

H. Neuroth, A. Oßwald, R. Scheffel, S. Strathmann, M. Jehn (Hrsg.)

# nestor Handbuch

Eine kleine Enzyklopädie  
der digitalen Langzeitarchivierung

Version 2.0

Kapitel 17.2  
Textdokumente

nestor 

vwth

nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung  
hg. v. H. Neuroth, A. Oßwald, R. Scheffel, S. Strathmann, M. Jehn  
im Rahmen des Projektes: 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: [editors@langzeitarchivierung.de](mailto:editors@langzeitarchivierung.de)  
c/o Niedersächsische Staats- und Universitätsbibliothek Göttingen,  
Dr. Heike Neuroth, Forschung und Entwicklung, Papendiek 14, 37073 Göttingen

Die Herausgeber danken Anke Herr (Korrektur), Martina Kerzel (Bildbearbeitung) und  
Jörn Tietgen (Layout und Formatierung des Gesamttextes) für ihre unverzichtbare  
Unterstützung bei der Fertigstellung des Handbuchs.

Bibliografische Information der Deutschen Nationalbibliothek  
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen  
Nationalbibliografie; detaillierte bibliografische Daten sind im Internet unter  
<http://www.d-nb.de/> abrufbar.

Die Inhalte dieses Buchs stehen auch als Onlineversion  
(<http://nestor.sub.uni-goettingen.de/handbuch/>)  
sowie über den Göttinger Universitätskatalog (<http://www.sub.uni-goettingen.de>) zur  
Verfügung.

Die digitale Version 2.0 steht unter folgender Creative-Commons-Lizenz:  
„Attribution-Noncommercial-Share Alike 3.0 Unported“  
<http://creativecommons.org/licenses/by-nc-sa/3.0/>



Einfache Nutzungsrechte liegen beim Verlag Werner Hülsbusch, Boizenburg.  
© Verlag Werner Hülsbusch, Boizenburg, 2009  
[www.vwh-verlag.de](http://www.vwh-verlag.de)  
In Kooperation mit dem Universitätsverlag Göttingen

Markenerklärung: Die in diesem Werk wiedergegebenen Gebrauchsnamen, Handelsnamen,  
Warenzeichen usw. können auch ohne besondere Kennzeichnung geschützte Marken sein und  
als solche den gesetzlichen Bestimmungen unterliegen.

Druck und Bindung: Kunsthaus Schwanheide

Printed in Germany – Als Typoskript gedruckt –

ISBN: 978-3-940317-48-3

URL für Kapitel 17.2 „Textdokumente“ (Version 2.0): [urn:nbn:de:0008-20090811846](http://nbn-resolving.de/urn/resolver.pl?urn=urn:nbn:de:0008-20090811846)  
<http://nbn-resolving.de/urn/resolver.pl?urn=urn:nbn:de:0008-20090811846>



*Gewidmet der Erinnerung an Hans Liegmann (†), der als Mitinitiator und früherer Herausgeber des Handbuchs ganz wesentlich an dessen Entstehung beteiligt war.*

## 17.2 Textdokumente

Karsten Huth

*Der langfristige Erhalt von Textdokumenten ist nur scheinbar einfacher als der Erhalt von anderen digitalen Objekten. In digitalen Textdokumenten vereint sich Fachwissen aus der Kunst des Buchdrucks, der Linguistik und der Semiotik. Nur der Aspekt, dass wir vom frühen Kindesalter ans Lesen, Schreiben und Textverstehen herangeführt werden, sodass Texte wie selbstverständlich unseren Alltag prägen, lassen uns die komplexen Kenntnisse und kognitiven Fähigkeiten, die der Mensch dafür benötigt, kaum mehr wahrnehmen. Das optische Erscheinungsbild eines digitalen Textes besteht auf der Datenebene aus zwei wesentlichen Komponenten. Die Wichtigste ist der Zeichensatz, über den numerische Werte Textzeichen zugewiesen werden. Die zweite Komponente ist der Font, d.h. ein kompletter Satz von Bildern der benötigten Schriftzeichen. Dieses Kapitel klärt in einführender Weise über die Abhängigkeiten dieser Komponenten auf.*

### 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<sup>1</sup>. 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,<sup>2</sup> 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.*

---

1 Vgl. Funk, Stefan, Kap 7.2 Digitale Objekte und Formate

2 Vgl. Funk, Stefan, Kap 8.3 Migration

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 Open Archive Information System (OAIS) größtenteils aus den Ansprüchen und Möglichkeiten der Archivnutzer.<sup>3</sup>

### **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“.<sup>4</sup> Bildformate werden in diesem Kapitel nicht weiter thematisiert.

---

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

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

## 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 Druckvorgang zu tun.<sup>5</sup>  
Beispiel: Beim American Standard Code for Information Interchange (ASCII) Zeichencode entspricht der Wert 65 (binär 01000001) dem Zeichen „A“.
- Abhängigkeit 2: Schriften (Font)  
Fonts geben den Zeichen eine Gestalt 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.

---

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>

## 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 empfehlen sind Dateiformate, die Unicode,<sup>6</sup> speziell Unicode Transformation Format (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, vorausgesetzt 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

---

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/>

Alle hier aufgeführten URLs wurden im April 2009 auf Erreichbarkeit geprüft .

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 mindestens 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 Fonts 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.

## Literatur

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