nestor
Kompetenznetzwerk Langzeitarchivierung

nestor Handbuch:

Eine kleine Enzyklopädie der digitalen Langzeitarchivierung

9.2 Dateiformate

Herausgeber:

Heike Neuroth
Hans Liegmann
Achim Oßwald
Regine Scheffel
Mathias Jehn

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

Dieser Artikel ist ein Auszug aus dem nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung

Dieser Artikel ist verfügbar unter der URL:
http://nestor.sub.uni-goettingen.de/handbuch/artikel/text_84.pdf

Die Online Version des Handbuches unter der URL:
http://nestor.sub.uni-goettingen.de/handbuch/

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

9.2 Dateiformate

von Stefan 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 Dokument 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[1] wird beispielsweise vom Projekt kopal[2] zur Erfassung von technischen Metadaten genutzt.

[1] JSTOR/Harvard Object Validation Environment http://hul.harvard.edu/jhove/index.html
[2] http://kopal.langzeitarchivierung.de