

2. DFG- Workshop 3.3. Sicherung

Sicherung und Erhaltung archivierter Websites

Antje Scheiding, 07.02.2006

Folie 1/ Folie 2

Herzlichen Dank Frau Dr. Höpfinger für die einleitenden Worte. Wie Sie an der Präsentation erkennen können, komme ich nicht von einem Archiv, das direkt an dem DFG-Projekt beteiligt ist. Ich bin im Rahmen der Arbeitsgruppe Archivterminologie in den DFG-Workshops eingebunden.

Deshalb an dieser Stelle ein herzliches Dankeschön, zu dem Thema Formate und Speicherkonzepte referieren zu dürfen.

Folie 3

Meine Damen und Herren, der Archivierungsprozess setzt im Regelfall am Ende des Publikationsprozesses von Webseiten ein. Wir, als Archiv, haben kaum bzw. keinen Einfluss auf die verwendete Webtechnologie oder gar Websitegestaltung. Ein wesentlicher Bestandteil für die Darstellung und Funktionalität der Website ist das Dateiformat. Es ist eine Konvention zur Interpretation der enthaltenen Informationen einer Datei. Da die Website jedoch für den Benutzer und nicht unter dem Aspekt der langfristigen Erhaltung und Sicherung erstellt wird, haben wir es mit einer Reihe von Risikofaktoren zu tun.

[Auf einiges ist bereits Frau Dr. Höpfinger eingegangen]

- Rasante Entwicklung der (Web)Technologie,
- Plattformabhängigkeit im allgemeinen – ob von Formaten, Software oder Hardware

Hervorzuheben sind in diesem Zusammenhang:

- verschiedenen HTML Versionen und Typen, aber auch nicht korrekt erstelltes HTML
- Die Ansicht des Archivobjektes unterscheidet sich von den verschiedenen Browsertypen und –versionen
- Nicht nur Hypertext/ Verlinkungsfunktion, sondern auch eingebettete und verlinkte Anwendungen, die für das look&feel – der Darstellung oft notwendig sind, bereiten große Schwierigkeiten

Folie 4

Trotzdem gibt es auch Grund zur Gelassenheit, so jedenfalls könnte man das Zitat von Mannheim aus dem Jahre 2000 interpretieren. Er gibt die Ergebnisse einer Formatanalyse wieder, die auch von anderen Webarchivierungsprojekten bestätigt worden sind. Wir haben es zu über 90% mit HTML und Bildformaten zu tun, die als de facto oder als offizielle Standards vorliegen.

Folie 5

Standardformate sind allgemein anerkannte Formate mit einer weiten Verbreitung. Im Rahmen eines Archivierungskonzeptes werden standardisierte Formate empfohlen, da sie eine Reihe von Vorteilen bieten: Ein wesentlicher Vorteil liegt darin, dass sie durch die breite Unterstützung von Herstellern und Programmierern weniger abhängig von Hard- und Softwarekomponenten sind. Mit Hilfe der meist offen liegenden Dokumentation wird nicht nur der Datenaustausch, sondern auch die Datenmigration erleichtert. Dagegen sollten gefährdete Formate in Standardformate umgewandelt bzw. migriert werden. Im Ergebnis müssen so weniger Formate verwaltet werden. Ein weiterer Vorteil von Standardformaten ist, dass durch ihre weite Verbreitung und meist offene Dokumentation notwendige Migrationsschritte in größeren zeitlichen Abständen erfolgen können.

Summiert man die angeführten Vorteile können auf lange Sicht die Personal- und Sachkosten für die Verwaltung des Webarchivs geringer ausfallen als bei einem Datenpool mit großer Formatvielfalt und anwenderbasierten Formaten.

Folie 6

In der kommenden Darstellung wurde versucht, die Standardformate in Form eines Dreiecks einzuteilen.

In der Spitze des Dreiecks sind die offiziellen Standards, im mittleren Bereich die De-facto-Standards und ganz unten die Anwenderformate angeordnet. Sie sehen, Standard ist nicht gleich Standard.

Offizielle Standards sind häufig genutzte Dateiformate, die durch ein internationale Standardisierungsinstitut (ISO, ANSI) genormt sind. Ferner sind sie offen dokumentiert und ohne patent- und lizenzrechtliche Gebühren zu benutzen.

Bei den De-facto-Standards haben wir es mit einer sehr häufigen Nutzung und großen Marktanteilen zu tun. Jedoch sind diese Formate nicht durch ein Standardisierungsinstitut genormt. Dazu zählen die nicht-proprietären, offen gelegten Formate, wie die Spezifikationen des W3C's sowie proprietäre Formate mit offenen oder geschlossenen Dokumentation.

Eine Anmerkung noch zu der Grafik selbst: Die Grafik wurde Anfang 2005 veröffentlicht. PDF/A wurde im Herbst letzten Jahres zum ISO-Standard und müsste zu den offiziellen Standards noch ergänzt werden.

Folie 7

Es stellt sich die Frage, ob man anhand der Grafik ableiten kann, welches Format für die Archivierung geeignet ist. Sicherlich sollte man auf Standardformate zurückgreifen, die im oberen bis mittleren Drittel zu finden sind. Im Rahmen der Arbeitskreises „Elektronische Archivierung“ der Vereinigung Deutscher Wirtschaftsarchive haben wir 10 Bewertungskriterien für Archivierungsformate festgelegt

1. **Lesbarkeit durch den Menschen** - der Grad, in der die digitale Darstellung offen ist für die unmittelbare Analyse durch Menschen mit Basiswerkzeugen

2. **Layouterhaltung** - Erhaltung von Layout, Struktur und Navigation, Erhaltung der Anzeige von Bild- und Textinformationen
3. **Freie Verarbeitbarkeit und Lesbarkeit durch die Maschine** - die Möglichkeit, den Inhalt der Dateien zu nutzen, an neue IT-Umgebung anzupassen ggf. zu konvertieren muss jederzeit gegeben sein (z. B. Formate mit technischen Verschlüsselungen oder sonstigen Nutzungseinschränkungen)
4. **Zugänglichkeit und Migrierbarkeit** - einfache Kodierung, möglichst unkomprimierte Formate zur unmittelbaren Lesbarkeiten von Informationen, Verwendung von internationalen Zeichenstandards und mitgelieferte Metadaten zum Format erleichtern die Zugänglichkeit und Migration
5. **Explizite Struktur/ Selbstdokumentation** - Formate, die sich selbst dokumentieren sind einfacher zu verstehen und daher weniger anfällig bei Katastrophen
6. **freie Nutzbarkeit** unabhängig von Software und Hardware, von lizenz- und patenrechtlichen Gebühren
7. **offene und vollständige Dokumentation** - freie Verfügbarkeit einer Dokumentation und von Tools inkl. Quellcode, nicht-proprietäre, offenen Standards sind üblicher Weise besser dokumentiert und besser durch Validier-Werkzeuge unterstützt
8. **großer Verbreitungsgrad** wird von einer breiten Palette von Werkzeugen unterstützt, ermöglicht den Austausch über Systemgrenzen hinweg
9. **eine gewisse Stabilität und Reife** des Formates sollte vorliegen (Formatdefinition sollte endgültige Version sein, und nicht zu viele Subtypen aufweisen)
10. **Verknüpfungen und dynamische Inhalte sind schlechter zu erhalten**, was sich aber bei unserer Archivgattung Website leider nicht vermeiden lässt, da es ja gerade ihr Charakteristikum darstellt

Im Arbeitskreis versuchen wir, die Standardformate anhand der Kriterien zu bewerten. Der Bewertungsprozess für jeden Dokumententyp ist noch nicht abgeschlossen. Bisher haben wir feststellen können, dass nicht jedes Format die Anforderungen zur vollsten Zufriedenheit erfüllt. Wir haben uns deshalb darauf geeinigt, keine endgültige Empfehlung für Archive zu geben. Jedes Archiv, jede Einrichtung muss für sich selbst entscheiden, welche Gewichtung die Kriterien einnehmen und mit welchen Abweichungen man leben kann.

Folie 8

Ich habe trotzdem versucht, Empfehlung herauszugeben. Für eingebundene Office-Dokumente und für Hyperlinkdokumente sind die angegebenen Formate empfehlenswert. PDF/A als neuer ISO-Standard und HTML 4.01 bzw. der xml-basierten Version XHTML.

In den übrigen Bereichen für Bild-, Audio- und Videoaufzeichnungen können keine klare Aussage getroffen werden.

Dazu einige Beispiele:

Das Tiff-Format ist als Archivierungsformat mittlerweile sehr weit verbreitet, da es ein verlustfreies Kompressionsverfahren unterstützt. Die Ansicht im Web ist jedoch nur mit Hilfe eines Plug-Ins möglich. Nachteilig ist zudem, dass die Dateien sehr groß sind, dadurch das Laden der Webseite erheblich verzögern.

Im Video- und Audibereich hat sich bisher noch kein Standard für die Archivierung durchgesetzt. Formate der MPEG-Gruppe sind offizielle Standards für digital kodierte Töne und Bilder, die mit verlustbehafteten Kompressionsverfahren arbeiten. Das WAV-Format ist ein nicht komprimierter Standard für kleine Tonaufzeichnungen. Es ist jedoch proprietär und nimmt viel Speicherplatz im Vergleich zu MP3-Dateien in Anspruch. Im Videobereich ist MPEG-4 sehr weit verbreitet, hat sich jedoch nicht überall durchgesetzt. Das Material Exchange Format ist eine Untergruppe des AAF-Austauschformats und kommt aus dem Bereich der professionellen Videoarchivierung. Es ist ein nicht-

proprietäres Standardformate, was sowohl eine verlustfreie als auch verlustbehaftete Kompression ermöglicht.

Es ist festzustellen, dass im Regelfall für die webbasierte Darstellung von Bild-, Audio- und Videoaufzeichnungen eine geringere Qualität in komprimierten Formaten zur Verfügung gestellt wird. Im Audio und Video-Bereich wäre zu prüfen, ob die Masteraufnahmen über den direkten Weg in das Archiv übernommen werden können, um auf risikobehaftete Formate im Erfassungsprozess zu verzichten.

AAC – Advanced Audio Coding (Nachfolger von MP3, sind nicht kompatibel)

AAF – Advanced Authoring Format (Austauschformat für bewegende Bilder mit dazugehörigen Metadaten. Das dazugehörige Abspielformat ist MXF.)

MXF – Material Exchange Format (Abspiel –und Streamingformat auf Basis von AAF)

Folie 9

Abschließend möchte ich kurz auf die unterschiedlichen Speicherkonzepte eingehen. Ich werde mich beschränken auf die Speicherformen und die Speicherstruktur.

Ich glaube, es ist selbstverständlich, dass zur Website-Archivierung eine Speicherform mit ausreichend Kapazität vorliegt, so dass der Erfassungsprozess nicht abgebrochen oder der Snapshot auf verschiedene Medien verteilt werden muss.

Als Online Speicherung wird in den meisten Fällen die Servervariante gewählt, weil da die Datensicherung und Kapazität eher gewährleistet ist als auf dem

Laufwerk. Für Archive gibt es bisher zwei Möglichkeiten, die für die Speicherung gewählt werden können.

- a) ausschließlich Speicherung auf einem Server
- b) Online-Speicherung und eine Sicherheitskopie auf einem Offline-Medium (begrenzte Lebensdauer und herstellerspezifische Formate der DVD sind zu beachten)

Eine andere Frage stellt sich bereits zu Beginn der Erfassung – nämlich wie und in welcher Form die Dateien heruntergeladen werden sollen.

- a) Die einfachste Variante ist die hierarchische Ablage nach der originalen Ablagestruktur in einem Verzeichnis.
- b) Einzelne Web-Archivierungsprojekte, die mit einem Harvester arbeiten, legen ihre Dateien automatisch im ARC-Format ab (z. B. Internet Archive)
- c) Eine weitere Variante bietet der Einsatz von XML, sowohl in Kombination mit der Ablage in der Verzeichnisstruktur als auch im ARC-Format.
- d) Eine Alternative zu der verzeichnisstrukturierten Speicherung ist die Ablage in einem Datenbanksystem. Es würde die Möglichkeit bieten, alle Arten von Daten incl. Metadaten mit abzuspeichern und komfortabel zu verwalten.

Welche Ablagestruktur gewählt wird, hängt von der Art und dem Umfang der zu verwaltenden Archivobjekte ab. Für kleine Web-Archivierungsprojekte ist die reine Ablage im Verzeichnis sicherlich ausreichend.

Folie 10

Damit bin ich auch am Ende meiner Ausführungen angelangt.

Ich bedanke mich für Ihre Aufmerksamkeit.