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

nestor Handbuch

Eine kleine Enzyklopädie der digitalen Langzeitarchivierung

Version 2.3

Kapitel 7.4

Formatcharakterisierung

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 7.4 „Formatcharakterisierung“ (Version 2.3): urn:nbn:de:0008-20100617150
http://nbn-resolving.de/urn/resolver.pl?urn:nbn:de:0008-20100617150

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

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.