Stellungnahme zur Anhörung zum Gesetzentwurf über die Ablieferung von Pflichtexemplaren in Nordrhein-Westfalen

LANDTAG
NORDRHEIN-WESTFALEN
16. WAHLPERIODE

STELLUNGNAHME
16/234

A12, A18

Tobias Steinke, Deutsche Nationalbibliothek, 16.11.2012

Langzeitverfügbarkeit

Die Langzeitarchivierung digitaler Publikationen stellt andere Herausforderungen an Bibliotheken und Archive als gedruckte Publikationen. Während die langfristige Bestandserhaltung von gedruckten Publikationen durch geeignete Maßnahmen (klimatisierte Lagerung, Papierentsäuerung, etc.) den dauerhaften Erhalt der physischen Originale zum Ziel hat, sind digitale Publikationen unabhängig vom physischen Datenträger. Die digitale Publikation ist ein binärer Datenstrom, der beim Kopieren von einem Datenträger zu einem anderen identisch bleibt und somit gibt es keinen Unterschied zwischen verschiedenen Kopien und kein physisches Original. Der Erhalt des Binärstroms kann somit durch rechtzeitiges Umkopieren auf einen anderen Datenträger dauerhaft gewährleistet werden.

Die besondere Herausforderung liegt in der dauerhaften Interpretierbarkeit des Binärstroms. Ein erhaltener Binärstrom kann nur mit technischen Hilfsmitteln (Systemumgebung aus Hardware und Software) zu einer für Menschen nutzbaren Information werden. Aufgrund des raschen technischen Wandels werden Systemumgebungen in wenigen Jahren durch neuere abgelöst, die oft nicht mit Daten früherer Generationen umgehen können. Ohne eine geeignete Systemumgebung zum Zugriff auf die enthaltenen Informationen kann auch ein durch rechtzeitiges Umkopieren dauerhaft erhaltener Binärstrom unbenutzbar werden und damit verloren sein. Zur Erhaltung der Langzeitverfügbarkeit sind zwei Strategien gängig:

1. *Migration*
   Dateien in Dateiformaten, die absehbar obsolet werden und somit nicht mehr nutzbar bleiben, werden in andere Dateiformate konvertiert, die absehbar nutzbar bleiben. Dies kann das Risiko mit sich bringen, dass inhaltliche Aspekte bei der Konvertierung verändert werden, aber die Alternative könnte ein totaler Verlust der Nutzbarkeit sein. Da es kein Dateiformat „für die Ewigkeit" gibt, ist Migration eine fortlaufende Aufgabe der Bestandserhaltung, die je nach technischer und Marktentwicklung immer wieder durchgeführt werden muss.

2. *Emulation*
   Mit bestimmter Software wird in einer aktuellen Systemumgebung eine frühere Systemumgebung virtuell nachgestellt und damit die Nutzung der obsoleten digitalen Objekte ermöglicht. Emulationen sind immer nur Annäherungen an die ursprünglichen Systemumgebungen und somit gibt es auch hierbei Risiken von veränderten Inhalten und Funktionalitäten. Da Emulationen selbst Programme sind, die abhängig von der aktuellen Systemumgebung sind, müssen diese bei Systemwechsel neu erstellt werden, so dass auch bei dieser Strategie ein fortwährender Aufwand entsteht.

Die beiden Strategien eignen sich unterschiedlich gut für verschiedene Publikationstypen. Statische Publikationen wie Texte oder Bilder lassen sich besser migrieren, während sich dynamische Publikationen wie Multimedia besser emulieren lassen. Eine archivierende Institution muss die dauerhaft nötigen Aufwände für diese Strategien für die Langzeitverfügbarkeit digitaler Publikation berücksichtigen.

Jegliche Form von technischem Kopierschutz bei digitalen Publikationen erschwert oder verhindert eine digitale Langzeitarchivierung. Daher ist es für archivierende Institutionen unumgänglich, eine kopierschutzfreie Version der digitalen Publikation zu bekommen. Zudem ist die Nutzung von Standards bei Dateiformaten wünschenswert (z. B. PDF/A).

Digitales Langzeitarchiv
Archivierende Institutionen sollten ein digitales Langzeitarchiv vorsehen, welches dem ISO-Standard OAIS-Referenzmodell folgt, der funktionale Komponenten für ein geeignetes Langzeitarchivierungssystem beschreibt. Ein solches System besteht aus Lieferwegen für digitale Publikationen (Ingest), Archivierung (Archival Storage), Metadaten (Data Management), Zugriff (Access) und Langzeitverfügbarkeit (Preservation Planning). Wichtig sind Metadaten, die nicht nur inhaltlichen Aspekte der Publikationen beschreiben, sondern auch die technischen, um gezielte Maßnahmen zur Langzeitverfügbarkeit ergreifen zu können. Der Ingest muss verschiedene sichere Wege zur Datenübernahme vorsehen, etwa per FTP, Webformular oder automatisierte Verfahren, bei denen die übermittelten digitalen Objekte validiert und technische Informationen generiert werden. Im Archival Storage muss der Binärstrom regelmäßig auf Fehler überprüft werden und durch geeignete Verfahren (verteilte Speicherung, Backup, Refreshing) die dauerhafte Datenkonsistenz sicher gestellt sein.

Die zu berücksichtigenden Aufwände sind nicht auf die Konzeption und Einrichtung eines digitalen Langzeitarchivs beschränkt, denn die schnell fortschreitende Entwicklung bei der Informationstechnik macht eine fortwährende Anpassung und Erneuerung der Systeme nötig. Die Bibliothek oder das Archiv muss daher dauerhaft geeignete technische Expertise vorhalten und Aufwände für Maßnahmen zur Langzeitverfügbarkeit und zur Weiterentwicklung des Archivsystems einplanen. Die tatsächlichen Aufwände hängen vom Umfang der Sammlung, der Art der Objekte und der technischen Entwicklung ab. Viele Aspekte der digitalen Langzeitarchivierung sind nach wie vor Gegenstand der Forschung, weshalb es etliche DFG- und EU-geförderte Projekte in dem Bereich gibt.

Webharvesting
Digitale Publikationen, die als Webseiten vorliegen, werden üblicherweise nicht aktiv abgeliefert, sondern durch das sogenannte Webharvesting von der archivierenden Institution selbst eingesammelt. Hierzu wird eine bestimmte Software genutzt, ein sogenannter Crawler oder auch Harvester, die ausgehend von einer URL eine Seite abspeichert und nacheinander alle Links auf der Seite verfolgt, um die dadurch beschriebenen Seiten aufzurufen und abzuspeichern. Dies wird so lange automatisiert weitergeführt, bis eine Abbruchbedingung auftritt, etwa ein Link zu einer Seite mit einer URL, die nicht in der gleichen Domain liegt (z. B. nicht mehr beginnend mit www.landtag.nrw.de).

Dabei ist zu beachten, dass Webseiten keine statischen Publikationen sind und sich jederzeit ändern können. Das Webharvesting kann immer nur eine Momentaufnahme der Seiten sammeln und archivieren. Üblicherweise wird daher die gleiche Seite in regelmäßigen Abständen (z. B. vierteljährlich) erneut geharvestet. Dies bedeutet einen dauerhaften technischen Aufwand und entsprechende Ressourcen für die zu speichernden Datenmengen. Zudem wandelt sich das Web zunehmen zu einem interaktiven Medium mit dynamischen Inhalten, was sich nicht in jedem Fall über dieses Verfahren erfassen lässt. Das automatisierte Verfahren kann daher nur eine lückenhafte Archivierung ermöglichen. Die Ergebnisse lassen sich teilweise durch aufwendige manuelle Nacharbeiten verbessern.

Webharvesting benötigt eigene Expertise, da die besonderen Eigenschaften von Webseiten zusätzliche Herausforderungen und entsprechende Lösungen bedeuten.