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

Abbildungsverzeichnis

Abkürzungsverzeichnis

CAD Computer-Aided Design. 8

CEPS Cisco Enterprise Print System. 14

CID Character Identifier Font. 9, 14

GDI Graphics Device Interface. 11

ICC International Color Consortium. 9, 14

ISO International Organization for Standardization. 2, 7, 9, 10

OPI Open Prepress Interface. 9, 14

PAdES PDF Advanced Electronic Signatures. 8

PCS Profile Connection Space. 14

PDF Portable Document Format. 2

PDL Page Description Language. 10–12

RIP Raster Image Processor. 10, 11

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

XFA XML Forms Architecture. 10

XML Extensible Markup Language. 10, 12

ZSA Zeitstempel-Anbieter. 13

Einleitung

Motivation

Aufbau der Arbeit

1 Grundlagen

1.1 PDF Vorstellung

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. [1] 90 % der Büroangestellten wollen auf das PDF Dateiformat nicht mehr verzichten. Drei Viertel aller archivierten Dokumente sind PDF Dokumente. [2] 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. 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. [3] 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. [3], [4] 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. [5] 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. [3] Royalty-free bedeutet hierbei, dass Computerherstellerfirmen pro verkauftes Endgerät keine Lizenzgebühr (royalties) bezahlen müssen, sowie keine fixe Jahrespauschale. [6] Heute wird PDF seit 2006 von der PDF Association weiterentwickelt. [4]

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. 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.

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. Dabei 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.

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 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. [7]

1.2.3 Bilder

Generell sollte für das Bearbeiten von Bildern ein externes Bildbearbeitungsprogramm verwendet werden, z.B. Adobe Photoshop oder 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. [7] 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. [8] 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. [4]

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. [9]

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. [9]

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 mit 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 verwendet werden mehrere Widgets verknüpft mit einem Feldobjekt. [9]

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 Art der Änderung. Folglich sollten mehrfach signierte Dokumente ebenfalls mit der Option inkrementelles Update gespeichert werden. [9]

1.2.9 Kompression

In PDF können die folgenden Kompressionsalgorithmen für Bilder verwendet werden: IP, 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. [9]

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, Ebenen gesperrt werden, Ebenen zusammengefügt werden, Ebenen aus anderen PDF-Dateien importiert werden und Ebenen für unterstützende Dateiformate für Adobeprogramme, z.B. InDesign, exportiert werden. Zusätzlich kann man eine Ebenennavigation aufbauen mit Hilfe von Links und Lesezeichen, um Ebenensichbarkeit zu steuern. [10]

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. [9]

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 Anwendererereignis. Dabei kann man als Aktion JavaScript-Code aufrufen, die mit Lesezeichen, Verweisen, Seiten und Dokumentereignisse verknüpft ist. Dies gilt auch für Formulare. [9] Diese JavaScript-Erweiterung für Acrobat ist eine proprietäre Technologie von Adobe. Viele andere nicht Adobe PDF-Programme bieten keine Unterstützung für JavaScript. [3]

1.3 PDF Dateiformate

1.3.1 PDF-X

Das PDF-X Dateiformat (ISO 15930) dient des simpleren Datenaustausches in der Druckvorstufe. Es beschreibt Eigenschaften von Druckvorlagen und vereinfacht die Datenübermittlung von der Druckvorstufe bis zum finalen Druck.

1.3.2 PDF-VT

Das PDF-VT Dateiformat stellt ein spezielles Austauschformat im variablen Datendruck und Transaktionsdruck dar.

1.3.3 PDF-A

Das PDF-A Dateiformat wurde zur gesetzteskonformen Langzeitarchivierung von digitalen Dokumenten entwickelt. Langzeitarchivierung von PDF-Dateien (als PDF/A-1 in ISO 19005-1:2005)

1.3.4 PDF-E

Das PDF-E Dateiformat wurde speziell für das Ingenieurwesen entworfen und kann interaktive 3D-Elemente darstellen. Im einzelnen können CAD-Dateien im 3D- und 2D-Format eingebettet werden.

1.3.5 PDF-UA

Das PDF-UA Dateiformat dient der Erstellung barrierefreier Dokumente.

1.3.6 Durchsuchbares PDF

Das Durchsuchbare PDF kann mit Suchfunktionalitäten eines PDF Readers durchsucht werden.

1.3.7 PAdES

PDF Advanced Electronic Signatures (PAdES) ergänzt den Funktionsumfang um Werkzeuge, um elektronische Signaturen anzupassen.

1.3.8 PDF-H

Das PDF-H Dateiformat soll im Gesundheitswesen Patientendaten erfassen, austauschen und archivieren.

1.4 PDF Dateiversionen

1.4.1 PDF 1.0

PDF 1.0 wurde 1992/1993 entwickelt und ist keine Norm. 1992 wurde die Spezifikation als Buch verkauft und 1993 wurde das der Spezifikation entsprechende digitale Format entwickelt, welches ausschließlich den RGB Farbraum darstellen konnte. Medien, die einen anderen Farbraum besaßen wurden in RGB umkonvertiert. Der RGB Farbraum ist nur für die Bildschirmdarstellung geeignet und beschreibt die für den Menschen 16,7 Mio. sichtbaren Farben mit Hilfe von additiver Farbmischung. In der Druckindustrie ist jedoch der CMYK Farbraum von Bedeutung und daher war PDF 1.0 nicht für den Printbereich ausgelegt. Damals war Adoba Acrobat 1.0 das einzige Programm, um mit dieser Dateiversion zu arbeiten. [11]

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 und Anmerkungen erschienen. [11]

1.4.3 PDF 1.2

Das ebenfalls 1996 erschienene PDF 1.2 wurde keine Norm, jedoch ermöglichte es erstmals den druckbaren CMYK Farbraum und Sonderfarben zu verwenden. Des weiteren wurden interaktive Formularfunktionen, Unicode Unterstützung, Multimedia Kompatibilität, Unterstützung der Open Prepress Interface (OPI) 1.3 Spezifikationen und eine Druckrasterfunktion implementiert. [11] In PDF 1.2 wurden erstmal AcroForms (Acrobat forms) vorgestellt.

1.4.4 PDF 1.3

1999 wurde PDF 1.3 auf den Markt gebracht und trug seinen Teil 2001 und 2002 bei zur Standardisierung des ISO PDF/X Standards bei. Es ist kompatibel mit PostScript 3 und bietet die Neuerungen der 2-Byte Character Identifier Font (CID) Schrifttypen, OPI 2.0 Unterstützung, Farbraumerweiterung für Sonderfarben durch International Color Consortium (ICC)-Profile, 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. [11]

1.4.5 PDF 1.5

2003 kam PDF 1.5 auf den Markt und hat sich nicht zur Norm entwickelt. In dieser Version wurden erstmals Ebenen implementiert, die erlauben dass man mehrere Elemente wie eine Gruppe auf einer Ebene speichern kann und diese Elemente auf einmal nach Bedarf ein- und ausblenden kann, sperren oder Operationen anwenden kann. Diese Funktionalität enthalten auch die Adobe Programme Photoshop, InDesign und Illustrator. Des weiteren wurden gesteigerte Kompressionstechniken einschließlich Objekt-Streams und JPEG 2000-Kompression, sowie eine verbesserte XRef-Tabelle und XRef-Streams implementiert. Die XRef-Tabelle enthält die Positionen der indirekten Objekte innerhalb der Datei. Streams binden Dateien ein. 12 weitere Seitenübergänge

für Präsentationen, verbesserte Unterstützung für Tagged PDF und die Adobe proprietäre Technologie XML Forms Architecture (XFA) wurden außerdem hinzugefügt. [11] XFAs Haupterweiterung zu Extensible Markup Language (XML) sind rechnergestützte, aktive Tags uns sein Datenformat ist kompatibel mit anderen Systemen, Anwendungen und Technologiestandards. [12]

1.4.6 PDF 1.4

Der erste PDF ISO-Standard ISO 16612-1:2005 wurde endlich verabschiedet.

1.4.7 PDF 1.6

1.4.8 PDF 1.7

Veröffentlichung am 1. Juli 2008 ist PDF in Version 1.7 als ISO 32000-1:2008 ein Offener Standard

1.4.9 PDF 2.0

XFA ist in PDF 2.0 vom ISO Gremium als veraltet markiert.

1.5 PDF Implementierung

PDF ist eine vektorbasierte Page Description Language (PDL) (Seitenbeschreibungssprache) und basiert auf dem PostScript-Format. 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. Das 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. 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 der Vorschau am Bildschirm um anklickbare Links (Hypertextfunktionalität), die die Navigation im Dokument erleichtern oder um URLs, die sich automatisch im Browser öffnen. [13] PDF-Dateien sind komprimiert und haben üblicherweise einen Bruchteil der Größe des Ursprungsformats oder von Bilddateien.

1.5.1 PostScript

Die PostScript PDL wurde in den 1980er Jahren von Adobe erfunden. [14] Hinzu wurden weiter PostScript Technologien entwickelt, die aus der stackorientierten, Turning-vollständigen, interpretierten Programmiersprache PostScript [15], 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. 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. 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. [14]

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, 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. [14] 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. [7]

1.5.3 Dateiformataufbau

PDF-Dateien enthalten Dokumentdaten in binärer Form. 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. Der Body enthält die Daten der Objekte, aus denen das Dokument besteht und die Cross-Reference Table deckt die Informationen über die Position der Objekte in der Datei ab. Zuletzt definiert der Trailer die Position der Cross-Reference Table und von speziellen Objekten im Body. 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. [9]

Metadaten werden durch den XMP Standart kodiert und als XML formatierte Daten in PDF-Dateien abgelegt. Unicode wird in den Metadaten unterstützt. [9]

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. [7]

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. [1]

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. [9]

1.7 Rolle von PDF in der Druckvorstufe und Designbranche

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. [16]
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. [17]
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. [18] 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. [19]
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. [20]

1.7.1 Preflight

1.7.2 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. [7]

2 PDF Programme auf dem Markt

Bis 2025 werden über 3 Millarden Dollar jährlich für PDF Editoren ausgegeben werden. [2]

2.1 Aktueller Stand von Forschung und Technik

2.2 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. [4]

2.2.1 PDFCreator

PDF Dokumente und Dateien erzeugen

2.2.2 LibreOffice

PDF Dokumente und Dateien erzeugen

2.2.3 OpenOffice

PDF Dokumente und Dateien erzeugen

2.2.4 ghostscript

2.3 Kostenpflichtige PDF Programme und Onlinedienste

2.3.1 Adobe Acrobat

Acrobat kann über JavaScript ferngesteuert werden. Dazu muss man die Berechtigung zur Ausführung von JavaScript erteilen. [7] 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. [21]

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 PDF umwandeln, unter anderem PDF in JPEG oder andere Bildformate, PDF Dateien bearbeiten und Kompression anwenden. 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. [21] Die kostenlose Version des PDF-Converters kann nur begrenzt oft genutzt werden.

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] 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).

[2] 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).

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

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

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

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

[7] 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.

[8] 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.

[9] 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).

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

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

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

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

[14] 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).

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

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

[17] 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).

[18] 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).

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

[20] 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).

[21] 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).

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