DIGITALE AUSSTELLUNG UND DANN ...? SICHERUNGSSTRATEGIEN FÜR DIGITALEN CONTENT UND DIGITALE OBJEKTE

Michael Steppes[a], Alexander Herschung[b]
[a]*startext GmbH, michael.steppes@startext.de;* [b]*GF der startext GmbH, alexander.herschung@startext.de*

KURZDARSTELLUNG: startext entwickelt seit 1980 modulare IT-Lösungen in den Bereichen Präsentation sowie Dokumentation für Erschließung, Verwaltung und Präsentation in Archiven, Museen, Sammlungen, Bibliotheken und Unternehmen. Zu den bewährten Produkten zählt der Hierarchische Datenbank-Administrator HiDA4 für die Inventarisation von Kulturgütern und die modulare Software-Gesamtlösung ACTApro für Archive. Ein produktunabhängiges OAIS-konformes Digitales Archiv und natürlich PABLO zur Sicherung von Webseiten bieten Lösungen für die digitale Langzeitarchivierung. Im Rahmen von Projektentwicklungen sind u.a. folgende IT-Lösungen entstanden: ManuscriptumXML für die Erfassung mittelalterlicher Handschriften, die archivische Gesamtlösung V.E.R.A., das Auskunfts- und Archivsystem CMS für Unternehmen, der Findbuch-Editor MidosaXML und der lernfähige XML-Editor MIDEX sowie das Archivportal ARGUS. Ergänzend zu diesen Produkten und Projektentwicklungen bietet startext individuell maßgeschneiderte Archivierungs- und Recherchekomponenten an. Zukunftssicherheit, offene Schnittstellen für den Datenimport und -Export in Standardformaten und Kompatibilität Ihrer Datenbestände sowie eine komfortable Benutzeroberfläche gehören zu den Grundanforderungen, die Kunden von den Softwareprodukten der startext GmbH erwarten können.

1. EINFÜHRUNG

Die Bewahrung von digitalem Content (z.B. Kunstwerke, die originär digital sind) kommt als Herausforderung in zunehmendem Maße auf Sammlungen und Museen zu. Es braucht Softwarelösungen für Fragestellungen wie diese: Wie gewährleisten wir die „Langzeitsicherung" digitaler Objekte? Wie kann man sicherstellen, dass z.B. Tonaufnahmen, Videos, Bilder noch in 50 Jahren zugreifbar und vor allem interpretierbar sein werden? Wie kann man solche digitalen Originale auf Dauer bewahren?

Die Inventarisierung in den Museen ist bislang ausgelegt auf die Erfassung von Metadaten zu physischen Objekten, nun sprechen wir hier aber von originär digitalen Objekten, deren Sicherung in einem Digitalen Archiv erfolgen muss. Im Vortrag wird die startext-Lösung eines digitalen,

OAIS-konformen Langzeitarchivsystems vorgestellt, das in dieser hybriden Umgebung zum Einsatz kommen kann. Startext REPOSITORY in Ergänzung zur Inventarisierungssoftware erlaubt die standardkonforme Sicherung der „born digitals".

Weiterhin bietet startext eine Lösung für die Herausforderung der Langzeitsicherung von Webpräsenzen: Webseiten sind einem steten Wandel unterworfen, komplette Neugestaltungen in kurzen Zeitabschnitten

sind keine Seltenheit. Die Vorversion der Webseite geht dabei häufig dauerhaft verloren. Was aber, wenn es sich z.B. um eine rein virtuelle Ausstellung handelt oder die Webseite eine Ausstellung begleitet? Diese thematische Präsentation geht mit Ablauf der Ausstellungsphase häufig verloren.
Herkömmliche Werkzeuge speichern Webpräsenzen als reine offline-Sammlung von HTML-, CSS-, JavaScript-Dateien und den verwendeten diversen Bildformaten etc. Dies ist im Kontext einer wirklichen digitalen Langzeitarchivierungsstrategie kritisch, denn aus dieser Dateisammlung wird erst durch die Interpretation durch einen heutigen Browser, in einem heutigen Betriebssystem die wirklich für den Nutzer erfahrbare Webpräsenz erzeugt.
Die Herausforderung ist, zu bewahren, wie sich eine Webseite heute in einem bestimmten Browser darstellt und verhält. Etablierte Standards und Verfahren finden sich zwar für die Speicherung (v.a. das WARC-Format), jedoch das Formatproblem der Archivierung ist hier nicht gelöst.
Die startext-Software PABLO bietet einen Lösungsvorschlag zu dem Formatproblem der digitalen Langzeitarchivierung von Webseiten.

2. STARTEXT REPOSITORY

startext REPOSITORY ist eine OAIS-konforme Softwarelösung zur digitalen Langzeitarchivierung.

Das OAIS-Referenzmodell hat sich weltweit als Standard im Bereich der digitalen Langzeitarchivierung, also der zeitlich unbegrenzten Bewahrung und Sicherstellung der Nutzbarkeit digitaler Inhalte, etabliert.

Im Rahmen von OAIS haben die folgenden Schlüsselstrategien zur digitalen Archivierung besonderes Gewicht:

1. Formatkontrolle – je genauer bei der Übernahme ins digitale Archiv Dateiformate kontrolliert, geprüft und begrenzt werden, desto besser sind die langfristigen Aussichten zur Nutzbarhaltung archivierter Daten.

   Insbesondere die Identifikation und Validierung von Dateiformaten, sowie die Umwandlung in geeignete Formate, die als langzeitstabil betrachtet werden, sind von besonderer Bedeutung.

2. Dokumentation – eine vollständige Dokumentation aller Metainformationen zu den archivierten digitalen Inhalten ist essentiell für die Vertrauenswürdigkeit des digitalen Archivs. Die Dokumentation umfasst Informationen zur Erstellung und Herkunft der digitalen Inhalte, sowie der Übergabe an bzw. Übernahme durch das Archiv und die in diesem Kontext durchgeführten Verarbeitungsschritte (z.B. wann und mit welchen Werkzeugen Formatwandlungen vorgenommen wurden).

   startext REPOSITORY besteht aus mehreren Komponenten, die wiederum verschiedene Aspekte des OAIS-Modells reflektieren:

- Ingest

  Der Ingest dient der Übernahme digitaler Inhalte in das Archivsystem. Hierbei werden eine oder mehrere Dateien in so genannten Übernahmepaketen (SIPs) gebündelt und durchlaufen einen mehrschrittigen Verarbeitungsprozess. Im Rahmen dieses Prozesses findet u.a. eine Formatkontrolle und gegebenenfalls Formatwandlungen statt.

  Technisch handelt es sich um eine konfigurierbare Verarbeitungskette mit austauschbaren Werkzeugen, die in Microservice-Architektur realisiert ist. Damit ist diese Komponente von vorneherein ausgelegt auf

Parallelisierbarkeit: Wenn die zu verarbeitenden Datenmengen es erfordern, kann ein zweiter (oder auch dritter,...) Ingest-Server installiert und dem Gesamtsystem hinzugefügt werden.

In der Verarbeitungskette selbst, kann jeder einzelne Schritt konfiguriert werden. Dies erstreckt sich insbesondere auf die hier verwendeten Werkzeuge z.B. zur Formaterkennung und -umwandlung.

Auch können der Verarbeitungskette eigene, zusätzliche Schritte hinzugefügt werden.

- Dokumentation

  Alle Verarbeitungsschritte werden standardkonform in PREMIS dokumentiert. Diese Dokumentation wird (als Teil einer METS-XML-Datei) mit im AIP abgespeichert.

  - Metadaten

    Während der Zusammenstellung der SIP werden auch Metainformationen erfasst:

    - automatische Metadatenermittlung aus den Primärdateien selbst. Hier werden Informationen, wie Erstelldatum, Autor, etc. aus den Dateien extrahiert und mit im Metadatensatz abgelegt. Auch der Volltext der Primärdateien wird hier ermittelt und im Metadatensatz mit abgespeichert.

    - Automatische Metadatenermittlungen aus (Teilen des) Datei- bzw. Ordnernamens. Hierbei können Informationen, die in Datei- bzw. Ordnernamen abgelegt sind, automatisch mit in den Metadatensatz übernommen werden.

    - Manuell erfasste Metadaten. Hier werden durch den Anwender manuell Informationen zur Provenienz und Übernahme, aber optional auch zu einzelnen AIPs eigegeben, die mit im Metadatensatz abgespeichert werden.

      Die so erfassten Metadaten werden als zusätzliche XML-Datei mit im AIP abgespeichert.

      Ein AIP umfasst somit:

    - Die Primärdateien

    - Der Prozessdokumentation in Form von PREMIS in METS-XM

    - Den beschreibenden Metadaten in Form einer XML-Datei

- Speichersystem

  Zur Unterstützung der physischen Speicherung integriert startext REPOSITORY die open source Komponente DSPACE (www.dspace.org). DSpace funktioniert mit allen Hardwarelösungen, die im Betriebssystem als Filesystem aufscheinen, unterstützt Versionierung von AIPs und identifiziert beim Speichern automatisch mehrfache Vorkommen einer Datei und speichert diese physisch nur einmalig ab.

- Data Management

  Als Datenmanager wird die startext Museumssoftware HiDA 4, oder die Archivsoftware ACTApro Desk genutzt. Hier werden alle Metadatensätze zu AIPs, die während des Ingest entstehen, in Form von XML gespeichert.

  Die Pflege der einzelnen Metadatensätze, insbesondere deren manuelle Anreicherung, erfolgt in

HiDA 4 bzw. ACTApro Desk. Aktualisierte Metadatensätze können zusätzlich redundant in das AIP gespeichert werden.

- Recherche

  Die Recherche von AIPs findet grundsätzlich in der Inventarisierungssoftware statt. Dadurch steht für Mitarbeiter und Nutzer eine übergreifende Recherche über soohl analoge, als auch digitale Objekte zur Verfügung.

  Neben allen Recherchemöglichkeiten, die herkömmlich für analoge Objekte verfügbar sind, ist für digitale Objekte die Recherche im Volltext der Primärdateien möglich.

  Unabhängig von der gewählten Vorgehensweise gelangt der Anwender so zu dem Metadatensatz eines AIPs. Hier kann er die Metadaten einsehen und ganze DIPs anfordern oder auch einzelne Primärdateien des AIPs direkt aufrufen.

  Sowohl ein vollständiges DIP als auch einzelne Dateien werden hierbei durch den DIP-Creator bereitgestellt.

- Outgest / Dip-Creator

  Der Zugriff auf ganze AIPs oder einzelne Primärdateien erfolgt grundsätzlich nicht direkt, sondern stets vermittelt über den so genannten DIP-Creator, der bedarfsangepasste Nutzungskopien der digitalen Inhalte erzeugt und bereitstellt.

  Der DIP-Creator funktioniert analog zum Ingest als konfigurierbare Verarbeitungskette mit austauschbaren Werkzeugen und ist ebenfalls in Microservicearchitektur realisiert und für parallelisierten Betrieb konzipiert.

- Preservation

  Auch die Durchführung von Erhaltungsmigrationen funktioniert analog zum Ingest als konfigurierbare Verarbeitungskette mit austauschbaren Werkzeugen, ist in Microservicearchitektur realisiert und für parallelisierten Betrieb konzipiert.

  Selbstverständlich werden auch hier alle Verarbeitungsschritte dauerhaft in PREMIS dokumentiert.

- Überblick

  Ein eigenes Web-Frontend gibt zum Einen den schnellen Überblick über archivierte Inhalte, Dateiformate, Mengengerüste und den Status von Ingestprozessen.

  Zum Anderen hat der Anwender hier den direkten Zugriff auf Detailinformationen einzelner Ingests oder AIPs, wie z.B. die Verarbeitungsprotokolle oder auch Fehlermeldungen.

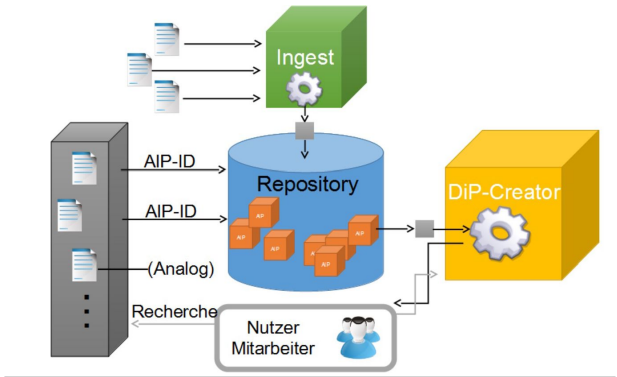

Abb. 1: Übersichtsschaubild startext REPOSITORY

3. WEBSEITENARCHIVIERUNG MIT PABLO

Webseitenarchivierung stellt im Bereich der digitalen Langzeitarchivierungen eine ganz eigene Problemklasse dar. Denn eine wesentliche Strategie der digitalen Langzeitarchivierung ist die Formatvereinfachung und –vereinheitlichung. Etablierte Formate und Verfahren existieren im Wesentlichen für Schriftdokumente und Bilder.

Für Schriftdokumente ist die Sicherung als PDF/A etablierter Standard. Bei einem PDF handelt es sich im Grunde um nichts anderes, als um eine besondere Form des Ausdruckens. Der Ausdruck erfolgt zwar in elektronischer Form, doch ein PDF bleibt letztendlich ein Druckformat. Die andere wichtige Objektklasse für die digitale Langzeitarchivierung sind Bilder, die in der Regel als unkomprimierte TIFFs abgespeichert werden.

Selbstverständlich ist auch der Bereich der audio-visuellen Datenformate Gegenstand von Archivierungsbemühungen. Doch etablierte Formate und Verfahren sind hier noch deutlich weniger entwickelt.

Aber darüber hinaus gibt es noch einen ganz eigenen Objektbereich: Webseiten.

Etablierte Standards und Verfahren finden sich zwar für die Speicherung (v.a. das WARC-Format), das Formatproblem der Archivierung ist hier jedoch nicht gelöst. Die Software PABLO implementiert einen Lösungsvorschlag zu dem Formatproblem der digitalen Langzeitarchivierung von Webseiten.

3.1 WAS SOLL DURCH DIE WEBSITEARCHIVIERUNG BEWAHRT WERDEN?

Hat man sich grundsätzlich zu einer Webseitenarchivierung entschlossen, stellt sich sogleich eine neue grundsätzliche Frage: Was soll durch die Websitearchivierung bewahrt werden?

Damit berühren wir den Punkt der signifikanten Eigenschaften, denn von einem digitalen Content kann häufig nicht alles bewahrt bzw. archiviert werden. Auch im Bereich der Schriftdokumente ist dies der Fall. Nehmen wir zum Beispiel ein Worddokument, in das ein dynamisches Feld – beispielsweise ein Tagesdatum - eingefügt wurde. Bei der Archivierung als PDF/A lässt sich lediglich eine Momentaufnahme – ein Snapshot des Dokumentes - sichern. Die Eigenschaft des Datumfeldes selbst, bei dem es sich um ein Element handelt, das sich dynamisch verändern bzw. aktualisieren kann, lässt sich dagegen nicht archivieren. Statt eines Worddokuments mit all seinen Eigenschaften wird also nur ein Ausdruck des betreffenden Worddokuments archiviert. Die gleiche Problematik stellt sich verschärft auch bei der Archivierung von Excel-Dateien, in denen beispielsweise Formeln hinterlegt sind.

Ein gewisser Verlust hinsichtlich der Dokumenteigenschaften wird in Kauf genommen, solange die jeweiligen signifikanten Eigenschaften nicht verfälscht werden.

Aber welche signifikanten Eigenschaften der Webseite sollen durch die Archivierung bewahren werden? Dies ist von Fall zu Fall hinsichtlich der jeweiligen Webseite individuell neu zu bewerten, grundsätzlich kann man aber vier signifikante Eigenschaften unterscheiden, auf die sich die Archivierung von Online-Inhalten fokussieren lässt:

- Text

Bei einigen Webseiten geht es praktisch ausschließlich um die dargestellte Information. Zum Beispiel, wenn Forschungsdaten archiviert werden sollen, die auf der Internetpräsenz einer Universität oder eines Forschungsinstituts

veröffentlicht wurden. In diesem Fall ist es in der Regel zweitrangig, wie diese Informationen dargestellt werden, wie das Layout der entsprechenden Webseite aussieht und wie die Webinhalte präsentiert werden.

- Darstellung

Wie stellt sich die Webseite dar? Wie sieht sie aktuell aus? Wie wirkt sie auf den Benutzer? Welche Funktionalitäten bietet sie dem Nutzer? In diesen Fällen muss die digitale Webseitenarchivierung eine Möglichkeit finden, die typischen Merkmale des entsprechenden Layouts bzw. des Designs zu speichern, also die durch heutige Browser erzeugte Darstellung.

- Verlinkung

Verlinkung, die Möglichkeit für den Nutzer, über Verknüpfungen von einer Seite zur nächsten zu navigieren, ist sicher eine Kerneigenschaft von Webpräsenzen, die in aller Regel zu bewahren ist.

- Interaktivität

Manche Webseiten leben davon, dass sie mit dem Benutzer interagieren und ausgefallene Interaktionsmöglichkeiten bieten. Zum Beispiel Inhalte, die darauf reagieren, wie sich der Benutzer auf der Webseite bewegt. Als Extrembeispiel für diesen Typ kann man auch Browserspiele in diese Gruppe mit einbeziehen. Die Bewahrung interaktiver Webseiten mit ihren Besonderheiten ist ein völlig offenes Problem in der digitalen Langzeitarchivierung.

PABLO bewahrt, bis auf die Interaktivität, alle signifikanten Eigenschaften von Webpräsenzen!

Zunächst einmal geht PABLO wie ein Webcrawler vor. Das Programm durchläuft einen Interauftritt von einer spezifischen Start-URL aus. Das kann eine komplette Webseite sein oder auch nur der Teil eines umfangreichen Internetauftritts. PABLO folgt dabei allen Links. Das ist an und für sich noch nichts Besonderes, denn es gibt auch noch andere Tools, die das können. Viel spannender ist vielmehr das Ausgabeformat, das PABLO liefert, denn PABLO erzeugt genau zwei Dateitypen: ein Bild und eine METS-XML Datei.

Abb. 2: Was macht PABLO?

3.2 PABLO VEREINFACHT DAS FORMAT RADIKAL

PABLO steuert einen Webbrowser (standardmäßig Mozilla Firefox) von außen an und macht von jeder einzelnen Webseite, die das Programm findet, ein Bild (sozusagen ein Foto). Diese Bilddatei - in einem individuell auswählbaren Format - bildet die Webseite genau so ab, so wie der Browser sie darstellt. Das Ausgabeformat ist bei PABLO alles andere als kleinteilig. Statt eines Sammelsuriums einzelner Teildateien, Gifs, JPGs, Textdateien etc. gibt es nur ein einziges Bild.

Die zweite Datei, die von PABLO erzeugt wird, ist eine XML-Datei. Diese Datei dient dazu, die digitalen Inhalte zu strukturieren und mit Metainformationen anzureichern. In der METS-Datei wird vor allem hinterlegt, wo in der Bilddatei der jeweiligen Webseite Verlinkungen sind und wo diese Verlinkungen hinführen.

Damit wird das Datei-Sammelsurium einer Webseite radikal vereinfacht und reduziert. Es gibt im Ergebnis nur noch zwei verschiedene Dateitypen: eine Bild- und eine METS-XML-Datei. Diese sind von der Struktur her so einfach, dass sie über jeden Technologiewechsel hinweg bewahrt werden können.

3.3 PABLO ERZEUGT EINE PRÄSENTATIONSFORM DER ARCHIVIERTEN WEBSEITE

Das erzeugte Format ist aber gleichzeitig so vollständig, dass daraus eine navigierbare Reproduktion der Webseite erstellt werden kann. Nutzer können durch diese Webseite surfen mit einem der ursprünglichen Seite vergleichbaren Nutzererlebnis. Und selbstverständlich kann man die Seite auch als Ganzes nutzen. Links lassen sich anklicken und man gelangt zu der entsprechenden Folgeseite.

SPRECHEN SIE UNS AN!

startext GmbH
Dottendorfer Straße 86
53129 Bonn
Tel: 0228 95996 0, Fax: 0228 95996 66
info@startext.de
www.startext.de