Langzeiterhaltung digitaler Daten in Museen

Tipps zur dauerhaften Bewahrung digitaler Daten

17

Digitalisierung von Textdokumenten

Für die Digitalisierung von Texten zum Zwecke der Langzeiterhaltung sind zwei grundsätzliche Aspekte von zentraler Bedeutung.

Art der Vorlage

Die Digitalisierung gedruckter oder handschriftlicher Textdokumente ähnelt in weiten Bereichen dem Verfahren zur Digitalisierung von Bildern (➜ Blatt 16).

Auch hierbei gilt, dass zunächst eine verbindliche Festlegung der Benennung der Dateien, der im Header (Dateikopf) der Datei zu speichernden Metadaten und des zu verwendenden langzeitarchivierungsfähigen Dateiformats am Beginn stehen muss.

Weitere Nutzung der Digitalisate

In welcher Form soll das langzeitarchivierte Dokument zukünftig genutzt werden? Steht die Wiedergabe des visuellen Erscheinungsbildes im Vordergrund der weiteren Nutzung oder ist der textliche Inhalt von vorrangiger Bedeutung? Hierbei ist festzustellen, ob die analoge Vorlage als finalisiertes Dokument zu betrachten ist oder ob das Dokument für eine weitere Verarbeitung zur Verfügung stehen soll.

Je nach Art der Vorlage und Nutzungsintention sind unterschiedliche Lösungswege oder Kombinationen derselben zu wählen.

Digitalisierung analoger Dokumentvorlagen

Geräte zur Digitalisierung

Als Geräte für die Digitalisierung stehen, neben handelsüblichen Scannern unterschiedlicher Qualitäten und Funktionsweisen, auch Spezialscanner und Spezialaufbauten von digitalen Kameras zur Verfügung. Drei Arten von Spezialscannern für die Digitalisierung von Texten seien hier kurz beschrieben.

- Einzugscanner

Im Gegensatz zu Flachbettscannern bewegt sich beim Einzugscanner die Vorlage über die Scaneinheit. Der Vorteil liegt in der wesentlich höheren Verarbeitungsgeschwindigkeit als bei der manuellen Dokumentenzuführung, da das Einlegen einzelner Vorlagen entfällt. Ein Nachteil ist die hohe mechanische Belastung der Vorlagen. Für Unikate und fragile Vorlagen ist dieser Scanner daher ungeeignet.

- Buch- oder Aufsichtscanner

Buch- oder Aufsichtscanner dienen dem Scannen gebundener Bücher, das Scanverfahren erfolgt mittels eines Lesekopfes von oben auf das Buch herab. Je nach Vorlage sind diese Scanner mit einer Buchwippe, für Bücher, deren Einband nicht vollständig geöffnet werden kann, und mit speziellen Vorrichtungen zum automatischen und dabei schonenden Umblättern der Seiten ausgestattet.

- Scanroboter

Besonders ausgefeilte Systeme werden als Scanroboter bezeichnet. Sie sind mit einer Software verbunden, die den Digitalisierungsprozess automatisiert, dokumentiert und die Digitalisate mit Metadaten versieht.

Inhaltliche Erfassung

Zentrales Element bei der Digitalisierung von Texten ist der Erhalt der im Text enthaltenen Informationen. Bei den analogen Vorlagen entsteht an dieser Stelle ein großes Problem. Das Scannen von Textseiten erzeugt keine Texte, sondern digitale Bilder. Die Textinformationen stehen somit für die Nutzung in Textverarbeitungen, beispielsweise für Suchvorgänge, nicht zur Verfügung. Hierzu ist als weiterer Arbeitsschritt zwingend die Extraktion des Textes erforderlich. Hierfür stehen zwei Möglichkeiten zur Auswahl:

- Die automatisierte Erfassung durch Texterkennungsprogramme (OCR = Optical Character Recognition)
- Die manuelle Erfassung der Texte.

Texterkennungsprogramme (OCR)

Die automatische Texterkennung mittels OCR ist seit langem im Einsatz. Es existieren eine Reihe leistungsfähiger Programme. Beim Erwerb eines Scanners werden häufig (wenig leistungsfähige) OCR-Programme mitgeliefert. Das Arbeitsprinzip basiert auf Analyse

‚der gescannten Pixel und dem Versuch durch Farbanalyse Muster zu identifizieren, denen Zeichen, i. d. R. Buchstaben, zugeordnet werden können. Durch den Einsatz von "intelligenten" Verfahren lässt sich die Erkennungsrate unterschiedlicher Schrifttypen verbessern. Sofern die Vorlage frei von Verschmutzungen ist, die Satztypen regelmäßig und einheitlich gedruckt sind, lassen sich Erkennungsquoten von bis zu 99% erreichen. Das bedeutet aber immerhin noch 30 Zeichenfehler auf einer Seite mit 3.000 Zeichen. Die meisten dieser Programme setzen in unterschiedlicher Qualität das ursprüngliche Layout inkl. Spalten, Grafiken und Textauszeichnungen um.

Bestimmte Einsatzbereiche zeigen aber auch die Grenzen der OCR auf. Sie ist – bislang – ungeeignet für alle handschriftlichen Texte, für Texte mit ungleichförmigem Schriftsatz (unter anderem auch ältere Schreibmaschinenseiten), seltene Schriftarten, Frakturdrucke und für Dokumente mit handschriftlichen Annotationen.

Manuelle Erfassung der Texte (Abschrift)

Eine andere Möglichkeit, Texte in ihrem Volltext zu erfassen, ist die Erzeugung einer Abschrift. Diese bietet insbesondere bei schwierigen Handschriften auch die Möglichkeit, "Leseunsicherheiten" zu markieren und die Abbildung des Originals mit der inhaltlichen Wiedergabe des Textes zu verbinden.

Neben der reinen Bereitstellung der Abbildungen und des Textes kann die Information über den Aufbau bzw. die Gliederung des Werkes hilfreich sein (➜ Blatt 12.).

Die Überführung analoger Textvorlagen in maschinenlesbaren Inhalt besteht in der Regel aus zwei Arbeitsschritten: dem Scannen und der Texterfassung. Für letzteres ist die weitere technische Entwicklung automatischer Verfahren zur Texterkennung zu beobachten.

LINKS

http://www.apsr.edu.au/publications/word_processing_preservation.pdf
Publikation von 2006 zu den grundsätzlichen Herausforderungen bei der Bewahrung textueller Inhalte, hrsg. von APSR (Australian Partnership for Sustainable Repositories).

http://ahds.ac.uk/preservation/presBinary_v4.rtf
Anleitung zur Bewahrung von digitalen Texten, hrsg. vom AHDS (arts and humantities data service).

http://www.kb.nl/hrd/dd/dd_links_en_publicaties/PDF_Guidelines.pdf
Richtlinien für die Erstellung von langzeitarchivierungsfähigen PDF-Dokumenten. Hrsg. von der Koninklijken Bibliotheek der Niederlande.

http://www.sub.uni-goettingen.de/ebene_2/vdf/endfas2.htm
Bericht der Arbeitsgruppe Technik zur Vorbereitung des Programms "Retrospektive Digitalisierung von Bibliotheksbeständen" im Förderbereich "Verteilte Digitale Forschungsbibliothek".

http://www.ub.uni-heidelberg.de/helios/digi/tech_workflow.html
Programm DWork – Heidelberger Digitalisierungsworkflow für die Arbeitsabläufe bei der Digitalisierung und der Webpräsentation der Bestände der UB Heidelberg.

Stand: Juni 2009