nestor
Kompetenznetzwerk Langzeitarchivierung

nestor Handbuch:

Eine kleine Enzyklopädie der digitalen Langzeitarchivierung

9 Formate

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

editors@langzeitarchivierung.de
c/o
Niedersächsische Staats- und Universitätsbibliothek Göttingen
Dr. Heike Neuroth
Forschung und Entwicklung
Papendiek 14
37073 Göttingen
Tel. +49 (0) 55 1 39 38 66

Der Inhalt steht unter folgender Creative Commons Lizenz:
http://creativecommons.org/licenses/by-nc-sa/2.0/de/

9 Formate

Einleitung

Stefan E. Funk

Dieser Text wurde vom Editor noch nicht freigegeben.

Ein Computer-Programm muss die Daten, die es verwaltet, auf einen permanenten Datenspeicher (zum Beispiel 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. Um sicher zu stellen, dass ein geladenes Dokument exakt dem Dokument entspricht, welches zuvor gespeichert wurde, ist es erforderlich, dass das Programm die gesicherten Daten (sprich die Folge von Nullen und Einsen) 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 in-

terpretieren 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. Die Bits, mit denen beispielsweise der Titel eines Dokuments gespeichert ist, müssen später exakt von derselben Stelle und semantisch gesehen auch als Titel wieder in unser Programm geladen werden, wenn das Dokument seine ursprüngliche Bedeutung behalten soll. Somit muss das Programm das Format genau kennen, muss wissen, welche Bits des Bitstreams welche Bedeutung haben, um diese richtig zu interpretieren und verarbeiten zu können.

Ein *Format-Spezifikation* ist nun eine Beschreibung der Anordnung der Bits und somit eine Beschreibung, wie die Daten 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 bekannt und bei offenen Formaten ist die Spezifikation frei zugänglich. Aus einer Datei, dessen Format und Spezifikation bekannt ist, kann die gespeicherte Information auch ohne das vielleicht nicht mehr verfügbare lesende Programm extrahiert werden. Ist die Spezifikation nicht verfügbar, ist die Gefahr sehr groß, dass die enthaltenen Daten nicht mehr korrekt interpretiert werden können und so Informationen verloren gehen. Aus diesem Grund sind dokumentierte Spezifikationen und standardisierte Formate für die Langzeitarchivierung digitaler Daten sehr wichtig.

Als *Standard* bezeichnet man ein Formate, das sich entweder aus dokumentierten proprietären Formaten etabliert hat, weil es von sehr vielen Nutzern/Programmen aufgegriffen wurde, oder das speziell als Standard entwickelt wurde mit dem Ziel, den Datenaustausch zwischen Programmen oder Plattformen zu vereinfachen oder gar erst zu ermöglichen. Als Beispiele seien hier das Open Document Format (ODF) sowie Grafik-Formate wie TIFF (Tagged Image File Format), GIF (Graphics Interchange Format) und JPEG (Joint Photographic Experts Group) oder auch PDF (Portable Document Format) genannt.

9.1 Digitale Objekte

Stefan E. Funk

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 OAIS[1]. 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", 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 bis hin zu kompletten Betriebssystemen.

Ein digitales Objektes kann auf drei Ebenen beschrieben werden, siehe Abbildung:

- als physisches Objekt,
- als logisches Objekt und schließlich
- als konzeptuelles Objekt.

Ein digitales Objekt kann beispielsweise eine Datei in einem spezifischen Dateiformat sein, z.B. eine einzelne Grafik, ein Word-Dokument oder eine PDF-Datei. Als ein digitales Objekt können auch komplexere Objekte bezeichnet werden, wie Anwendungsprogramme wie Word oder Mozilla, eine komplette Webseite inkl. Text und Grafik, eine durchsuchbare Datenbank auf CD inklusive einer Suchoberfläche oder ein Betriebssystem wie Linux, Windows oder Mac OS .

Das physische Objekt - Daten auf einem Speichermedium

Als physisches Objekt sieht man die Menge der Zeichen an, die auf einem Informationsträger gespeichert sind. Die Art und Weise der physischen Be-

1 Open Archival Information System

schaffenheit 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.

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.

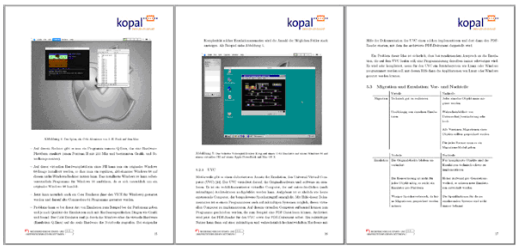
Das logische Objekt - Eine Folge von Bits als Einheit

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 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.

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 oder Emulation.

Das konzeptuelle Objekt - Das Objekt „zum Begreifen“

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. Dies sind zunächst die Objekte, Zeichen und Töne, die der Mensch über seine Sinne wahrnimmt. Auch interaktive Dinge wie das



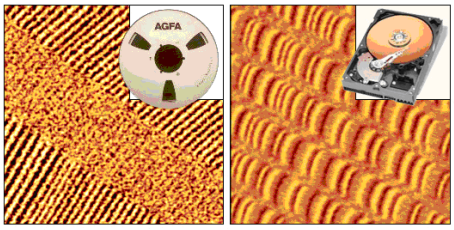
Wahrnehmung durch den Nutzer mit Hilfe von Soft- und Hardware

Konzeptuelles Objekt

0010010001111001
1001000011001101
0011100110001100
1001001001010101
1001000100001111
1001110000011010

Der Bitstrom wird durch Software als Dateiformat erkannt

Logisches Objekt



physische Datenstruktur

Physisches Objekt

Die drei Ebenen eines digitalen Objekts

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 „Digital Preservation“ zu schützen gilt.

Die Erhaltung des konzeptuellen Objekts

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 zu Beginn der Archivierung wiederspiegeln. Dies ist nicht möglich, wenn sich 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.

Literatur

- Reference Model for an Open Archival Information System (OAIS) <http://ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents/pdf/CCSDS-650.0-B-1.pdf> (letzter Zugriff: 7. Juni 2006)
- Huth, Karsten, Andreas Lange: Die Entwicklung neuer Strategien zur Bewahrung und Archivierung von digitalen Artefakten für das Computerspiele-Museum Berlin und das Digital Game Archive (2004) <http://www.ichim.org/ichim04/contenu/PDF/2758_HuthLange.pdf> (letzter Zugriff: 7. Juni 2006)
- Thibodeau, K.: Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years. In The State of Digital Preservation: An International Perspective. Washington D.C.: Council on Library and Information Resources. 4-31 (2001) <http://www.clir.org/PUBS/abstract/pub107abst.html> (letzter Zugriff: 12. Juli 2006)

9.2 Dateiformate

Stefan E. Funk

Dateiformate, in denen ein digitales Objekt vorliegt, um von Anwendungsprogrammen verarbeitet werden zu können, spielen bei der Archivierung eine große Rolle. Diese Formate sind mehr oder weniger klar spezifiziert, einige sind offene Standards und andere sind proprietäre Formate einzelner Firmen. Als Beispiele lassen sich hier Formate nennen wie PDF (Portable Document Format), XML (eXtensive Markup Language), HTML (HyperText Markup Language), DOC (Windows Document Format), verschiedene Bildformate wie TIF (Tagged Image Format) oder GIF (Graphic Interchange Format).

Formaterkennung
Will man solche Dokumente für die Nachwelt erhalten und den Zugriff auf deren Inhalte sichern, besteht die dringende Notwendigkeit, diese verschiedenen Formate zu kennen und zu erkennen. Es ist sehr wichtig zu wissen, welches Dateiformat ein digitales Dokument hat und ob das Format dieses Dokuments auch korrekt ist. Die Korrektheit dieser Daten stellt sicher, dass ein Dokument genutzt bzw. angezeigt und später im Sinne von Migration und Emulation bearbeitet werden kann. Bevor ein Objekt in ein Langzeitarchiv eingespielt wird, müssen spezifische Informationen über dieses Objekt vorhanden sein, sogenannte Metadaten, die genaue Aussagen darüber machen, welches Dateiformat in welcher Version vorliegt. Die Spezifikationen der unterschiedlichen Formate müssen hinreichend bekannt sein, um eine spätere Migration zu ermöglichen. Es reicht unter Umständen nicht aus, ein Dokument mit Hilfe eines Programmes anzeigen zu können, es sollte auch möglich sein, anhand der Spezifikationen ein Anzeige- oder Konvertierungsprogramm zu entwickeln.

Validation

Für die Langzeitarchivierung reicht es nicht aus zu wissen, dass eine Datei in einem bestimmten Format und in einer bestimmten Version dieses Formats vorliegt. Eine weitere wichtige Information ist die Korrektheit des Dokument im Sinne der Spezifikation dieses Formats. Nur so ist ein späteres Bearbeiten der Dokumente möglich, denn die Tools zur Konvertierung (oder Migration) bauen auf den Formatspezifiktionen auf. Habe ich beispielsweise ein Doku-

ment im PDF-Format der Version 1.2 vorliegen und prüfe nicht eingehend, ob dieses Format auch den Spezifikationen entspricht, könnte es sein, dass spätere Migrations- und Konvertierungs-Tools, die aus PDF 1.2 ein neueres Format (zum Beispiel PDF 1.6) erstellen sollen, das Dokument nicht richtig oder im schlimmsten Fall gar nicht verarbeiten können. Selbst wenn eine Datei korrekt dargestellt wird, ist noch nicht sichergestellt, dass sie auch der Formatspezifikation entspricht, da viele Anzeigeprogramme sehr fehlertolerant sind. Informationsverlust bis hin zum Verlust des gesamten Dokuments kann die Folge sein.

Metadaten

Zur Verwaltung von digitalen Objekten innerhalb eines Archivsystems werden Metadaten benötigt. Dies sind Daten über ein digitales Objekt. Zur Bestandserhaltung von digitalen Objekten werden zunächst technische Metadaten benötigt. Dies sind Daten wie Dateiformat und Version, Dateigröße, Dateiname, Checksumme zur Kontrolle der Integrität, MIME-Type, Erstellungsprogramm, Anzeigeprogramm, etc. Zur Dokumentation der Migrationsschritte dienen Provenance Metadaten. Diese beschreiben die Herkunft des Dokuments, beispielsweise die Art der Migration, den Zeitpunkt, die einzelnen durchgeführten Schritte und bei der Migration genutzte Programme. Deskriptive Metadaten beschreiben das Objekt inhaltlich, hierzu gehören unter anderem der Titel des Dokuments, der Name der Autoren, Abstract, Erscheinungsdatum und -Ort sowie Verlag. Rechtliche Metadaten schließlich beinhalten rechtliche Daten über das Dokument wie Eigentümer, Zugriffserlaubnis, etc.

Hilfsmittel

Es gibt Möglichkeiten, einige Metadaten maschinell zu erfassen. Die deskriptiven Metadaten zum Beispiel können aus den digitalen Katalogsystemen entnommen werden, sofern dafür geeignete Schnittstellen existieren. Die technischen Metadaten automatisch zu erfassen, ist in gewissen Grenzen ebenfalls möglich. Einige Programmier-Tools können technische Metadaten aus den digitalen Objekten extrahieren, zum Beispiel das Dateiformat und die Version desselben. Wie umfangreich die erhaltenen Metadaten sind, hängt von der Qualität des Tools ab. Im Einzelfall wird man solche Tools an die einzelnen Anforderungen anpassen müssen. Das Metadaten-Extraktions-Tool JHOVE[2] wird beispielsweise vom Projekt kopal[3] zur Erfassung von technischen Metadaten genutzt.

2 JSTOR/Harvard Object Validation Environment <http://hul.harvard.edu/jhove/index.html>

3 <http://kopal.langzeitarchivierung.de>

9.4 Formaterkennung und Validierung

Matthias Neubauer

Die Archivierung von digitalen Objekten steht und fällt mit der Erkennung 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. Dies beinhaltet vor allem zwei große Bereiche:

a) Die Formaterkennung

Zunächst muss das genaue Format ermittelt werden, in welchem die fragliche Datei vorliegt. Unterschiedliche Formate verwenden auch sehr unterschiedliche Identifizierungsmerkmale, was ein generell anwendbares Verfahren erschwert. Ein Merkmal, das zunächst nahe liegend erscheint, ist die so genannte Dateiendung oder File Extension. Dies 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[4] 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

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

Format abgespeichert wurde.

b) Die Validierung gegen eine Formatspezifikation

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 interpretierenden 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.

Sowohl für die aktuelle Bereitstellung der archivierten Dateien, als auch für spätere Migrations- und Emulationsszenarien ist demnach sowohl die Erkennung als auch die Validierung von Dateiformaten eine notwendige Voraussetzung. Ein Versäumnis dieser Aktionen kann einen erheblich höheren Arbeitsaufwand oder sogar einen vollkommenen Datenverlust zu einem späteren Zeitpunkt bedeuten.

9.5 File Format Registries

Andreas Aschenbrenner, Thomas Wollschläger

1. 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 ist, 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 seiner 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) die File Format Encyclopedia,
http://pipin.tmd.ns.ac.yu/extra/fileformat/
(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/omar
(VIIb) DCC RI RegRep,
http://dev.dcc.rl.ac.uk/twiki/bin/view/Main/DCCRegRepV04
(VIII) FCLA Data Formats,
http://www.fcla.edu/digitalArchive/pdfs/recFormats.pdf

2. 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.

3. Bibliografie

- Abrams, Seaman: Towards a global digital format registry. 69th IFLA 2003. http://www.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

9.6 Tools

Matthias 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[5] 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-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.

a) 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. Link: 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.

b) Metadatengewinnung

Da es für die Langzeitarchivierung, insbesondere für die Migrationsbemü-

5 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).

hungen, 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, Link: 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.

c) Validierung

Validierungstools für Dateiformate stellen sicher, dass eine Datei, welche in einem fraglichen Format vorliegt, dessen Spezifikation auch vollkommen entspricht. 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., Link: http://www.oxygenxml.com) oder „XMLSpy" (Altova GmbH, Link: http://www.altova.com/XMLSpy) integriert sein können.

d) 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, Link: http://www.intarsys.de/produkte/dokumententechnologien/pdf-a-live), welches zur Validierung und Korrektur von PDF/A konformen Dokumenten dient.

e) 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, Link: http://www.adobe.com/de/products/acrobat/), welches viele Formate in PDF[6] überführen kann.

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.

---

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