nestor
Kompetenznetzwerk Langzeitarchivierung

nestor Handbuch:

Eine kleine Enzyklopädie der digitalen Langzeitarchivierung

15.1 Textdokumente

Herausgeber:

Heike Neuroth
Hans Liegmann †
Achim Oßwald
Regine Scheffel
Mathias Jehn
Stefan Strathmann

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Im Auftrag von:

nestor – Kompetenznetzwerk Langzeitarchivierung und Langzeitverfügbarkeit digitaler Ressourcen für Deutschland
nestor – Network of Expertise in Long-Term Storage of Digital Resources
http://www.langzeitarchivierung.de

Kontakt:

Niedersächsische Staats- und Universitätsbibliothek Göttingen
Dr. Heike Neuroth
Forschung und Entwicklung
Papendiek 14
37073 Göttingen
neuroth@sub.uni-goettingen.de
Tel. +49 (0) 55 1 39 38 66

15.1 Textdokumente

Karsten Huth

Definition

Die Definition des Begriffs Textdokument im Bereich der Langzeitarchivierung bzw. die Antwort auf die Frage: "Was ist ein Textdokument?", ist nicht einfach zu beantworten. Kommen doch zwei Ebenen eines digitalen Objekts für eine Definitionsgrundlage in Frage[1]. Auf der konzeptuellen Ebene liegt ein Textdokument genau dann vor, wenn das menschliche Auge Text erkennen, lesen und interpretieren kann. Diese Anforderung kann auch eine Fotografie, bzw. das Bild eines Textes erfüllen. Auf der logischen Ebene eines digitalen Objektes, der Ebene der binären Codierung und Decodierung liegt ein Textdokument genau dann vor, wenn innerhalb des Codes auch Textzeichen codiert sind und dadurch Gegenstand von Operationen werden (z.B. Kopieren und Verschieben, Suchen nach bestimmten Worten und Wortfolgen, Ersetzen von bestimmten Zeichenfolgen usw.).

Da ein Archiv seine Archivobjekte generell auf der konzeptuellen Ebene betrachten muss, insbesondere da sich die technikabhängige logische Ebene im Laufe der Zeit durch Migration grundsätzlich ändert[2], soll für dieses Kapitel die erste Definition zur Anwendung kommen:

> *Ein Textdokument liegt genau dann vor, wenn das menschliche Auge Text erkennen, lesen und interpretieren kann.*

Diese Definition ermöglicht die Verwendung von Dateiformaten zur Speicherung von Bildinformationen ebenso wie die speziell auf Textverarbeitung ausgerichteten Formate. Welchen Formattyp ein Archiv zur Speicherung wählt, hängt von den wesentlichen Eigenschaften des Archivobjekts ab. Die wesentlichen Eigenschaften eines digitalen Archivobjekts müssen vom Archiv bei oder bereits vor der Übernahme des Objekts in das Archiv festgelegt werden und ergeben sich gemäß den Vorgaben des OAIS größtenteils aus den Ansprüchen und Möglichkeiten der Archivnutzer.[3]

---

1 vgl. Funk, Stefan, Kap 9.1 Digitale Objekte

2 vgl. Funk, Stefan, Kap 12.2 Migration

3 Consaltative Committee for Space Data Systems (Hrsg.) (2002): *Reference Model for an Open Archive Information Sytem: Blue Book*. Washington, DC. Page 3-4

Archivierung von Textdokumenten mit Bildformaten:

Die Archivierung von Textdokumenten in Bildformaten empfiehlt sich genau dann, wenn der optische Eindruck eines Textdokuments eine der wesentlichen Eigenschaften des Archivobjekts ist, welches auf das Genaueste erhalten werden muss. Solche Fälle ergeben sich z.B. bei der Digitalisierung von amtlichem Schriftgut, bei der anschließend das Original aus Papier vernichtet wird, während man die digitale Fassung archiviert. Da bei diesen Objekten das originale Schriftbild sowie von Hand aufgetragene Zeichen (z.B. Anmerkungen, Unterschriften und Paraphen) für die dauerhafte korrekte Interpretation des Archivobjektes unbedingt erhalten werden müssen, ist die Speicherung als Bild der beste Weg. In einem Bildformat sind in der Regel nur Informationen über Bildpunkte und ihre jeweiligen Farb- und Helligkeitswerte in einem Raster verzeichnet (Bitmap-Grafik). Diese Formate beinhalten von sich aus keinerlei Informationen über den abgebildeten Text. Deshalb kann man in einer solchen Datei nicht nach bestimmten Textstellen suchen, Textinhalte herauskopieren oder verschieben. Die Unveränderlichkeit der inhaltlichen und optischen Darstellung ist für ein Archiv von Vorteil.

Eine Abhandlung zu möglichen Bildformaten im Archiv befindet sich im Kapitel 15.2 „Bilddokumente“.[4] Bildformate werden in diesem Kapitel nicht weiter thematisiert.

Archivierung von Textdokumenten mit Textformaten:

Die Archivierung von Textdokumenten in Textformaten empfiehlt sich genau dann, wenn die Erhaltung der Textinformation des Objektes im Vordergrund steht. Bei der Archivierung von Textformaten sind grundsätzliche technische Abhängigkeiten zu beachten.

- Abhängigkeit 1: der Zeichensatz (Character Set)

Einen Zeichensatz kann man sich als Tabelle vorstellen, in der ein numerischer einem Zeicheninhalt zugeordnet wird. Die Maschine nimmt den Wert der Zahl und sieht in der Zeichensatztabelle an der entsprechenden Stelle nach, in welches Zeichen die Zahl decodiert werden muss. Dieser Vorgang hat noch nichts mit der Darstellung eines Zeichens auf dem Bildschirm oder beim Druckvor-

4 für eine kurze Übersicht über Bildformate s. Rohde-Enslin, Stefan (2004): *nestor - Ratgeber - Nicht von Dauer: Kleiner Ratgeber für die Bewahrung digitaler Daten in Museen*. Berlin: nestor, IfM . S. 12ff : urn:nbn:de:0008-20041103017

gang zu tun.[5]
Beispiel: Beim ASCII Zeichencode entspricht der Wert 65 (binär 01000001) dem Zeichen „A“.

- Abhängigkeit 2: Schriften (Font)

Fonts geben den Zeichen eine Gestallt auf dem Bildschirm oder beim Druck. Dem Zeichen eines Zeichensatzes ist innerhalb eines Fonts ein Bild (oder mehrere Bilder) zugeordnet. Bekannte Schrifttypen sind z.B. Arial, Times New Roman usw.

Die korrekte Darstellung eines Textes ergibt sich demnach aus einer Kette von Abhängigkeiten. Um ein Textdokument mitsamt dem Schriftbild (d.h. Formatierungen, Absätze und Font) erhalten zu können, benötigt ein Archiv den korrekten Zeichensatz und den korrekten Font. Dies ist ein Problem für den dauerhaften Erhalt, denn die meisten Dateiformate, die im Bereich der Textverarbeitung verwendet werden, sind von Zeichensätzen und Fonts abhängig, die außerhalb der Textdatei gespeichert werden. Insbesondere die Zeichensätze sind oft ein Teil des Betriebssystems. Das Textverarbeitungsprogramm leistet die Verknüpfung von Code-Zeichen-Schriftzeichen und sorgt für die korrekte Darstellung des Textdokuments.

Konsequenzen für das Archiv

Für die langfristige Darstellbarkeit eines Textes muss das Archiv zumindest den oder die verwendeten Zeichensätze kennen. Die Informationen über die Zeichensätze sollten mit Bezug auf die jeweiligen Dateien in den Metadaten des Archivs fest verzeichnet sein.

Bei Neuzugängen sollte das Archiv Dateiformate wählen, die weit verbreitete und standardisierte Character Sets unterstützen. Der älteste (seit 1963) Zeichensatz ASCII kann beinahe auf allen Plattformen decodiert und dargestellt werden. Leider wurde dieser Zeichensatz allein für den amerikanischen Raum entwickelt, so dass er keinerlei Umlaute und kein „ß“ enthält. Damit ist ASCII für deutsche Textdokumente nicht ausreichend. Für Archive besonders zu

5 für eine gelungene Einführung in das Gebiet der Zeichensätze s. Constable, Peter (2001): *Character set encoding basics. Understanding character set encodings and legacy encodings.* In: Implementing Writing Systems: An introduction. 13.06.2001. <http://scripts.sil.org/IWS-Chapter03> (Abrufdatum: 12.12.2007)

empfehlen sind Dateiformate, die Unicode[6], speziell UTF-8 (Unicode encoding Form neben UTF-16 und UTF-32) unterstützen. UTF-8 ist auch der empfohlene Zeichensatz für Dokumente im HTML, XML oder SGML-Format. Weit verbreitet und für Archive geeignet ist der Zeichensatz „Latin-1, Westeuropäisch" ISO 8859-1, der auch ASCII-Texte darstellen kann.

Die gewissenhafte Dokumentation der verwendeten Zeichensätze sollte ein Archiv zumindest vor dem Verlust der reinen Textinformation bewahren. Um auch die ursprüngliche optische Form zu erhalten, sollten die technischen Informationen über die verwendeten Schriftsätze (Fonts) ebenso vom Archiv in den Metadaten nachgewiesen werden.

Bei bereits bestehenden Beständen an Textdokumenten, sollte mit geeigneten technischen Werkzeugen der zugrundeliegende Zeichensatz ermittelt werden. Sollte der ermittelte Zeichensatz nicht den oben erwähnten weit verbreiteten Standards entsprechen, empfiehlt sich auf lange Sicht wahrscheinlich eine Migration, vorrausgesetzt die geeigneten technischen Werkzeuge sind im Archiv vorhanden.

Besonders geeignete Dateiformate für Archive

Da das Archiv alle Informationen über die verwendeten Zeichensätze und Fonts sammeln und erschließen muss, sollten nur Dateiformate verwendet werden, aus denen diese Informationen auch gewonnen werden können. Dies ist bei Dateiformaten der Fall, wenn ihr technischer Aufbau öffentlich (entweder durch Normung oder Open Source) beschrieben ist. Ein Archiv sollte Textformate meiden, deren technischer Aufbau nicht veröffentlicht wurde (proprietäre Formate), da dann der Zugriff auf die für die Langzeitarchivierung wichtigen technischen Informationen kompliziert ist.

Ein Beispiel für ein offenes Dokumentformat ist das „Open Document Format" (ODF). Der gesamte Aufbau einer ODF-Datei ist öffentlich dokumentiert. Eine Datei besteht im wesentlichen aus mehreren komprimierten XML-Dateien, die alle mit dem Zeichensatz UTF-8 gespeichert wurden. Die von ODF-Dateien verwendeten Schriftsätze sind kompatibel zu UTF-8 und in den XML-Dateien angegeben. Sollte eine ODF-Textdatei im Archiv mit den vorhandenen technischen Mitteln nicht mehr darstellbar sein, dann kann min-

6 Whistler, Ken/ Davis, Mark/ Freytag, Asmus (2004): *Unicode Technical Report #17. Character Encoding Model. Revision 5*. In: Unicode Technical Reports. 09.09.2004. < http://www.unicode.org/reports/tr17/> (Abrufdatum: 12.12.2007)

destens der Textinhalt und die Struktur des Dokuments über die zugrundeliegenden XML-Dateien zurückgewonnen werden.

Ein Textformat, das speziell für die Archivierung entwickelt wurde, ist das PDF/A-Format. Das Dateiformat wurde so konzipiert, dass Zeichensatz und die verwendeten Fonds in der jeweiligen Datei gespeichert werden. Ein Textdokument im PDF/A Format ist somit unabhängiger von der jeweiligen Plattform, auf der es dargestellt werden soll.