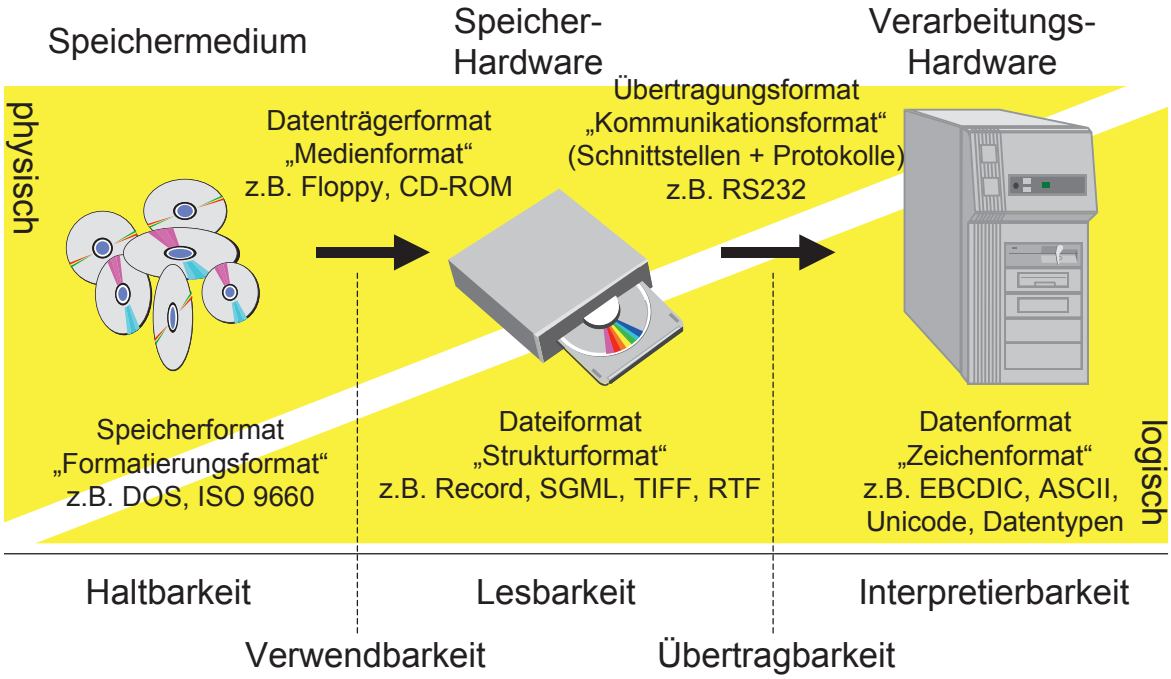
Potentielle Dateiformate zur Langzeitarchivierung von Dokumenten unter Berücksichtigung von Primär- und Metainformationen*

RAPHAEL OSTERMANN

1 Formatvielfalt

Auf dem Weg vom Speichermedium über das Lesegerät bis in den Arbeitsspeicher des Rechners und von dort durch die Programmauswertung auf den Bildschirm – und umgekehrt bei der Erstellung einer Datei und deren Speicherung auf entsprechenden Medien – sind eine Reihe von Formaten zu berücksichtigen, die sowohl die Hardware als auch die Software betreffen. Das System muß „zusammenpassen“, die unterschiedlichen Formate müssen durch die diversen Systemkomponenten unterstützt werden.

Abbildung 1: Speicherformate, Dateiformate und Datenformate



Der Schwerpunkt dieses Beitrages gilt den Formaten zur Ablage von Dokumentinformationen in Dateien, also den *Dateiformaten* zur Speicherung der Dokumentabbilder, den „Grafikformaten“, die zudem als geeignet für die Langzeitarchivierung von Schriftgut erscheinen.

Digital gespeichertes Schriftgut besteht neben den Primärinformationen der Dokumente auch aus Metainformationen, die beispielsweise die Dokumente kurzgefaßt beschreiben und die ebenfalls in Formate verpackt werden müssen. Dazu diskutiert der Beitrag einen Ansatz, wie die verschiedenen Informationen kombiniert und archiviert werden könnten.

Die trotz der Fokussierung umfangreiche Thematik kann an dieser Stelle lediglich überblicksartig und in Stichworten dargestellt werden, zur Vertiefung sei auf die Literatur verwiesen.

* Der Aufsatz konnte aufgrund einer Verhinderung des Verfassers nicht auf der Tagung des Arbeitskreises vorgetragen werden.

2 Klassifikation der Dateiformate

Zur Unterscheidung der Vielzahl von Grafikformaten können die nachstehenden Kriterien herangezogen werden. Zu allen Punkten sind als Beispiele Formate mit ihren gebräuchlichen Kurzbezeichnungen angegeben.

- „Kodierung“ der elektronischen Dokumente:
  - *NCI-Dokumente (non coded information):* Darstellung des Dokuments als eine Folge von Bildpunkten, z. B. als Bitmap-Grafik: TIFF, BMP, PCX, GIF, JPEG, PNG. Diese Formate erlauben keine sofortige Volltextrecherche, sondern erfordern die vorhergehende OCR[1]-Wandlung.
  - *CI-Dokumente (coded information):* Interpretierbare Darstellung des Dokuments als Codefolge, z. B. durch eine Seitenbeschreibungssprache: PS, RTF, SGML, HTML, XML, DOC, ODA/ODIF. Diese Formate können (mit Einschränkungen) direkt im Volltext recherchiert werden.

- Art der Grafikdarstellung (Dateityp) als erweiterte Unterscheidung von CI/NCI:
  - *Bitmap* (entspricht NCI): TIFF, BMP, PCX, GIF, JPEG, PNG.
  - *Vektor*: Beschreibung der Lage und des Aussehens von Grafikobjekten (Text, Form) in einem Koordinatensystem: XLS (Tabelle), POV.
  - *Metafile:* Bitmap- und Vektor-Daten in derselben Datei: PS, EPS
  - *Hybrid-Text:* Unstrukturierter Text[2] und Bitmap-Daten sind in einer Datei gemischt: RTF, DOC.
  - *Hypertext:* Unstrukturierter Text mit „Links“ (Verknüpfungen, Referenzen) zu anderen (Bild-/Text-) Dateien und zu Textstellen in derselben oder anderen Dateien: SGML, HTML, XML.
  - Grafiken können weiterhin als *„Szenen“* (z. B. POV), *Animationen* (z. B. auch GIF) oder multimedial (z. B. AVI) abgelegt oder als „BLOB“ (Binary Large Object) in einer *Hybrid-Datenbank* zusammen mit strukturierten Daten verwaltet werden.

- Dateielemente:
  - *„Raw“-Formate:* ohne Formatelemente (Strukturinformationen), diese sind in den Lese-/Schreib-Programmen festgelegt, die Dateien enthalten ausschließlich Bildinformationen.
  - Datenstrukturen:
    - *Felder (fields):* Mit fester Größe oder beigefügter Größenangabe und fester relativer oder absoluter Position in der Datei, daher kann auf eine explizite Kennzeichnung verzichtet werden.
    - *Marken (tags):* Kennzeichnungen mit variabler Größe und Position, können selbst wieder Marken oder Felder enthalten.
    - *Ströme (streams)*: Nur Beginn und Endpunkt sind bekannt; streams können in Pakete variabler Größe eingeteilt sein.
    - Beispiele: TIFF kombiniert tags und fixed fields, GIF fixed fields und streams; Vektordaten sind in streams organisiert; Hybrid-Text und Hypertext verwenden in der Regel tags.

[1] OCR = Optical Character Recognition.

[2] Unter *unstrukturiertem Text* werden Zeichenfolgen verstanden, die keiner Struktur aus Längenbeschränkung und Position in der Datei unterworfen sind. – Unstrukturierter Text kann *formatiert* sein, so daß Zeichenfolgen mit Eigenschaften wie fett, kursiv etc. ausgezeichnet sind. – Begrifflich ist unstrukturierter Text von der *Dokumentenstruktur* (Textgliederung durch Adresse, Absender, Haupttext, ...) zu unterscheiden.

  - Im Gegensatz zu den selteneren Raw-Formaten enthalten die anderen, strukturierten Formate einen Header, der das Bild beschreibt (z. B. Formatversion, Größe), und gegebenenfalls die benutzte Farbpalette, die eigentlichen Bildinformationen und unter Umständen weitere Informationsblöcke.

- Kompressionsverfahren (insbesondere bei Bitmap-Formaten)
  - *Verlustlos:* RLE, Packed Bits, LZW, CCITT (Huffman).
  - *Verlustbehaftet:* JPEG, Fraktale, Wavelet. Verluste entstehen durch Reduzierung der Qualität oder durch Beschreibung der Grafik anhand von Berechnungsformeln, die der tatsächlichen Darstellung angenähert werden. Beispiele für Artefakte sind Farbverfälschungen, Weichzeichnung vormals scharfer Kanten (insbesondere bei Schriften und ihrer Lesbarkeit wichtig), Moiré-Muster, Blockmuster. Der Umfang der Verluste kann frei gewählt werden, beeinflußt aber die Komprimierungsdichte (Dateigröße).
  - Allgemein bringen alle Formate Verluste mit durch:
    - *Farbreduktion:* Die natürliche, „kontinuierliche" Farbe wird digital abgebildet auf eine begrenzte Anzahl Farben: Farbtiefe (1 bis 24 Bit, d. h. schwarz/weiß bis 16 Millionen Farben), Palette, Farbmodelle (YUV, RGB).
    - Bei Bitmap-Formaten ist die *Auflösung* des Bildes begrenzt, d. h. das Bild wird in Punkte (Pixel) zerlegt, deren Anzahl entweder vom Grafikformat her begrenzt ist oder vom Benutzer eingeschränkt wird (z. B. auf Bildschirmauflösung (ca. 75 Bildpunkte pro Zoll (dots/pixel per inch)) oder Druckerauflösung (z. B. 300 dpi). Bitmap-Formate können daher als „diskret" bezeichnet werden, während Vektor-Formate „kontinuierliche" Verläufe beschreiben.

- Plattform(un)abhängigkeit
  - *Big-endian* (Motorola MC680xx) und Little-endian (Intel): Anordnung zweier Bytes (Wort) zueinander (Low-Byte und High-Byte).
  - *Fließkommadarstellung* (Punkt oder Komma, Zeichenposition).
  - *Bit-Anordnung* in einem Byte (Leserichtung).
  - *Dateinamenkonventionen* (z. B. MS-DOS: 8 Zeichen für Dateinamen, 3 Zeichen für Dateierweiterung).

- Hersteller(un)abhängigkeit
  - *Standardisierte Formate:* JPEG (ISO[3] 10918), HTML 4.0 (W3C[4]), XML 1.0 (W3C), PNG 1.0 (W3C), ODA/ODIF (ISO 8613), SGML (ISO 8879).
  - *Defacto-Standards:* TIFF CCITT[5] Gruppe 4, PS, PDF, RTF.
  - *Proprietäre Formate:* DOC, XLS, PCX, GIF, BMP.

Von besonderem Interesse sind einige *Speicherungstechniken* der Dateitypen, wobei im folgenden mit „Bild" eine Dokumentseite gemeint ist:

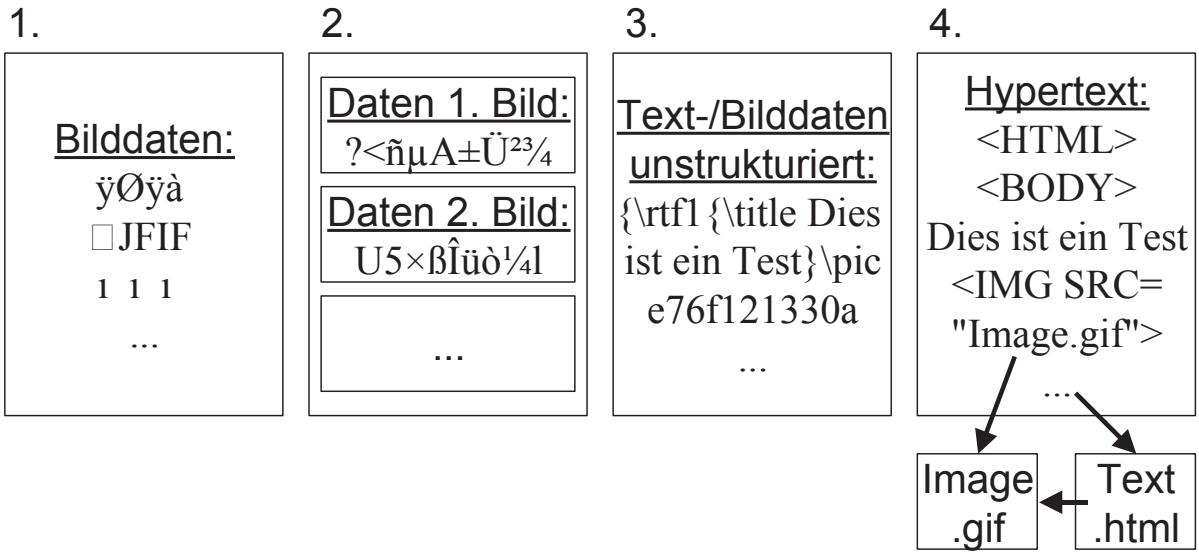
1. Bilddateien, die ein einziges Bild enthalten, z. B. PCX, BMP, JPEG, PNG.
2. Bilddateien mit der Möglichkeit zur strukturierten Speicherung von mehreren Bildern in der Datei, z. B. TIFF, GIF, XLS (mit Bild = Tabellenblatt).
3. Bilddateien mit unstrukturiert gemischten Bildelementen (Text/Grafik) in der Datei, z. B. PS, RTF,

---

[3] *ISO = International Standardization Organization*, beteiligt sind Normausschüsse aus mehr als 50 Ländern.
[4] *W3C = WorldWideWeb-Consortium*, Gremium mit Vertretern u.a. von IBM, Microsoft, Netscape, Novell, Sun, das sich die *Empfehlung (Recommendation)* von Standards im Bereich des WWW zur Aufgabe gesetzt hat.
[5] *CCITT/TSS = Consultative Committee for International Telegraphy and Telephony*. Vorläufer der *Telecommunications Standardization Sector (TSS)* in der *ITU (International Telecommunication Union)*.

PDF, DOC.

4. Bilddateien mit Verweisen auf externe Bilddateien, wobei die Verweise erst bei der Ansicht der Ausgangsdatei aufgelöst und in das Ursprungsbild eingebunden werden, z. B. HTML, XML.
5. Bilddateien mit Dokumentstrukturierung, d. h. nicht das Layout, sondern die *Struktur eines Dokuments* wird beschrieben (Kopf, Absender, Adressat, Textbausteine, ...), z. B. SGML, XML, eingeschränkt auch HTML.



Während die Formen 1. bis 3. für die Speicherung eines Dokuments nur eine einzige Datei anlegen, wird das Dokument im Hypertextformat (4.) für die Speicherung zerlegt bzw. verweist auf mehrere Dateien, wobei diese Verknüpfungen unterschiedlich gehandhabt werden: Image-Dateien werden bei der Anzeige des Textes in die Darstellung eingebaut, Verweise auf andere Hypertext-Dateien hingegen werden nur gekennzeichnet (z. B. durch Unterstreichung). Wie Verweise in Hypertext behandelt werden, ist abhängig vom Typ der referenzierten Datei.

Ein Beispieltest: Microsoft WORD 97 erlaubt die Speicherung in diversen Formaten, neben dem originären (DOC), beispielsweise in RTF oder HTML, über den Umweg des Druckens auch PS. Wird die DOC-Datei, die sowohl Text als auch Bilder enthält, als HTML gespeichert, so erzeugt WORD einzelne Image-Dateien (GIFs) für die Bilder und verweist in der HTML-Datei auf diese. Beim Archivieren ist darauf zu achten, daß sämtliche zusammengehörige, d. h. referenzierende und referenzierte Dateien abgelegt werden müssen, um die Konsistenz des Dokuments zu erhalten, ganz im Gegensatz zu der „allumfassenden" einzelnen DOC-Datei. Die Umwandlung ins HTML-Format stellte sich im Test als fatal heraus, denn eine einfache Strichzeichnung, die mit den Bordmitteln von WORD erstellt wurde, ging bei der Konvertierung verloren und tauchte weder in HTML-Text noch als Grafikdatei auf!

3 Konvertierungsmöglichkeiten zwischen NCI und CI

Generell sollten Konvertierungen vermieden werden, da mit ihnen in der Regel ein Qualitätsverlust bezüglich der Darstellung (z. B. Farbverfälschungen, Unschärfen) oder der Datenkorrektheit – beispielsweise bei fehlerhafter OCR-Erkennung von Bitmap-Schriften – einhergeht.

Die Fehlerquellen für die Konvertierungsmöglichkeiten sind im einzelnen:

- *Bitmap zu Bitmap:*

- Farbverlust, wenn das Zielformat eine geringere Farbtiefe hat (z. B. TIFF bis 24 Bit Farbtiefe nach GIF mit 8 Bit)
- Farbverfälschungen durch Farbverlust (s.o.) oder unterschiedliche Farbdarstellung (Farbmodelle/Paletten)

◆ *Vektor zu Vektor:*

- Deckungsungleiche Bildobjekte: Aussehen und Anzahl, z. B. Formen, Bezier-Kurven, Hintergrundmuster, Schriften; Verlust von Bildobjekten oder Formatierungen
- Positionierung der Bildobjekte: Relative/absolute Koordinaten, Granularität der Einheiten, unterschiedliche Maßeinheiten – sogar der „typografische" Punkt kann unterschiedlich groß bemessen sein (1/72 Zoll, 0,375 mm, ...).

◆ *Metafile zu Metafile:* Siehe Bitmap zu Bitmap und Vektor zu Vektor

◆ *Vektor/Metafile zu Bitmap:*

- Auflösung, Größe, Farbenzahl etc. – und damit Qualität – des Ziel-Bitmaps müssen festgelegt werden.
- Qualitätsverlust bei Diagonalen, Kreisen und Schriften durch „Treppenstufen", Größe des Verlusts ist abhängig von der festgelegten Auflösung.

◆ *Bitmap/Metafile zu Vektor:*

- Umwandlung von Grafikelementen in Bildobjekte (Vektorisierung).
- Möglicher Farbverlust.
- Bei Umwandlung von Schriften aus dem Bitmap in Vektordaten sind OCR-Routinen erforderlich, die je nach Schriftfamilie (z. B. Schreibschriften, Zierschriften) und dargestellter Größe und Qualität (Auflösung) fehleranfällig oder unbrauchbar sind.

◆ *Bitmap/Vektor zu Metafile:* Siehe Bitmap/Metafile zu Vektor

◆ *Hybrid-Text und Hypertext:* Siehe Metafile-Konvertierungen

4 Vorstellung einzelner Dateiformate

Vorab sei bemerkt, daß *ASCII* (American Standard Code for Information Interchange, ISO 646) *kein Dateiformat* ist, sondern ein Zeichensatz-Code auf 7-Bit-Basis, bei dem jedes Zeichen in einem Byte (8 Bit, zusätzliches Bit mit Wert Null) abgelegt wird. Daher sind nur die ersten 128 Zeichen standardisiert. Nationale Sprachzeichen wie die deutschen Umlaute, die sich im Zeichencodebereich 128 bis 255 befinden, sind in ASCII nicht festgelegt! – ASCII-Text ist prinzipiell plattform-/systemabhängig. Unter Unix und Linux werden Zeilenumbrüche als Zeichen „Linefeed" gekennzeichnet, unter DOS, OS/2 und WINDOWS jedoch als „Carriage return" + „Linefeed" (zwei Zeichen)! Viele Programme ignorieren diesen Makel automatisch.

TIFF (Tag Image File Format) – Bitmap

- Entwickler/Initiator: Aldus
- Zweck: Speicherung und Austausch von Bilddaten
- Versionen: 3.0 (1986), 4.0 (1987), 5.0 (1988), 6.0 (1992)
- Unterstützte Komprimierungsverfahren: Umkomprimiert, RLE, LZW, CCITT Gruppe 3 und 4,

JPEG
- Farbtiefe: 1 bis 24 Bit
- Mehrere Bilder pro Datei speicherbar
- Pixel- (bildpunkt-) orientiert
- Erstellung und Bearbeitung durch Bildbearbeitungs-Software (zahlreich)
- Betrachtung mit einer Vielzahl von Viewern möglich
- Keine Unterstützung in WWW-Browsern
- Speicherung binär

Durch die Speichermöglichkeit von mehreren Bildern pro Datei können auch mehrseitige Dokumente zusammenhängend abgelegt werden. Dadurch wird der Zusammenhalt der Seiten gesichert und die Dateiverwaltung vereinfacht. Um die Dateigröße zu verringern, können die Daten komprimiert werden, wobei zwischen verschiedenen verlustfreien (RLE, LZW, CCITT) und verlustbehafteten (Qualitätsverlust; JPEG) Verfahren gewählt werden kann.

TIFF ist durch die Verwendung von *tags* erweiterbar – alle Versionen bauen aufeinander auf und erweitern die Vorgänger – und wird daher als kompliziert und teilweise „mysteriös" betrachtet, obwohl die Spezifikation aller *tags* verfügbar ist. Häufig wird die Spezifikation nicht korrekt in Software umgesetzt, so daß Programme fehlerhafte TIFF-Dateien erstellen oder „saubere" TIFFs nicht richtig lesen können. Der verbreitetste Fehler ist, daß ein Programm nur die erste Seite einer mehrere Seiten umfassenden TIFF-Datei anzeigt.

GIF (Graphics Interchange Format) – Bitmap
- Entwickler: CompuServe
- Mehrere Bilder pro Datei möglich
- Multimedia-Fähigkeiten (z. B. animierte Bildsequenzen)
- Maximal 256 Farben (8 Bit)
- Komprimierung: LZW
- Little-endian
- Unterstützung durch WWW-Browser

GIF war ein auch durch das Internet sehr weit verbreitetes Format, bis CompuServe Lizenzen von den Software-Herstellern verlangte, die den Komprimierungsalgorithmus (nicht Dekomprimierung) implementierten. Um den lizenzpflichtigen Algorithmus zu umgehen und wegen der geringen Farbtiefe wird heutzutage eher das JPEG-Format oder PNG verwendet.

JPEG (Joint Photographic Experts Group; JFIF (JPEG File Interchange Format)) – Bitmap
- Entwickler/Initiator: C-Cube Microsystems
- Nur ein Bild pro Datei
- Farbtiefe: Bis zu 24 Bit
- Big-endian
- Unterstützung durch WWW-Browser

Mit JPEG wird sowohl das Dateiformat bezeichnet als auch der verwendete verlustbehaftete (De-) Komprimierungsalgorithmus.

PNG (Portable Network Graphics) – Bitmap
- Entwickler: PNG Development Group des W3C

- Zweck: Erweiterbares Dateiformat für die verlustlose, portable und gut komprimierte Speicherung von Rasterbildern (Bitmaps)
- Nur ein Bild pro Datei
- Farbtiefe: Bis 16 Bit für Graustufen, bis 48 Bit für Truecolor-Bilder; unterstützt auch Farbpaletten
- Komprimierung mit LZ77-Algorithmus (verwandt mit ZIP-Packer-Format), nicht lizenzpflichtig, sehr effektiv
- Beliebige Textdaten können zum Bild gespeichert werden (`tEXt`- und `zTXt`-Chunks).
- Unterstützung erst durch neuere Browser-Versionen

PNG ist vorgesehen als patentfreier Ersatz für GIF und kann ebenso TIFF in vielen allgemeinen Einsatzbereichen ersetzen.

PS (PostScript) – Metafile
- Entwickler: Adobe
- Seitenbeschreibungssprache, d. h. das Format einer Seite ist festgelegt und jede Seite wird einzeln beschrieben; die Seiteneinteilung des Textes ist fixiert
- Dokumente werden nur dann identisch dargestellt, wenn der Rechner, der zur Darstellung eingesetzt wird, auf die identischen Schriftfonts zugreifen kann, die auch der Rechner zur Erstellung der Dokumente verwendet hat
- Nicht vorhandene Schriften werden durch Standardschriften ersetzt, wodurch die Formatierung verändert wird
- Speicherung in 7-Bit ASCII
- Versionen (für Druck-/Bildschirmausgaben): Postscript Level 1, Postscript Level 2 (aktuell 3.0)

In die Dateien können Kommentare eingefügt werden, die nicht ausgewertet werden. Ausnahmen sind beispielsweise Systemkommentare zur Identifizierung der Postscript-Version.

EPS (Encapsulated PS) – Metafile
- Entwickler: Adobe
- Seitenbeschreibungssprache
- Zweck: Für Illustrations- und DTP-Anwendungen, zum Austausch von Bitmap- und Vektor-Daten
- Daten einer EPS-Datei sind in einer Untermenge des Befehlssatzes der Postscript-Seitenbeschreibungssprache kodiert.
- EPS-Dateien können Bitmaps der Seiten als Preview enthalten, die binär in 8-Bit (TIFF, WMF, PICT) oder in geräteunabhängigem 7-Bit ASCII (EPSInterchange-Format) eingebunden werden.

PDF (Portable Document Format) – Metafile
- Entwickler: Adobe
- Zweck: Geräte- und systemunabhängige Darstellung von Dokumenten im PDF-Format
- Versionen: 1.0 (1993), 1.1 (1996), 1.2 (1996)
- Wiedergabe von Text, Bildern und Grafiken
- Komprimierte Ablage von Elementen (Text, Bilder) innerhalb der PDF-Datei nach JPEG, CCITT Gruppe 3 und 4, LZW, ZIP; Daten kodiert nach ASCII-base-85-Verfahren
- Nicht vorhandene Schriften werden durch zwei Multi Master Fonts (Serif, Sans Serif) ersetzt, die die Metrik, aber nicht den Schriftschnitt einhalten, so daß die Formatierung erhalten bleibt. Spezialschriften wie besondere Symbolschriften (Ausnahme Symbol und Zapf Dingbats) werden dabei nicht berücksichtigt und müssen vollständig in die PDF-Datei aufgenommen werden (ca. 200 KB zusätzlicher Speicherbedarf pro Schrift).
- Beliebige Anzahl Seiten mit festem Seitenformat (z. B. DIN-A-4)

- Darstellung mittels Seitenbeschreibungssprache, Abbildungsmodell gemäß Postscript
- Erstellung durch spezielle Programme (z. B. Adobe Acrobat/PDF Writer, Adobe Distiller; noch nicht sehr zahlreich)
- Bearbeitung problematisch und nur in geringem Umfang möglich (z. B. Korrektur von Einzelzeichen); Adobe Exchange erlaubt Einfügen von Verknüpfungen, Erzeugen von Seitenminiaturen, Definition von Lesezeichen, Anbringen von Notizen zum Text, Löschen und Hinzufügen ganzer Seiten.
  Änderungen werden an die PDF-Datei angehängt und anhand von Verweisen referenziert. Der Gesamttext kann nicht frei bearbeitet werden.
- Einbindung von technischen Informationen wie Hypertext-Links und Verweise auf Objekte in der Dateistruktur möglich
- Betrachtung mit Adobe Acrobat Reader, Adobe Exchange oder auch Ghostview, wenige andere Programme
- Speicherung als 7-Bit ASCII oder binär
- Volltextsuche möglich, Leistung ist abhängig von verwendetem Anzeigeprogramm; binäre PDF-Daten müssen für die Volltextsuche umgewandelt werden, beispielsweise durch die Anzeige des Dokuments.

PDF dient zur Generierung von geräteunabhängig darstellbaren Dokumenten aus elektronischen Vorlagen, die in einer anderen Seitenbeschreibungssprache wie Postscript, DOC oder RTF unter Umständen mit Einbindung von Bitmap-Grafiken verfaßt wurden. Die erforderliche Konvertierungs-Software, die hauptsächlich von Adobe entwickelt und angeboten wird, wandelt lediglich die Daten von Postscript direkt oder quasi als Postscript-Druckertreiber nach PDF um. PDF ist ein reines Ausgabe- und kein Bearbeitungsformat. Eine nachträgliche Bearbeitung oder Erweiterung wie etwa das Hinzufügen (Einblenden) eines Eingangsstempels oder das Ergänzen/Überarbeiten kompletter Textpassagen und die damit verbundene neue Seitenaufteilung ist nicht möglich.

RTF (Rich Text Format) – Metafile

- Entwickler: Microsoft
- Zweck: Kodierung von unstrukturiertem, aber formatiertem Text und Grafiken für den Austausch zwischen Anwendungen
- Little-endian
- Dokumente werden nur dann identisch dargestellt, wenn die bei der Erstellung verwendeten Schriftfonts auf dem Anzeigerechner installiert sind
- Nicht vorhandene Schriften werden durch Standardschriften ersetzt
- Speicherung der Daten in 7-Bit ASCII unter Verwendung eines erweiterten Zeichensatzes (ANSI, MS-DOS, Macintosh)
- Bitmaps werden in binärer oder hexadezimaler Form unkomprimiert eingelagert, weshalb RTF-Dokumente sehr speicherintensiv werden können.
- Versionen: Spezifikation 1 (wird auch noch von MS WORD 97 verwendet)
- Farben: 256

Für Kommentare kann der `\doccom`-Befehl verwendet werden (hier könnten Metainformationen abgelegt werden). Da RTF aus einem festen Befehlssatz besteht, muß auf einen bestehenden Befehl zurückgegriffen werden, um zusätzliche Daten auswertbar abzulegen. Zwar können mit `\def` Befehlssequenzen angelegt werden, diese müssen aber von der Textverarbeitungs-Software ausgewertet und bei Änderung des Dokuments wieder zu den Dokumentdaten hinzugefügt werden, da sie sonst nicht gespeichert werden.

Nicht jede RTF-lesende/-schreibende Software interpretiert das Format korrekt und vollständig.

SGML (Standard Generalized Markup Language) – Hypertext
- ISO-Standard zur Beschreibung von Dokumentstrukturen: ISO 8879 (1986)
- Zweck: Standardisiertes Dateiformat zum Austausch von Dokumenten, das Mittel zur Beschreibung der Dokumentstruktur bereitstellt
- Verwendung des ASCII-Zeichensatzes (7 Bit; ISO 646)
- In WWW-Browsern ist nicht SGML, sondern lediglich HTML implementiert, auch wenn SGML für die Beschreibung des HTML-Standards verwendet wurde

Es gibt spezielle *SGML-Datenbanken*, die Daten und auch die Struktur der Dokumente speichern und wieder auslesen, strukturbezogene Suchfunktionen unterstützen und Ergebnisse als SGML-Dokument liefern. Nach Modifizierung von Daten oder Strukturen gewährleisten die Datenbanken Konsistenz (Integritätsbedingungen).

SGML ist im eigentlichen Sinne eine „Metabeschreibungssprache, die definiert, wie Dokumentenauszeichnungssprachen [wie HTML und XML] auszusehen haben. SGML ist also keine solche [Auszeichnungssprache]“ [6]. Durch die Standardisierung ist es aber möglich, eigene Strukturen, beispielsweise für Metainformationen, mit den Sprachelementen von SGML zu definieren, so daß Software-Werkzeuge, die SGML unterstützen, auch die neudefinierten Elemente auswerten können. Metainformationsfelder können somit als Strukturelemente festgelegt werden.

HTML (Hypertext Markup Language) – Hypertext
- Spezifizierung durch das W3-Consortium
- Zweck: HTML ist eine mit SGML erstellte DTD[7], die leicht zu handhaben ist und für die einfache Darstellung von Inhalten im WWW entworfen wurde
- Versionen: 2.0 (1994, Proposed Standard IETF[8], RFC[9] 1866), 3.0 (1995, Draft wurde nicht standardisiert, repräsentiert nicht mehr die Sichtweise von W3C und IETF), 3.2 (1997, W3C Recommendation), 4.0 (Dezember 1997)
- Bilddaten werden im HTML-Dokument als Referenz auf eigenständige Dateien verwaltet. Dies führt zu einer Aufsplittung des Dokuments in eine Reihe von Dateien. Für den Erhalt des Dokuments müssen alle Einzeldateien archiviert werden

Die Definition von eigenen Befehlen ist in HTML nicht vorgesehen. Unbekannte Befehlsfolgen werden von den Browsern nicht angezeigt. Die Browser der verschiedenen Hersteller stellen HTML-Dokumente häufig unterschiedlich dar, wenn die Auswertung von bestimmten Befehlsfolgen nicht im Standard definiert ist. Die Browser-Hersteller haben HTML meist fest im Programm kodiert und greifen nicht auf SGML zurück.

Die Version 4.0 ist der letzte Standard, der zu HTML entstand, da das W3-Consortium HTML durch XML ablösen möchte.

XML (Extensible Markup Language) – Hypertext
- Vereinfachte Form von SGML
- Entwickler: W3C
- Zweck: Anwendungsprofil für SGML, optimiert für die Nutzung im Internet
- Version: 1.0 (1998)

---

[6] iX 2/1999, S. 37.
[7] *DTD = Document Type Definition*.
[8] *IETF = Internet Engineering Task Force*, internationaler Zusammenschluß von Wissenschaftlern, Herstellern u.a. mit dem Ziel, das Internet weiterzuentwickeln.
[9] RFC = Request for Comments.

- Einfache Implementierbarkeit
- Interoperabilität mit SGML und HTML
- Wird noch nicht von den aktuellen WWW-Browser-Versionen unterstützt.

XML soll der Nachfolger von HTML werden, da sich gezeigt hat, daß der Entwurf von HTML zu eng gesteckt war und nicht den Anforderungen einer vielgestaltig publizierenden Web-Gemeinde genügen konnte, was auch durch die proprietären Ergänzungen von HTML durch die Browser-Hersteller Netscape und Microsoft verdeutlicht wurde. XML ist darauf ausgelegt, erweitert werden zu können, wobei die Erweiterungen als DTD-Elemente auf dem Standard basieren und somit von jedem validierenden XML-Browser ausgewertet werden können.

Zu beachten ist, daß zusätzliche DTDs in separaten Dateien abgelegt sein können, die bei der Anzeige eines XML-Dokuments von den Browsern gelesen werden. Die „Zersplitterung“ des eigentlichen Dokuments ist daher unter Umständen wesentlich stärker als bei HTML-Dokumenten.

Der Standard ist noch sehr jung, so daß nicht vor dem Jahr 2000 mit marktgängigen Produkten zu rechnen ist, obschon XML-Werkzeuge in der Entwicklung und als Beta-Versionen verfügbar sind.

5 Schlußbemerkung

In diesem Beitrag konnten nur wenige Dateiformate und ihre Besonderheiten angesprochen werden, die für die Langzeitarchivierung interessant sind. In vielen Bereichen, wie beispielsweise den Hypertext-Formaten, ist die Entwicklung noch nicht abgeschlossen, und neue Bitmap-Formate wie PNG entstehen. Es kann keine Aussage getroffen werden, welches Format sich durchsetzen und langfristig Unterstützung finden wird.

6 Literaturhinweise

- Henning Behme: Kunst der Stunde. Wozu die Extensible Markup Language gut ist. In: iX. Magazin für professionelle Informationstechnik, Heft 2, 1999, S. 36–41.
- Konzept zur Aussonderung elektronischer Akten (Schriftenreihe der KBSt, Band 40). Bonn/Köln 1998.
- Stefan Middendorf: Wohlgeformte Bohnen. XML-Verarbeitung mit Java. In: iX. Magazin für professionelle Informationstechnik, Heft 2, 1999, S. 42–49.
- James D. Murray, William van Ryper: Encyclopedia of Graphics File Formats. Sebastopol (USA) 1994.
- Wolfgang Rieger: SGML für die Praxis. Ansatz und Einsatz von ISO 8879. Mit einer Einführung in HTML. Berlin u.a. 1995.
- W3C: PNG (Portable Network Graphics) Specification Version 1.0. W3C Recommendation 01-October-1996. Siehe http://www.w3.org/TR/REC-png.html.
- W3C: Extensible Markup Language (XML) 1.0. Empfehlung des W3C, 10. Februar 1998. Deutsche Übersetzung von Henning Behme und Stefan Mintert.
  Siehe http://www.mintert.com/xml/ REC-xml-19980210-de.html.
  Originaltexte (englisch) siehe http://www.w3.org/TR/1998/REC-xml.