

## Überblick - Was ist PDF?

PDF ist ein Format, das ein Dokument geräteunabhängig und plattformübergreifend beschreibt. Es hat Ähnlichkeit mit PostScript und wurde wie dieses von Adobe Systems Incorporated entwickelt. Eine PDF-Datei enthält eine oder mehrere Seiten mit Text, Graphik, Bildern, Hypertext-Links und/oder Formulare. Eine kostenlose Software von Adobe, der Acrobat Reader, ermöglicht das Betrachten von PDF-Dokumenten.

## PDF im Browser

Findet ein Browser im Web eine am MIME-Typ `application/pdf` erkenntliche PDF-Ressource vor, so kann er auf verschiedene Weise reagieren. Normalerweise ruft er ein externes Programm (helper application) auf oder speichert die Datei nur auf Platte. Mit einem Plugin oder ActiveX-Control dagegen kann der Browser das Dokument direkt in seinem Fenster anzeigen. Noch einen Schritt weiter geht die Einbettung von PDF in HTML-Seiten. Mit den Tags `<embed>` (Netscape) bzw. `<object>` (Internet Explorer) kann ein PDF-Dokument in eine Teilfläche der HTML-Seite eingefügt werden. Anders als im Acrobat Reader fehlen hier aber die Menüs und eventuell auch die Werkzeugleiste zum Navigieren im Dokument. Beim Drucken der Web-Seite bleibt die entsprechende Fläche derzeit noch weiß. Anders als bei GIF- oder JPEG-Bildern gibt es

auch keine einfache Möglichkeit zum Abspeichern eingebetteter PDF-Dateien. Des Weiteren kann das Plugin bzw. ActiveX-Control nur wenige PDF-Dokumente gleichzeitig anzeigen und mit PDF-Verweise und Artikelflüsse nicht umgehen.

Das Plugin und das ActiveX-Control bieten aber auch Vorteile gegenüber dem Acrobat Reader. Eine mit `#xml=` an den URL angehängte Datei kann benutzt werden, um dynamisch Textstellen im Dokument hervorzuheben. (Siehe Abschnitt Volltextindizierung.) Des Weiteren kann das PDF-Dokument inkrementell geladen werden, das heißt der Benutzer kann Teile der Seite sehen noch bevor die ganze Datei geladen ist. Dazu muß die Datei optimiert sein (Option in Exchange bei Speicher unter ...) und das Byterange-Protokoll (ab HTTP 1.1) unterstützt werden. Das progressive Rendering der Seite läuft dann wie folgt ab. Zuerst werden Hypertextelemente und der Artikelfluß geladen. Damit löst ein Mausklick sofort die vorgesehene Funktion aus. Dann werden die Font-Deskriptoren (Abmaße der Zeichen) eingelesen, die nötig sind, um anschließend den Text an die richtige Stelle setzen zu können. Der Text wird erst danach zusammen mit etwaigen Vektorgraphiken angefordert und auf die Seite plaziert. Mit sogenannten Multiple Master Fonts, die stufenlose Parameter wie Stärke, Stil und Breite haben, werden die gewünschten Schriftarten gemäß den Font-Deskriptoren angenähert. Erst nachdem

auch noch Pixelgraphik geladen wurde werden die Fonts abgefragt (Aussehen der Zeichen) und der Text neu gezeichnet. Zuletzt werden Thumbnails, kleine Bilder der Seiten des Dokuments, geladen. Das stückweise Laden der PDF-Datei geschieht über das Byterange-Protokoll. Bei jeder normalen GET-Anfrage kann mit `Range: bytes=` eine Komma-separierte Liste von Dateibereichen angegeben werden, die dann vom Server als `multipart/byteranges` bzw. `multipart/x-byteranges` (alt, vor HTTP 1.1) übertragen werden.

## Beispiel für eingebettetes PDF

Mit einem Trick ist es unnötig, per Skript den Browsertyp festzustellen, um dann das richtige Tag, `<embed>` oder `<object>`, zu verwenden. `<embed>` eingeschachtelt von `<object>` wird ignoriert, wenn `<object>` unterstützt wird. Im anderen Fall werden die Parameter von `<object>` überlesen und das `<embed>`-Tag ausgewertet:

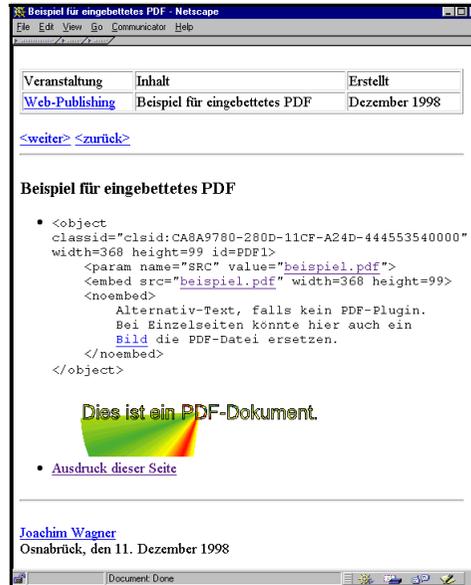
```
<object
classid="clsid:CA8A9780-
280D-11CF-A24D-
444553540000" width=368
height=99 id="PDF1">
  <param name="SRC" value="
  beispiel.pdf">
  <embed
src="
beispiel.pdf"
width=368 height=99>
```

```

<noembed>
  Alternativ-Text,
falls kein PDF-Plugin.
  Bei Einzelseiten
könnte hier auch ein
  Bild die PDF-Datei
ersetzen.
</noembed>
</object>

```

Das Attribut classid gibt eine eindeutige Kennung des ActiveX-Moduls an. Mit width und height wird der zu verwendende Platz festgelegt. id gibt einen Namen an, unter dem das Objekt referenziert werden kann, z.B. in einem VBScript-Programm. Mit <param> werden die Parameter für das ActiveX-Control gesetzt. Weitere Daten bis zu </object> werden ignoriert, außer natürlich wenn <object> nicht unterstützt wird. Dann werden die Attribute von <embed> ausgewertet. Nur wenn beide Tags nicht verstanden werden wird der Text zwischen <noembed> und </noembed> angezeigt. Da der Ausdruck eingebetteter PDF-Dokumente noch nicht funktioniert, hier ein Screenshot:



## Gestaltungsmöglichkeiten

Ein PDF-Dokument besteht nicht nur aus beliebig angeordneten Text in verschiedenen Schriftarten, sondern kann auch weitere Elemente enthalten, die die Arbeit mit dem Dokument erleichtern. Mit Lesezeichen kann zu einer vorher festgelegten Stelle gesprungen werden. Zwar werden die Lesezeichen häufig als Inhaltsverzeichnis verwendet. Sie sind aber nicht Bestandteil einer Seite des Dokuments, sondern werden außerhalb, im Acrobat Reader links vom Dokument, angezeigt. Die Lesezeichen sind hierarchisch angeordnet und lassen sich im Acrobat Reader entsprechend zusammenfallen und expandieren. Da das Sprungziel eines Lesezeichens auch in einem anderen Dokument liegen darf, läßt sich so ein ganzer

Dokumentsatz zusammenfassen. Thumbnails bieten eine zweite Möglichkeit sich im Dokument zurechtzufinden. Die verkleinerten Abbilder der Seiten helfen allerdings nur, wenn das Dokument stark strukturiert ist, da sonst alle Seiten gleich aussehen.

Eher für die Weiterverarbeitung sind Notizen gedacht, da sie nicht im Ausdruck erscheinen. In Acrobat werden sie als gelbe „Klebezettel“ angezeigt, die kleine Mengen Text enthalten können. Jede Notiz hat eine in der PDF-Datei festgelegte Position im Dokument. Notizen eignen sich daher gut für Ergänzungen und Korrekturen. Allgemeine Informationen zum Dokument bringt man besser im Dokumentinfo unter. Neben den vordefinierten Felder wie Titel und Autor können auch eigene Felder definiert werden. Letztere sind bei Suchanfragen in großen Dokumentenbeständen besonders nützlich, wenn eine Zuordnung zu einem Volltextindex hergestellt wurde. (Siehe Volltextindizierung.)

Ein Artikelfluß beschreibt, wo sich am Ende einer „Spalte“ der Text fortsetzt. Im Acrobat Reader erscheint dann jeweils unten rechts ein Pfeil zum Weiterklicken. Bei nicht für die Bildschirm Betrachtung optimierten Layouts erleichtert dies das Lesen sehr. Natürlich gibt es auch richtige Hyperlinks. Man unterscheidet bei PDF zwei Sorten: Web-Links starten automatisch den Web-Browser zur Anzeige des Dokuments. Das Sprungziel muß dabei nicht

ein PDF-Dokument sein. Es kann ein beliebiger URL angegeben werden. PDF-Links dagegen verweisen auf eine Stelle im selben oder in einem anderen Dokument. Schließlich kann ein PDF-Dokument Formularelemente wie z.B. Optionsfelder und Schaltflächen enthalten. (Siehe Abschnitt Formulare.)

## Viele PostScript-Eigenschaften

PDF übernimmt viele PostScript-Eigenschaften (Level 1 und 2). Da Text beliebig positioniert, rotiert, skaliert und eingefärbt werden kann, lassen sich beliebig komplexe Formatierungen realisieren. Neun verschiedene Farbmodelle können verwendet werden, darunter RGB, CMYK und Lab. Mit Schwarz- und Weißpunkt im CIE 1931 Farbraum, Gamma-Werten und linearen Abbildungen wird angegeben, wie die Farbwerte zu interpretieren sind. Über das Open Prepress Interface (OPI, ursprünglich von Aldus entwickelt) können externe Bilder eingebunden werden. Neben verlustfreier Kompression, mit der jedes Objekt im PDF verpackt werden kann, steht auch JPEG zur Verfügung.

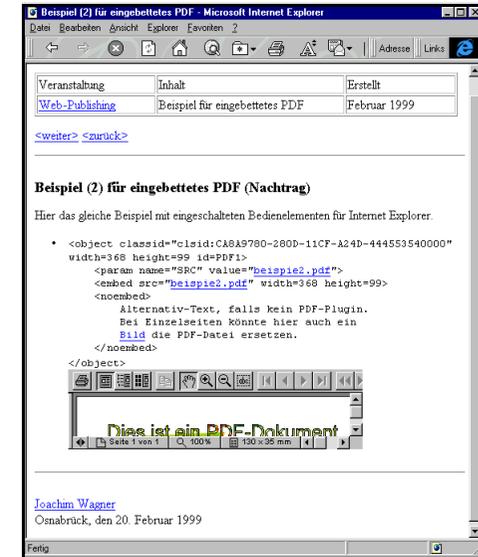
## Bedingtes Anzeigen von Dokumentbestandteilen

Ab PDF 1.2 (Acrobat 3.0) können Dokumentbestandteile bedingt angezeigt werden. Es gibt die Möglichkeit, Anzeige und Drucken oder auch nur das Drucken zu verhindern. Mit Ak-

tionen (s.u.) kann der Zustand gewechselt werden. So kann z.B. ein Hilfetext erscheinen, sobald der Mauszeiger in einen bestimmten Bereich des Dokuments eintritt. Obwohl diese Technik in die Attribute von Notizen eingebaut wurde, was etwas geflickt aussieht, wird sie wohl auch in späteren Versionen unterstützt, da im Reference Manual selbst der Tip gegeben wird, damit eine kontextsensitive Hilfefunktion zu realisieren.

## Einfluß auf angezeigte Bedienelemente

Der Browser betrachtet das PDF-Dokument als eine Einheit. Einzelne Seiten können mit seinen Navigationsknöpfen nicht angesprungen werden. In Netscape und Internet Explorer wird daher zusätzlich die Acrobat Werkzeugleiste eingeblendet, wenn PDF alleine angezeigt wird. Bei in HTML eingebetteten PDF-Dokumenten ist dagegen das Verhalten nicht gleich. Netscape verzichtet auf weitere Bedienelemente. Der Internet Explorer entscheidet entsprechend den Vorgaben in der PDF-Datei, was angezeigt wird. Eingestellt werden sie in Acrobat Exchange unter Datei - Dokumentinfo - Öffnen. Dort kann auch der Skalierungsfaktor und die Anzeige von Lesezeichen und Thumbnails festgelegt werden. Hier ein Screenshot mit eingeschalteten Bedienelementen:



Im Netscape Navigator würde die gleiche Seite in etwa so aussehen wie es der Screenshot weiter oben zeigt. An diesem Beispiel sieht man auch, daß genügend Platz für die Bedienelemente vorgesehen werden muß, da sie sonst abgeschnitten werden.

## Aktionen

In PDF-Dokumenten kann eine Vielzahl von Ereignissen Aktionen auslösen. Dazu gehören das Bewegen der Maus in eine Fläche, das Auswählen eines Lesezeichens, das Öffnen des Dokuments und vieles mehr. Als Aktionen kommen u.a. Sprünge zu anderen Seiten, Dokumenten oder URLs, Auslösen von Menübefehlen, Versenden von Formular-