

# PDF (Portable Document Format)

*Annette Maile / Frank Scholze*

PDF ist ein Datenformat, das auf der Seitenbeschreibungssprache Postscript basiert. Das PDF-Format wurde ebenso wie Postscript von der Firma Adobe Systems Inc. entwickelt. Ähnlich dem Postscript-Format, das im grafischen Gewerbe und auch in weiten Bereichen der Druckindustrie zum Standarddokumentformat wurde, ermöglicht PDF eine hochwertige und plattformunabhängige Bildschirm- und Druckausgabe. Die äußere Form des am Textverarbeitungssystem seitenorientiert entwickelten Dokuments bleibt - anders als beispielsweise bei einer Konvertierung in HTML - authentisch gewährleistet. Verschiedene Tools erlauben die Volltextindizierung und -recherche von PDF-Dokumenten. Im Unterschied zu Postscript, daß lediglich in unkomprimierter ASCII-Darstellung vorliegt, werden PDF-Dokumente einschließlich enthaltener Grafiken standardmäßig komprimiert. Dieses ressourcenschonende Format eignet sich daher für den Einsatz in Netzwerken, also auch dem WWW.

## Erzeugen von PDF-Dokumenten

Aus allen gängigen Macintosh bzw. Windows-Anwendungen können PDF-Dateien durch Drucken in eine Datei mit Hilfe des PDF-Writer erzeugt werden. Zur Zeit in Entwicklung befindet sich der PDF-Maker für Word 97, der als Beta-Version bereits frei verfügbar ist. Dieses Makro stellt unter anderem die Seitenzahlen des Inhaltsverzeichnisses als Links auf die jeweiligen Kapitel dar.

Weitgehend plattformunabhängig können alle Postscript-Dateien durch das Programm Acrobat Distiller nach PDF konvertiert werden. Es fehlen noch Versionen des Distiller für Linux, SGI und Digital UNIX. Der Konverter ist netzwerkfähig und kann zur Überwachung eines Verzeichnisses (beispielsweise auf einem FTP-Server) eingesetzt werden, in dem eingehende Postscript-Dateien dann automatisch konvertiert und in ein anderes Verzeichnis zur Abholung bereitgestellt werden. Diese Möglichkeit soll im Rahmen des Projekts Online Publikationsverbund an der Universität Stuttgart (OPUS) getestet werden.<sup>1</sup> Eine weitere Möglichkeit zur Konvertierung von Postscript-Dateien nach PDF ist die Verwendung des Shell Scripts ps2pdf, das im frei verfügbaren Postscript-Interpreter Ghostscript (ab Version 3.x) enthalten ist.

## Sonderfall TeX bzw. LaTeX

Dokumente, die von TeX oder LaTeX über Postscript nach PDF konvertiert werden, enthalten oft nicht skalierbare Fonts, die am Bildschirm oder beim Ausdruck zu schlechten Resultaten führen.

---

<sup>1</sup> Vgl. den entsprechenden Artikel „Online Publikationsverbund der Universität Stuttgart (OPUS)“ in dieser Ausgabe.

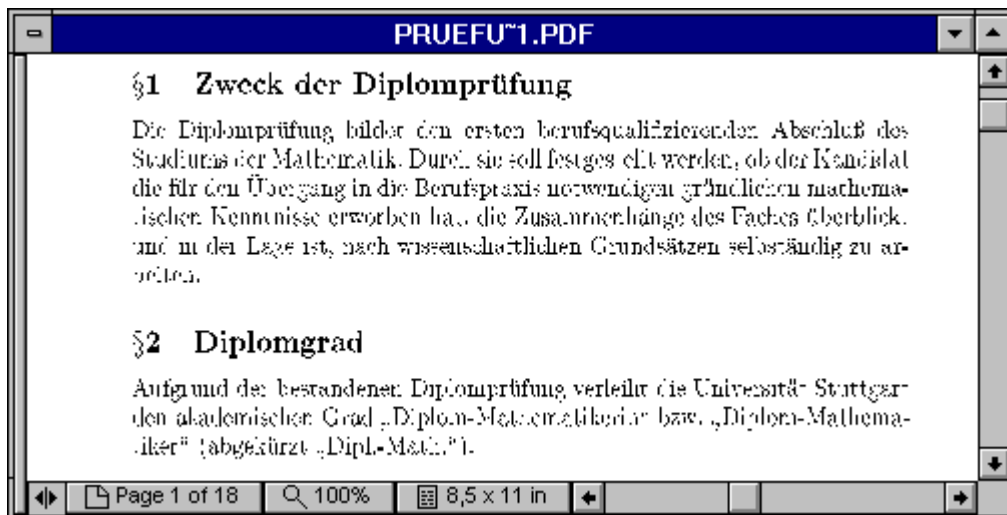


Abb. 1: Beispiel für schlecht lesbares PDF

Dies liegt daran, daß Postscript die Fonts in unterschiedlichen Formaten darstellen kann. Postscript Typ 3 Fonts können entweder als fixed-resolution bitmaps oder als skalierbare Vektor-Fonts beschrieben werden. Postscript Typ 1 Fonts sind auflösungsunabhängige Vektor-Fonts, die an geringe Auflösungen und kleine Fontgrößen angepaßt werden können.

Die „Computer Modern Fonts“, die üblicherweise mit Donald Knuths TeX verwendet werden, werden als fixed-resolution bitmaps ausgegeben und im DVI-Format (Device independent) gespeichert. Das Programm DVIPS, welches aus DVI Postscript erstellt, konvertiert normalerweise diese Fonts in Typ 3 fixed-resolution bitmap fonts.

Wenn ein solches PS-File nach PDF konvertiert wird, so ist das Resultat am Bildschirm sehr schlecht. Da PDF dafür konzipiert wurde, daß es an unterschiedlichen Monitoren und Druckern mit unterschiedlichen Auflösungen angesehen bzw. ausgedruckt wird, müssen skalierbare Vektor-Fonts benutzt werden.

Die Lösung dieses Problems besteht darin, die Typ 3 Fonts mit fester Auflösung durch skalierbare PS-Typ 1 Fonts zu ersetzen. Eine Anleitung hierzu wurde im Rahmen des Projekts OPUS erstellt und ist unter [http://www.uni-stuttgart.de/Rus/Projekte/EIPub/dvi\\_ps.html](http://www.uni-stuttgart.de/Rus/Projekte/EIPub/dvi_ps.html) zugänglich.

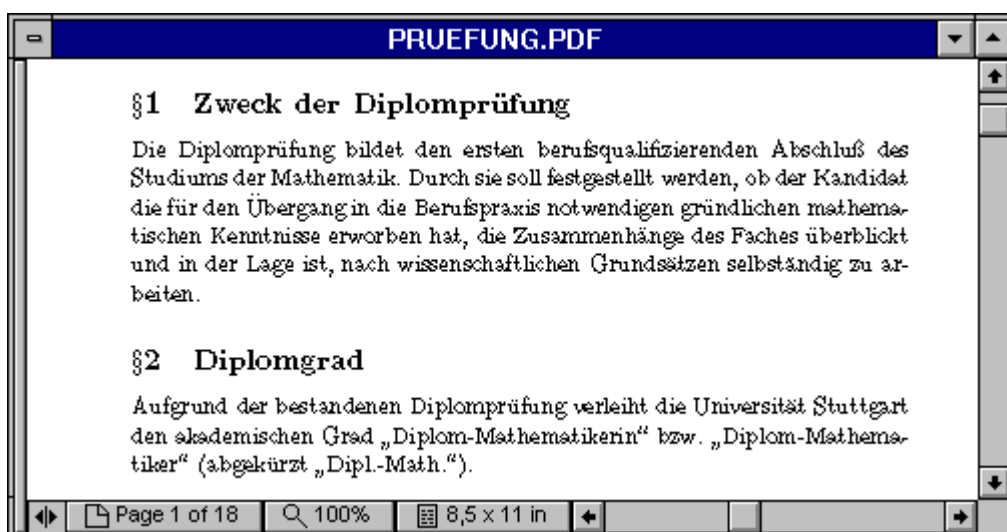


Abb. 2: Beispiel für gut lesbares PDF mit PS-Typ 1 Fonts

## Ausgabe und Bearbeiten von PDF-Dokumenten

Auch für die Nutzung von PDF-Dokumenten gibt es verschiedene Möglichkeiten. Ausgabe am Bildschirm und Drucker lassen sich mit dem frei erhältlichen Postscript-Viewer Ghostview ebenso realisieren wie mit dem etwas komfortableren, ebenfalls kostenlosen Acrobat Reader.<sup>2</sup> Dieser ermöglicht Navigation, Volltextsuche und im PC-Bereich das Abspielen von im Dokument eingebetteten Ton oder Videosequenzen mittels Quick Time. Der Reader kann als Plug-In zu einem Web-Browser installiert werden. In diesem Fall erscheint ein PDF-Dokument innerhalb des Browsers. Ansonsten kann der Browser so konfiguriert<sup>3</sup> werden, daß der Acrobat Reader automatisch mit dem Dokument als externes Fenster geöffnet wird.

Weitere Möglichkeiten eröffnet das (allerdings kostenpflichtige) Programm Acrobat Exchange. Mit ihm können PDF-Dokumente nachträglich bearbeitet, mit Links versehen, verbunden und durchsucht werden. Ein Vorteil von Exchange besteht darin, Dokumente mit Sicherheitsmechanismen versehen zu können. Dazu gehört das Sperren bestimmter Funktionen, wie z.B. Ausdruck, Bearbeiten, Kopieren von Text bzw. Grafik und Hinzufügen bzw. Ändern von Annotationen. Dies ermöglicht einerseits Dokumente authentisch elektronisch zu archivieren und erlaubt andererseits auch einen besseren Copyright-Schutz für Autoren und Verlage.

Das Problem etwaiger fehlender Fonts bei der Darstellung eines Dokuments am Bildschirm wurde bei PDF nicht dadurch gelöst, daß die Fonts zusammen mit dem Dokument gespeichert werden. Statt dessen werden alle im Dokument benutzten Fonts in sogenannten „font descriptors“ beschrieben. Diese *descriptors* enthalten den Fontnamen und Informationen zu Art und Abmessung der Fonts. Fehlende Fonts werden beim Laden des Dokuments durch generische Fonts ersetzt, die sich aufgrund der in den „font descriptors“ enthaltenen Informationen automatisch dem Aussehen der fehlenden Fonts anpassen.<sup>4</sup>

## Byteserving-Technologie

Durch das "Byte Range Retrieval Extension to HTTP" - Protokoll ist es möglich PDF-Dokumente auch seitenweise von einem Web-Server zu laden bzw. das Dokument „im Hintergrund“ zu laden, während die erste Seite betrachtet wird. Damit dieser Vorgang erfolgreich ablaufen kann, müssen jedoch verschiedene technische Voraussetzungen erfüllt sein.

Sowohl eingesetzter Web-Browser wie auch der Web-Server müssen die Byte-Range-Technologie unterstützen. Dies ist bei den Browsern von Netscape (ab Version 3.x) und Microsoft (ab Version 3.x), sowie bei den gängigsten Web-Servern (Apache ab Version 1.2, Netscape Enterprise Server ab Version 2.0, OpenMarket Secure WebServer ab Version 2.0, Microsoft Internet Information Server u.a.) der Fall. Die meisten anderen Web-Server können durch ein entsprechendes Plug-In oder ein frei verfügbares Perl-Skript nachgerüstet werden.<sup>5</sup> Zu guter letzt müssen auch die PDF-Dokumente so strukturiert sein, daß sie ein seitenweises Laden ermöglichen. Eine derartige Optimierung der Dokumente kann mit Acrobat Exchange ab Version 3.0 realisiert werden.

Diese sehr sinnvolle und zeitsparende Technik ist derzeit aufgrund der geschilderten Voraussetzungen noch nicht weit verbreitet, sie wird jedoch voraussichtlich in Zukunft an Bedeutung gewinnen.

---

<sup>2</sup> Erhältlich z.B. unter <http://www.adobe.com/prodindex/acrobat/readstep.html>

<sup>3</sup> Netscape Version 3.x: Menü Options - General Preferences - Helpers

Internet Explorer Version 3.x: Menü View - Options - Programs - File Types

<sup>4</sup> Vgl. *Portable Document Format Reference Manual Version 1.2*, S. 31

<sup>5</sup> <http://www.adobe.com/prodindex/acrobat/byteserve.html>

## Weitere Acrobat-Komponenten

Mit dem Tool Acrobat Catalog kann ein Volltextindex aus mehreren PDF-Dateien erstellt werden. Dieser kann mit dem in Acrobat Exchange und Acrobat Reader (ab Version 3.01) enthaltenen Plug-In Search durchsucht werden. Mit Hilfe von Search können indexierte PDF-Dokumente nach einzelnen Wörtern, Begriffen oder Zeichenfolgen durchsucht werden. Einzelne unbekannte Zeichen innerhalb eines Begriffes können durch Platzhalter ersetzt werden. Search ist dabei auf die Suche in lokalen Netzwerken begrenzt. Es existieren jedoch verschiedene Software-Produkte mit denen eine Volltextrecherche in PDF-Dokumenten über das WWW möglich ist. Dazu gehören Mediasphere (Cascade Systems), Retrievalware EFS (Excalibur Technologies Corp.), Excite (Excite), Surfboard (Fulcrum), Livelink Search (Open Text Corp.), PL Web (Personal Library Software) und SearchPDF (Verity).

Acrobat Capture ermöglicht das Konvertieren von Papierdokumenten in PDF-Dateien. Mit Hilfe des Acrobat Capture Reviewer können Dokumente vor dem Speichern überprüft und korrigiert werden.

## Sicherheitsaspekte

Es soll nicht verschwiegen werden, daß immer wieder auf bestehende Sicherheitsmängel bei PDF hingewiesen wurde. Als gravierendster Mangel gilt dabei die Tatsache, daß unter Windows (3.x, 95 und NT) von PDF-Dateien aus Programme und Systemaufrufe ausgeführt werden können ohne daß der Benutzer darüber informiert wird.<sup>6</sup>

Als Lösung diese Problems wird ab Version 3.01 des Acrobat Readers bzw. Acrobat Exchange dann eine Meldung ausgegeben, wenn ein Programm oder eine ausführbare Datei gestartet werden soll.

## Projekte im Bereich Web-Publishing mit PDF

PDF hat während des letzten Jahres eine rasche Verbreitung erfahren. Immer mehr Projekte im Bereich des elektronischen Publizierens setzen auf dieses Format, unter anderem der amerikanische Hochschulschriftenverlag University Microfilms<sup>7</sup> und die in partieller Zusammenarbeit damit entstehende Networked Digital Library of Thesis and Dissertations (NDLTD).<sup>8</sup>

In Europa sind die Bestrebungen des WebDoc-Projekts der niederländischen Pica-Stiftung, der Universität Groningen,<sup>9</sup> des Bibliothekssystems Oldenburg<sup>10</sup> und der Kassel University Press<sup>11</sup> zu nennen. Auch das bereits erwähnte Volltextprojekt der Universität Stuttgart (OPUS)<sup>12</sup> wird PDF als Dokumentformat verwenden.

---

<sup>6</sup> Jürgen Kuri: „Und Äktschn! - Sicherheitsrisiko Acrobat PDF“, in *c't* 1997, Heft 6, S. 48 ([http://www.heise.de/ct/art\\_ab97/9706048/](http://www.heise.de/ct/art_ab97/9706048/))

<sup>7</sup> <http://www.lib.umi.com/solutions/>

<sup>8</sup> <http://www.ndltd.org/> Vgl. auch <http://www.dlib.org/dlib/september97/theses/09fox.html>

<sup>9</sup> <http://www.ub.rug.nl/html/eldoc.html>

<sup>10</sup> <http://www.bis.uni-oldenburg.de/bisverlag/bisverle.html>

<sup>11</sup> <http://www.upress.de/>

<sup>12</sup> <http://www.uni-stuttgart.de/Rus/Projekte/EIPub/opus1.html>

## Software

- Acrobat Reader: Freeware, erhältlich unter:  
<http://www.uni-stuttgart.de/Rus/Projekte/EIPub/viewer.html>
- Aladdin Ghostscript und Ghostview: Freeware, erhältlich unter:  
<http://www.uni-stuttgart.de/Rus/Projekte/EIPub/viewer.html>
- Acrobat PDF-Writer, Distiller, Exchange, Catalog und Capture: Lizenzsoftware, zu beziehen bei Adobe: <http://www.adobeshop.com/de/> bzw. für Hochschulen über Steckenborn Computer GmbH, Gießen: <http://www.uni-stuttgart.de/Rus/Lizenzen/rus/adobe.html>

## Literatur

- Merz, Thomas: *PostScript & Acrobat, PDF: applications, troubleshooting, and cross-platform-publishing*, Berlin: Springer, 1997
- Müller, Thomas: *Acrobat & PDF: vom Papier zur digitalen Information*, Heidelberg: dpunkt, Verl. für Digitale Technologie, 1996
- Page, Bruce ; Holm, Diana: *Web publishing with Adobe Acrobat and PDF*, New York: Wiley, 1996
- *Portable Document Format Reference Manual Version 1.2*, November 27, 1996, <http://www.adobe.com/prodindex/acrobat/details.html>

Annette Maile, 685-4468  
RUS, Abt. Informationsdienste  
E-Mail: [maile@rus.uni-stuttgart.de](mailto:maile@rus.uni-stuttgart.de)

Frank Scholze, 685-4731  
Universitätsbibliothek  
E-Mail: [Frank.Scholze@ub.uni-stuttgart.de](mailto:Frank.Scholze@ub.uni-stuttgart.de)