nestor
Kompetenznetzwerk Langzeitarchivierung

nestor Handbuch:

Eine kleine Enzyklopädie der digitalen Langzeitarchivierung

9.4 Formaterkennung und Validierung

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

9.4 Formaterkennung und Validierung

von 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 Bitstroms 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:

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 naheliegend erscheint, ist die sogenannte 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 sogenannte "Magic Number". Dies ist ein Wert, welcher in einer Datei des entsprechenden Formats immer an einer in der Spezifikation bestimmten Stelle[1] 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.

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-

[1] Eine bestimmte Stelle in einer Datei wird oft als "Offset" bezeichnet und mit einem hexadezimalen Wert adressiert.

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.