H. Neuroth, A. Oßwald, R. Scheffel, S. Strathmann, K. Huth (Hrsg.)

nestor Handbuch

Eine kleine Enzyklopädie der digitalen Langzeitarchivierung

Version 2.3

Kapitel 7

Formate

nestor

nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung
hg. v. H. Neuroth, A. Oßwald, R. Scheffel, S. Strathmann, K. Huth
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

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.

Neben der Online Version 2.3 ist eine Printversion 2.0 beim Verlag Werner Hülsbusch, Boizenburg erschienen.

URL für Kapitel 7 „Formate“ (Version 2.3): urn:nbn:de:0008-2010062462
http://nbn-resolving.de/urn/resolver.pl?urn:nbn:de:0008-2010062462

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

7 Formate

7.1 Einführung

Jens Ludwig

Bereits in der alltäglichen Nutzung elektronischer Daten und Medien sind sich die meisten Nutzer der Existenz von Formaten und ihrer Schwierigkeiten bewusst. Es gehört zum digitalen Alltag, dass nicht jedes Videoformat mit jeder Software abspielbar ist, dass dasselbe Textverarbeitungsdokument manchmal von verschiedenen Programmen verschieden dargestellt wird und dass Programme im Speicherdialog eine Vielzahl von Formaten anbieten, von deren Vor- und Nachteilen man keine Ahnung hat. Für die langfristige Erhaltung von Informationen stellen sich diese Probleme in verschärfter Form. Formate sind ein wesentlicher Faktor für die Gefahr des technologischens Veraltens digitaler Informationen.

Dieses Kapitel soll dabei helfen, die wesentlichen Aspekte für den Umgang mit Formaten für die Langzeitarchivierung zu verstehen. In „Digitale Objekte und Formate“ werden dafür zuerst die begrifflichen Grundlagen gelegt: Was sind die digitalen Objekte, mit denen wir alltäglich umgehen, und welche Rol-

le spielen Formate? Der Abschnitt „Auswahlkriterien“ bietet Hilfestellung für eine der meist gestellten Fragen bezüglich der Langzeitarchivierung: Welches Format soll ich verwenden? Leider gibt es hier weder eine allgemeingültige Lösung, nicht ein Format, das alle anderen überflüssig macht, noch sind mit der sinnvollen Wahl eines Formates alle Aufgaben gelöst, die im Zusammenhang mit Formaten anfallen. „Formatcharakterisierung“ beschreibt zusammen mit den Aufgaben der Identifizierung von Formaten, der Validierung und der Extraktion von technischen Metadaten einige technische Werkzeuge, die dafür genutzt werden können. Den Abschluss bildet „File Format Registries“, das einige zentrale Verzeichnisse beschriebt, in denen Referenzinformationen über Formate gesammelt werden.

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.

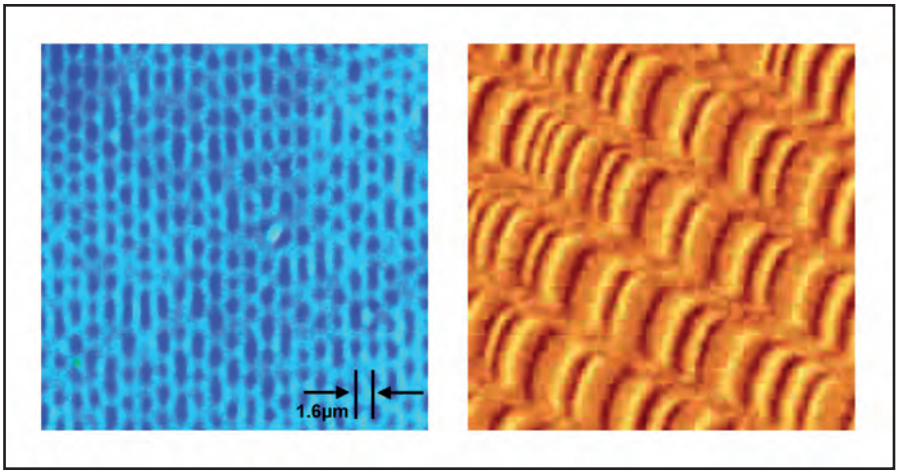

Abbildung 1: Das physische Objekt: „Nullen“ und „Einsen“ auf der Oberfläche einer CD-Rom (blau) und einer Festplatte (gelb) [1].

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 Mai 2010 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.

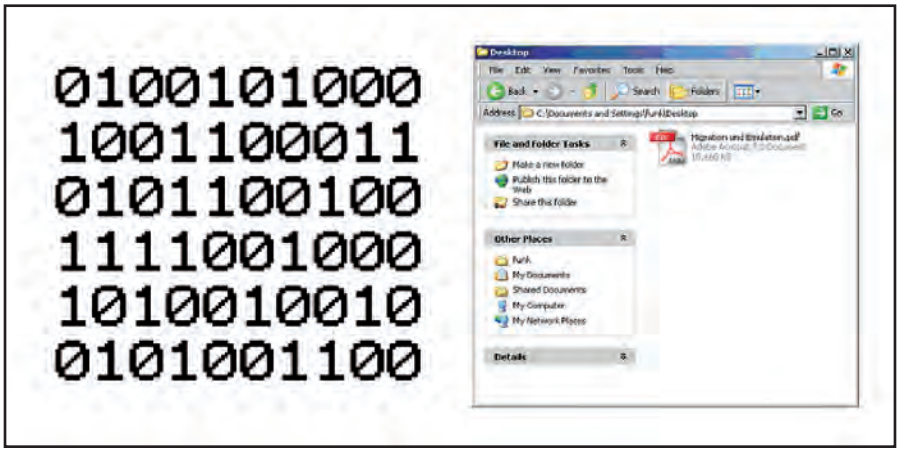

Abbildung 2: Das logische Objekt: Eine Bit-Folge als Repräsentation eines PDF-Dokuments

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

7.3 Auswahlkriterien

Jens Ludwig

Formate sind in unterschiedlichem Maße dem Risiko zu veralten ausgesetzt. Daher ist es naheliegend die langfristige Nutzbarkeit der digitalen Objekte durch die Verwendung eines geeigneten Formates zu unterstützen. Bevor man aber versucht zu beantworten, welches Format theoretisch am besten für die Langzeitarchivierung geeignet ist, muss man sich klarmachen, was die begrenzenden Faktoren der Formatwahl sind.

Die wichtigste und in gewissem Sinne triviale Einschränkung der Formatauswahl ist, dass ein Format auch die benötigte Funktionalität aufweisen muss. Es gibt Formate mit identischen Funktionen, die leicht durcheinander ersetzt werden können, aber genauso Formate für Spezialzwecke, die man dann leider nicht mit für die Langzeitarchivierung besser geeigneten austauschen kann, weil diese Spezialfunktionen eben benötigt werden. Um ein Format auswählen zu können, muss man sich also bewusst sein, was für Funktionalitäten benötigt werden.

In diesem Zusammenhang gilt es auch die Position des „Langzeitarchivierers" in der Verarbeitungskette zu berücksichtigen: Muss schon bei der Bearbeitung und Erstellung des digitalen Objekts das richtige Format ausgewählt werden, weil z.B. ein Dokument genauso wiederverwendet und bearbeitet werden soll? Dann muss man selbst der Ersteller sein oder hinreichenden Einfluss auf die Erstellung haben, sonst muss man hinnehmen, was man bekommt. Oder reicht ggf. eine statische Version, die nur den visuellen Eindruck erhält, und es ist deshalb möglich, das Objekt in ein neues, selbst ausgewähltes Format zu überführen?

Und selbst wenn die digitalen Objekte in den nach bestem Wissen und Gewissen ausgesuchten Formaten vorliegen, heißt das nicht, dass das Problem gelöst ist. Quasi jedes Format kann veralten, auch wenn es sich einmal als die beste Wahl dargestellt hat, Anforderungen können sich ändern und der technische Fortschritt kann neue Formate ermöglichen und erfordern. Aus all diesen Gründen kann man keine dauerhafte Festlegung auf ein bestimmtes Format treffen.

Kriterien

Trotz dieser Einschränkungen und Absicherungen lassen sich aber eine Reihe von allgemeinen Faktoren aufführen, was für Formate für digitale Objekte sinnvoll sind, die langfristig genutzt werden sollen. Und es gibt eine Vielzahl von Katalogen, die solche Faktoren aufführen: z.B. eher klassische Aufstellungen wie Lormant et al. (2005), Stanescu (2004) oder Arms, Fleischhauer (2005), deren Autoren auch die informative Seite der Library of Congress zum Thema Formate pflegen (Arms, Fleischhauer 2007), aber auch spezialisierte wie Barkstrom, Folk (2002), die besonders Erwägungen für naturwissenschaftliche Forschungsdaten berücksichtigen, oder Christensen et al. (2004), die Kriterien für Kontainerformate zur Internetarchivierung aufstellen. So interessant die unterschiedlichen Perspektiven und Kriterien im Detail sein mögen, auf einer abstrakten Ebenen lassen sich die Kriterien zusammenfassen. Angelehnt an Rog, van Wijk (2008) sind zu nennen:

- Offenheit: Ist die Spezifikation des Formates frei verfügbar oder ist sie ein Betriebsgeheimnis eines Herstellers? Ist das Format durch Normungsinstitutionen standardisiert worden? Mit der Spezifikation besteht die Möglichkeit, dass Format zu verstehen und ggf. selbst Nutzungssoftware zu entwickeln, auch wenn es keinen Anbieter mehr gibt.
- Verbreitung: Wie verbreitet ist das Format? Wie häufig wird es genutzt, wieviel unabhängige Anbieter von Nutzungssoftware gibt es? Eine hohe Verbreitung ist ein Indiz dafür, dass das Format noch lange und von vieler Software unterstützt wird, da ein großer Markt dafür vorhanden ist.
- Komplexität: Wie kompliziert ist das Format? Technische Komplexität erschwert die fehlerfreie Entschlüsselung bzw. Nutzung. Je mehr Wissen zum Verständnis eines Formates notwendig ist, desto eher kann ein Teil des notwendigen Wissens verloren gehen.
- Schutzmechanismen: Kopierschütze und Verschlüsselungen mögen für bestimmte Anwendungen sinnvoll sein, für die Langzeitarchivierung sind sie es nicht. Die langfristige Erhaltung setzt das Kopieren der digitalen Objekte voraus und eine Verschlüsselung erfordert als Minimum die zusätzliche Kenntnis des Schlüssels und Verschlüsselungsverfahrens.
- Selbstdokumentation: Wenn ein Format die Integration von Metadaten ermöglicht, dann erleichtert das voraussichtlich das Verständnis des digitalen Objekts und verringert die Abhängigkeit von externen Metadatenquellen.
- Robustheit: Je robuster ein Format ist, desto weniger wirken sich Veränderungen aus. Wie stark wirken sich Fehler einzelner Bits auf die Nutz-

barkeit des gesamten Objekts aus? Gibt es nur einen kleinen, vernachlässigbaren Darstellungsfehler oder lässt es sich ggf. überhaupt nicht mehr nutzen? Wie kompatibel sind verschiedene Versionen bzw. Weiterentwicklungen des Formats untereinander?

- Abhängigkeiten: Formate, die weniger von spezieller Hard- oder Software oder anderen Ressourcen (z.B. Internetzugang) abhängig sind als andere, sind zu bevorzugen.

Wie bereits erwähnt wurde, sind über diese Kritierien hinaus die spezifisch benötigten Funktionalitäten zu erwägen. Diese selbst nur für bestimmte Medientypen auszuführen, würden den Umfang dieses Kapitels sprengen. Gute weiterführende Quellen für bestimmte Medientypen sind neben dem Kapitel „Vorgehensweise für ausgewählte Objekttypen“ dieses Handbuchs auch Arms, Fleischhauer (2007) und AHDS (2006).

Literatur

AHDS (arts and humanities data service): Preservation Handbooks. 2006. http://ahds.ac.uk/preservation/ahds-preservation-documents.htm

Arms, Caroline/ Fleischhauer, Carl: *Digital Formats: Factors for Sustainability, Functionality, and Quality.* 2005. Paper for IS&T Archiving 2005 Conference, Washington, D.C. http://memory.loc.gov/ammem/techdocs/digform/Formats_IST05_paper.pdf

Arms, Caroline/ Fleischhauer, Carl (Hrsg.): Sustainability of Digital Formats. Planning for Library of Congress Collections. 2007. http://www.digitalpreservation.gov/formats/index.shtml

Barkstrom, Bruce R./ Folk, Mike: *Attributes of File Formats for Long Term Preservation of Scientific and Engineering Data in Digital Libraries.* 2002. http://www.ncsa.uiuc.edu/NARA/Sci_Formats_and_Archiving.doc

Christensen, Steen S. et al.: *Archival Data Format Requirements.* 2004. http://netarkivet.dk/publikationer/Archival_format_requirements-2004.pdf

Lormant, Nicolas et al.: *How to Evaluate the Ability of a File Format to Ensure Long-Term Preservation for Digital Information?* 2005. Paper for PV 2005, The Royal Society, Edinburgh. http://www.ukoln.ac.uk/events/pv-2005/pv-2005-final-papers/003.pdf

Rog, Judith/ van Wijk, Caroline: *Evaluating File Formats for Long-term Preservation.* 2008. http://www.kb.nl/hrd/dd/dd_links_en_publicaties/publicaties/KB_file_format_evaluation_method_27022008.pdf

Stanescu, Andreas: *Assessing the Durability of Formats in a Digital Preservation Environment.* In: D-Lib Magazine, November 2004, Volume 10 Number 11. doi:10.1045/november2004-stanescu

7.4 Formatcharakterisierung

Stefan E. Funk und Matthias Neubauer

Die Archivierung von digitalen Objekten steht und fällt mit der Charakterisierung und Validierung der verwendeten Dateiformate. Ohne die Information, wie die Nullen und Einsen des Bitstreams einer Datei zu interpretieren sind, ist der binäre Datenstrom schlicht unbrauchbar. Vergleichbar ist dies beispielsweise mit der Entzifferung alter Schriften und Sprachen, deren Syntax und Grammatik nicht mehr bekannt sind. Daher ist es für die digitale Langzeitarchivierung essentiell, die Dateien eines digitalen Objektes vor der Archivierung genauestens zu betrachten und zu kategorisieren.

Eine nach oben genannten Kriterien erfolgte Auswahl geeigneter Formate ist ein erster Schritt zu einer erfolgreichen Langzeitarchivierung. Eine automatisierte Charakterisierung der vorliegenden Formate ist ein weiterer Schritt. Die Speicherung der digitalen Objekte und deren Archivierung sollte unabhängig voneinander geschehen können, daher muss davon ausgegangen werden, dass außer dem zu archivierenden Objekt selbst keinerlei Daten zu dessen Format vorliegen.

Ziel einer Charakterisierung ist es, möglichst automatisiert das Format einer Datei zu identifizieren und durch Validierung zu kontrollieren, ob diese Datei auch deren Spezifikationen entspricht – bei einer sorgfältigen Auswahl des Formats ist diese ja bekannt. Eine einer Spezifikation entsprechende Datei kann später, beispielsweise für eine Format-Migration, nach dieser Spezifikation interpretiert werden und die Daten in ein aktuelleres Format umgewandelt werden. Außerdem sollen möglichst viele technische Daten über das Objekt (technische Metadaten) aus dem vorliegenden Objekt extrahiert werden, so dass eine Weiterverwendung auch in ferner Zukunft hoffentlich wahrscheinlich ist.

7.4.1 Identifizierung

Bei der *Identifizierung* eines digitalen Objekts geht es in erster Linie um die Frage, welches Format nun eigentlich vorliegt. Als Anhaltspunkte können zunächst interne oder externe Merkmale einer Datei herangezogen werden, zum Beispiel ein *HTTP Content-Type Header* oder ein *Mimetype* – zum Beispiel „text/xml“ für eine XML-Datei oder „application/pdf“ für eine PDF-Datei, die *Magic Number* oder als externes Merkmal eine *File Extension* (Dateiendung).

Die Dateiendung oder File Extension bezeichnet den Teil des Dateinamens, welcher rechts neben dem letzten Vorkommen eines Punkt-Zeichens liegt (wie

beispielsweise in „Datei.ext“). Dieses Merkmal ist jedoch meist nicht in einer Formatspezifikation festgelegt, sondern wird lediglich zur vereinfachten, oberflächlichen Erkennung und Eingruppierung von Dateien in Programmen und manchen Betriebssystemen genutzt. Vor allem aber kann die Dateiendung jederzeit frei geändert werden, was jedoch keinerlei Einfluss auf den Inhalt und damit auf das eigentliche Format der Datei hat. Daher ist es nicht ratsam, sich bei der Formaterkennung allein auf die Dateiendung zu verlassen, sondern in jedem Fall noch weitere Erkennungsmerkmale zu überprüfen, sofern dies möglich ist.

Einige Dateiformat-Spezifikationen definieren eine so genannte Magic Number. Dies ist ein Wert, welcher in einer Datei des entsprechenden Formats immer an einer in der Spezifikation bestimmten Stelle[2] der Binärdaten gesetzt sein muss. Anhand dieses Wertes kann zumindest sehr sicher angenommen werden, dass die fragliche Datei in einem dazu passenden Format vorliegt. Definiert ein Format keine Magic Number, kann meist nur durch den Versuch der Anwendung oder der Validierung der Datei des vermuteten Formats Klarheit darüber verschafft werden, ob die fragliche Datei tatsächlich in diesem Format abgespeichert wurde.

7.4.2 Validierung

Die *Validierung* oder auch Gültigkeitsprüfung ist ein wichtiger und notwendiger Schritt vor der Archivierung von Dateien. Auch wenn das Format einer zu archivierenden Datei sicher bestimmt werden konnte, garantiert dies noch nicht, dass die fragliche Datei korrekt gemäß den Formatspezifikationen aufgebaut ist. Enthält die Datei Teile, die gegen die Spezifikation verstoßen, kann eine Verarbeitung oder Darstellung der Datei unmöglich werden. Besonders fragwürdig, speziell im Hinblick auf die digitale Langzeitarchivierung, sind dabei proprietäre und gegebenenfalls undokumentierte Abweichungen von einer Spezifikation oder auch zu starke Fehlertoleranz eines Darstellungsprogrammes.

Ein gutes Beispiel hierfür ist HTML, bei dem zwar syntaktische und grammatikalische Regeln definiert sind, die aktuellen Browser jedoch versuchen, fehlerhafte Stellen der Datei einfach dennoch darzustellen oder individuell zu interpretieren. Wagt man nun einmal einen Blick in die „fernere“ Zukunft – beim heutigen Technologiewandel etwa 20-30 Jahre – dann werden die proprietären Darstellungsprogramme wie beispielsweise die unterschiedlich interpre-

2 Eine bestimmte Stelle in einer Datei wird oft als „Offset“ bezeichnet und mit einem hexadezimalen Wert addressiert

tierenden Web-Browser Internet Explorer und Firefox wohl nicht mehr existieren. Der einzige Anhaltspunkt, den ein zukünftiges Bereitstellungssystem hat, ist also die Formatspezifikation der darzustellenden Datei. Wenn diese jedoch nicht valide zu den Spezifikationen vorliegt, ist es zu diesem Zeitpunkt wohl nahezu unmöglich, proprietäre und undokumentierte Abweichungen oder das Umgehen bzw. Korrigieren von fehlerhaften Stellen nachzuvollziehen. Daher sollte schon zum Zeitpunkt der ersten Archivierung sichergestellt sein, dass eine zu archivierende Datei vollkommen mit einer gegebenen Formatspezifikation in Übereinstimmung ist.

Weiterhin kann untersucht werden, zu welchem Grad eine Formatspezifikation eingehalten wird – dies setzt eine erfolgreiche Identifizierung voraus. Als weiteres Beispiel kann eine XML-Datei beispielsweise in einem ersten Schritt *well-formed* (wohlgeformt) sein, so dass sie syntaktisch der XML-Spezifikation entspricht. In einem zweiten Schritt kann eine XML-Datei aber auch noch *valid* (valide) sein, wenn sie zum Beispiel einem XML-Schema entspricht, das wiederum feinere Angaben macht, wie die XML-Datei aufgebaut zu sein hat.

Da Format-Spezifikationen selbst nicht immer eindeutig zu interpretieren sind, sollte eine Validierung von Dateien gegen eine Spezifikation für die digitale Langzeitarchivierung möglichst konfigurierbar sein, so dass sie an lokale Bedürfnisse angepasst werden kann.

7.4.3 Extraktion, technische Metadaten und Tools

Mathias Neubauer

Wie bei jedem Vorhaben, das den Einsatz von Software beinhaltet, stellt sich auch bei der Langzeitarchivierung von digitalen Objekten die Frage nach den geeigneten Auswahlkriterien für die einzusetzenden Software-Tools.

Besonders im Bereich der Migrations- und Manipulationstools kann es von Vorteil sein, wenn neben dem eigentlichen Programm auch der dazugehörige Source-Code[3] der Software vorliegt. Auf diese Weise können die während der Ausführung des Programms durchgeführten Prozesse auch nach Jahren noch nachvollzogen werden, indem die genaue Abfolge der Aktionen im Source-

3 Der Source- oder auch Quellcode eines Programmes ist die les- und kompilierbare, aber nicht ausführbare Form eines Programmes. Er offenbahrt die Funktionsweise der Software und kann je nach Lizenzierung frei erweiter- oder veränderbar sein (Open Source Software).

Code verfolgt wird. Voraussetzung dafür ist natürlich, dass der Source-Code seinerseits ebenfalls langzeitarchiviert wird.

Nachfolgend werden nun einige Tool-Kategorien kurz vorgestellt, welche für die digitale Langzeitarchivierung relevant und hilfreich sein können.

Formaterkennung

Diese Kategorie bezeichnet Software, die zur Identifikation des Formats von Dateien eingesetzt wird. Die Ergebnisse, welche von diesen Tools geliefert werden, können sehr unterschiedlich sein, da es noch keine global gültige und einheitliche Format Registry gibt, auf die sich die Hersteller der Tools berufen können. Manche Tools nutzen jedoch schon die Identifier von Format Registry Prototypen wie PRONOM (beispielsweise „DROID", eine Java Applikation der National Archives von Großbritannien, ebenfalls Urheber von PRONOM (http://droid.sourceforge.net). Viele Tools werden als Ergebnis einen so genannten „Mime-Typ" zurückliefern. Dies ist jedoch eine sehr grobe Kategorisierung von Formattypen und für die Langzeitarchivierung ungeeignet, da zu ungenau.

Metadatengewinnung

Da es für die Langzeitarchivierung, insbesondere für die Migrationsbemühungen, von großem Vorteil ist, möglichst viele Details über das verwendete Format und die Eigenschaften einer Datei zu kennen, spielen Tools zur Metadatengewinnung eine sehr große Rolle. Prinzipiell kann man nie genug über eine archivierte Datei wissen, jedoch kann es durchaus sinnvoll sein, extrahierte Metadaten einmal auf ihre Qualität zu überprüfen und gegebenenfalls für die Langzeitarchivierung nur indirekt relevante Daten herauszufiltern, um das Archivierungssystem nicht mit unnötigen Daten zu belasten. Beispiel für ein solches Tool ist „JHOVE" (das JSTOR/Harvard Object Validation Environment der Harvard University Library, http://hul.harvard.edu/jhove/), mit dem sich auch Formaterkennung und Validierung durchführen lassen. Das Tool ist in Java geschrieben und lässt sich auch als Programmier-Bibliothek in eigene Anwendungen einbinden. Die generierten technischen Metadaten lassen sich sowohl in Standard-Textform, als auch in XML mit definiertem XML-Schema ausgeben.

Validierung

Validierungstools für Dateiformate stellen sicher, dass eine Datei, welche in einem fraglichen Format vorliegt, dessen Spezifikation auch vollkommen ent-

spricht. Dies ist eine wichtige Voraussetzung für die Archivierung und die spätere Verwertung, Anwendung und Migration beziehungsweise Emulation dieser Datei. Das bereits erwähnte Tool „JHOVE" kann in der aktuellen Version 1.1e die ihm bekannten Dateiformate validieren; verlässliche Validatoren existieren aber nicht für alle Dateiformate. Weit verbreitet und gut nutzbar sind beispielsweise XML Validatoren, die auch in XML Editoren wie „oXygen" (SyncRO Soft Ltd., http://www.oxygenxml.com) oder „XMLSpy" (Altova GmbH, http://www.altova.com/XMLSpy) integriert sein können.

Formatkorrektur

Auf dem Markt existiert eine mannigfaltige Auswahl an verschiedensten Korrekturprogrammen für fehlerbehaftete Dateien eines bestimmten Formats. Diese Tools versuchen selbstständig und automatisiert, Abweichungen gegenüber einer Formatspezifikation in einer Datei zu bereinigen, so dass diese beispielsweise von einem Validierungstool akzeptiert wird. Da diese Tools jedoch das ursprüngliche Originalobjekt verändern, ist hier besondere Vorsicht geboten! Dies hat sowohl rechtliche als auch programmatische Aspekte, die die Frage aufwerfen, ab wann eine Korrektur eines Originalobjektes als Veränderung gilt, und ob diese für die Archivierung gewünscht ist. Korrekturtools sind üblicherweise mit Validierungstools gekoppelt, da diese für ein sinnvolles Korrekturverfahren unerlässlich sind. Beispiel für ein solches Tool ist „PDF/A Live!" (intarsys consulting GmbH, (http://www.intarsys.de/de/produkte/pdfa-live), welches zur Validierung und Korrektur von PDF/A konformen Dokumenten dient.

Konvertierungstools

Für Migrationsvorhaben sind Konvertierungstools, die eine Datei eines bestimmten Formats in ein mögliches Zielformat überführen, unerlässlich. Die Konvertierung sollte dabei idealerweise verlustfrei erfolgen, was jedoch in der Praxis leider nicht bei allen Formatkonvertierungen gewährleistet sein kann. Je nach Archivierungsstrategie kann es sinnvoll sein, proprietäre Dateiformate vor der Archivierung zunächst in ein Format mit offener Spezifikation zu konvertieren. Ein Beispiel hierfür wäre „Adobe Acrobat" (Adobe Systems GmbH, http://www.adobe.com/de/products/acrobat/), welches viele Formate in PDF[4] überführen kann.

4 Portable Document Format, Adobe Systems GmbH, Link: http://www.adobe.com/de/products/acrobat/adobepdf.html

Für Langzeitarchivierungsvorhaben empfiehlt sich eine individuelle Kombination der verschiedenen Kategorien, welche für das jeweilige Archivierungsvorhaben geeignet ist. Idealerweise sind verschiedene Kategorien in einem einzigen Open Source Tool vereint, beispielsweise was Formaterkennung, -konvertierung und -validierung betrifft. Formatbezogene Tools sind immer von aktuellen Entwicklungen abhängig, da auf diesem Sektor ständige Bewegung durch immer neue Formatdefinitionen herrscht. Tools, wie beispielsweise „JHOVE", die ein frei erweiterbares Modulsystem bieten, können hier klar im Vorteil sein. Dennoch sollte man sich im Klaren darüber sein, dass die Archivierung von digitalen Objekten nicht mittels eines einzigen universellen Tools erledigt werden kann, sondern dass diese mit fortwährenden Entwicklungsarbeiten verbunden ist. Die in diesem Kapitel genannten Tools können nur Beispiele für eine sehr große Palette an verfügbaren Tools sein, die beinahe täglich wächst.

7.5 File Format Registries

Andreas Aschenbrenner und Thomas Wollschläger

Zielsetzung und Stand der Dinge

Langzeitarchive für digitale Objekte benötigen aufgrund des ständigen Neuerscheinens und Veraltens von Dateiformaten aktuelle und inhaltlich präzise Informationen zu diesen Formaten. File Format Registries dienen dazu, den Nachweis und die Auffindung dieser Informationen in einer für Langzeitarchivierungsaktivitäten hinreichenden Präzision und Qualität zu gewährleisten. Da Aufbau und Pflege einer global gültigen File Format Registry für eine einzelne Institution so gut wie gar nicht zu leisten sind, müssen sinnvollerweise kooperativ erstellte und international abgestimmte Format Registries erstellt werden. Dies gewährleistet eine große Bandbreite, hohe Aktualität und kontrollierte Qualität solcher Unternehmungen.

File Format Registries können verschiedenen Zwecken dienen und dementsprechend unterschiedlich angelegt und folglich auch verschieden gut nachnutzbar sein. Hinter dem Aufbau solcher Registries stehen im Allgemeinen folgende Ziele:

- Formatidentifizierung
- Formatvalidierung
- Formatdeskription/-charakterisierung
- Formatlieferung/-ausgabe (zusammen mit einem Dokument)
- Formatumformung (z.B. Migration)
- Format-Risikomanagement (bei Wegfall von Formaten)

Für Langzeitarchivierungsvorhaben ist es zentral, nicht nur die Bewahrung, sondern auch den Zugriff auf Daten für künftige Generationen sicherzustellen. Es ist nötig, eine Registry anzulegen, die in ihrer Zielsetzung alle sechs genannten Zwecke kombiniert. Viele bereits existierende oder anvisierte Registries genügen nur einigen dieser Ziele, meistens den ersten drei.

Beispielhaft für derzeit existierende File Format Registries können angeführt werden:

(I) file-format.net,
http://file-format.net/articles/
(II) FILExt,
http://filext.com/
(III) Library of Congress Digital Formats,
http://www.digitalpreservation.gov/formats/fdd/browse_list.shtml
(IV) C.E. Codere's File Format site,
http://magicdb.org/stdfiles.html
(V) PRONOM,
http://www.nationalarchives.gov.uk/pronom/
(VI) das Global Digital Format Registry,
http://hul.harvard.edu/gdfr/
(VIIa) Representation Information Registry Repository,
http://registry.dcc.ac.uk:8080/RegistryWeb/Registry/
(VIIb) DCC RI RegRep,
http://twiki.dcc.rl.ac.uk/bin/view/OLD/DCCRegRepV04
(VIII) FCLA Data Formats,
http://www.fcla.edu/digitalArchive/pdfs/recFormats.pdf

Bewertung von File Format Registries

Um zu beurteilen bzw. zu bewerten, ob sich spezielle File Format Registries für eine Referenzierung bzw. Einbindung in das eigene Archivsystem eignen, sollten sie sorgfältig analysiert werden. Sinnvoll können z.B. folgende Kriterien als Ausgangspunkt gewählt werden:

- Was ist der Inhalt der jeweiligen Registry? Wie umfassend ist sie aufgebaut?
- Ist der Inhalt vollständig im Hinblick auf die gewählte Archivierungsstrategie?
- Gibt es erkennbare Schwerpunkte?
- Wie werden Beschreibungen in die Registry aufgenommen? (Governance und Editorial Process)
- Ist die Registry langlebig? Welche Organisation und Finanzierung steckt dahinter?
- Wie kann auf die Registry zugegriffen werden? Wie können ihre Inhalte in eine lokale Archivierungsumgebung eingebunden werden?

Künftig werden File Format Registries eine Reihe von Anforderungen adressieren müssen, die von den im Aufbau bzw. Betrieb befindlichen Langzeit-Archivsystemen gestellt werden. Dazu gehören u.a. folgende Komplexe:

I) Vertrauenswürdigkeit von Formaten

Welche Rolle spielt die qualitative Bewertung eines Formats für die technische Prozessierung? Braucht man beispielsweise unterschiedliche Migrationsroutinen für Formate unterschiedlicher Vertrauenswürdigkeit? Wie kann dann ein Kriterienkatalog für die Skalierung der *confidence* (Vertrauenswürdigkeit) eines Formats aussehen und entwickelt werden? Unter Umständen müssen hier noch weitere Erfahrungen mit Migrationen und Emulationen gemacht werden, um im Einzelfall zu einem Urteil zu kommen. Es sollte jedoch eine Art von standardisiertem Vokabular und Kriteriengebrauch erreicht werden und transparent sein.

II) Persistent Identifier

Wie können *Persistent Identifier* (dauerhafte und eindeutige Adressierungen) von File Formats sinnvoll generiert werden? So kann es bestimmte Vorteile haben, Verwandtschafts- und Abstammungsverhältnisse von File Formats bereits am Identifier ablesen zu können. Die Identifizierung durch „Magic Numbers" scheint zu diesem Zweck ebenso wenig praktikabel wie die anhand eventueller ISO-Nummern. Die vermutlich bessere Art der Identifizierung ist die anhand von Persistent Identifiers wie URN oder DOI.

III) ID-Mapping

Wie kann ein Mapping verschiedener Identifikationssysteme (Persistent Identifier, interne Identifier der Archivsysteme, ISO-Nummer, PRONOM ID, etc.) durch Web Services erreicht werden, um in Zukunft die Möglichkeit des Datenaustausches mit anderen File Format Registries zu ermöglichen?

IV) Integration spezieller Lösungen

Wie kann in die bisherigen nachnutzbaren Überlegungen anderer Institutionen die Möglichkeit integriert werden, spezifische Lösungen für den Datenaustausch bereit zu halten? Dies betrifft beispielsweise die Möglichkeit, lokale Sichten zu erzeugen, lokale *Preservation Policies* zuzulassen oder aber mit bestimmten Kontrollstatus von eingespielten Records (z.B. „imported", „approved", „deleted") zu arbeiten.

Literatur

Abrams, Seaman: *Towards a global digital format registry.* 69th IFLA 2003. http://archive.ifla.org/IV/ifla69/papers/128e-Abrams_Seaman.pdf

Representation and Rendering Project: *File Format Report.* 2003. http://www.leeds.ac.uk/reprend/

Lars Clausen: *Handling file formats.* May 2004. http://netarchive.dk/publikationer/FileFormats-2004.pdf