nestor
Kompetenznetzwerk Langzeitarchivierung

nestor Handbuch:

Eine kleine Enzyklopädie der digitalen Langzeitarchivierung

9.6 Tools

Herausgeber

Heike Neuroth
Hans Liegmann †
Achim Oßwald
Regine Scheffel
Mathias Jehn
Stefan Strathmann

GEFÖRDERT VOM

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

Niedersächsische Staats- und Universitätsbibliothek Göttingen
Dr. Heike Neuroth
Forschung und Entwicklung
Papendiek 14
37073 Göttingen
neuroth@sub.uni-goettingen.de
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/

9.6 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>