Gutachten

Konventionelle Dateiformate für die Archivierung im Kontext klinischer Studien

im Auftrag der

TMF - Technologie- und Methodenplattform für die vernetzte medizinische Forschung

Version 1.0 vom 28. Oktober 2007

TMF-Produktnummer P042021

Gutachter:

Antje Brandner, Dr. Ralf Brandner

TMF - Technologie- und Methodenplattform für die vernetzte medizinische Forschung e. V., Neustädtische Kirchstr. 6, 10117 Berlin / Tel. 030 31011950 / info@tmf-ev.de

Inhaltsverzeichnis

1 Einleitung

Die Durchführung klinischer Studien unterliegt vielfältigen Regeln und gesetzlichen Vorgaben. Sowohl nach der Beendigung als auch nach dem Abbruch einer klinischen Studie müssen die Studienunterlagen archiviert werden. Die Archivierungszeiträume für Studienunterlagen betragen nach gesetzlichen Vorschriften bis zu 30 Jahre und mehr [Häber 2005; Semler 2005]. Archivierung bedeutet in diesem Zusammenhang nicht nur die langfristige Speicherung der Studienunterlagen, sondern auch den Zugriff auf deren Inhalte. Daraus ergibt sich bei der elektronischen Archivierung der Studienunterlagen die Herausforderung, dass die elektronischen Studiendaten über sehr lange Zeiträume zugreifbar und verarbeitbar sein müssen.

Im Rahmen dieses Gutachtens im Auftrag der TMF (Telematikplattform für Medizinische Forschungsnetze e.V.) sollen konventionelle Dateiformate untersucht und hinsichtlich ihrer Eignung für die Archivierung im Kontext klinischer Studien bewertet werden. Das Dateiformat, welches auch als Dateityp bezeichnet wird, ist eine Konvention, die angibt, auf welche Weise die in der Bitfolge einer Datei enthaltene Information zu interpretieren ist. Unter konventionellen Dateiformaten werden die Formate verstanden, die bereits seit mehreren Jahren im Rahmen der elektronischen Archivierung verwendet werden.

Eine Umfrage bei 24 TMF-Verbünden vom Januar 2007 zeigte einen hohen Bedarf an folgenden konventionellen Dateiformaten:

- PDF (Portable Document Format) und TIFF (Tagged Image File Format) als „klassische“ Dateiformate für die Archivierung,
- Office Formate wie z.B. Microsoft Word, OpenDoc und EPS (Encapsulated PostScript) als Textformate,
- ASCII (American Standard Code for Information Interchange) und CSV (Character Separated Values) als Dateiformate für Rohdaten,
- JPEG (Joint Photographic Experts Group) und DICOM (Digital Imaging and Communications in Medicine) für die Archivierung von Bilddaten.

Darüber hinaus werden HTML (Hypertext Markup Language) als Dateityp des Internets und S/MIME (Secure Multipurpose Internet Mail Extensions) als Dateiformat für gesicherte E-Mails untersucht.

Im folgenden Kapitel werden zunächst die Bewertungskriterien dargestellt, anhand welcher die konventionellen Dateiformate analysiert und bewertet wurden. Im Anschluss daran werden die Dateiformate beschrieben und hinsichtlich der Kriterien bewertet. Im letzten Kapitel des Gutachtens werden die im Pflichtenheft für diese Gutachten formulierten Fragen beantwortet.

2 Kriterien zur Bewertung der Dateiformate

Je nach Anwendungszweck werden verschiedene Anforderungen an ein Dateiformat gestellt und es können unterschiedliche Bewertungskriterien abgeleitet werden. Um die Eignung eines Dateiformats für die Archivierung im Kontext Klinischer Studien beurteilen zu können, werden in diesem Dokument die Kriterien

1. Zweck und Verwendung,
2. Transparenz und Standardisierung,
3. Stabilität,
4. Präsentation in den Ausprägungen Präsentationswerkzeuge, Darstellung und Struktur sowie
5. Sicherheit in den Ausprägungen Sicherheitsmechanismen und elektronische Signatur

genauer untersucht und bewertet.

Diese Bewertungskriterien, welche im Rahmen umfangreicher Untersuchungen des Forschungsprojektes „ArchiSig – Beweiskräftige und sichere Langzeitarchivierung digital signierter Dokumente“ aufgestellt wurden, haben sich zur Bewertung der Eignung von Dateiformaten im Kontext der sicheren und beweiskräftigen Langzeitspeicherung im Gesundheitswesen bewährt [Hollerbach 2003]. Die Gesamtergebnisse des Projektes „ArchiSig“ sind in [Roßnagel 2005] zusammengefasst. Bezüglich der Inhalte decken die hier verwendeten Bewertungskriterien die laut Maßnahmenkatalog[1] der IT-Grundschutz-Kataloge des BSI beschriebenen Kriterien ab und gehen darüber hinaus [BSI 2006].

Da die verwendeten Kriterien zur Analyse und Bewertung von Dateiformaten für Texte, Bilder, Tonaufzeichnungen und Videos entwickelt wurden, sind diese abstrakt und allgemein formuliert.

Die Ergebnisse der nachfolgenden Untersuchungen werden auf einer Skala eingeordnet. Die Werte der Skala zur Beurteilung reichen von + + bis – – und haben folgende übergeordnete Bedeutung:

+ + *Das Dateiformat ist sehr gut geeignet.*

+ *Das Dateiformat ist geeignet.*

– *Das Dateiformat ist mit Einschränkungen geeignet.*

– – *Das Dateiformat ist nicht zu empfehlen.*

Kann ein Dateiformat in einem Kriterium nicht beurteilt werden, so wird es mit „o“ gekennzeichnet.

Im Folgenden werden die Kriterien genauer beschrieben. Die Kriterien werden auf der Skala eingeordnet und den Skalenteilen wird semantische Bedeutung zugewiesen.

2.1 Zweck und Verwendung

Anhand dieses Kriteriums soll dargestellt werden, für welchen Zweck das Dateiformat konzipiert wurde und wofür das Format eingesetzt wird. Für den medizinischen Bereich und somit auch für den Bereich klinischer Studien ist interessant, ob es Anwendungen des Dateiformats im Gesundheitswesen gibt oder ob es explizit für diesen Bereich entwickelt wurde.

Skala für die Beurteilung von Zweck und Verwendung:

+ + *Das Dateiformat wurde für den Einsatz im Gesundheitswesen konzipiert und wird in diesem Bereich weit verbreitet angewendet.*

\+ *Das Dateiformat wurde nicht explizit für den Einsatz im Gesundheitswesen entwickelt, wird jedoch im medizinischen Bereich häufig verwendet.*

– *Das Dateiformat wurde für den Einsatz im medizinischen Bereich entwickelt, ist jedoch kaum verbreitet in der Anwendung oder befindet sich noch in der Entwicklung.*

– – *Das Dateiformat wurde nicht für den Einsatz im Gesundheitswesen konzipiert und wird in diesem Bereich kaum verwendet.*

Erläuternde Beispiele:

Als Standard für bildgebende Verfahren und Bildkommunikation wurde beispielsweise DICOM speziell für den Einsatz im Gesundheitswesen entwickelt und hat sich weltweit als Standard für die Übermittlung von Bildmaterial etabliert. Darüber hinaus finden nicht-medizinische, weit verbreitete Dateiformate auch verschiedene Anwendungen im Gesundheitswesen, z.B.

[1] Siehe M 4.170 „Auswahl geeigneter Datenformate für die Archivierung von Dokumenten“

ASCII für Labordatenübermittlung, Microsoft Word zur Arztbriefschreibung oder TIFF für gescannte Papierdokumente.

2.2 Transparenz und Standardisierung

Ein wichtiges Kriterium für die Archivierung von Studienunterlagen ist sicherlich die Transparenz eines Dateiformats. Man spricht von einem transparenten Dateiformat, wenn seine Spezifikation vollständig offen gelegt ist. Das ist beispielsweise bei einem Standard gegeben, der durch ein öffentlich bekanntes Standardisierungsgremium festgelegt wurde. Teilweise gibt es auch Industriestandards, deren Spezifikationen offen gelegt sind. Die Beschreibung einer Spezifikation durch Dritte ist jedoch nicht äquivalent mit einem offenen Standard. Von den meisten Industriestandards gibt es keine vollständige Spezifikation.

Einen Überblick über Standards und Standardisierungsaktivitäten im Gesundheitswesen geben [Wirsz 2000; Binder 2001; Schug 2001].

Skala zur Beurteilung der Transparenz und Standardisierung:

+ + *Das Dateiformat ist ein De-jure-Standard, die Spezifikation ist offen gelegt.*

+ *Das Dateiformat ist ein De-facto-Standard, die Spezifikation ist offen gelegt.*

– *Das Dateiformat ist ein De-facto-Standard ohne offen gelegte Spezifikation.*

– – *Das Dateiformat ist nicht standardisiert, die Spezifikation des Standards ist nicht offen gelegt.*

Wurde das Dateiformat von einem internationalen Standardisierungsgremium standardisiert, z.B. von der ISO, so wird es zusätzlich mit „*“ gekennzeichnet.

Erläuternde Beispiele:

ASCII ist ein Beispiel für ein Dateiformat, das von einem Normierungsgremium standardisiert und offen gelegt wurden. PDF und PostScript sind Industriestandards mit offen gelegten Spezifikationen. Die meisten industriellen Hersteller legen die Spezifikationen ihrer Dateiformate jedoch nicht offen. Die Spezifikation von Microsoft Word war lange Zeit nicht offen gelegt, so dass keine anderen Hersteller Werkzeuge zum direkten Verarbeiten von Word-Dokumenten entwickeln bzw. Entwicklungen nur in direkter Abstimmung mit der Microsoft Corporation vornehmen können.

2.3 Stabilität

Um ein Format für die Langzeitspeicherung einsetzen zu können wird gefordert, dass das Format keine häufigen Versionswechsel und Änderungen der Nutzdatenstruktur besitzt. Dies wird im Folgenden als Stabilität des Formats bezeichnet.

Die Stabilität eines Dateiformats ist oftmals schwer einzuschätzen, gerade wenn es sich um ein relativ neues Format handelt. Die hier verwendete Möglichkeit, die Stabilität durch einen berechneten Wert einzuschätzen, ist:

$$\textit{Stabilität } S = \frac{\textit{Alter } A}{\textit{Anzahl der Versionen}} \quad \text{mit}$$

Alter A = Jahr der Bewertung B – Jahr der ersten Veröffentlichung des Dateiformats E

Für B wird im Rahmen dieses Gutachtens das Jahr 2007 eingesetzt.

Der Wert für die Stabilität S <= 1 bedeutet einen häufigen, im Durchschnitt mindestens jährlichen Versionswechsel. Je größer der Wert S wird, desto weniger Versionswechsel gibt es bei dem Dateiformat.

Skala zur Beurteilung der Stabilität:

+ + *Wert für* $S > X_3$,

+ *Wert für* $S > X_2$ *und* $S \leq X_3$,

– *Wert für* $S > X_1$ *und* $S \leq X_2$,

– – *Wert für* $S \leq X_1$,

mit $X_1 = 1$, $X_2 = 10$ *und* $X_3 = 30$.

Für X_1 wird im Folgenden der Wert 1 angenommen, weil dieser durchschnittlich einen jährlichen Versionswechsel bedeutet. Die Werte für X_2 und X_3 lassen sich aus den gesetzlich vorgeschriebenen Archivierungsfristen ableiten, die zwischen 10 und 30 Jahren liegen.

Da ein kleiner Wert von S nichts über die Versionsänderungen und die Stabilität in den letzten Jahren aussagt, sollte noch ein zusätzlicher Wert betrachtet werden, der als Zeitraum seit der letzten Änderung bezeichnet wird.

Zeitraum seit der letzten Änderung Z = B – Jahr der Veröffentlichung der letzten Version L

Es sollte auch untersucht werden, ob es bei Versionswechseln gravierende Änderungen in der Nutzdatendefinition gab, so dass die Kompatibilität zu vorherigen Versionen nicht mehr gegeben ist.

Erläuternde Beispiele:

Ein Dateiformat muss auch noch nach Jahren interpretiert werden können. Häufige Versionswechsel erschweren die Verarbeitung von Dateien in älteren Versionen eines Dateiformats. Werden bei neuen Formatversionen nicht nur Erweiterungen der Formatspezifikation vorgenommen sondern auch Änderungen in der Definition der Nutzdaten, so wird das Lesen von alten Dateien mit neuen Anwendungssoftwareprodukten teilweise unmöglich.

2.4 Präsentation

Als Präsentationsproblem wird analog zu [Pordesch 2000] der Fall bezeichnet, dass mindestens zwei Präsentationen derselben Daten so voneinander abweichen, dass sie von Menschen unterschiedlich interpretiert werden. Das Präsentationsproblem ist abhängig von der jeweiligen Interpretation der Daten durch die genutzten Programme, die im Betriebssystem vorhandenen Zeichensätze (Fonts), der vom Benutzer getätigten Hard- und Softwareeinstellungen und weiteren Faktoren.

Die Präsentation soll im Folgenden hinsichtlich der Verfügbarkeit von Präsentationswerkzeugen, der Darstellung eines Dokuments und der Strukturiertheit der Daten analysiert werden.

Skala zur Verfügbarkeit von Präsentationswerkzeugen:

+ + *Es existieren Präsentationswerkzeuge von verschiedenen Herstellern, die bzgl. der Eindeutigkeit der Darstellung validiert sind.*

+ *Es existieren Präsentationswerkzeuge von verschiedenen Herstellern, die jedoch nicht bzgl. der Eindeutigkeit der Darstellung validiert sind.*

– *Es existieren nur wenige, evt. an einen Hersteller gebundene Präsentationswerkzeuge.*

– – *Es existieren keine Präsentationswerkzeuge.*

Erläuternde Beispiele:

Je nach Interpretation der Spezifikation kann es zu werkzeugspezifischen Ausprägungen in der Anzeige desselben Dokuments kommen. Beispielsweise gibt es Unterschiede in der Darstellung von HTML-Dokumenten, abhängig davon, ob man Browser von Microsoft oder Mozilla benutzt. Bei Industriestandards ohne offen gelegte Spezifikation sind meist auch die Werkzeuge zur Darstellung des Dokuments an einen Hersteller gebunden, dies ist z.B. bei Microsoft Word der Fall.

Skala zur Beurteilung der Darstellung:

+ + *Die Interpretation der Inhalte ist erwartungskonform.*

+ *Die Darstellung aktiver Elemente kann unterbunden werden. Externe Informationen können eingebunden werden, so dass das Dokument weitgehend eindeutig dargestellt werden kann.*

– *Die Anzeige aktiver Elemente ist nicht kontrollierbar. Externe Informationen können nicht ins Dokument eingebunden werden.*

– – *Es ist nicht nachvollziehbar, ob aktive Elemente im Hintergrund verarbeitet werden.*

Erläuternde Beispiele:

Besonders im medizinischen Bereich ist die Kongruenz der Darstellung eines Dokuments unabhängig von Zeitpunkt, Ort und technischer Umgebung der Präsentation sehr wichtig. DICOM macht deshalb Vorschriften dazu, wie z.B. ein Bild interpretiert wird. Falls in einem Format aktive Elemente eingefügt werden können, so müssen diese kontrollierbar sein.

Skala zur Beurteilung der Struktur:

+ + *Das Dateiformat ist bzgl. Aufbau und Inhalt strukturiert.*

+ *Das Dateiformat ist nur bzgl. des Aufbaus strukturiert.*

– *Das Dateiformat ist nur bzgl. des Inhalts strukturiert.*

– – *Das Dateiformat ist unstrukturiert bzgl. Aufbau und Inhalt.*

Erläuternde Beispiele:

Die meisten Dateiformate sind bezüglich ihres Aufbaus strukturiert. Dabei werden die Daten in einen Format-„Header“ und eine Inhaltkomponente unterteilt. Im „Header“ sind Informationen über das Format gespeichert, z.B. Versionsnummer und verwendeter Zeichensatz oder Farbtabellen. Die eigentlichen inhaltlichen Daten der Inhaltskomponente

können wiederum strukturiert vorliegen. Ein strukturierter Inhalt ermöglicht eine multiple Verwendbarkeit der Daten.

Ein Beispiel für ein komplett unstrukturiertes Dateiformat stellt ein ASCII-Text dar. Der Großteil der Dateiformate ist im Hinblick auf den Aufbau strukturiert. Nur wenige, so genannte Meta-Dateiformate, wie beispielsweise die Extensible Markup Language (XML), strukturieren auch den Inhalt.

2.5 Sicherheit

In den wenigsten Dateiformaten wird bisher auf Sicherheitseinstellungen Wert gelegt. Auch die Anwendung der elektronischen Signatur in Dateiformaten ist noch nicht sehr weit verbreitet. Daher soll bei der Betrachtung dieses Kriteriums zunächst untersucht werden, ob und auf welche Weise die Integrität, Authentizität und Vertraulichkeit eines Dokuments gewährleistet wird.

Skala zur Beurteilung der Sicherheitsmechanismen:

+ + *Die Integrität, Authentizität und Vertraulichkeit eines Dokuments wird mit Hilfe kryptographischer Verfahren gewährleistet.*

+ *Integrität, Authentizität und Vertraulichkeit werden zum Teil mit kryptographischen Verfahren sichergestellt.*

– *Integrität, Authentizität und Vertraulichkeit werden teilweise gewährleistet, jedoch ohne kryptographische Verfahren.*

– – *Es gibt keine Sicherheitsmechanismen für die Gewährleistung der Integrität, Authentizität und Vertraulichkeit.*

Erläuternde Beispiele:

Die elektronische Signatur ist in Formaten wie PDF bereits integriert. Auch Standards im Gesundheitswesen wie DICOM bieten diese Funktionalität an. Inwieweit die in den Formaten eingebundene elektronische Signatur standardisierte Signaturformate unterstützt ist mit der folgenden Skala zu analysieren.

Skala zur Beurteilung der elektronischen Signatur:

+ + *Innerhalb des Dateiformats ist die elektronische Signatur in einem standardisierten Signaturformat enthalten, in welchem alle notwendigen Verifikationsdaten vollständig gespeichert werden können.*

\+ *Innerhalb des Dateiformats ist die elektronische Signatur in einem standardisierten Signaturformat enthalten, in welchem die notwendigen Verifikationsdaten jedoch nicht vollständig gespeichert werden können.*

– *Innerhalb des Dateiformats ist die elektronische Signatur in einem speziellen Signaturformat enthalten, in welchem die notwendigen Verifikationsdaten vollständig gespeichert werden können.*

– – *Innerhalb des Dateiformats ist die elektronische Signatur in einem speziellen Signaturformat enthalten, in welchem die notwendigen Verifikationsdaten jedoch nicht vollständig gespeichert werden können.*

Erläuternde Beispiele:

Als Dateiformat für elektronische Signaturen hat sich die Cryptographic Message Syntax (CMS) etabliert, welche u.a. von S/MIME verwendet wird. CMS ist ein von der IETF standardisiertes Dateiformat und bietet die Möglichkeit, eine beliebige Anzahl an Zertifikaten und zugehörigen Zertifikatstatusinformationen abzuspeichern.

3 Bewertung konventioneller Dateiformate

3.1 Klassische Dateiformate zur Archivierung

3.1.1 Portable Document Format (PDF)

Das Portable Document Format ist ein von der Firma Adobe Systems Incorporated entwickeltes Dateiformat, dessen erste Version 1993 veröffentlicht wurde.

Zweck und Verwendung: +

PDF ist eine vektorbasierte Seitenbeschreibungssprache, welche als plattformübergreifendes Dateiformat für die Anzeige und den Austausch fertig gestellter Dokumente konzipiert wurde. Ein PDF-Dokument kann die Inhalte des Ausgangsdokumentes einschließlich aller Farben, Raster- und Vektorgrafiken sehr präzise wiedergeben. Dies gilt grundsätzlich ebenfalls für Schriften. Die Layouttreue ist einer der wesentlichen Vorteile von PDF gegenüber anderen Beschreibungssprachen wie SGML, HTML und XML.

PDF ist hinsichtlich des Grafikmodells und einiger Entwicklungen eng mit PostScript (Kapitel 3.2.3) verwandt, setzt jedoch unterschiedliche Schwerpunkte. Das Portable Document Format wurde für die Anzeige und den Austausch von Dokumenten über alle Betriebssysteme hinweg optimiert. Im Gegensatz zu PostScript bietet PDF zusätzliche Funktionen, die über eine reine Beschreibung des Seiteninhaltes hinausgehen. Hypertextelemente, wie z.B. Verweise und Lesezeichen, erleichtern die Navigation innerhalb eines Dokuments [Borghoff 2003].

Metainformationen enthalten Angaben über das ganze Dokument oder einzelne Teile daraus. Strukturinformationen ermöglichen die Wiederverwendung der im Dokument enthaltenen Inhalte. Aus PDF-Dokumenten lassen sich Textpassagen, Tabellen und Grafiken leicht kopieren und durch Einfügen in anderen Programmen weiter verarbeiten. Die Bearbeitungsfunktionalität in PDF ist aber nicht mit der von herkömmlichen Textverarbeitungsprogrammen zu vergleichen. Neben der Notiz- und Kommentarfunktion können kleinere Änderungen wie die Korrektur von Tippfehlern vorgenommen werden.

Durch die Textsuche im einzelnen Dokument oder die Volltextrecherche innerhalb einer Dokumentensammlung lassen sich sehr einfach Detailinhalte auffinden. Der Umfang eines PDF-Dokuments kann mehrere 100.000 Seiten umfassen und auch die Seitengröße ist durch

das Format nicht begrenzt. Einen Überblick über diese und weitere Möglichkeiten von PDF gibt [Martins 2003; Reich 2005].

Basierend auf der PDF Version 1.4 wurde in der ISO-Norm 19005-1 Anforderungen festgelegt, um PDF als Dateiformat für die Langzeitarchivierung von Dokumenten nutzen zu können [PDF-Tools 2007]. ISO 19005-1 definiert "ein Dateiformat basierend auf PDF, genannt PDF/A (Archive), welches einen Mechanismus zur Verfügung stellt, um elektronische Dokumente auf eine Weise darzustellen, so dass das visuelle Erscheinungsbild über die Zeit erhalten bleibt, unabhängig von den Werkzeugen und Systemen zur Herstellung, Speicherung und Reproduktion". PDF/A präzisiert im Wesentlichen spezifische Eigenschaften der PDF Version 1.4 und definiert ob sie obligatorisch, empfohlen, eingeschränkt oder verboten sind. Verboten sind u.a. Referenzen auf Ressourcen wie z.B. Schriften, die nicht in der Datei selbst enthalten sind und damit irgendwann nicht mehr zugänglich sein könnten. Auch das Sperren von Funktionen wie Drucken und Kopieren sowie die Verschlüsselung sind untersagt. Die Mindestanforderungen der visuell eindeutigen Reproduzierbarkeit definiert PDF/A-1b. Die Lesbarkeit des enthaltenen Textes und Textextrahierung wird erst mit der höheren Konformanzebene PDF/A-1a garantiert [PDF-Tools 2007].

PDF wurde nicht für den Einsatz im Gesundheitswesen entwickelt, findet dort aber als Austausch- und Archivierungsformat mittlerweile weite Verbreitung. Im Bereich klinischer Studien wird es für nahezu alle Dokumenttypen verwendet und ist weit verbreitet. In den Forschungsverbünden wird PDF zur langfristigen Speicherung von TMFs, Labordaten, E-Mails, SOPs, Prüfplänen, Patienteneinwilligungen, CRFs und vielen anderen Dokumenttypen verwendet.

Mit PDF/H (Health care) wird darüber hinaus an einer internationalen Norm für den Austausch von Gesundheitsdaten basierend auf PDF gearbeitet. Das von der AIIM (Association for Information and Image Management) durchgeführte Projekt soll noch im Jahr 2007 erste Ergebnisse und Empfehlungen liefern.

Transparenz und Standardisierung: +

PDF hat sich trotz der Bindung an einen Hersteller zu einem De-facto-Standard für den elektronischen Dokumentenaustausch entwickelt. Die Sprachspezifikation des Dateiformates wurde von Adobe offen gelegt (siehe [Adobe 2006]). Als ISO 19005-1 wurde PDF/A im Jahr 2005 für den Bereich der Langzeitarchivierung von Dokumenten international normiert. Darüber hinaus versucht Adobe die gesamte Formatspezifikation von der ISO normieren zu lassen.

Stabilität: –

Erste Veröffentlichung E = 1993

Alter A = B – E = 14

Anzahl Versionen V = 8

Stabilität S = A / V = 1,75

Letzte Veröffentlichung L = 2006

Zeitraum seit der letzten Änderung Z = 1

Von der ersten Version des PDF-Formats, PDF 1.0 im Jahr 1993, bis zum Jahr 2007 gab es 7 Versionswechsel. Die derzeitig aktuelle Version ist PDF 1.7 aus dem Jahr 2006 ([Adobe 2006]). Die jeweils zugehörigen Acrobat Versionen sind einfach von 1 bis 8 durchnummeriert. Die erste Zeile eines PDF-Dokuments enthält die Versionsnummer, die der Datei zugrunde liegt.

Präsentation:

Werkzeuge: +

PDF-Dokumente können auf verschiedene Arten erzeugt werden. Eine Möglichkeit stellen die verschiedenen Komponenten der Acrobat-Familie dar, wie z.B. Acrobat Distiller oder Acrobat PDFWriter. Neben diesen lizenzkostenpflichtigen Programmen sind aber auch lizenzkostenfreie Programme zur Erstellung von PDF-Dateien wie z.B. PDFCreator verfügbar. Manche Anwendungssoftwareprodukte haben einen eigenen Programmcode für die PDF-Ausgabe, wobei sich die Qualität der erzeugten PDF-Dokumente jedoch unterscheidet.

Zur Anzeige von PDF-Dateien existieren verschiedene Werkzeuge. Am weitesten verbreitet ist der kostenlose Acrobat Reader der Firma Adobe. Ein weiteres Produkt ist xPDF der Firma Glyph & Cog, LLC.

Darüber hinaus stehen auch zur Erzeugung und Validierung von PDF/A Werkzeuge von verschiedenen Anbietern zur Verfügung. Neben der Adobe Acrobat Version 8 unterstützen auch Microsoft Office 2007 und verschiedene Werkzeuge von Drittanbietern wie z.B. die PDF Tools AG die Erzeugung von PDF/A [PDF-Tools 2007].

Darstellung: +

Die Layouttreue ist eine der großen Stärken von PDF, die durch PDF/A bis hin zur langfristigen visuell eindeutigen Reproduzierbarkeit gewährleistet werden soll. Dennoch sollten folgende Aspekte berücksichtigt werden, die bei normalem PDF auftreten können.

Ein PDF-Dokument soll auf jedem Rechner gleich aussehen, selbst wenn die verwendete Schrift nicht vorhanden ist. Dazu werden in PDF die Schriftabmessungen, auch Font Metrics genannt, gespeichert. Diese Informationen können dann dazu verwendet werden, andere Schriftarten entsprechend anzupassen. Dadurch ist es jedoch auch möglich, die Anzeige eines Dokumenteninhalts durch manipulierte Fonts zu beeinflussen, wenn nicht alle Schriften in der PDF-Datei eingebettet sind ([Merz 2001]).

Durch den Einsatz von Formularfeldern ist es möglich, Dokumente zu erstellen, deren Bildschirmdarstellung vom Ausdruck abweicht. Es lassen sich Formularfelder erstellen, deren Inhalte entweder nur gedruckt oder nur am Bildschirm angezeigt werden. Auf diese Weise könnte man leicht ein PDF-Dokument gestalten, welches im Ausdruck einen anderen Inhalt erzeugt, z.B. geänderte Summen bei einem Vertrag ([Merz 2001]).

Es ist ebenfalls denkbar, dass die Präsentation einer PDF-Datei durch eine eingebettete Datei oder JavaScript-Implementierungen verändert werden. Der Start solcher Dateien kann deaktiviert werden, ebenso die Ausführung von JavaScript-Code. Im Allgemeinen erhalten JavaScript-Programme in PDF keinen direkten Zugriff auf das Dateisystem.

Bei der Schwärzung von PDF-Dokumenten wird davor gewarnt, dass die Inhalte auch wirklich gelöscht und nicht nur überschrieben werden sollen, da diese sonst wiederhergestellt werden können.

Struktur: + +

PDF-Dokumente sind im Hinblick auf ihren Aufbau strukturiert. Mit Acrobat 5 wurde der Ansatz des „Tagged PDF“ erweitert, der eine hierarchische, inhaltliche Gliederung von PDF-Dokumenten erlaubt. Ein PDF-Dokument kann als „Tagged PDF“ bezeichnet werden, wenn verschiedene Bedingungen erfüllt sind. So muss es z.B. für sämtliche Textinhalte eine korrekte Unicode-Zuordnung geben und das Dokument muss einen PDF-Strukturbaum enthalten (vgl. [Merz 2002]). Um ein Tagged PDF-Dokument erzeugen zu können, ist die Unterstützung in den jeweiligen Anwendungssoftwareprodukten erforderlich, mit denen die

Dokumente vor der PDF-Konvertierung erzeugt werden wie es z.B. in Microsoft Office und Adobe FrameMaker der Fall ist.

Sicherheit:

Sicherheitsmechanismen: + +

Seit Acrobat 2.0 (PDF 1.1) gibt es verschiedene Möglichkeiten, geschützte PDF-Dateien zu erstellen, was in den folgenden Versionen noch verfeinert wurde. Ist der Inhalt eines Dokuments vertraulich und soll nur bestimmten Benutzern zugänglich sein, so kann der Ersteller die Datei über symmetrische Verschlüsselungsverfahren unter Angabe eines Passwortes verschlüsseln. Eine zweite Schutzvariante ist die Vergabe bestimmter Nutzungsbeschränkungen beim Erstellen der PDF-Datei. Auf diese Weise kann man z.B. das Ausdrucken oder Verändern der Datei verhindern. Bei der Einstellung derartiger Berechtigungen ist die Vergabe eines Haupt- und eines Benutzerkennwortes unbedingt notwendig, da die Einstellungen sonst relativ einfach umgangen werden können. Grundsätzlich ist davon auszugehen, dass die in einer PDF-Datei eingestellten Berechtigungen keinen ernsthaften Schutz bieten, da die Implementierung des Viewers die Einhaltung der Nutzungsbeschränkungen gewährleistet. Um die Zugänglichkeit der Inhalte in PDF-Dokumenten zu gewährleisten, sind sowohl die Verschlüsselung als auch die Verwendung der Nutzungsbeschränkungen in PDF/A untersagt.

Zu der Vergabe von Passwörtern ist zu sagen, dass die kurzen symmetrischen Schlüssel von Acrobat 4 unsicher sind. Einen sicheren Schutz bieten die 128-Bit-Schlüssel ab Acrobat 5 in Verbindung mit ausreichend langen Passwörtern (siehe [Merz 2002]).

Elektronische Signatur: +

Acrobat 4 führte die elektronische Signatur für PDF-Dokumente ein, welche bis heute ständig erweitert wurde. Das Signieren von Dokumenten und das Validieren von Signaturen sowie die Verwaltung der zugehörigen Zertifikate ist mit der Acrobat-Vollversion möglich und seit Version 5.1 über entsprechende PlugIns auch mit dem kostenlosen Acrobat Reader.

Es werden „SelfSign“-Unterschriften mit dem von Adobe mitgelieferten „SelfSign“-Plugin sowie Produkte von Drittherstellern durch die Spezifikation einer Schnittstelle unterstützt (siehe [Merz 2002]). Die Signatur-Plugins sind u.a. von CSC Ploenzke, SignCubes und Signature Perfect verfügbar [Adobe 2003].

In PDF können Signaturen sichtbar und unsichtbar erzeugt werden. Im Gegensatz zur Verschlüsselung sind Signaturen auch in PDF/A erlaubt. Die Signatur bezieht sich auf ausgewählte Objekte eines PDF-Dokumentes oder auf einen definierten Byteumfang. Signierte PDF-Dokumente können auch noch überarbeitet werden, da die Signatur mit der Version des Dokumentes gespeichert wird und durch die Änderung eine neue Dokumentversion entsteht. Gemäß PDF Spezifikation [Adobe 2006] werden zwei unterschiedliche Formate unterschieden. Neben einer internen Struktur zur Ablage der Signaturinformationen unterstützt PDF auch das international normierte „PKCS#7 Signature Format". Das interne Signaturformat stellt ein proprietäres Format für Signaturen von Adobe dar, bei welchem neben signaturbeschreibenden Informationen (Signaturgrund, -ort, -zeit), den Zertifikaten einer Zertifizierungshierarchie (Benutzer-, CA-, Root-Zertifikat) der verschlüsselte Signaturwert angegeben wird. Verifikationsdaten wie Zeitstempel, Sperrlisten, Zertifikatstatusinformationen (OCSP-Responses) und Signaturerneuerungen können nicht gespeichert werden. Das „PKCS#7 Signature Format" gibt die Möglichkeit Signaturen im standardisierten PKCS#7-Format [RFC2315 1998] zu speichern. Dieses diente als Grundlage des Datentyps „SignedData" der Cryptographic Message Syntax (CMS), welche auch Bestandteil von S/MIME ist und in Kapitel 3.5.2 beschrieben ist. In das PKCS#7 Element können Verifikationsdaten wie Zertifikate [RFC3280 2002], Attributzertifikate [RFC3281 2002], Zeitstempel [RFC3161 2001], Sperrlisten [RFC3280 2002] und Zertifikatstatusinformationen [RFC2560 1999] eingebettet werden. Da CMS [RFC3852 2004] weiter entwickelt wurde, Adobe aber an der mittlerweile in die Jahre gekommenen PKCS#7 Spezifikation von 1998 festhält, ist die Integration von Verifikationsdaten nicht mehr komplett standardkonform.

3.1.2 Tagged Image File Format (TIFF)

Das Tagged Image File Format wurde bereits 1986 ursprünglich von der Firma Aldus zur Speicherung hochauflösender grauskalierter Scanner-Bilder entwickelt. Die Bildinformationen werden im Rasterformat abgelegt, wobei große Bilder in „Stripes" (Streifen) oder als „Tiled" (Kacheln) organisiert sein können. TIFF unterstützt neben Schwarz-Weiß und Graustufen auch Paletten- und Echtfarben. Es unterstützt verschiedene Farbmodelle und eine Farbtiefe bis 24 Bit pro Pixel, womit maximal 16,7 Millionen Farben darstellbar sind. Es besteht die Möglichkeit, so genannte „Multi-Image-TIFFs" zu speichern, dabei werden in einer TIFF-Datei mehrere Bilder abgelegt. Außerdem können in standardisierten (z.B. für Höhe und Breite des Bildes) und privaten Tags Metainformationen gespeichert werden.

Der größte Nachteil von TIFF ist seine Komplexität. Die Vielfalt möglicher gültiger TIFF-Dateien kann von keinem einzelnen Programm unterstützt werden. In der Spezifikation des Dateiformats ist deswegen eine Untermenge gültiger TIFF-Dateien definiert, das so genannte Baseline TIFF, das jedes TIFF-fähige Programm verarbeiten können sollte.

Zweck und Verwendung: +

TIFF wird zum Scannen und Reproduzieren bei allen digitalen Faxgeräten benutzt und ist somit weltweit verbreitet. Zum Einsatz kommt bei Faxgeräten der Gruppe 4 beispielsweise TIFF G4 als Schwarz-Weiß Ausprägung von TIFF.

Von TIFF 6.0 werden die verschiedene Kompressionsalgorithmen wie LZW[2], JPEG, Packbits, Group-III-Fax, Group-IV-Fax und CCITT unterstützt, welche auch von professionellen Grafikprogrammen unterstützt werden [Borghoff 2003]. TIFF-Dateien können jedoch auch unkomprimiert gespeichert werden.

Aufgrund der internationalen Standardisierung und der Stabilität hat sich TIFF auch als Archivierungsformat für gescannte Papierdokumente etabliert. Das Format wird im Bereich klinischer Studien beispielsweise beim Scannen Versicherungsnachweisen verwendet. TIFF ist in elektronischen Archiven des Gesundheitswesens weit verbreitet. Nach [Lehmann 2002] gehört TIFF zu den Standard-Dateiformaten für medizinische Bilddaten.

Transparenz und Standardisierung: + + *

TIFF ist ein weit verbreiteter Industriestandard. Die Spezifikation, momentan Version 6.0 ([Adobe 1992]), ist offen gelegt. TIFF ist nach ISO für die medienunabhängige Bildverarbeitung standardisiert.

Stabilität: –

Erste Veröffentlichung E = 1986

Alter A = B – E = 21

[2] LZW ist ein Kompressionsalgorithmus, der nach seinen Erfindern Abraham Lempel, Jakob Ziv und Terry Welch benannt wurde. Der LZW-Algorithmus ist ein gut dokumentiertes, verlustfreies, eindimensionales Kompressionsverfahren. Das im Internet stark verbreitete GIF-Format beruht zum Beispiel auf dem LZW-Algorithmus. Der LZW-Algorithmus war lange Zeit mit Patenten der Firma UNISYS belegt. Das letzte Patent ist im Juni 2004 ausgelaufen. Seit dieser Zeit kann die LZW-Codierung als „nicht proprietär“ bezeichnet werden.

Anzahl Versionen V = 6

Stabilität S = A / V = 3,5

Letzte Veröffentlichung L = 1992

Zeitraum seit der letzten Änderung Z = 15

Die erste Version des Tagged Image File Formats wurde 1986 von der Firma Aldus veröffentlicht. An den folgenden Spezifikationen waren unter anderem auch Hewlett Packard und Microsoft beteiligt. Seit 1994 hat die Firma Adobe Systems Incorporated die Pflege des Dateiformats übernommen. Revision 6.0 des TIF-Formats liegt seit dem Jahr 1992 vor und wurde seither nicht mehr geändert.

Präsentation:

Präsentationswerkzeuge: +

Das Format wird von allen Herstellern von Dokumenten-Mangement-, Archiv- und Workflow-Systemen sowie von Text-, Grafik- und Präsentationsprogrammen unterstützt ([Kampffmeyer 2000]). Werkzeuge für TIFF-Bilder sind für verschiedene Plattformen verfügbar. Das Sharewareprogramm ACDSee ist beispielsweise für die Betriebssysteme Apple MacOS und Microsoft Windows verfügbar. Manche Viewer können TIFF-Dateien, die aus mehreren Seiten bestehen, so genannte „Multi-Image-TIFFs“, nicht anzeigen.

Darstellung: +

Die Darstellung von TIFF-Dateien sollte sich mit verschiedenen Präsentationswerkzeugen kaum unterscheiden. Anzeigedifferenzen können sich jedoch durch unterschiedliche Hardware, Kalibrierung, sowie durch Softwareparameter wie Auflösung, Farbtiefe etc. ergeben. Wurden private Tags definiert, so ist bei der Darstellung darauf zu achten, dass der Viewer diese interpretieren kann.

Struktur: +

Die Struktur einer TIFF-Datei besteht aus einem Header und den eigentlichen Bilddaten. Im Header sind Informationen zur Auflösung und zur Kompression enthalten. Die Bildinformationen sind im Rasterformat abgelegt.

Sicherheit:

Sicherheitsmechanismen: – –

Es sind keine Sicherheitsmechanismen zur Gewährleistung von Integrität, Authentizität und Vertraulichkeit in der Formatspezifikation vorgesehen. Es gibt jedoch die Möglichkeit, sich private Tags reservieren zu lassen, wo evt. sicherheitsrelevante Informationen gespeichert werden können, was jedoch ein proprietärer Ansatz wäre.

Elektronische Signatur: o

In TIFF lassen sich keine elektronischen Signaturen einbinden. Die Bewertung ist daher nicht möglich.

Obwohl in TIFF notwendige Sicherheitsmechanismen wie elektronische Signaturen fehlen, kann die Integrität digitaler Medien z.B. durch die Verwendung von Wasserzeichen gesichert werden. Durch fragile Wasserzeichen können Änderungen an digitalem Bildmaterial erkannt werden. Generell verfälschen diese fragilen Wasserzeichen durch das direkte Integrieren der Informationen das Original und können zu Fehlinterpretationen der medizinischen Inhalte führen. Nach [Pharow 2003; Steinebach 2007] sollten zur Sicherstellung von Authentizität und Integrität digitaler Medien deshalb fragile invertierbare Wasserzeichen unter Verwendung elektronischer Signaturen eingesetzt werden, die auch die Wiederherstellung des Originals ermöglichen.

Sonstige Beurteilungen:

Das TIF-Format ist heute einer der Standards der digitalen Bildverarbeitung. TIFF wird von [Kampffmeyer 2000] in komprimierter Form als geeignetes Format für die Langzeitspeicherung empfohlen. In [SAGA 2006] wird TIFF für Grafikinformationen empfohlen, die keinerlei Informationsverlust erlauben. Um eine einwandfreie Portabilität zu gewährleisten, sollte laut den [DLM-Forum 1997] zusätzlich zum TIF-Format keine weitere Komprimierung vorgenommen werden.

3.2 Dateiformate für Dokumente

3.2.1 Microsoft Word[3]

Microsoft Word ist ein von der Firma Microsoft Corporation hergestelltes sowie kommerziell vertriebenes Textverarbeitungsprogramm und bildet den De-facto-Standard für Textverarbeitung auf der Windows-Betriebssystemplattform. Das Microsoft Word-Dokument mit der Dateiendung „.doc“ ist wohl das am häufigsten auftretende Dateiformat unter den Textformaten. Der Markanteil dürfte bei ca. 80 – 90 Prozent liegen. Seit der Entwicklung des Microsoft Word-Formats in den achtziger Jahren haben sich häufige Versionswechsel vollzogen, wobei das Dateiformat mehrmals abgewandelt wurde.

Zweck und Verwendung: +

Mit dem Textverarbeitungsprogramm der Firma Microsoft Corporation lassen sich verschiedenartige Texte, beispielsweise Geschäftsbriefe, Formulare und Berichte, erstellen und bearbeiten. Als Grundlage bei der Erstellung von Dokumenten können Vorlagen benutzt werden, die teilweise schon mit dem Produkt mitgeliefert werden, die aber auch selbst angefertigt werden können. Neben normalem Text kann ein Dokument auch mit Tabellen, Grafiken oder Diagrammen ausgestattet werden.

Im Gesundheitswesen wird Microsoft Word unter anderem bei der Arztbriefschreibung verwendet und ist in verschiedene Anwendungssysteme direkt integriert. Im Bereich klinischer Studien werden u.a. Prüfpläne, Lebensläufe, Datenmanagementhandbücher und die Dokumentation im Rahmen der Pharmakovigilanz mit Microsoft Word erstellt.

Transparenz und Standardisierung: +

Microsoft Word ist ein weit verbreiteter Industriestandard. Bis zur Version aus dem Jahr 2007 war die Spezifikation des Dateiformats nicht offen gelegt. Die Binärdateiformate von Microsoft Word und anderen Microsoft Office Anwendungen können gebührenfrei nach Unterzeichnung einer Erklärung direkt bei Microsoft angefordert werden.

Die aktuellste Microsoft Word Version von 2007 nutzt das von Microsoft entwickelte Dateiformat Open Office XML, welches von der ECMA international (ehemals European

Computer Manufacturers Association), einer privaten Normungsorganisation, standardisiert wurde. Darüber hinaus wird die Standardisierung durch ISO angestrebt, deren Ausgang aber noch offen ist.

Neben Open Office XML von Microsoft existiert mit dem OpenDocument Format (siehe Abschnitt 3.2.2) ein weitere Dateiformat für Büroanwendungen auf der Basis von XML. Ein Vergleich beider Spezifikationen wurde u.a. durch [Macnaghten 2007] durchgeführt.

Stabilität: –

Erste Veröffentlichung E = 1983

Alter A = B – E = 24

Anzahl Versionen V = 13[4]

Stabilität S = A / V = 1,8

Letzte Veröffentlichung L = 2007

Zeitraum seit der letzten Änderung Z = 0

Das Format hat seit seinem Entstehen mehrmals das Speicherformat geändert. Nach Wikipedia (Eng.) existieren mit Microsoft Word 2007 mittlerweile 5 Versionen des Dateiformats die jeweils von einem Teil der Microsoft Word Versionen unterstützt werden. Daraus ergibt sich das Problem, dass verschiedene Versionen des Textformats zueinander inkompatibel sind. Um dieses Problem beherrschen zu können bietet Microsoft jeweils entsprechende Konvertierungspakete wie z.B. das Microsoft Office Compatibility Pack for Word, Excel, and PowerPoint 2007 File Formats an.

[3] Microsoft Word ist Teil der Büroanwendung Microsoft Office, in welcher u.a. noch Programme zur Tabellenkalkulation, Präsentation und Datenbankentwicklung enthalten sind. Als Vertreter für Microsoft Office wird in diesem Dokument Microsoft Word analysiert und bewertet.

[4] Bei der Anzahl der Versionen wurden sowohl die Versionen für Microsoft DOS als auch die Versionen für Microsoft Windows (16- und 32 Bit) gezählt, wobei im gleichen Jahr erschienene Versionen nicht doppelt gezählt wurden.

Präsentation:

Präsentationswerkzeuge: –

Es existieren Microsoft Word Versionen für die Microsoft Betriebssysteme DOS (Microsoft Word Versionen von 1983 bis 1993) und Windows (Microsoft Word Versionen von 1989 bis heute) sowie für Apple Macintosh (Microsoft Word Versionen von 1985 bis 2004) und IBM OS/2 (Microsoft Word Versionen von 1989 bis 1991). Darüber hinaus bietet Microsoft kostenlose Anzeigeprogrammen für Microsoft Word Dokumente an.

Produkte von anderen Herstellern wie z.B. Adobe Framemaker oder OpenOffice.Org verfügen über Konverter, mit welchen Microsoft Word Dokumente in das interne Dateiformat umgewandelt werden können. Neben Importmöglichkeiten unterstützen manche Hersteller auch den Export erzeugter Dokumente im Microsoft Word Dateiformat.

Darstellung: –

Wie schon erwähnt, ergeben sich durch die Versionswechsel oftmals Probleme bei der Darstellung von Dokumenten in verschiedenen Word-Versionen. [Fox 1998] gibt ein Beispiel, wie sich die Anzeige von einem Vertrag, erstellt in Microsoft Word für Windows und angezeigt in Word für Macintosh, unterscheiden kann. Auch die neue Spezifikation des Dateiformats lässt befürchten, dass derartige Probleme auftauchen werden.

Struktur: + +

Die Spezifikationen von Microsoft Word waren bis zur Version von 2007 nicht offen gelegt, weshalb für die früheren Versionen keine Aussagen zur Struktur getroffen werden können. Das neue Dateiformat Open Office XML ist im Hinblick auf den Aufbau und den Inhalt strukturiert. Ein Microsoft Office Dokument besteht aus einem ZIP-Archiv, in welchem verschiedene Dateien zusammengefasst werden, u.a. der Dokumentinhalt, die Formatierungen, die Dokument-Metadaten, die Verknüpfungen zwischen den Dateien sowie ggf. enthaltene Bilder.

Sicherheit:

Sicherheitsmechanismen: + +

In den neueren Versionen von Microsoft Word können Integrität und Authentizität der Dokumente mit Hilfe von kryptographischen Verfahren gesichert werden.

Gegen unberechtigtes Öffnen oder Schreiben können Microsoft Word Dokumente mit einem Passwortschutz versehen werden. Dieser Passwortschutz konnte in früheren Versionen leicht umgangen werden und auch die aktuelle Nutzung des symmetrischen Verschlüsselungsverfahren RC4 weist Mängel auf.

Da Microsoft Word in der neuesten Version ein XML basiertes Dateiformat verwendet, werden zur Sicherung von Integrität und Authentizität der Dokumente die Standards zur XML Verschlüsselung ([Imamura 2002]) und XML Signatur ([Bartel 2002]) verwendet [ECMA 2006].

Elektronische Signatur: +

In Microsoft Word Version von 2002 existiert neben dem Schutz von Makros auch die Möglichkeit der digitalen Signatur des Dokumentinhaltes. Elektronische Signaturen in Microsoft Word basieren auf kryptographischen Verfahren, wobei bis zur Version aus dem Jahr 2007 weder die internen Mechanismen noch die Formate offen gelegt waren. Signierte Dokumente können nach Veränderungen nur gespeichert werden, wenn auch die Signatur gelöscht wird. Werden signierte Worddokumente manipuliert, können diese nicht mehr mit Microsoft Word geöffnet werden, wobei die angezeigte Fehlermeldung aber nicht auf Manipulationen schließen lässt.

Wie oben bereits beschrieben wird für Open Office XML in der aktuellsten Microsoft Word Version von 2007 die standardisierte XML Signature [Bartel 2002] eingesetzt.

3.2.2 OpenDocument Format (ODF)

OpenDocument ist ein von der OASIS (Organization for the Advancement of Structured Information Standards) spezifiziertes Austauschformat für Büroanwendungen aus dem Jahr 2005 ([OASIS 2005]).

Zweck und Verwendung: – –

ODF wurde von der OASIS auf der Basis der Dateiformate von OpenOffice.org entwickelt. OpenOffice.org ist ein lizenzkostenfreies Programmpaket, welches u.a. Programme zur Textverarbeitung, Tabellenkalkulation und Präsentation und Datenbankentwicklung enthält. ODF wurde spezifiziert, um als Gegenpol für Microsoft Office vor allem ein standardisiertes Dateiformat für Textverarbeitungsprogramme zu etablieren. Unter dem ODF existieren u.a. Dateiformate für Texte (.odt), Tabellen (.ods), Präsentationen (.odp) und Zeichnungen (.odg). Die Spezifikation für Datenbanken ist derzeit nicht im ODF enthalten. Eine ODF-Datei

besteht aus einer oder mehreren XML-Dateien und ggf. zusätzlichen Binärdaten z.B. für eingebettete Grafiken. Die Dateien und Ordner werden über ein ZIP-Archiv zusammengefasst und komprimiert.

In Deutschland ist ODF derzeit noch wenig verbreitet. Im Bereich der TMF wird ODF derzeit vereinzelt für Patienteninformationen und -einwilligungen, Investigators Broschüren sowie für Prüfpläne und -bögen eingesetzt. Im europäischen Ausland ist ODF weiter verbreitet. In Belgien wurde ODF für den Austausch von Bürodokumenten im Bereich der Regierung vorgeschlagen und ab September 2007 muss jede Bundesbehörde ODF unterstützen. Auch in Frankreich und England wird ODF zur Verwendung im öffentlichen Bereich vorgeschlagen.

Transparenz und Standardisierung: + + *

ODF wurde von der OASIS entwickelt und 2006 als internationale Norm ISO/IEC 26300 veröffentlicht.

Stabilität: – –

Erste Veröffentlichung E = 2005

Alter A = B – E = 2

Anzahl Versionen V = 2

Stabilität S = A / V = 1

Letzte Veröffentlichung L = 2007

Zeitraum seit der letzten Änderung Z = 0

ODF 1.0 wurde am 1. Mai 2005 als OASIS Standard verabschiedet. Am 19. Juli 2006 wurde die so genannte Second Edition der ODF Version 1.0 veröffentlicht, die jedoch keine neue Version darstellt sondern nur die redaktionellen Änderungen des Standardisierungsprozesses beinhaltet. Am 2. Februar 2007 wurde die Version 1.1 von der OASIS verabschiedet.

Präsentation:

Präsentationswerkzeuge: +

Obwohl ODF ein relativ neues Dateiformat ist, gibt es bereits eine Reihe von Werkzeugen für dessen Bearbeitung. Mit OpenOffice.Org und dem auf dem gleichen Quellcode basierenden StarOffice der Firma Sun Microsystems existieren zwei Büroanwendungen, die ODF als Standarddateiformat nutzen. Für andere Büroanwendungen wie Microsoft Office und

Softmaker Office existieren Import- und Exportschnittstellen und auch die Serveranwendung Google Docs & Spreadsheets unterstützt ODF.

Darstellung: +

In ODF werden die Inhalte des Dokumentes strickt von der Definition der Darstellung getrennt. Die Inhalte werden in einer XML-Datei namens content.xml abgelegt, während die Definition der Schriften etc. in der Datei styles.xml gespeichert werden.

ODF-Dateien können aber externe Objekte oder Hyperlinks enthalten, die die Darstellung ggf. verändern können.

Struktur: + +

Ebenso wie die Trennung von Dokumentinhalt und -darstellung in unterschiedlichen XML-Dateien werden auch die Dokument-Metadaten wie Autor, Erzeugungsdatum u.a. in einer separaten Metadatendatei, der Datei meta.xml gespeichert. Demnach sind ODF-Dokumente im Hinblick auf den Aufbau und den Inhalt strukturiert.

Sicherheit:

Sicherheitsmechanismen: +

Zu Sicherung von ODF-Dateien ist in der Formatspezifikation die Verschlüsselung von Dateien vorgesehen. Die Verschlüsselung erfolgt passwortbasiert unter Verwendung kryptographischer Algorithmen. Die notwendigen Informationen für die Entschlüsselung sind in der Datei manifest.xml abgelegt.

Elektronische Signatur: o

Elektronische Signaturen werden von ODF bislang noch nicht unterstützt. Jedoch realisieren verschiedene Werkzeuge wie OpenOffice.Org bereits elektronische Signaturen für ODF auf Basis von XML Signaturen ([Bartel 2002]). Elektronische Signaturen sollen in Version 1.2 von ODF in den Standard aufgenommen werden.

3.2.3 PostScript

PostScript ist eine von der Firma Adobe Systems Incorporated in den achtziger Jahren entwickelte Programmiersprache zur Beschreibung von Seiten und graphischen Elementen innerhalb eines Dokuments. In PostScript können Texte und Grafiken geräte- und

auflösungsunabhängig definiert werden, d.h. die Ausgabe eines Dokuments auf dem Bildschirm und auf dem Drucker ist identisch.

Zweck und Verwendung: – –

PostScript ist eine zum Industriestandard gewordene Seitenbeschreibungssprache, die alle auf einer Visualisierungseinheit auszugebenden Daten und ihre Position auf eine einheitliche Datenstruktur bringt. Text-, Bild- und Grafikelemente können völlig unabhängig vom jeweils benutzten Ausgabegerät bearbeitet und positioniert werden. PostScript wird zum Austausch elektronischer Dokumente verwendet, sei es für die Verwendung am Bildschirm oder die spätere Ausgabe auf Papier oder Film.

Zum Austausch einzelner Grafiken zwischen verschiedenen Programmen und Betriebssystemen wird das Format Encapsulated PostScript (EPS) eingesetzt. Das EPS-Format gehört einerseits zur Gruppe der Vektorgrafikformate, andererseits ist es auch ein Metagrafikformat, da es außer Vektorinformationen auch Rastergrafiken enthalten kann.

Neben dem Grafikmodell ist die Programmierbarkeit eine der wichtigsten Eigenschaften von PostScript und macht das Format sehr leistungsfähig und erweiterbar. Allerdings ist diese Funktion auch eine Fehlerquelle, da die korrekte Verarbeitung von PostScript nicht immer gewährleistet wird.

Im Bereich des Gesundheitswesens und klinischer Studien wird PostScript nur vereinzelt eingesetzt, da mittlerweile in vielen Bereichen PDF verwendet wird.

Transparenz und Standardisierung: +

PostScript ist ein Industriestandard. Die Spezifikationen wurden von dem Hersteller Adobe Systems Incorporated offen gelegt und sind unter [Adobe 1999] erhältlich.

Stabilität: –

Erste Veröffentlichung E = 1985

Alter A = B – E = 22

Anzahl Versionen V = 3

Stabilität S = A / V = 7,3

Letzte Veröffentlichung L = 1997

Zeitraum seit der letzten Änderung Z = 10

Die erste Version von PostScript, PostScript Level 1, wurde im Jahr 1985 veröffentlicht. Diese ursprüngliche PostScript-Sprache bildet den Kern der nachfolgenden Versionen. Wesentliches Merkmal von Level 1 ist die noch fehlende Farbunterstützung. Diese und weitere Verbesserungen, wie z.B. eine standardisierte Schnittstelle für den Zugriff auf verschiedene Druckerfähigkeiten, wurden 1991 in PostScript Level 2 integriert. Die neueste Version des Formats, PostScript 3, erschien 1997. Hier wurden diverse Verbesserungen im Zusammenhang mit der Verarbeitung von Schriften durchgeführt und alte Programmierkonstrukte wurden durch effizientere Varianten ersetzt.

Präsentation:

Präsentationswerkzeuge: +

Zur Anzeige von PostScript-Dateien stehen z.B. der Adobe Acrobat Reader und GSview der Firma Ghostgum Software Pty Ltd. zur Verfügung. Viele Textverarbeitungs- und Desktop-Publishing-Anwendungen unterstützen PostScript, da das EPS-Format hauptsächlich zur Übertragung in Layoutprogramme verwendet wird.

Darstellung: +

PostScript verwendet ein geräteunabhängiges Ausgabeverfahren. Das bedeutet, dass eine PostScript-Datei nur das Aussehen einer Seite beschreibt, jedoch keine Annahmen über den Drucker oder den Bildschirm macht, auf dem sie dargestellt wird.

Struktur: +

Eine PostScript-Datei ist bezüglich des Aufbaus strukturiert. Da PostScript ein layoutorientiertes Format ist, genauer gesagt eine Seitenbeschreibungssprache, werden inhaltlich relevante Daten nicht strukturiert. Das Format beschreibt das Aussehen der Daten.

Sicherheit:

Sicherheitsmechanismen: – –

In PostScript sind keine Sicherheitsmechanismen integriert. Die Firma Adobe hat sich auf die Sicherheit von PDF-Dateien konzentriert, was im Kapitel über das Portable Document Format beschrieben ist.

Elektronische Signatur: o

Die elektronische Signatur wird in PostScript nicht realisiert und kann daher nicht bewertet werden.

3.3 Dateiformate für Rohdaten

3.3.1 American Standard Code for Information Interchange (ASCII)

ASCII ist die Abkürzung für American Standard Code for Information Interchange, einen standardisierten Zeichencode. Der ursprüngliche ASCII-Code ist die nationale Variante des ISO-7-Bit-Codes (ISO 646) in den USA und wird deshalb auch als US-ASCII bezeichnet. Daneben gibt es verschiedene 8-Bit Zeichensätze, welche auf US-ASCII aufbauen, z.B. der ANSI-Code oder die ISO 8859-Zeichensatzfamilie.

ASCII steht jedoch nicht nur für den Zeichensatz, sondern auch für ein Textformat. Ein ASCII-Text beschreibt ein Dokument, das nur aus Zeichen des ASCII-Zeichensatzes besteht, jedoch keine Steuerzeichen, z.B. für das Ansteuern eines Druckers, oder Layoutinformationen beinhaltet.

Ein Zeichensatz oder Zeichencode ist eine Sammlung verschiedener Zeichen wie etwa Buchstaben, Zahlen, Satz- und Sonderzeichen (siehe [Matzer 2000]). Der eingangs beschriebene US-ASCII-Zeichensatz (ISO 646) enthält 2^7 = 128 Zeichen. Dies sind die lateinischen Grundbuchstaben, Ziffern und Satzzeichen, die im Amerikanischen gebräuchlich sind, jedoch keine durch diakritische oder sonstige Zeichen erweiterten Buchstaben, also auch keine deutschen Umlaute. Die ersten 32 Zeichen wurden als Steuerzeichen reserviert. Aufbauend auf den US-ASCII-Zeichensatz wurden später verschiedene 8-Bit-Zeichensätze entwickelt.

Die amerikanische Standardisierungsorganisation schuf den ANSI-Code. Dieser übernimmt den US-ASCII Zeichensatz für die Werte 0 bis 127 und definiert für die Werte zwischen 128 und 255 Sonderzeichen, wie z.B. wichtige Alphabetzeichen verbreiteter Sprachen, kaufmännische, wissenschaftliche und diakritische Zeichen. Auch den Zeichensätzen der ISO-8859-Familie, die von der European Computer Manufacturer‘s Association (ECMA) entwickelt wurde, liegt der US-ASCII-Zeichensatz zugrunde. Hierbei handelt es sich um verschiedene Zeichensätze für alphabetische Schriften, wie z.B. die lateinischen und die kyrillischen Schriften.

EBCDIC, auch IBM-Code genannt, ist ein weiterer, von der Firma IBM für Großrechneranlagen entwickelter 8-Bit-Zeichensatz. Der EBCDI-Code ist dem US-ASCII-Code sehr ähnlich, es gibt jedoch keine identische Teilmenge. Heute findet der Zeichensatz außer in IBM-Großrechnern kaum noch Verwendung und wird deshalb in der weiteren Bewertung nicht mehr betrachtet.

Neben den 7- bzw. 8-Bit-Codes gibt es das Unicode-System. Ziel dieses Systems ist es, alle bekannten Zeichen aus gegenwärtigen und vergangenen Schriftkulturen möglichst vollständig zu erfassen. Ursprünglich wurde das Unicode-System als 16-Bit-Code entworfen, wurde jedoch im März 2001 durch ein 32-Bit-Schema abgelöst.

Bei den beschriebenen Zeichensätzen handelt es sich nicht um verschiedene Versionen von ASCII, sondern um verschiedene Standards. Bei der Bewertung der Stabilität weichen die Standards voneinander ab, deshalb wird dort die Berechnung für die verschiedenen Codes getrennt durchgeführt.

Zweck und Verwendung: +

Der ASCII-Standard wurde ursprünglich für den Datenaustausch in der Datenfernübertragung (DFÜ) entwickelt. Jedes Betriebssystem schreibt vor, mit welchem Zeichensatz es arbeitet. So arbeiten ältere DOS-Programme mit dem IBM-ASCII-Zeichensatz. Microsoft Windows hingegen verwendet die neuere ANSI-Norm, um Zeichen zu codieren. Neuere Programme unterstützen meist auch das UNICODE-System.

Im Gesundheitswesen werden ASCII-Texte beispielsweise zur Übermittlung von Labordaten verwendet und sind in diesem Bereich weit verbreitet. Im Bereich der TMF wird ASCII hauptsächlich im Bereich der Studiendaten eingesetzt.

Transparenz und Standardisierung: + + *

Der ASCII-Code ist ursprünglich ein nationaler Standard, der vom American National Standards Institute (ANSI) veröffentlicht wurde. Die Bemühungen zur Standardisierung eines Zeichensatzes gehen auf den Anfang der sechziger Jahre zurück. Diese nordamerikanische Version des ASCII-Codes und einige weitere nationale Varianten wurden im Jahr 1972 von der International Standards Organisation als ISO 646 spezifiziert.

Das Unicode-System ist seit Version 2.0 mit der internationalen Norm ISO/IEC 10646 synchronisiert.

Stabilität:

Um die Stabilität der verschiedenen Zeichensätze beurteilen zu können, werden die Berechnungen beispielhaft für den 7-Bit US-ASCII-Zeichensatz, für den 8-Bit Latin1-Zeichensatz der ISO 8859-Familie und für das Unicode-System vorgenommen.

Stabilität des US-ASCII-Zeichencodes: + +

Erste Veröffentlichung E = 1972

Alter A = B – E = 35

Anzahl Versionen V = 1

Stabilität S = A / V = 35

Letzte Veröffentlichung L = 1972

Zeitraum seit der letzten Änderung Z = 35

Stabilität für ISO 8859-1 (Latin 1): +

Erste Veröffentlichung E = 1986

Alter A = B – E = 21

Anzahl Versionen V = 1

Stabilität S = A / V = 21

Letzte Veröffentlichung L = 1986

Zeitraum seit der letzten Änderung Z = 21

Stabilität für das Unicode-System: – –

Erste Veröffentlichung E = 1991

Alter A = B – E = 16

Anzahl Versionen V = 18

Stabilität S = A / V = 0,89

Letzte Veröffentlichung L = 2006

Zeitraum seit der letzten Änderung Z = 1

1972 wurde der US-ASCII-Code als internationaler Standard in ISO-646 spezifiziert. Die erweiterten 8-Bit Zeichencodes, z.B. die der ISO-8859-Familie, gehen auf die achtziger Jahre zurück. Standards der Serie ISO 8859 werden nie geändert. Deshalb enthält der weit verbreitete ISO 8859-1 (Latin 1) Zeichensatz beispielsweise kein Eurozeichen. Fehlende Zeichen werden in Erweiterungen eingearbeitet. So ist das Eurozeichen ist in ISO 8859-15 enthalten.

Das neuere Unicode-System, dessen erste Version Anfang der neunziger Jahre veröffentlicht wurde, wird ständig weiterentwickelt und es werden immer neue Zeichen aufgenommen. Deshalb werden häufig überarbeitete und erweiterte Versionen des Systems veröffentlicht.

Präsentation:

Präsentationswerkzeuge: +

Editoren für die Anzeige von ASCII-Textdokumenten sind in jedem Betriebssystem standardmäßig vorhanden. Es sind keine validierten Präsentationswerkzeuge bekannt.

Darstellung: –

Ein ASCII-Text lässt sich prinzipiell auf jedem Rechner gleich darstellen. Dabei stellt ein Programm, dass eine ASCII-Datei anzeigen soll, einfach jedes Zeichen der Datei der Reihe nach am Bildschirm dar. Voraussetzung für eine korrekte Darstellung ist die Verwendung des richtigen Zeichensatzes. Da in einem reinen ASCII-Textdokument jedoch keine Zeichensätze eingebunden werden, ist die korrekte Darstellung nicht gewährleistet. So kann es passieren, dass das im ANSI- Zeichensatz geschriebene „ü“ in der Darstellung mit dem erweiterten DOS-Zeichensatz eine hochgestellte ³ ergibt. [Pordesch 2000] gibt ein weiteres Beispiel für die Beeinflussung der Darstellung von ASCII-Texten in Abhängigkeit von der Interaktion mit dem Betriebssystem. Durch den vom Nutzer gewählten Zeilenumbruch und die Fenstergröße kann der folgende ASCII-Text auf verschiedene Arten dargestellt werden und so zu Missverständnissen führen.

Struktur: – –

Ein ASCII Text-Dokument, auch Plain Text genannt, benutzt den ASCII-Zeichensatz und kennt weder Layout- noch Strukturinformationen. Texte werden als beliebig lange Zeichenketten aufgefasst.

Sicherheit:

Sicherheitsmechanismen: – –

Ein ASCII-Textdokument enthält keine Sicherheitsmechanismen. Die Vertraulichkeit, Authentizität und Integrität des Inhaltes eines Dokuments kann durch das Format nicht gewährleistet werden.

Elektronische Signatur: o

Die elektronische Signatur kann nicht bewertet werden. Es werden keine elektronischen Signaturen in einem ASCII-Text realisiert.

3.3.2 Comma Separated Values (CSV)

Zweck und Verwendung: +

CSV steht für Comma Separated Values und beschreibt ein Dateiformat, für den Austausch einfach strukturierter Daten und Tabellen auf der Basis von ASCII. Die einzelnen Werte einer CSV-Datei sind, wie der Name schon sagt, durch Komma voreinander getrennt. Manchmal wird der Begriff CSV auch für Colon Separated Values (Semikolon als Trennzeichen) oder Character Separated Values (u.a. Doppelpunkt oder Tabulator als Trennzeichen) verwendet.
Nach RFC 4180 steht jeder Datensatz einer CSV-Datei in einer Zeile, die durch einen Zeilenumbruch beendet wird. Optional kann eine Kopfzeile mit den Spaltenüberschriften existieren, die genau so viele Werte wie eine Datenzeile enthält. Wenn Kommas, Zeilenumbrüche oder Anführungszeichen in den Werten vorkommen, müssen die Werte durch Anführungszeichen eingeschlossen werden. Weiter gehende Regeln zum Aufbau von CSV-Dateien sind u.a. in [Repici 2002] beschrieben.
Im Bereich des Gesundheitswesens wird CSV zur Speicherung von Tabellen- und Datenbankinhalten verwendet. Im Bereich des TMF werden u.a. Studiendatenbanken im Dateiformat CSV gespeichert.

Transparenz und Standardisierung: +

CSV ist kein allgemeiner Standard, sondern nur informell durch [RFC4180 2005] beschrieben.

Stabilität: –

Die Bewertung der Stabilität bezieht sich auf den RFC 4180. Die erste Version des RFC wurde 2005 verabschiedet und seither nicht verändert.

Erste Veröffentlichung E = 2005

Alter A = B – E = 2

Anzahl Versionen V = 1

Stabilität S = A / V = 2

Letzte Veröffentlichung L = 2005

Zeitraum seit der letzten Änderung Z = 2

Präsentation:

Präsentationswerkzeuge: +

Das Dateiformat CSV kann mit beliebigen Texteditoren, die in allen Betriebssystemen vorhanden sind, bearbeitet und angezeigt werden. Auch Büroanwendungen wie Microsoft Excel oder OpenOffice.Org können CSV-Dateien importieren und exportieren.

Darstellung: +

Grundsätzlich treffen die Aussagen, die bzgl. der Darstellung für ASCII getroffen wurden auch für CSV zu. Darstellungsprobleme bei CSV-Dateien können durch die Wahl des falschen Zeichensatzes entstehen. Darüber hinaus können z.B. Dezimalzahlen die ein Komma enthalten, falsch interpretiert werden, wenn die Verwendung von Anführungszeichen für derartige Werte nicht beachtet wird.

Struktur: –

Das Dateiformat CSV ist nur bzgl. des Inhalts strukturiert, nicht aber bzgl. des Aufbaus.

Sicherheit:

Sicherheitsmechanismen: – –

Eine CSV-Datei enthält keine Sicherheitsmechanismen. Die Vertraulichkeit, Authentizität und Integrität des Inhaltes einer CSV-Datei kann durch das Format nicht gewährleistet werden.

Elektronische Signatur: o

3.4 Dateiformate für Bilddaten

3.4.1 Joint Photographic Experts Group (JPEG)

Die Joint Photographic Experts Group ist eine interdisziplinäre Expertengruppe im Bereich der Bildverarbeitung. JPEG steht jedoch auch für ein von dieser Expertengruppe entwickelte Kompressionsverfahren für Bilder. Mit dem Dateiformat „JPEG File Interchange Format“ (JFIF), welches umgangssprachlich ebenfalls als JPEG bezeichnet wird, wurde ein Austauschformat für JPEG-komprimierte Bilder definiert.

Zweck und Verwendung: +

Laut [Matzer 2000] ist das JPEG-Format nach dem GIF-Format das zweithäufigste im World Wide Web verwendete Dateiformat für Bilddaten. Der Unterschied zu Bildformaten wie TIFF liegt darin, dass JPEG verlustbehaftet komprimiert. Mit dem Kompressionsverfahren JPEG wird die Dateigröße durch selektives Löschen von Daten komprimiert, indem gleiche Bildpunkte eingespart werden. Die Größe der Datei hängt vom gewählten Komprimierungsgrad ab. Das Verfahren eignet sich deshalb nicht für scharfe Kanten und Übergänge, wie sie z.B. in Liniengrafiken und Cartoons auftreten. JPEG erhält bei der Komprimierung alle Farbinformationen und bewahrt damit die für saubere Darstellung von Fotografien notwendigen Farbverläufe.

JPEG ist einer der Standards für medizinische Bilddaten ([Lehmann 2002]), für welche es auch im Rahmen der TMF eingesetzt wird.

Transparenz und Standardisierung: + + *

Die Joint Photographic Experts Group wurde von zwei internationalen Standardisierungsgremien, der International Telecommunications Union (ITU) und der International Organization for Standardisation (ISO), ins Leben gerufen. Das Kompressionsverfahren JPEG wurde 1992 veröffentlicht und 1994 als ISO Standard 10918-1 angenommen. Da die Definition der Struktur von JPEG-Dateien viele Freiheiten erlaubt, wurde mit dem „JPEG File Interchange Format“ (JFIF) 1992 ein minimales Dateiformat definiert, das den Austausch von JPEG-komprimierten Bilddaten ermöglicht. JFIF hat sich zum De-facto-Standard für den Austausch von JPEG-komprimierten Bilddaten entwickelt, wird aber nicht mehr weiter entwickelt.

Stabilität: +

Erste Veröffentlichung E = 1992

Alter A = B – E = 15

Anzahl Versionen V = 1

Stabilität S = A / V = 15

Letzte Veröffentlichung L = 1992

Zeitraum seit der letzten Änderung Z = 15

Sowohl das Komprimierungsverfahren JPEG als auch das definierte Dateiformat JFIF sind seit 1992 stabil.

Präsentation:

Präsentationswerkzeuge: +

Durch die weite Verbreitung wird das Format von den meisten Bildverarbeitungsprogrammen, beispielsweise vom Microsoft Photo Editor, der mit dem Microsoft Office Paket mitgeliefert wird, oder dem Jasc Paint Shop Pro, unterstützt.

Darstellung: +

Die Darstellung einer JPEG-Datei sollte auf verschiedenen Ausgabemedien gleich sein. Dabei ist die Qualität einer Grafik in erster Linie von dem bei der Erstellung gewählten Komprimierungsgrad abhängig.

Struktur: +

Der Aufbau einer JPEG-/ JFIF-Datei besteht aus verschiedenen Marker-Segmenten. Diese Marker beschreiben den Beginn und das Ende einer Datei sowie Segmente, in denen die komprimierten Bilddaten und zusätzliche Informationen zu den gespeicherten Bilddaten, wie z.B. Version oder Auflösung, gespeichert sind. Die inhaltlichen Daten selbst sind nicht strukturiert.

Sicherheit:

Sicherheitsmechanismen: – –

Im JPEG-Dateiformat sind keine Sicherheitsmechanismen zur Gewährleistung von Integrität, Authentizität und Vertraulichkeit beinhaltet.

Elektronische Signatur: o

Die elektronische Signatur wird nicht im Format realisiert und kann daher nicht bewertet werden.

Sonstige Beurteilungen:

[Kampffmeyer 2000] weist darauf hin, dass nur eine verlustfreie Kompression zur revisionssicheren Archivierung zu Einsatz kommen darf.

In [SAGA 2006] wird JPEG als bewährter Standard für den Austausch von Bildern beschrieben und für die Langzeitarchivierung empfohlen. In [DLM-Forum 1997] wird JPEG als stabiler, anerkannter Standard bezeichnet, der hinsichtlich Speicherplatz und Langlebigkeit empfohlen wird.

Bemerkung: JPEG 2000

Seit Ende der 90iger Jahre arbeitet die Joint Photographic Experts Group an der Entwicklung des neuen Standards JPEG 2000. Dieser besteht aus derzeit aus 13 Teilen, die zum Großteil bereits in der Standardisierungsreihe ISO/IEC 15444 standardisiert sind. Das in JPEG 2000 angewendete Wavelet-Verfahren besitzt bessere Kompressionsraten bei besserer Bildqualität. Diese JPEG-Kompression ist variabel und kann verlustfrei sowie verlustbehaftet sein. Des Weiteren können unterschiedliche Bildbereiche (region of interest, ROI) unterschiedlich komprimiert werden. So können wichtige, detaillierte Bildelemente nicht oder wenig komprimiert werden, während unwichtige Inhalte stark komprimiert werden können. Gerade diese Eigenschaft ist ein interessanter Aspekt für die Speicherung medizinischer Bilddaten. Im Gegensatz zu JPEG werden bei JPEG 2000 auch Mechanismen zur Gewährleistung von Integrität, Authentizität und Vertraulichkeit der Bilddaten beschrieben, die im Jahr 2007 als ISO/IEC 15444-8 standardisiert wurden. Trotz vielfältiger Möglichkeiten von JPEG 2000 hat sich der Standard bis heute kaum als Dateiformat in der Praxis durchgesetzt.

3.4.2 Digital Imaging and Communications in Medicine (DICOM)

Zweck und Verwendung: + +

DICOM wurde für die Anwendung im medizinischen Bereich konzipiert, da einfache Bitmap-Formate wie TIFF oder JPEG bei der Übertragung medizinischer Bilddaten unzureichend sind. Zusätzlich zum Bild müssen Informationen zum Patienten, über die Modalität oder die Organisation standardisiert übertragen werden können. Der Standard definiert Formate und

Informationsobjekte und den Mechanismus zur Speicherung und Übertragung der Informationen.

DICOM konnte sich in Deutschland als Standard für die Übermittlung von Bildmaterial etablieren und wird vor allem in der Radiologie angewendet ([Horsch 2002]). Aber auch in der Strahlentherapie setzt sich der Standard zunehmend durch ([Neumann 2002]). [Eichelberg 2002] ist der Meinung, dass radiologische Bilddaten mit den Sicherheitserweiterungen auch über Plattformgrenzen hinweg sicher ausgetauscht werden können.

Transparenz und Standardisierung: + + *

DICOM wird von der National Electrical Manufacturers Association (NEMA) und dem American College of Radiology (ACR) standardisiert. 1995 wurde DICOM unter der Bezeichnung MEDICOM ([ENV12052 1995]) europäische Norm.

Stabilität: –

Erste Veröffentlichung E = 1985

Alter A = B – E = 22

Anzahl Versionen V = 3

Stabilität S = A / V = 7,3

Letzte Veröffentlichung L = 1993

Zeitraum seit der letzten Änderung Z = 0

Version 1.0 des DICOM-Standards erschien im Jahr 1985. Seit 1993 ist die dritte Version, ACR-NEMA 3.0, im Einsatz. Der Standard ist jedoch nicht statisch und wird ständig weiterentwickelt. Mittlerweile besteht DICOM aus 18 Teilen („parts“). Durch ständige Ergänzungen („supplements“) und Korrekturen („correction proposals“) wird der Standard an neue bildgebende Modalitäten angepasst und Schwächen werden ausgebessert.

Präsentation:

Präsentationswerkzeuge: +

Der DICOM-Standard ist von nahezu allen Herstellern von Bildgebungsprodukten übernommen worden: Modalitäten, PACS, Diagnose-Workstations und Archive. Darüber hinaus existieren verschiedene lizenzkostenfreie Werkzeuge, u.a. vom OFFIS e.V..

Darstellung: + +

Besonders in der Teleradiologie und der Telediagnostik ist es wichtig, dass die Darstellung der Bilder auf verschiedenen Systemen gleich ist. Das in DICOM spezifizierte Bildformat enthält unter anderem Zusatzinformationen zur Art der Bildakquisition und aufnahmespezifische Eigenschaften der Bilddaten, die standardisiert dokumentiert werden. Aus dem Bild-Header können die für die Visualisierung wesentlichen Bildzusatzinformationen über entsprechende Zugriffsfunktionen extrahiert werden ([Horsch 2002]). Durch die Möglichkeit der Definition von privaten Objekten kann es dennoch zu Inkompatibilitäten kommen.

Struktur: + +

Die Idee der Struktur von DICOM entstand aus der objekt-orientierten Programmierung. DICOM definiert Objekte und ordnet diesen Objekten unterschiedliche Methoden zur Verwaltung ihrer Daten zu (siehe [Vorwerk 2001]). Die inhaltlichen Daten liegen in strukturierter Form vor.

Sicherheit:

Sicherheitsmechanismen: + +

Die Datensicherheit kann in DICOM durch vier verschiedene Sicherheitserweiterungen gewährleistet werden. Eine Möglichkeit der Absicherung der DICOM-Netzwerkkommunikation ist die Verwendung des TLS -Protokolls, welche unter der Bezeichnung „Security enhancements one“ als DICOM Supplement 31 ([DICOM_Suppl31 2000]) Teil des DICOM Standards ist. Eine weitere Möglichkeit stellt die selektive Verschlüsselung einzelner Datenfelder in einem DICOM Bild dar. Durch diesen Ansatz, welcher als DICOM Supplement 55 ([DICOM_Suppl55 2001]) veröffentlicht wurde, können einzelne Patientenidentifikationsdaten verschlüsselt werden. Ohne den entsprechenden kryptographischen Schlüssel sind die Bilder ohne Patientenbezug lesbar. Die dritte Sicherheitserweiterung beschreibt die Verwendung elektronischer Signaturen für Befunde und Bilder im DICOM-Standard, welcher als DICOM Supplement 41 ([DICOM_Suppl41 2001]) verabschiedet wurde. Die vierte Sicherheitserweiterung ist die Integration von elektronischen Signaturen in strukturierte Befunde, die als DICOM Supplement 86 ([DICOM_Suppl86 2004]) im Jahr 2004 verabschiedet wurde

Elektronische Signatur: – –

Die in den Supplements 41 und 86 ([DICOM_Suppl41 2001; DICOM_Suppl86 2004]) des DICOM Standards beschriebene Spezifikation elektronischer Signaturen entspricht keinem standardisierten Signaturformat. Es wird angegeben, wie die Signatur zusammen mit dem Zertifikat des Unterzeichners nach X.509 Spezifikation und einem Zeitstempel nach Time-Stamp Protocol im „Header" des DICOM-Objekts abgelegt werden kann. Zusätzliche Zertifikate der Zertifizierungshierarchie sowie Sperrlisten, Zertifikatstatusinformationen und Signaturerneuerungen können in der spezifizierten Struktur nicht gespeichert werden.

3.5 Dateiformate für Webseiten und E-Mail

3.5.1 Hypertext Markup Language (HTML)

Die Hypertext Markup Language ist die „lingua franca" für Dokumente innerhalb des WWW und wurde im Zuge der Verbreitung des Internets zum erfolgreichsten und verbreitetsten Dateiformat der Welt. Die Entwicklung der Sprache geht auf Web-Gründer Tim Berners-Lee Anfang der neunziger Jahre zurück und wird seid 1995 vom W3C standardisiert.

Ursprünglich war die Hypertext Markup Language als reine Dokumentenbeschreibungssprache geplant und auf die Anforderungen elektronisch übertragener Dokumente ausgelegt. Mit der Entwicklung des WWW kam die Einbindung von Grafiken, Hyperlinks und Multimedia-Anwendungen.

Zweck und Verwendung: +

HMTL wird für die Publikation im Internet eingesetzt und unterstützt unter anderem Text, Grafiken und Tabellen. Die Popularität von HTML lässt sich zu einem großen Teil auf den einfachen Aufbau der Sprache zurückführen. Außerdem kann HTML mit jedem Texteditor bearbeitet werden, der Daten als reine Textdokumente abspeichern kann, man bezeichnet HTML auch als ein so genanntes Klartext-Format. Im Internet werden Leitlinien oder sonstiges medizinisches Wissen in HTML zur Verfügung gestellt. Im Rahmen des TMF wird HTML u.a. für Statusberichte eingesetzt.

Transparenz und Standardisierung: +

HTML wird seit Version 2.0 vom W3-Konsortium (W3C) standardisiert. Die Spezifikationen sind unter [Raggett 1999] nachzulesen.

Stabilität: –

Erste Veröffentlichung E = 1992

Alter A = B – E = 15

Anzahl Versionen V = 6

Stabilität S = A / V = 2,5

Letzte Veröffentlichung L = 1999

Zeitraum seit der letzten Änderung Z = 8

Aufgrund der verschiedenen HTML-Spezifikationen des W3-Konsortiums und herstellerspezifischer Erweiterungen existieren mehrere HTML-Versionen. Die Urversionen von HTML aus den Jahren 1992 und 1993 sind heute nicht mehr erwähnenswert. HTML 2.0 wurde im November 1995 offizieller Sprachstandard und gilt als kleinster gemeinsamer Nenner der Sprache. Einige der Sprachbestandteile, die in HTML 3.2 Anfang des Jahres 1997 eingeführt wurden, sollten in den folgenden Versionen wieder aus dem HTML-Standard entfallen, da sie durch andere, ergänzende Technologien wie Cascading Stylesheets (CSS) realisierbar sind. HTML 4.0 wurde erstmals im Dezember 1997 als Sprachstandard verabschiedet. Diese Sprachversion wurde jedoch mehrfach überarbeitet und liegt seit Dezember 1999 in Version 4.01 vor. Aufgrund der Etablierung von XML wurde HTML in Gestalt der „Extensible HyperText Markup Language“ (XHTML) neu definiert. So existiert heute neben dem SGML-basierten, klassischen HTML das XML-basierte XHTML. Version 1.0 wurde im Januar 2000 veröffentlicht, die überarbeitete Version ist seit August 2002 erhältlich. XHTML 2.0 liegt als Draft-Version vor, die aktuellste Version ist aus dem Jahr 2006.

Präsentation:

Präsentationswerkzeuge: +

HTML-Dokumente werden durch HTML-Browser, auch Webbrowser genannt, wie z.B. Microsoft Internet Explorer oder Mozilla Firefox, dargestellt. Im Gegensatz zu SGML sind keine Styledefinitionen notwendig, da viele Tags schon Layoutinformationen besitzen, die der Browser umsetzt. Unter Verwendung von Styledefinitionen kann diese Standardformatierung jedoch überschrieben werden.

Darstellung: –

Die Inhalte einer HTML-Seite können mit verschiedenen Browsern unterschiedlich aussehen. Außerdem besteht das Problem der aktiven Links, die häufig fehlerhafte Verweise liefern können.

Struktur: + +

HTML ist bezüglich des Aufbaus und des Inhaltes strukturiert. Obwohl HTML auf der Beschreibungssprache SGML basiert, kann die Sprache nicht durch individuelle Definitionen erweitert werden, denn die Markup-Befehle sind vorgegeben. Im Laufe der Entwicklung von HTML konzentrierte man sich außerdem mehr auf die Kontrolle des Layouts als auf die Struktur. Deshalb ist in einem HTML-Dokument die Trennung von Layout und eigentlichem Inhalt nicht möglich.

Sicherheit:

Sicherheitsmechanismen: – –

Wie auch in der SGML-Definition sind in HTML keine Sicherheitsmechanismen vorgesehen. Die Sprache kann nicht durch individuelle Definitionen erweitert werden, da die Markup-Befehle vorgegeben sind.

Elektronische Signatur: o

Die elektronische Signatur wird in HTML nicht realisiert und kann daher nicht bewertet werden.

3.5.2 Secure Multipurpose Internet Mail Extensions (S/MIME)

Zweck und Verwendung: +

S/MIME ist ein von der Firma RSA Data Security, Inc. entwickelter offener Standard zur Sicherung von E-Mails im MIME-Format, wobei die MIME-Komponenten sowohl verschlüsselt als auch signiert werden können.

MIME ist ein Internet-Standard für den Aufbau von E-Mails, der das Anhängen von beliebigen Binärinformationen an die Nachricht erlaubt. In so genannten MIME-Types werden verschiedene Nachrichtenformate für Text, Grafik, Audio und Video sowie anwendungsspezifische Datendateien vom Typ „application“ vordefiniert. Der MIME-Standard

wurde in den RFC-Dokumenten RFC 2045 bis RFC 2049 spezifiziert, welche zum Teil überarbeitet und erweitert wurden.

S/MIME erweitert die MIME-Types um Konstrukte für signierte und verschlüsselte Nachrichten und wird als Verschlüsselungsstandard von verschiedenen E-Mail-Programmen wie Microsoft Outlook oder IBM Lotus Notes direkt unterstützt.

S/MIME baut auf der Cryptographic Message Syntax (CMS) auf und verwendet die dort spezifizierten Mechanismen zur Verschlüsselung und Signatur. S/MIME erfordert eine Public Key Infrastruktur auf Basis von X.509 Zertifikaten, wobei E-Mails im MIME-Format gemäß [RFC2015 1996] auch über Pretty Good Privacy (PGP) gesichert werden können [Kirsch 2001].

Da die Nutzdaten von E-Mails auch im Gesundheitswesen kryptographisch gesichert werden müssen, kommt S/MIME auch in diesem Bereich zum Einsatz.

Transparenz und Standardisierung: +

S/MIME ist ursprünglich eine Entwicklung der Firma RSA Data Security Inc., die im Rahmen der IETF weiterentwickelt wurde. S/MIME Version 1 wurde 1995 und Version 2 1998 veröffentlicht. Nennenswerte Verbreitung fand S/MIME erst mit Version 2 aus dem Jahr 1998, die in zwei informativen RFC-Dokumenten RFC 2311 und RFC 2312 vorgestellt wurde. S/MIME Version 3, welche 1999 veröffentlicht wurde, ist in RFC 2632 bis RFC 2634 dokumentiert. Die aktuellste Version 3.1 erschien im Jahr 2004 und wurde als RFC 3850 und RFC 3851 veröffentlicht. RFC 2634 wurde im Jahr 2007 durch RFC 5035 erneuert.

Stabilität: –

Erste Veröffentlichung E = 1995

Alter A = B – E = 12

Anzahl Versionen V = 4

Stabilität S = A / V = 3

Letzte Veröffentlichung L = 1999

Zeitraum seit der letzten Änderung Z = 3

Präsentation:

Präsentationswerkzeuge: +

S/MIME wird von mehreren Firmen in ihren E-Mail-Systemen verwendet, darunter auch die Firmen Microsoft und IBM. Die E-Mail-Programme Outlook und Outlook Express der Firma Microsoft und Lotus Notes der Firma IBM unterstützen S/MIME.

Darstellung: o

Der Standard präsentiert in erster Linie Austauschformate, jedoch nicht die Präsentation der Daten.

Struktur: + +

S/MIME nutzt die MIME-Struktur von Nachrichten und baut darauf mit kryptographischen Elementen auf. Die Nachrichten sind bezüglich ihres Aufbaus und des Inhalts strukturiert, wobei der Inhalt nicht nur aus Texten, sondern auch Bildern, Videos oder sonstigen Anhängen bestehen kann.

Sicherheit:

Sicherheitsmechanismen: + +

Die Verwendung kryptographischer Verfahren zur Sicherung der Integrität, Authentizität und Vertraulichkeit bei signierten und verschlüsselten S/MIME-Nachrichten wird durch die CMS spezifiziert. Der CMS-Standard basiert auf PKCS#7. Version 2 von S/MIME baute ebenfalls auf PKCS#7 auf, während in S/MIME Version 3 auf die aktuellere CMS gewechselt wurde.

Eine S/MIME-Nachricht besteht aus einer Kombination von so genannten MIME-Bodies und geschützten CMS-Objekten. Es gibt verschiedene Typen für signierte und verschlüsselte Nachrichten, welche miteinander kombiniert sein können.

Tabelle 1: Nachrichtentypen in S/MIME

	Type	Subtype	S/MIME-Type
Signierte Nachrichten	multipart	signed	
	application	Pkcs7-mime	Signed-data
Verschlüsselte Nachrichten	application	Pkcs7-mime	Enveloped-data

Elektronische Signatur: + +

Wie in Tabelle 1 beschrieben gibt es für elektronische Signaturen zwei unterschiedliche Nachrichtentypen:

- application/pkcs7-mime/signed-data und
- multipart/signed

Beim application/pkcs7-Nachrichtentyp wird der MIME-Body direkt im Content-Feld des CMS-Objekts abgelegt. Beim multipart/signed-Nachrichtentyp stehen der MIME-Body und das CMS-Objekt in verschiedenen Teilen der Nachricht, wobei das Content-Feld des CMS-Objekts leer bleibt. Dadurch kann der Inhalt der MIME-Nachricht auch ohne CMS-Unterstützung gelesen werden.

CMS bietet in der von S/MIME referenzierten Version die Möglichkeit, eine beliebige Anzahl an Zertifikaten der Zertifizierungshierarchie sowie zugehörige Sperrlisten nach X.509-Spezifikation oder andere Zertifikatsstatusinformationen zu speichern ([RFC3852 2004]). Mit der Evidence Record Syntax (ERS) ist eine Spezifikation zur Langzeitsicherung elektronisch signierter Daten erarbeitet und als [RFC4998 2007] standardisiert wurden. Die in der ERS spezifizierten Signaturerneuerungen können für beliebige signierte Daten erzeugt und in CMS Objekte integriert werden.

Sonstige Beurteilungen:

Die Arbeitsgruppe „Internet“ der GMDS empfiehlt S/MIME für die klinische Nutzung von E-Mail.

4 Zusammenfassung

Im Anschluss an die detaillierte Analysen und Bewertungen der einzelnen Dateiformate werden in diesem Kapitel die Ergebnisse zusammengefasst und die im Pflichtenheft für dieses Gutachten gestellten Fragen beantwortet.

Tabelle 2: Bewertung der Dateiformate

	PDF	TIFF	Word	ODF	PS	ASCII	CSV	JPEG	DICOM	HTML	S/MIME
Verwendung	+	+	+	– –	– –	+	+	+	+ +	+	+
Standardisierung	+	+ +	+	+ +	+	+ +	+	+ +	+ +	+	+
Stabilität	–	–	–	– –	–	+ +	–	+	–	–	–
Werkzeuge	+	+	–	+	+	+	+	+	+	+	+
Darstellung	+	+	–	+	+	–	+	+	+ +	–	o
Struktur	+ +	+	+ +	+ +	+	– –	–	+	+ +	+ +	+ +
Sicherheit	+ +	– –	+ +	+	– –	– –	– –	– –	+ +	– –	+ +
Signatur	+	o	+	o	o	o	o	o	– –	o	+ +

+ + Das Dateiformat ist sehr gut geeignet. + Das Dateiformat ist geeignet. – Das Dateiformat ist mit Einschränkungen geeignet. – – Das Dateiformat ist nicht zu empfehlen. o Keine Bewertung

Welche konventionellen Dateiformate eignen sich für die elektronische Archivierung?

PDF und TIFF als klassische Dateiformate für die elektronische Archivierung sind auch im Kontext klinischer Studien geeignet. Durch die Standardisierung von PDF/A im Rahmen der ISO im Jahr 2005 hat sich PDF zum allgemein anerkannten Dateiformat für die Archivierung von Dokumenten entwickelt. Neben der eindeutigen visuellen Reproduzierbarkeit gewährleistet PDF/A-1a auch die Textextrahierung. Werkzeuge für die Erzeugung und Visualisierung sind von verschiedenen Anbietern auch lizenzkostenfrei verfügbar.

TIFF ist ein Grafikformat welches ebenfalls von der ISO standardisiert. Die Spezifikation des Dateiformats ist seit 1992 stabil. Auch TIFF ermöglicht die eindeutige visuelle Reproduzierbarkeit aber enthaltene Texte können durch die Speicherung als Rastergrafik

nicht extrahiert werden. Das Komprimierungsverfahren LZW ist seit 2004 nicht mehr durch Patente geschützt und kann für die verlustfreie Kompression verwendet werden. Werkzeuge für die Erzeugung und Visualisierung sind von verschiedenen Anbietern auch lizenzkostenfrei verfügbar.

Bei Microsoft Word und den anderen Microsoft Office Formaten handelt es sich weiterhin nur um De-facto-Standards, deren Spezifikationen in der neuesten Version aber offen gelegt sind. Durch die teilweise Inkompatibilität zwischen den verschiedenen Versionen ergeben sich oftmals Probleme bei der eindeutigen Visualisierung von Word-Dokumenten, weshalb die Eignung als Archivierungsformat begrenzt ist. Das Dateiformat ist in seiner neuesten Version bzgl. Aufbau und Inhalt strukturiert. Für die Erzeugung und Visualisierung sind in der Regel Microsoft Programme notwendig, da andere Hersteller zumeist nur begrenzte Import- und Exportfunktionalitäten anbieten.

ODF ist ein auf XML basierendes, sehr neues Dateiformat für Büroanwendungen welches 2006 von der ISO standardisiert wurde. Da bei ODF die Dokumentinhalte, Formatierungen und Metadaten in verschiedenen Dateien gespeichert und über ein ZIP-Archiv zusammengefasst und komprimiert werden, muss darauf geachtet werden, dass für die eindeutige Reproduzierbarkeit eines Dokumentes alle Dateien vorhanden sind. Durch die Speicherung der Dokumentinhalte in XML ist die Textextrahierung möglich. Werkzeuge für die Erzeugung und Visualisierung sind von verschiedenen Anbietern auch lizenzkostenfrei verfügbar.

PostScript ist eine zum De-facto-Standard gewordene Seitenbeschreibungssprache die seit 1997 stabil ist. PostScript ist ein layoutorientiertes Dateiformat ohne inhaltliche Strukturierung. Werkzeuge für die Erzeugung und Visualisierung sind von verschiedenen Anbietern auch lizenzkostenfrei verfügbar.

ASCII ist bereits sehr lange vorhandenes Textformat für unstrukturierte Dokumente welches international standardisiert ist. ASCII ist als 7-Bit US-ASCII Zeichensatz bereits seit 1972 unverändert und als 8-Bit ISO 8859-1 Zeichensatz seit 1986 stabil. Vorraussetzung für die eindeutige Darstellung ist die Wahl des richtigen Zeichensatzes. Texteditoren für die Erzeugung und Visualisierung sind von verschiedenen Anbietern auch lizenzkostenfrei verfügbar.

CSV ist ein auf ASCII basierendes Dateiformat für einfach strukturierte Daten und Tabellen. CSV ist ein De-facto-Standard, der seit 2005 über einen informellen RFC spezifiziert ist.

Probleme bei der Darstellung von CSV können durch die Fehlinterpretation der separierenden Kommas bei Dezimalzahlen und wie bei ASCII durch die Verwendung des falschen Zeichensatzes auftreten. CSV Dateien können mit beliebigen Texteditoren erstellt und visualisiert werden und werden von verschiedenen Büroanwendungen unterstützt.

JPEG ist ein Kompressionsverfahren und ein Grafikformat welches ebenfalls von der ISO standardisiert und seit 1992 stabil ist. Auch JPEG ermöglicht die eindeutige visuelle Reproduzierbarkeit aber enthaltene Texte können durch die Speicherung als Rastergrafik nicht extrahiert werden. Im Gegensatz zu LZW Komprimierung ist die JPEG Komprimierung aber verlustbehaftet. Werkzeuge für die Erzeugung und Visualisierung sind von verschiedenen Anbietern auch lizenzkostenfrei verfügbar. JPEG 2000 unterstützt auch verlustfreie Kompression, wird aber von vielen Werkzeugen noch nicht unterstützt.

DICOM hat sich in den letzten Jahren zum allgemein anerkannten Dateiformat für die Speicherung und Übertragung medizinischer Bilddaten entwickelt. DICOM ist international standardisiert und wird stetig weiterentwickelt. Bei standardkonformen Bilddaten ist die eindeutige visuelle Reproduzierbarkeit gewährleistet. Neben den Bilddaten können auch strukturierte Befunde in DICOM abgelegt werden. DICOM wird von nahezu allen Herstellern von Modalitäten, PACS, Diagnose-Workstations und Archiven unterstützt und auch lizenzkostenfreie Werkzeuge sind verfügbar.

HTML hat sich als Standard vor allem für Webseiten durchgesetzt. HTML wird vom W3C standardisiert und liegt in der Version 4.01 seit 1999 vor. Da in HTML oftmals aktive Elemente und externe Informationen verwendet werden, ist die eindeutige Reproduzierbarkeit oft nicht gewährleistet, weshalb eine Archivierung derartiger Dokumente nicht zu empfehlen ist. Die Inhalte eines HTML-Dokuments können mit verschiedenen Browsern unterschiedlich aussehen.

S/MIME ist das Standardformat für sichere E-Mails, welches von der IETF standardisiert wird. Neben der elektronischen Signatur von E-Mails ermöglicht S/MIME auch deren Verschlüsselung. Bei der verschlüsselten Archivierung von E-Mails müssen Schlüssel hinterlegt werden, um die dauerhafte Zugreifbarkeit der Inhalte zu gewährleisten. S/MIME wird von den meisten E-Mail-Programmen unterstützt.

Welche Normen existieren und welche Rolle spielen Sie?

Der Großteil der analysierten und bewerteten Dateiformate ist international standardisiert. Die detaillierten Normen der einzelnen Dateiformate sind in Kapitel 3 jeweils unter dem

Kriterium Transparenz und Standardisierung aufgeführt. An dieser Stelle wird nur noch auf die klassischen Dateiformate PDF und TIFF eingegangen.

PDF wird von der Firma Adobe Systems Incorporated entwickelt und ist nicht als komplettes Dateiformat standardisiert. Trotz der Bindung an einen Hersteller hat sich PDF zu einem De-facto-Standard für den elektronischen Dokumentenaustausch entwickelt und die Spezifikationen sind frei verfügbar. Als ISO 19005-1 wurde PDF/A im Jahr 2005 für den Bereich der Langzeitarchivierung von Dokumenten international normiert. Auch in anderen Anwendungsbereichen bildet PDF den Ursprung für internationale Normen. So definiert PDF/X eine Untermenge von PDF speziell für die Anforderungen der Druckindustrie. PDF/A definiert Anforderungen für den Bereich der Langzeitarchivierung von Dokumenten und legt verschiedene Konformanzebenen auf der Basis der PDF Version 1.4 fest. Mittlerweile existieren verschiedene Werkzeuge von unterschiedlichen Herstellern zum Erzeugen und Visualisieren von PDF/A, die u.a. im PDF/A Competence Center (www.pdfa.org) zu finden sind.

TIFF wurde von der Firma Aldus entwickelt, wird aber mittlerweile ebenfalls von der Firma Adobe Systems Incorporated gepflegt und veröffentlicht. Darüber hinaus wurde TIFF nach ISO für die medienunabhängige Bildverarbeitung standardisiert und als ISO 12639 unter dem Namen „Tag image file format for image technology“ (TIFF/IT) veröffentlicht.

Welche Vor- und Nachteile bieten die Formate PDF und TIFF im Hinblick auf Datensicherheit (Vermeidung von Änderung oder Verlust) und Datenschutz (Vermeidung von unbefugtem Zugriff)?

PDF bietet sehr umfangreiche Möglichkeiten zur Gewährleistung von Datenschutz und Datensicherheit. Ist der Inhalt eines Dokuments vertraulich und soll nur bestimmten Benutzern zugänglich sein, so kann der Ersteller die Datei über symmetrische Verschlüsselungsverfahren unter Angabe eines Passwortes verschlüsseln.

Eine zweite Schutzvariante ist die Vergabe bestimmter Nutzungsbeschränkungen beim Erstellen der PDF-Datei. Auf diese Weise kann man z.B. das Ausdrucken oder Verändern der Datei verhindern. Bei der Einstellung derartiger Berechtigungen ist die Vergabe eines Haupt- und eines Benutzerkennwortes unbedingt notwendig, da die Einstellungen sonst relativ einfach umgangen werden können. Grundsätzlich ist davon auszugehen, dass die in einer PDF-Datei eingestellten Berechtigungen keinen ernsthaften Schutz bieten, da die Implementierung des Viewers die Einhaltung der Nutzungsbeschränkungen gewährleistet.

Bei der Schwärzung von PDF-Dokumenten wird davor gewarnt, dass die Inhalte auch wirklich gelöscht und nicht nur überschrieben werden sollen, da diese sonst wiederhergestellt werden können.

Um die Zugänglichkeit der Inhalte in PDF-Dokumenten zu gewährleisten, sind sowohl die Verschlüsselung als auch die Verwendung der Nutzungsbeschränkungen in PDF/A untersagt.

TIFF dagegen bietet keine Möglichkeiten zur Gewährleistung von Datenschutz und Datensicherheit in der Formatspezifikation. Unabhängig davon können aber andere Mechanismen angewandt werden, um den Datenschutz und Datensicherheit von TIFF-Dokumenten zu gewährleisten. Zur Sicherung der Vertraulichkeit können TIFF-Dokumente beispielsweise gemäß CMS verschlüsselt werden.

Welche Möglichkeiten bieten PDF und TIFF hinsichtlich Authentizität und Integrität, wo liegen die Nachteile oder Risiken?

In PDF können elektronische Signaturen zur Nachprüfbarkeit der Integrität und Authentizität sichtbar und unsichtbar angebracht werden. Im Gegensatz zur Verschlüsselung sind elektronische Signaturen auch in PDF/A erlaubt. Die Signatur bezieht sich auf ausgewählte Objekte eines PDF-Dokumentes oder auf einen definierten Byteumfang. Signierte PDF-Dokumente können auch noch überarbeitet werden, da die Signatur mit der Version des Dokumentes gespeichert wird und durch die Änderung eine neue Dokumentversion entsteht. In PDF werden zwei unterschiedliche Signaturformate unterschieden. Neben einer internen Struktur zur Ablage der Signaturinformationen unterstützt PDF auch das international normierte „PKCS#7 Signature Format“. Das interne Signaturformat stellt ein proprietäres Format für Signaturen von Adobe dar, bei welchem neben signaturbeschreibenden Informationen (Signaturgrund, -ort, -zeit), den Zertifikaten einer Zertifizierungshierarchie (Benutzer-, CA-, Rootzertifikat) der verschlüsselte Signaturwert angegeben wird. Das „PKCS#7 Signature Format“ gibt die Möglichkeit Signaturen im standardisierten PKCS#7-Format zu speichern. In das PKCS#7 Element können zusätzliche Verifikationsdaten wie Zertifikate, Attributzertifikate, Zeitstempel, Sperrlisten und Zertifikatstatusinformationen eingebettet werden. Problematisch ist, dass PKCS#7 nicht mehr weiter entwickelt wird, und die Nachfolgespezifikation CMS andere Möglichkeiten zur Integration von Verifikationsdaten vorsieht, welche Adobe aber nicht unterstützt.

TIFF bietet auch keine Möglichkeiten zur elektronischen Signatur in der Formatspezifikation. Unabhängig davon können aber andere Mechanismen angewandt werden, um die Datensicherheit von TIFF-Dokumenten zu gewährleisten.
Zur Sicherung der Integrität und Authentizität können TIFF-Dokumente beispielsweise gemäß CMS signiert werden. CMS bietet die Möglichkeit, eine beliebige Anzahl an Zertifikaten der Zertifizierungshierarchie sowie zugehörige Sperrlisten nach X.509-Spezifikation oder andere Zertifikatsstatusinformationen sowie Zeitstempel zu speichern. Mit der Evidence Record Syntax (ERS) ist eine Spezifikation zur Langzeitsicherung elektronisch signierter Daten erarbeitet und standardisiert wurden. Die in der ERS spezifizierten Signaturerneuerungen können für beliebige signierte Daten erzeugt und in CMS Objekte integriert werden.
Eine andere Möglichkeit zur Sicherung der Integrität digitaler Medien stellt die Verwendung von Wasserzeichen dar. Durch fragile Wasserzeichen können Änderungen an digitalem Bildmaterial erkannt werden. Generell verfälschen diese fragilen Wasserzeichen durch das direkte Integrieren der Informationen das Original und können zu Fehlinterpretationen der medizinischen Inhalte führen. Zur Sicherstellung von Authentizität und Integrität digitaler Medien sollten nach [Pharow 2003; Steinebach 2007] deshalb fragile invertierbare Wasserzeichen unter Verwendung elektronischer Signaturen eingesetzt werden, die auch die Wiederherstellung des Originals ermöglichen. Diese Sicherheitstechnologien befinden sich aber noch im Bereich der Forschung und müssen noch in anwendbare Technologien überführt werden.

Bei der Sicherung von Integrität und Authentizität über elektronische Signaturen muss berücksichtigt werden, dass die Konvertierung von Daten in andere Dateiformate dazu führt, dass die elektronische Signatur des Ausgangsdokumentes für das konvertierte Dokument nicht als Sicherungsmittel verwendet werden kann. Dass bedeutet, dass z.B. eine elektronisch signierte E-Mail im Dateiformat S/MIME nach der Konvertierung nach PDF als unsigniertes Dokument vorliegt. Zur Erhaltung des Beweiswertes von signierten Dokumenten im Rahmen einer Transformation müssen neben der eigentlichen Konvertierung zusätzliche Maßnahmen wie die Prüfung und Speicherung von Signaturdaten und die Sicherung des transformierten Dokumentes vorgenommen werden. Konzepte und Lösungen zur rechtssicheren Transformation signierter Dokumente wurden im Projekt „TransiDoc“ erarbeitet und sind u.a. in [Kunz 2005] sowie über www.transidoc.de verfügbar.

Wie kann der Zugriff auf Archivdateien (konventioneller Formate) optimiert werden?

Grundsätzlich sollten zu allen Dokumenten, egal in welchem Dateiformat sie vorliegen, Metadaten gespeichert werden. Über diese Metadaten, wie z.B. Dokumenttyp, Erstellungsdatum etc. können die Dokumente gesucht und gefunden werden.

Bei PDF ermöglichen Strukturinformationen die Wiederverwendung der im Dokument enthaltenen Inhalte. Durch die Textsuche im einzelnen Dokument oder die Volltextrecherche innerhalb einer Dokumentensammlung lassen sich sehr einfach Detailinhalte auffinden.

Zur Optimierung derartiger Volltextsuche kann ein Index aufgebaut und durchsucht werden. In diesem Index werden Schlagwörter gespeichert, die auf die Dokumente, in denen diese enthalten sind, referenzieren. Zum Aufbau derartiger Volltextindizes und zur Suche von Dokumenten über diesen Index existieren mittlerweile auch Open Source Bibliotheken wie z.B. Lucene ([Gospodnetic 2004]). Derartige Mechanismen bieten sich für alle textorientierten Dateiformate an.

Bei Rastergrafiken wie TIFF und JPEG sind derartige Suchmechanismen nicht möglich. Hier müssen die Suchbegriffe entweder bei der Archivierung der Dateien mitgegeben oder durch Texterkennung ermittelt werden. Die Texterkennung, welche auch Optical Character Recognition (OCR) bezeichnet wird, ist jedoch sehr aufwendig und teilweise fehleranfällig.

Welche Kostentreiber sind bei der Archivierung mit konventionellen Formaten besonders zu beachten?

Neben den Investitionskosten für Hard- und Software müssen vor allem die Betriebskosten für die elektronische Archivierung berücksichtigt werden.

Werden die Dokumente elektronisch erstellt, entstehen keine zusätzlichen Kosten für das Scannen und Indexieren der Dokumente. Nach [Schmücker 1996] wird die elektronische Archivierung dann zum günstigsten Verfahren, wenn ein Digitalisierungsgrad von mindestens 30 Prozent insbesondere durch die automatische Übernahme von elektronisch erstellten Dokumenten erreicht ist.

Darüber hinaus muss berücksichtigt werden, dass elektronisch signierte Dokumente im Laufe der Zeit erneut elektronisch signiert werden müssen, um ihrer Beweiskraft zu erhalten. Durch die Anwendung der im Projekt ArchiSig entwickelten Konzepte kann die Signaturerneuerung auch für große elektronische Archive wirtschaftlich gestaltet werden, da sehr viele Dokumente gemeinsam erneut elektronisch signiert werden können ([Brandner 2003]).

Gibt es neue aussichtsreiche Entwicklungen hinsichtlich alternativer Formate?

Im Rahmen der Analyse wurden mit ODF und Open Office XML sehr neue auf XML basierende Dateiformate untersucht und bewertet.

Darüber hinaus existieren verschiedene neue Dateiformate für die unterschiedlichsten Anwendungsbereiche. Dazu zählen unter anderem die XML Paper Specification (XPS) von Microsoft, die als zukünftige Alternative zu PDF/X im Bereich der Druckindustrie angesehen wird. Des Weiteren ist in dieser Kategorie das von den AT&T Laboratories entwickelte Dateiformat DjVu zu nennen, welches für gescannte Dokumente eingesetzt werden kann, da es neben dem Seitenhintergrund als Rastergrafik auch Text enthalten kann.

All diese Dateiformate sind aber noch am Beginn Ihrer Entwicklung und werden noch weiter überarbeitet. Ihre Erstellung und Visualisierung ist noch auf wenige Werkzeuge begrenzt, was einen flächendeckenden Einsatz über Anwendungssystemgrenzen hinweg erschwert.

5 Abkürzungsverzeichnis

ACR	American College of Radiology
AIIM	Association for Information and Image Management
ANSI	American National Standards Institute
ASCII	American Standard Code for Information Interchange
BSI	Bundesamt für Sicherheit in der Informationstechnik
CA	Certification Authority
CCITT	Comité Consultatif Internationale de Téléphones et Télégraphes
CMS	Cryptographic Message Syntax
CRF	Case Record Forms
CSS	Cascading Style Sheet
CSV	Character Separated Values
DICOM	Digital Imaging and Communications in Medicine
DOS	Disk Operating System
EBCDIC	Extended Binary Coded Decimal Interchange Code
ECMA	European Computer Manufacturer‘s Association
EPS	Encapsulated PostScript
ERS	Evidence Record Syntax
GIF	Graphics Interchange Format
HTML	Hypertext Markup Language
IEC	International Electrotechnical Commission
IETF	Internet Engineering Task Force
ISO	International Organization for Standardization
ITU	International Telecommunications Union

JFIF	JPEG File Interchange Format
JPEG	Joint Photographic Experts Group
LZW	Lempel-Ziv-Welch-Algorithmus
MIME	Multipurpose Internet Mail Extensions
NEMA	National Electrical Manufacturers Association
OASIS	Organization for the Advancement of Structured Information Standards
OCSP	Online Certification Status Protocol
ODF	Open Document Format
PDF	Portable Document Format
PGP	Pretty Good Privacy
PKCS	Public-Key Cryptography Standards
RFC	Request For Comments
S/MIME	Secure Multipurpose Internet Mail Extensions
SGML	Standard Generalized Markup Language
SOP	Standard Operating Procedures
TIFF	Tagged Image File Format
TLS	Transport Layer Security
TMF	Telematikplattform für Medizinische Forschungsnetze e.V.
TMF	Trial Master Files
W3C	World Wide Web Consortium
WWW	World Wide Web
XHTML	Extensible HyperText Markup Language
XML	Extensible Markup Language
XPS	XML Paper Specification

6 Literaturverzeichnis

[Adobe 1992] Adobe (1992). TIFF Revision 6.0. Seattle, Adobe Developers Association.

[Adobe 1999] Adobe (1999). PostScript LANGUAGE REFERENCE third edition. Reading, Massachusetts, Addison-Wesley Publishing Company.

[Adobe 2003] Adobe (2003). Elektronische Signaturen in Adobe Acrobat, Adobe Systems Incorporated.

[Adobe 2006] Adobe (2006). PDF Reference sixth edition. Adobe Portable Document Format Version 1.7, Adobe Systems Incorporated.

[Bartel 2002] Bartel, M., Boyer, J., et al. (2002). XML-Signature Syntax and Processing D. Eastlake, J. Reagle und D. Solo, W3C.

[Binder 2001] Binder, B. (2001). Interoperabilität in der Telemedizin - Wie wird aus vielen Einzelsystemen ein medizinisches Netzwerk? Bad Nauheim.

[Borghoff 2003] Borghoff, U. M., Rödig, P., et al. (2003). Langzeitarchivierung - Methoden zur Erhaltung digitaler Dokumente. Heidelberg, dpunkt.verlag.

[Brandner 2003] Brandner, R. und Pordesch, U. (2003). "Konzept zur signaturgesetzkonformen Erneuerung qualifizierter Signaturen." Datenschutz und Datensicherheit 27(6): 354-359.

[BSI 2006] BSI, Ed. (2006). IT-Grundschutz-Kataloge. Köln, Bundesanzeiger-Verlaggesellschaft mbH.

[DICOM_Suppl31 2000] DICOM_Suppl31 (2000). Security Enhancements One. NEMA Standards Publication PS 3, Supplement 31, DICOM Standards Committee.

[DICOM_Suppl41 2001] DICOM_Suppl41 (2001). Supplement 41: Digital Signatures, DICOM Standards Committee, Working Group 14.

[DICOM_Suppl55 2001] DICOM_Suppl55 (2001). Supplement 55: Attribute Level Confidentiality (including De-identification), DICOM Standards Committee, Working Group 14 - Security.

[DICOM_Suppl86 2004] DICOM_Suppl86 (2004). Supplement 86: Digital Signatures in Structured Reports, DICOM Standards Committee, Working Group 14 - Security.

[DLM-Forum 1997] DLM-Forum (1997). Leitlinien für den Umgang mit elektronischen Informationen - Maschinenlesbare Daten und elektronische Dokumente. Luxemburg: Europäische Gemeinschaften.

[ECMA 2006] ECMA (2006). OFFICE OPEN XML OVERVIEW. T. NGO.

[Eichelberg 2002] Eichelberg, M., Riesmeier, J., et al. (2002). "Standards für den sicheren Datenaustausch in der Teleradiologie am Beispiel der Bild- und Befundverteilung." Der Radiologe 42(2): 94-100.

[ENV12052 1995] ENV12052 (1995). European prestandard on medical informatics - Medical imaging communication (MEDICOM) prENV 12052. E. C. f. Standardization.

[Fox 1998] Fox, D. (1998). "Zu einem prinzipiellen Problem digitaler Signaturen." Datenschutz und Datensicherheit 22(7): 386-388.

[Gospodnetic 2004] Gospodnetic, O. und Hatcher, E. (2004). Lucene in Action. Greenwich, CT Manning Publications.

[Häber 2005] Häber, A., Dujat, C., et al. (2005). Leitfaden Dokumentenmanagement und digitale Archivierung im Gesundheitswesen. Darmstadt, GIT Verlag.

[Hollerbach 2003] Hollerbach, A. und Brandner, R. (2003). "Kriterien und Bewertung von Datenformaten für die beweiskräftige und sichere Langzeitspeicherung medizinischer Dokumente." Forum der Medizin_Dokumentation und Medizin_Informatik 5(4): 105-109.

[Horsch 2002] Horsch, A. und Handels, H. (2002). Telematik im Gesundheitswesen. Handbuch der Medizinischen Informatik. T. Lehman und E. Meyer zu Bexten. München, Hanser.

[Imamura 2002] Imamura, T., Dillaway, B., et al. (2002). XML Encryption Syntax and Processing. D. Eastlake und J. Reagle, W3C.

[Kampffmeyer 2000] Kampffmeyer, U. und Rogalla, J. (2000). Grundsätze der elektronischen Archivierung. Darmstadt, VOI Verband Organisations- und Informationssysteme e.V.

[Kirsch 2001] Kirsch, C. (2001). S/MIME vs. OpenPGP: Eine Entscheidungshilfe. Management und Wissen E-Mail-Verschlüsselung.

[Kunz 2005] Kunz, T., Schmidt, A., et al. (2005). "Konzepte für rechtssichere Transformationen signierter Dokumente " Datenschutz und Datensicherheit 29(5): 279-285.

[Lehmann 2002] Lehmann, T. M., Hiltner, J., et al. (2002). Medizinische Bildverarbeitung. Handbuch der Medizinischen Informatik. T. Lehman und E. Meyer zu Bexten. München, Hanser.

[Macnaghten 2007] Macnaghten, E. (2007). Technical Distinctions of ODF and OOXML. ODF/OOXML Technical White Paper**:** 1-34.

[Martins 2003] Martins, F. P. und Kobylinska, A. (2003). Adobe Acrobat 6 Standard und Professional. Berlin, Springer.

[Matzer 2000] Matzer, M. und Lohse, H. (2000). Dateiformate - Bedeutung, Einsatz und Konvertierung. München, Deutscher Taschenbuch Verlag.

[Merz 2001] Merz, T. (2001). "Verschlüsselung mit Acrobat - Fast sicher." iX 2001(9): 56-59.

[Merz 2002] Merz, T. und Drümmer, O. (2002). Die PostScript & PDF-Bibel. Heidelberg, dpunkt-Verlag.

[Neumann 2002] Neumann, M. (2002). "DICOM - Current status and future developments for radiotherapy." Zeitschrift für Medizinische Physik 12(3): 171-176.

[OASIS 2005] OASIS (2005). Open Document Format for Office Applications (OpenDocument) v1.1. P. Durusau, G. Edwards, D. Faure, T. Magliery und D. Vogelheim, OASISOpen.

[PDF-Tools 2007] PDF-Tools (2007). PDF/A – Ein neuer Standard für die Langzeit-Archivierung.

[Pharow 2003] Pharow, P., Dittmann, J., et al. (2003). Sicherstellung von Integrität und Verbindlichkeit in digitalen Medien. GMDS 2003.

[Pordesch 2000] Pordesch, U. (2000). "Der fehlende Nachweis der Präsentation signierter Daten." Datenschutz und Datensicherheit 24(2): 89-95.

[Raggett 1999] Raggett, D., Hors, A. L., et al. (1999). HTML 4.01 Specification, W3C.

[Reich 2005] Reich, M. und Wilczek, M. (2005). Das Praxisbuch zu Adobe Acrobat 7 und PDF im Unternehmensalltag. Kilchberg CH, Smart Books Publishing AG.

[Repici 2002] Repici, J. (2002). The Comma Separated Value (CSV) File Format Creativyst, Inc.

[RFC2015 1996] RFC2015 (1996). MIME Security with Pretty Good Privacy (PGP).

[RFC2315 1998] RFC2315 (1998). PKCS #7: Cryptographic Message Syntax Version 1.5.

[RFC2560 1999] RFC2560 (1999). X.509 Internet Public Key Infrastructure Online Certificate Status Protocol - OCSP.

[RFC3161 2001] RFC3161 (2001). Internet X.509 Public Key Infrastructure Time-Stamp Protocol (TSP).

[RFC3280 2002] RFC3280 (2002). Internet X.509 Public Key Infrastructure Certificate and Certificate Revocation List (CRL) Profile.

[RFC3281 2002] RFC3281 (2002). An Internet Attribute Certificate Profile for Authorization.

[RFC3852 2004] RFC3852 (2004). Cryptographic Message Syntax (CMS).

[RFC4180 2005] RFC4180 (2005). Common Format and MIME Type for Comma-Separated Values (CSV) Files.

[RFC4998 2007] RFC4998 (2007). Evidence Record Syntax (ERS).

[Roßnagel 2005] Roßnagel, A. und Schmücker, P., Eds. (2005). Beweiskräftige elektronische Archivierung - Bieten elektronische Signaturen Rechtssicherheit? Heidelberg, Economica.

[SAGA 2006] SAGA (2006). SAGA - Standards und Architekturen für E-Government-Anwendungen. R. I. K. Bundesministerium des Innern.

[Schmücker 1996] Schmücker, P. und Dujat, C. (1996). "Rechnerunterstützte Dokumentenverwaltung und Optische Archivierung: Der Weg zur digitalen Krankenakte." Das Krankenhaus 3/1996: S.98-105.

[Schug 2001] Schug, S. und Sembritzki, J. (2001). Standards für die sichere Kommunikation und das Management verteilter elektronischer Patientendaten - Aktivitäten, Gremien und Projekte. Bad Nauheim.

[Semler 2005] Semler, S. C. und Ripkens-Reinhardt, A. (2005). Archivierung von klinischen Forschungsunterlagen. Telemedizinführer Deutschland. A. Jäckel. Ober-Mörlen, Medizin-Forum. 2006**:** 353-356.

[Steinebach 2007] Steinebach, M., Croce-Ferri, L., et al. (2007). Digitale Wasserzeichen in eHealth-Anwendungen als Schutzmechanismus für Multimedia-Dateien. Inovationsmotor IT-Sicherheit, Bonn, Bundesamt für Sicherheit in der Informationstechnik.

[Vorwerk 2001] Vorwerk, L. und Meinel, C. (2001). Die Bedeutung des DICOM Standards für das europäische Gesundheitswesen. Trier.

[Wirsz 2000] Wirsz, N. (2000). "IT-Standards im Gesundheitswesen." Electromedica 68(1).