Dokumentenformate bei der Archivierung

Pfaffhausen 10.1.2008: Als wir vor mehr als 15 Jahren mit der Entwicklung unserer Archivista-Produkte begonnen haben, war ein digitales Archiv mit hohen Hürden verbunden. Die ersten Flachbettscanner kosteten mehrere tausend Franken und das Scannen einer Seite dauerte doch eher eine Ewigkeit (Minuten!), von einem erschwinglichen Einzug ganz zu schweigen. Zwar gab es damals schon Disketten im 3.5"-Format, doch *der Daten-Austausch scheitert(e) — damals wie heute — daran, dass Textverarbeitungssoftware XYZ nicht mit der Software ABC kompatibel war bzw. ist.*

Vor zehn Jahren haben wir begonnen, Archivista zu vertreiben. Zwischen dem Entschluss, ein Produkt haben zu wollen und der "Geburt" von Archivista vergingen satte fünf Jahre. In dieser Zeit haben wir viel gelernt. Darunter so banale Dinge, dass es sich nicht lohnt, Dokumente zu verwenden, die mit Patenten behaftet sind, weil damit zu einem späteren Zeitpunkt unter Umständen bereits archivierte Dokumente plötzlich nicht mehr gelesen werden können, weil der Patentinhaber die Nutzungsbedingungen massiv verändert (Beispiel GIF-Dateien).

Standardisierte Archivista-Formate seit 1993

Weiter haben wir (insbesondere als Anbieter) gelernt, nur Formate zu verwenden, die wir in- und auswendig kennen. Das hatte schon früh Konsequenzen. So verzichteten wir von Anfang an auf die Microsoft Office-Dokumente als Archivformat. Natürlich wäre es lukrativ gewesen, diese Dokumente einfach ins Archiv zu stellen. Nur, wie lesen wir Office-Dokumente zehn Jahre später ohne Microsoft Office wieder?

Die PDF-Falle kann jederzeit zuschlagen

Das gleiche gilt im übrigen auch für die um die Jahrtausendwende aufkommenden PDF-Dokumente. Wie öffnen wir diese ohne den Acrobat-Reader von Adobe? Wer nun einwendet, der Acrobat Reader laufe unter allen Betriebssystemen, dem sei entgegnet, dass die Lesbarkeit von archivierten Daten nicht von einem bestimmten Betriebssystem abhängig sein sollte, mal abgesehen davon, dass die aktuellen Acrobat-Versionen unzählige Abhängigkeiten haben, die zu erfüllen nicht immer so einfach sein dürfte.

Bleiben wir noch etwas beim PDF-Format. *Die derzeit gültige **PDF-Spezifikation** lässt sich mit einer älteren Version des Acrobat-Readers nicht mehr öffnen. Es erscheint einzig noch die erste Seite, den Rest (die übrigen 1400 Seiten) kriegen wir nur zu Gesicht, wenn wir updaten.* Ehrlich gesagt habe ich (nicht nur als Anbieter einer DMS-Lösung), sondern insbesondere auch als Anweder ein Problem, wenn ich bei standardisierten Formaten updaten soll. War das nicht etwa so, dass ein Standard ja geradezu darum geschaffen wurde, damit es immer gleich läuft, und ich folglich nicht upzudaten brauche?

"Formatkrieg" zwischen Office- und PDF-Dateien

Ich habe mich lange Zeit gewundert, warum nur die Microsoft Office Produkte keinen PDF-Export anbieten. Ich glaubte auch, dass es wohl mehr daran liegen muss, dass Microsoft partout nicht will, dass ich Office-Dokumente ins PDF-Format übertragen kann; letztlich braucht mein Gegenüber ja dann keine Office-Suite mehr, um das Dokument betrachten zu können.

Dann, im letzten Jahr (2007), sollte es plötzlich einen **PDF-Export von Microsoft (Quelle: golem.de)** geben. *Pech gehabt, dass passte nun Adobe nicht in dem Kram, weil die um ihren lukrativen Markt mit den Acrobat-Produkten fürchteten.* Natürlich wird dies zuweilen auch wieder dementiert, aber in den Kram passt es Adobe dennoch nicht. Mit der Konsequenz freilich, dass Microsoft nun versucht mit XPS ein Konkurrenzprodukt anzubieten. Und dabei bemühen sich beide Anbieter in letzter Zeit in einer Weise um Standards, ich traue meinen Ohren kaum. Und soll ich nun meinen Ohren oder doch lieber meinem Gedächtnis trauen?

Ein Anbieter, der mit jeder Office-Version automatisch ein neues Dokumentenformat einführte, will nun plötzlich verantwortlich zeichnen, dass es DAS standartisierte Format für Textdokumente gibt. Ich würde ihm entgegnen wollen, er soll doch bitte erst mal 10 Jahre das gleiche Dokumentenformat verwenden, und danach können wir die Sache in aller Ruhe nochmals überdenken.

Der andere Anbieter betont immer die ganze Zeit wie sehr doch PDF ein standardisiertes Format sei. So, und warum muss ich denn alle Wochen wieder eine neue Version des Acrobat-Readers herunterladen? Davon, dass die aktuelle PDF-Spezifikation mit einer älteren Version des Acrobat Readers nicht mehr gelesen werden kann, davon war schon die Rede. Dass dazu aber die ca. 16 MByte grosse Datei auch noch erst extrahiert und in 32 MByte umgewandelt werden muss, ehe ich sie ansehen kann, dass hab ich erst feststellen dürfen, nachdem ich die neuste Version des Acrobat Readers installiert habe; soviel zur Effizienz von PDF-Dateien.

OpenDokument-Format und XML

Was heisst das für die Archivierung, für ein langfristig ausgerichtetes DMS? *Ehrlich gesagt kann die offene Antwort nur lauten, wir verzichten auf beide Formate, denn weder Office- noch die PDF-Dokumente eignen sich für die Archivierung.* Gibt es Alternativen? Im Prinzip schon, mit dem OpenDocument-Format stünde ein Format zur Verfügung, dass wenigstens offen dokumentiert ist.

Doch auch hier gilt, genauer hinsehen lohnt sich. Eine OpenDokument-Datei ist **(gemäss wikipedia)** *entweder eine XML-Datei oder eine Sammlung verschiedener XML-Dateien und anderer Objekte (z.B. eingebundene Bilder), die zu einer komprimierten Datei (z.B. Zip-Datei) zusammengefasst werden.* Ehrlich gesagt habe ich etwas Mühe mit der Definition. Warum entweder oder und weshalb oder anderer Objekte.

Ein Standard sollte doch irgendwo ein Anfang und Ende haben, sonst nützt der Standard wenig bzw. ist kein Standard mehr. Und wie ist das noch mit XML-Dateien? Kurz zur Erinnerung, *XML steht für Extensible Markup Language (erweiterbare Auszeichnungssprache).* Damit ist meines Erachtens bereits auch gesagt, dass XML für die Archivierung nicht wirklich was taugt, denn XML bedeutet ja eben gerade, dass das Format beliebig erweitert werden kann. Und *beliebige Erweiterbarkeit wird früher oder später zu Problemen führen müssen.*

Ich würde es so formulieren wollen: Mit XML-Formaten ist es den Programmen zwar grundsätzlich möglich zu erkennen, das ist ein Satz und das ist ein Wort und wenn es gut geht, wissen beide Programme auch, wir sollten uns in der Sprache XY unterhalten, aber ob sie den gleichen Wortschatz haben (mitunter das Gleiche so aufbereiten können, dass der Mensch die gleiche Information kriegt), diese Aufgabe vermag XML nicht zu erfüllen. Und schon gar nicht, wenn die "grossen" Anbieter XML als Dokumentenstandard in je einer anderen Ausrichtung proklamieren. Und wer das nicht glaubt, soll einmal eine XML-Datei auf nicht lesbare Zeichen durchforsten; es ist die reinste Schande, hat aber auch damit zu tun, dass XML als Textdatei nicht wirklich für grössere Daten (z.B. Bilder) geeignet ist. *Folglich ist das OpenDokument-Format zwar ein guter Ansatz, aber kein für die Archivierung wirklich tauglicher.*

Und was sagen die Spzeialist/Innen auf dem Gebiet?

Der **Verein Schweizerischer Archivarinnen und Archivare** beschäftigt sich auf der Hauptseite mit der **Konservierung der Mikrofilme.** Richtig, den Mikrofilmen hat man eine Lebensdauer von 500 Jahren vorausgesagt. Nach einigen Jahrzehnten sind sie allesamt Sanierungsfälle. Und was sagt der Verein heute zur Langzeitarchivierung. Die **Archivierungsempfehlungen** werden als Word-Dokument veröffentlicht (immerhin ist das Inhaltsverzeichnis als HTML-Datei abrufbar). Dann viel Spass beim Lesen in 10 Jahren. Damit sei explizit gesagt, dass der Verein sehr wohl einen wichtigen Beitrag zur Sicherstellung von wichtigen Daten leistet (viele Informationen sind nach wie vor nur auf Mikrofilm verfügbar), dass er aber nicht unbedingt einen Beitrag zur gegenwärtigen Problematik leistet, wenn er selber Word-Dateien zum Download anbietet.

Die wichtigsten zehn Eckpunkte für langfristige Datenbestände

Erwarten Sie nun bitte nicht, dass ich sage, mit der ArchivistaBox wäre das alles nicht passiert. So einfach ist es leider nicht. Und doch kümmern wir uns bei Archivista seit ca. 15 Jahren um die Archivierung digitaler Daten mit den folgenden zehn Eckpunkten:

- Jede Archivista-Lösung besteht aus einem oder mehreren Archiven.
- Ein Archiv besteht aus Akten und Seiten, wobei eine Akte eine oder mehrere Seiten hat.
- Die Archive werden mit Datenbanken verwaltet, welche SQL-Abfragen verstehen.
- Jede Akte hat eine eindeutige Identifikationsnummer (ID).

- Jede Akte kann mehrere Beschlagwortungsfelder (z.B. Datum, Titel) haben.
- Der Inhalt der Akten und Seiten kann über Volltextrecherchen abgerufen werden.
- Die Recherche ist beliebig kombinierbar (virtuelle Archivmappen anstatt starre Ordner).
- Die Anzeige erfolgt mit quelloffenen Bitmat-Dateien und herstellerunabhängig.
- Archive sind nicht an einen bestimmten Datenträger bzw. Technologie gebunden.
- Erstellte Datenträger (ISO 9660) können mit jedem Betriebssystem gelesen werden.

In den letzten Jahren haben wir diese Grundsätze verfeinert:

- Wir verwenden nur noch Software, die quelloffen vorliegt (OpenSource).
- Die gesamte Lösung wird IP-basierend betrieben (ArchivistaBox-Konzept).
- Vertrauliche Dokumente können verschlüsselt übermittelt werden.
- Das Erstellen der Datenträger bzw. des Backups erfolgt automatisiert.
- Jedes Archiv ist beliebig duplizierbar (Klonen von Archiven im laufenden Betrieb).

Seit mehr als 12 Jahren keinen Datenträger konvertiert

Die Bilanz dieser Eckpunkte ist erstaunlich: Seit mehr als 12 Jahren haben wir keinen Datenträger mehr konvertiert. Die einzige Konvertierung wurde notwendig, weil wir 1994 das Gif-Format einführten, um 1996 feststellen zu müssen, dass die Firma Unisys plötzlich hohe Patentgebühren einfordert. Nebenbei erwähnt, dass war zwei Jahre, bevor wir das Produkt Archivista überhaupt verkauft haben.

Im Maximum 10 Milliarden Seiten und 0,4 Cents Kosten pro Seite/Jahr.

Wir haben mit kleineren Archiven (einige tausend Belege begonnen). Technisch ist es uns möglich, ca. 150 Millionen Akten pro Archiv bzw. ca. 10 Milliarden Seiten pro Archiv zu verwalten. Derzeit haben unsere Kunden Archive mit einem höheren Millionenbestand an Seiten am Laufen. Auf den derzeitigen *ArchivistaBoxen (Typ Eiger) können redundant für weniger als 5000 Euro ca. 5 Mio Seiten verwaltet werden.* Die jährliche Wartungsgebühr beträgt (inkl. Ersatz der Hardware) 15 Prozent, dies ergibt über 10 Jahre weitere 7500 Euro an Kosten. Dies ergibt einen Seitenpreis von geradeamal 4 Cents für die Wartung einer Seite über 10 Jahre bzw. 0,4 Cents pro Jahr. Nicht eingerechnet sind hier Betriebskosten (insbesondere Stromkosten), diese dürften aber beim Betrieb der stromsparenden ArchivistaBoxen (Embedded-Produkte) bei einem zweistelligen Euro-Betrag pro Jahr liegen und dürften den Seitenpreis nur im Promillebereich beeinflussen.

Mindestens 30 Jahre Garantie auf Lesbarkeit der Daten

Wir garantieren für sämtliche mit der ArchivistaBox *archivierten Informationen 30 Jahre Investitionsschutz auf die Bildinformationen (Seiten) und Datenstrukturen (Akten),* d.h. wenn Sie heute eine ArchivistaBox erwerben, können diese Daten auch in 30 Jahren

noch mit jedem handelsüblichen Rechner gelesen werden. Wir garantieren weiter, dass ihre Daten keine ausführbaren Datenprogramme enthalten, d.h. dass die Datenstrukturen zu 100 Prozent von den Programmen der ArchivistaBox getrennt verwaltet werden. Mit all diesen Punkten und viel Engagement tragen wir dazu bei, Informationen langfristig verfügbar zu halten.

Fazit: Ja, es geht…

Seit nunmehr 15 Jahren beschäftigen wir uns mit der Archivierung digitaler Daten. Seit 10 Jahren profitieren unsere Kunden von unseren Lösungen. Seit mehr als zwei Jahren kann die ArchivistaBox komplett als OpenSource betrieben werden. Die Kosten für den Betrieb eines digitalen Archivs sind — so denn die Lösung richtig implementiert wird — minim. Und noch etwas, wer glaubt, alles extern in Auftrag vergeben zu können und vom Wissen der externen Fachkräfte profitieren zu können, der läuft immer Gefahr, dass diese externe Fachkraft auch falsch liegt. In diesem Sinne laden wir herzlich zum **Download der ArchivistaBox** ein.

Archivista GmbH, Urs Pfister