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

nestor Handbuch

Eine kleine Enzyklopädie der digitalen Langzeitarchivierung

Version 2.0

Kapitel 7.2
Digitale Objekte und Formate

nestor

vwh

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
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
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 7.2 „Digitale Objekte und Formate“ (Version 2.0):
urn:nbn:de:0008-20090811338
http://nbn-resolving.de/urn/resolver.pl?urn=urn:nbn:de:0008-20090811338



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

7.2 Digitale Objekte und Formate

Stefan E. Funk

Digitale Objekte

Die erste Frage, die im Zusammenhang mit der digitalen Langzeitarchivierung gestellt werden muss, ist sicherlich die nach den zu archivierenden Objekten. Welche Objekte möchte ich archivieren? Eine einfache Antwort lautet hier zunächst: digitale Objekte!

Eine Antwort auf die naheliegende Frage, was denn digitale Objekte eigentlich sind, gibt die Definition zum Begriff „digitales Objekt" aus dem Open Archival Information System (OAIS). Dieser Standard beschreibt ganz allgemein ein Archivsystem mit dessen benötigten Komponenten und deren Kommunikation untereinander, wie auch die Kommunikation vom und zum Nutzer. Ein digitales Objekt wird dort definiert als

An object composed of a set of bit sequences

(CCSDS 2001), also als ein aus einer Reihe von Bit-Sequenzen zusammengesetztes Objekt. Somit kann all das als ein digitales Objekt bezeichnet werden, das mit Hilfe eines Computers gespeichert und verarbeitet werden kann. Und dies entspricht tatsächlich der Menge der Materialien, die langzeitarchiviert werden sollen, vom einfachen Textdokument im .txt-Format über umfangreiche PDF-Dateien mit eingebetteten Multimedia-Dateien bis hin zu kompletten Betriebssystemen. Ein digitales Objekt kann beispielsweise eine Datei in einem spezifischen Dateiformat sein, zum Beispiel eine einzelne Grafik, ein Word-Dokument oder eine PDF-Datei. Als ein digitales Objekt können allerdings auch komplexere Objekte bezeichnet werden wie Anwendungsprogramme (beispipelsweise Microsoft Word und Mozilla Firefox), eine komplette Internet-Seite mit all ihren Texten, Grafiken und Videos, eine durchsuchbare Datenbank auf CD inklusive einer Suchoberfläche oder gar ein Betriebssystem wie Linux, Mac OS oder Windows.

Ein digitales Objektes kann auf drei Ebenen beschrieben werden, als *physisches Objekt*, als *logisches Objekt* und schließlich als *konzeptuelles Objekt*.

Als *physisches Objekt* sieht man die Menge der Zeichen an, die auf einem Informationsträger gespeichert sind – die rohe Manifestation der Daten auf dem Speichermedium. Die Art und Weise der physischen Beschaffenheit dieser Zeichen kann aufgrund der unterschiedlichen Beschaffenheit des Trägers

sehr unterschiedlich sein. Auf einer CD-ROM sind es die sogenannten „Pits“ und „Lands“ auf der Trägeroberfläche, bei magnetischen Datenträgern sind es Übergänge zwischen magnetisierten und nicht magnetisierten Teilchen. Auf der physischen Ebene haben die Bits keine weitere Bedeutung außer eben der, dass sie binär codierte Information enthalten, also entweder die „0“ oder die „1“. Auf dieser Ebene unterscheiden sich beispielsweise Bits, die zu einem Text gehören, in keiner Weise von Bits, die Teil eines Computerprogramms oder Teil einer Grafik sind.

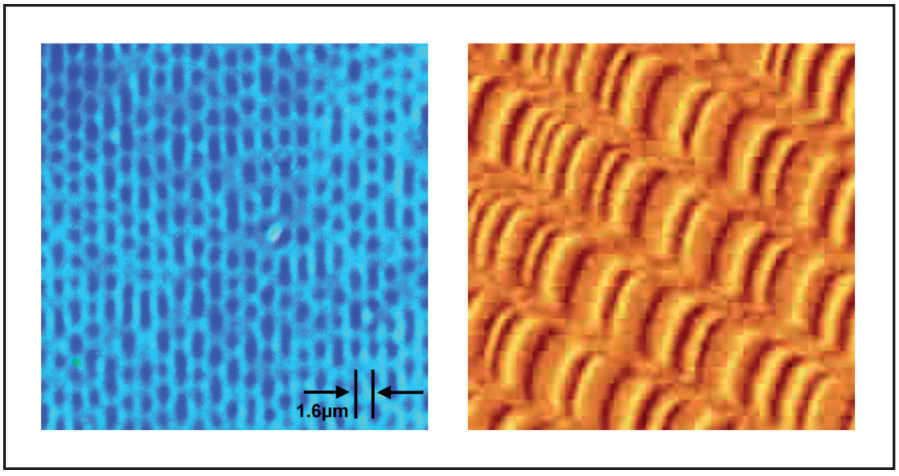


Abbildung 1: Das physische Objekt: „Nullen“ und „Einsen“ auf der Oberfläche einer CD-Rom (blau) und einer Festplatte (gelb) [1].

Die Erhaltung dieses Bitstreams (auch Bitstreamerhaltung) ist der erste Schritt zur Konservierung des gesamten digitalen Objekts, er bildet sozusagen die Grundlage aller weiteren Erhaltungs-Strategien.
Unter einem *logischen Objekt* versteht man eine Folge von Bits, die von einem Informationsträger gelesen und als eine Einheit angesehen werden kann. Diese können von einer entsprechenden Software als Format erkannt und verarbeitet werden. In dieser Ebene existiert das Objekt nicht nur als Bitstream, es hat bereits ein definiertes Format. Die Bitstreams sind auf dieser Ebene schon sehr viel spezieller als die Bits auf dem physischen Speichermedium. So müssen diese zunächst von dem Programm, das einen solchen Bitstream zum Beispiel

1 Bildquelle CD-Rom-Oberfläche: http://de.wikipedia.org/wiki/Datei:Compactdiscar.jpg, Bildquelle Festplatten-Oberfläche: http://leifi.physik.uni-muenchen.de/web_ph10/umwelt-technik/11festplatte/festplatte.htm
Alle hier aufgeführten URLs wurden im April 2009 auf Erreichbarkeit geprüft .

als eine Textdatei erkennen soll, als eine solche identifizieren. Erst wenn der Bitstream als korrekte Textdatei erkannt worden ist, kann er vom Programm als Dateiformat interpretiert werden.

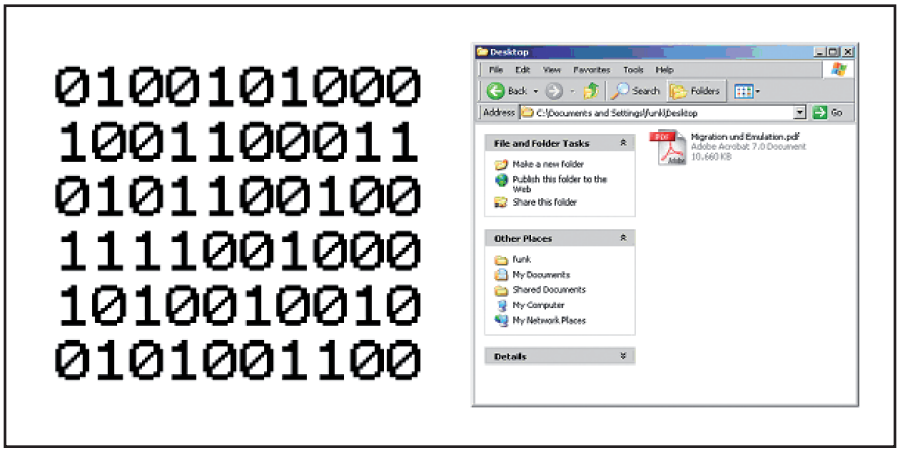


Abbildung 2: Das logische Objekt: Eine Bit-Folge als Repräsentation eines PDF-Dokuments

Will man diesen logischen Einheiten ihren Inhalt entlocken, muss das Format dieser Einheit genau bekannt sein. Ist ein Format nicht hinreichend bekannt oder existiert die zu dem Format gehörige Software nicht mehr, so wird die ursprüngliche Information des logischen Objektes sehr wahrscheinlich nicht mehr vollständig zu rekonstruieren sein. Um solche Verluste zu vermeiden, gibt es verschiedene Lösungsansätze, zwei davon sind Migration und Emulation.
Das *konzeptuelle Objekt* beschreibt zu guter Letzt die gesamte Funktionalität, die dem Benutzer des digitalen Objekts mit Hilfe von dazu passender Soft- und Hardware zur Verfügung steht – es ist das Objekt „zum Begreifen". Dies sind zunächst die Objekte, Zeichen und Töne, die der Mensch über seine Sinne wahrnimmt. Auch interaktive Dinge wie das Spielen eines Computerspiels oder eine durchsuchbare Datenbank zählen dazu, denn die Funktion eines Computerspiels ist es, gespielt werden zu können. Ein weiteres Beispiel ist eine komplexe Textdatei mit all ihren Editierungsmöglichkeiten, Tabellen und enthaltenen Bildern, die das verarbeitende Programm bietet.

Dieses konzeptuelle Objekt ist also die eigentliche, für den Betrachter bedeutungsvolle Einheit, sei es ein Buch, ein Musikstück, ein Film, ein Computerprogramm oder ein Videospiel. Diese Einheit ist es, die der Nachwelt erhalten bleiben soll und die es mit Hilfe der digitalen Langzeitarchivierung zu schützen gilt.

Das Ziel eines Langzeitarchivs ist es also, das konzeptuelle Objekt zu archivieren und dem Nutzer auch in ferner Zukunft Zugriff auf dessen Inhalte zu gewähren. Die Darstellung bzw. Nutzung des digitalen Objekts soll so nahe wie möglich den Originalzustand des Objekts zur Zeit der Archivierung wiederspiegeln. Dies ist nicht möglich, wenn sich bereits Probleme bei der Archivierung auf den unteren Ebenen, der logischen und der physischen Ebene, ergeben. Gibt es eine unbeabsichtigte Veränderung des Bitstreams durch fehlerhafte Datenträger oder existiert eine bestimmte Software nicht mehr, die den Bitstream als Datei erkennt, ist auch eine Nutzung des Objekts auf konzeptueller Ebene nicht mehr möglich.

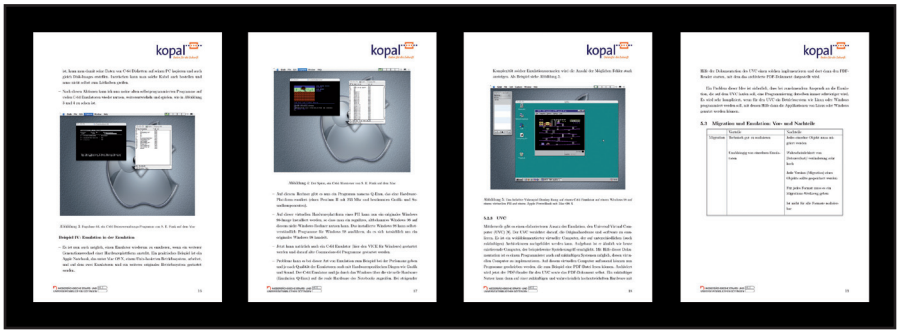


Abbildung 3: Das konzeptuelle Objekt: Die PDF-Datei mit allen ihren Anzeige- und Bearbeitungsmöglich-keiten

Formate

Ein Computer-Programm muss die Daten, die es verwaltet, als Bit-Folge auf einen dauerhaften Datenspeicher (zum Beispiel auf eine CD oder eine Festplatte) ablegen, damit sie auch nach Ausschalten des Computers sicher verwahrt sind. Sie können so später erneut in den Rechner geladen werden. Damit die gespeicherten Daten wieder genutzt werden können, ist es erforderlich, dass das ladende Programm die Bit-Folge exakt in der Weise interpretiert, wie es beim Speichern beabsichtigt war.

Um dies zu erreichen, müssen die Daten in einer Form vorliegen, die sowohl das speichernde als auch das ladende Programm gleichfalls „verstehen“ und interpretieren können. Ein Programm muss die Daten, die es verwaltet, in einem definierten *Dateiformat* speichern können. Dies bedeutet, alle zu speichernden Daten in eine genau definierte Ordnung zu bringen, um diese dann als eine Folge von Bits zu speichern, als sogenannten *Bitstream*. Die Bits, mit denen beispielsweise der Titel eines Dokuments gespeichert ist, müssen später auch wie-

der exakt von derselben Stelle und semantisch als Titel in das Programm geladen werden, damit das Dokument seine ursprüngliche Bedeutung behält. Somit muss das Programm das Format genau kennen und muss wissen, welche Bits des Bitstreams welche Bedeutung haben, um diese korrekt zu interpretieren und verarbeiten zu können.

Formate sind also wichtig, damit eine Bit-Folge semantisch korrekt ausgewertet werden kann. Sind zwei voneinander unabhängige Programme fähig, ihre Daten im selben Format zu speichern und wieder zu laden, ist ein gegenseitiger Datenaustausch möglich. Für die digitale Langzeitarchivierung sind Formate sehr relevant, weil hier zwischen dem Schreiben der Daten und dem Lesen eine lange Zeit vergehen kann. Die Gefahr von (semantischen) Datenverlusten ist daher sehr groß, denn ein Lesen der Daten ist nicht mehr möglich, wenn das Format nicht mehr interpretiert werden kann.

Eine *Format-Spezifikation* ist eine Beschreibung der Anordnung der Bits, das heißt eine Beschreibung, wie die Daten abgelegt und später interpretiert werden müssen, um das ursprüngliche Dokument zu erhalten. Grob kann zwischen proprietären und offenen *Dateiformaten* unterschieden werden. Bei proprietären Dateiformaten ist die Spezifikation oft nicht oder nicht hinreichend bekannt, bei offenen Formaten hingegen ist die Spezifikation frei zugänglich und oft gut dokumentiert. Aus einer Datei, deren Format und Spezifikation bekannt ist, kann die gespeicherte Information auch ohne das vielleicht nicht mehr verfügbare lesende Programm extrahiert werden, da mit der Spezifikation eine Anleitung zur semantischen Interpretation vorhanden ist.

Zum *Format-Standard* kann eine Format-Spezifikation dann werden, wenn sich das durch sie beschriebene Format weithin als einheitlich für eine bestimmte Nutzung durchgesetzt hat – auch und gerade gegenüber anderen Formaten – und es von vielen beachtet und genutzt wird. Ein solcher Vorgang kann entweder stillschweigend geschehen oder aber gezielt durch einen Normungsprozess herbeigeführt werden, indem eine möglichst breite Anwendergruppe solange an einer Spezifikation arbeitet, bis diese von allen Beteiligten akzeptiert wird und anwendbar erscheint. Als Ergebnis eines solchen Normungsprozesses wird die erarbeitete Format-Spezifikation als Norm von einer Behörde oder Organisation veröffentlicht und dokumentiert. Als Beispiel ist hier auf nationaler Ebene das Deutsches Institut für Normung e.V. (DIN) zu nennen, auf europäischer und internationaler Ebene das Europäisches Komitee für Normung (CEN) und die Internationale Organisation für Normung (ISO).

Literatur

Consultative Committee for Space Data Systems (2001): *Reference Model for an Open Archival Information System (OAIS)*, CCSDS 650.0-B-1, BLUE BOOK, http://public.ccsds.org/publications/archive/650x0b1.pdf

Huth, Karsten, Andreas Lange (2004): *Die Entwicklung neuer Strategien zur Bewahrung und Archivierung von digitalen Artefakten für das Computerspiele-Museum Berlin und das Digital Game Archive*, http://www.archimuse.com/publishing/ichim04/2758_HuthLange.pdf

Thibodeau, Kenneth (2002): Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years, In: *Council on Library and Information Resources: The State of Digital Preservation: An International Perspective,* http://www.clir.org/PUBS/reports/pub107/thibodeau.html

Abrams, Steffen, Sheila Morrissey, Tom Cramer (2008): *What? So what? The Next-Generation JHOVE2 Architecture for Format-Aware Characterization*, http://confluence.ucop.edu/download/attachments/3932229/Abrams_a70_pdf.pdf?version=1