse: https://www.techtarget.com/whatis/definition/PPD-file-Postscript-Printer-Description-file (besucht am 29. 12. 2023).

[20] Wikipedia. „Seitenbeschreibungssprache.“ (2021), Adresse: https://de.wikipedia.org/wiki/Seitenbeschreibungssprache (besucht am 20. 12. 2023).

[21] Adobe Systems Incorporated. „PostScript, LANGUAGE REFERENCE third edition.“ (1999), Adresse: https://web.archive.org/web/20090419181826/http://www.adobe.com/devnet/postscript/pdfs/PLRM.pdf (besucht am 20. 12. 2023).

[22] Wikipedia. „PostScript.“ (2023), Adresse: https://de.wikipedia.org/wiki/PostScript (besucht am 19. 12. 2023).

[23] Ange Albertini. „PDF 101 & PDF Secrets, Learning the PDF basics, applying it to hide/reveal informations in documents.“ (2014), Adresse: https://media.ccc.de/v/MRMCD2014_-_6007_-_en_-_grossbaustelle_ber_-_201409051830_-_pdf_101_pdf_secrets_-_ange_albertini (besucht am 29. 12. 2023).

[24] Fabian Ising and Vladislav Mladenov. „How to Break PDFs, Breaking PDF Encryption and PDF Signatures.“ (2019), Adresse: https://media.ccc.de/v/36c3-10832-how_to_break_pdfs (besucht am 29. 12. 2023).

[25] Ido Solomon. „BADPDF, Stealing Windows Credentials via PDF Files.“ (2019), Adresse: https://media.ccc.de/v/gpn19-45-badpdf-stealing-windows-credentials-via-pdf-files (besucht am 29. 12. 2023).

[26] Julia Wolf. „OMG WTF PDF, What you didn't know about Acrobat.“ (2011), Adresse: https://media.ccc.de/v/27c3-4221-en-omg_wtf_pdf (besucht am 29. 12. 2023).

[27] Mehmet Bayram, formilo. „Popularität und Statistiken der PDF.“ (o. D.), Adresse: https://www.formilo.com/pdf-formulare/einfuehrung/popularitaet-statistiken/ (besucht am 19. 12. 2023).

[28] Oliver Helfrich, KOFAX. „30 Jahre PDF, Ein Geschenk, das uns immer wieder neu überrascht.“ (2023), Adresse: https://www.kofax.de/learn/blog/30-years-of-pdf (besucht am 19. 12. 2023).

[29] Adobe Systems Incorporated. „Dokumentenformate: Alles, was du wissen musst.“ (o. D.), Adresse: https://www.adobe.com/de/acrobat/resources/document-files.html (besucht am 20. 12. 2023).

[30] BenQ. „ICC-Profil Grundlagen.“ (2021), Adresse: https://www.benq.eu/de-de/knowledge-center/knowledge/icc-profile-basics.html (besucht am 20. 12. 2023).

[31] PREPRESS Secrets. „Die Rolle des Profile Connection Space.“ (2015), Adresse: `https://www.prepress-secrets.at/index_files/profile-connection-space.html` (besucht am 20. 12. 2023).

[32] HELIOS. „Welche Vorteile hat DeviceN für die Druckvorstufe?“ (o. D.), Adresse: `https://www.helios.de/web/DE/news/deviceN_prepress.html` (besucht am 20. 12. 2023).

[33] PrintWiki, The Free Encyclopedia of Print. „Open Prepress Interface.“ (o. D.), Adresse: `http://printwiki.org/Open_Prepress_Interface` (besucht am 20. 12. 2023).

[34] Typografie.info. „PostScript Type 0, Bedeutung/Definition.“ (o. D.), Adresse: `https://www.typografie.info/3/wiki.html/p/postscript-type-0-r43/` (besucht am 20. 12. 2023).

Anhang

Erklärung

Ich versichere, die von mir vorgelegte Arbeit selbstständig verfasst zu haben. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten oder nicht veröffentlichten Arbeiten anderer oder der Verfasserin/des Verfassers selbst entnommen sind, habe ich als entnommen kenntlich gemacht. Sämtliche Quellen und Hilfsmittel, die ich für die Arbeit benutzt habe, sind angegeben. Die Arbeit hat mit gleichem Inhalt bzw. in wesentlichen Teilen noch keiner anderen Prüfungsbehörde vorgelegen.

Anmerkung: In einigen Studiengängen findet sich die Erklärung unmittelbar hinter dem Deckblatt der Arbeit.

Köln, 04.03.2024 Unterschrift