

nestor Handbuch:
**Eine kleine Enzyklopädie
der digitalen Langzeitarchivierung**

9 Formate
Einführung / Überblick

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-

interpretieren 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.