Kathrin Schroeder

Format fürs Archiv

Das Bundesarchiv lagert inzwischen auch elektronische Dokumente. Dazu wurde ein Digitales Archiv aufgebaut. Das Format PDF/A spielt dabei eine wichtige Rolle.

Der Umgang mit Formaten im Kontext der dauerhaften Archivierung ist eine Herausforderung für das Bundesarchiv, da elektronische Daten in sehr heterogenen Strukturen und Formaten übergeben werden. Eine Lieferung aus einem Fachverfahren einer Behörde beinhaltet etwa 65 unterschiedliche Dateiformate. Erhaltungsstrategien im Sinne des OAIS-Referenzmodells (Open Archival Information System) für diese Dateiformate zu realisieren, wäre nicht sinnvoll. Aus diesem Grund wird im Digitalen Archiv die Begrenzung auf möglichst wenige Archivierungsformate verfolgt, um sicher gewährleisten zu können, dass Daten dauerhaft interpretierbar und zugriffsfähig bleiben. PDF/A nimmt dabei eine zentrale Rolle ein.

Empfehlungen für PDF/A

Mittlerweile wird PDF/A als Archivierungsformat in sehr unterschiedlichen Projekten und Branchen wie dem Airbus Projekt A380 oder beim Europäischen Patentamt, in der Versicherungsbranche, Automobilindustrie oder im Gesundheitswesen angewendet. Die Office-Suite OpenOffice.org unterstützt ab Version 2.4 PDF/A als Exportformat. Das Bundesministerium des Innern empfiehlt auch PDF/A als Format für den Dokumentenaustausch in der Bundesverwaltung. In Version 4.0 der Standards und Architekturen für E-Government-Anwendungen (SAGA 4.0) wird PDF/A für die dauerhafte Archivierung empfohlen. Eine Herausforderung bleibt dennoch die Konvertierung aus verschiedenen Ausgangsformaten in das Archivierungsformat PDF/A.

Situation im Bundesarchiv

Das Bundesarchiv mit Hauptsitz in Koblenz unterhält acht Standorte mit rund 800 Mitarbeitern, die dafür sorgen, dass auch künftige Generationen zu historisch wertvollen Dokumenten der Bundesrepublik, der DDR und des Deutschen Reichs Zugang haben. In der Abteilung B (Bundesrepublik Deutschland) laufen seit über 20 Jahren digitale Unterlagen der verschiedenen Bundesbehörden zusammen. Der Datenbestand an originärem digitalen Archivgut, das bereits in elektronischer Form entstanden ist, umfasst über neun Millionen Dateien, die zu 203 digitalen Archivobjekten, also inhaltlich zusammengehörenden Datenkomplexen, zusammengestellt wurden.

Die vorhandene Infrastruktur war jedoch nicht für die Übernahme elektronischer Akten ausgelegt und stieß zudem an die Grenzen ihrer Kapazität. Deshalb wurde ein Projekt initiiert, welches das Ziel hatte, ein Digitales Archiv aufzubauen, um dadurch in der Lage zu sein, der gesetzlichen Verpflichtung in effizienter Weise nachzukommen. Gefragt war ein neuer Ansatz für die digitale Langzeitarchivierung.

Mitte Oktober 2008 ging ein OAIS-konformes Archivinformationssystem im Bundesarchiv in den Echtbetrieb. Diese Lösung wurde mit Hewlett-Packard (HP) als Generalunternehmer und der Partner-Firma SER geschaffen. Damit ist es möglich, digitale Dokumente zu bewerten, ins Archivsystem zu übernehmen sowie dauerhaft zu sichern und zu nutzen.

Die Praxis, dass Bundesbehörden ihre digitalen Unterlagen dem Bundesarchiv zur dauerhaften Aufbewahrung übergeben, hat gerade erst begonnen. Um in Zukunft re-

Kurz gefasst

Bis vor Kurzem war das Bundesarchiv in Koblenz für die Übernahme elektronischer Akten nicht gerüstet. Jetzt gibt es ein Digitales Archiv, das bereits neun Millionen Dateien an originärem digitalen Archivgut enthält. Der Beitrag stellt das Projekt vor und nennt die Vorteile des Formats PDF/A bei der Langzeitarchivierung.

gelmäßig aus verschiedenen Quellen elektronische Akten in großem Umfang übernehmen zu können, ist das System so angelegt, dass es eine hohe Skalierbarkeit bietet, sowohl in Bezug auf die Datenmengen als auch auf die Einbindung weiterer Behörden.

Die Autorin: Kathrin Schroeder

Foto: Privat

Kathrin Schroeder arbeitet seit 2006 im Bundesarchiv vorrangig für das Projekt Digitales Archiv. Sie ist Mitglied des Kompetenzteams BBEA – Bundesarchiv Behördenberatung elektronische Akten. Zuvor war Schroeder bei der Deutschen Nationalbibliothek im Bereich digitaler Langzeitarchivierung beschäftigt.

Konvertierung nach PDF/A

Daten aufbereiten, archivieren und den Zugriff sichern: Dies ist der wesentliche Ablauf, den die Architektur des Systems mit einer Workflow- und Archivkomponente abbildet. Dabei ist die Formatkonvertierung nach PDF/A als Teilprozess an die Eingangsbearbeitung (Ingest-Prozess) gekoppelt, der in der Workflow-Komponente implementiert ist. Dadurch kann eine hohe Qualität der Daten sichergestellt werden, wenn diese durch einen definierten Eingangsprozess aufbereitet werden.

Endet die Aufbewahrungsfrist für elektronische Akten der einzelnen Bundesbehörden, bieten diese ihre Daten dem Bundesarchiv an. Gesteuert werden die Prozesse durch das Standard-Archivierungsmodul (SAM), dessen Entwicklung im Rahmen der E-Government-Initiative des Bundes finanziell gefördert wurde. Das Datenangebot wird automatisch in den Workflow übernommen und dem zuständigen Referat zur Bewertung zugesendet. Welche Daten archivwürdig sind und welche sofort vernichtet werden können, prüfen die zuständigen Mitarbeiter aus den Fachreferaten auf Basis der Metadaten. Die abgebende Behörde liefert auf der Grundlage der archivischen Vorbewertung daraufhin die angeforderten Daten, die in den Workflow übernommen werden und dort den Prozess der Eingangsbearbeitung durchlaufen. Die Aussonderung besteht aus Metadaten im XML- oder CSV-Format und den Primärdaten. Diese liegen im Idealfall im Format PDF/A vor. Sollten die Primärdaten in einem nicht archivfähigen Format übergeben werden, wird automatisch ein Konvertierungsprozess ausgelöst, in dessen Rahmen diese Daten nach PDF/A konvertiert werden.

Für eine größtmögliche Flexibilität war die Konzeption einer Architektur für die Konvertierungskomponente erforderlich. Diese sieht vor, dass Daten zur Konvertierung in ein definiertes Verzeichnis aus dem laufenden Prozess heraus exportiert werden. Anschließend werden je nach Ausgangsformat die entsprechenden Konverter automatisch angesteuert. Das Ergebnis der Konvertierung wird in dem Archivierungsformat XBARCH dokumentiert. Nach Abschluss der Konvertierung werden die Daten wieder dem laufenden Prozess übergeben. Diese Architektur erlaubt es, auf zukünftige Formate zu reagieren sowie neue Konverter mit wenig Aufwand einfach einzubinden.

Sind die Daten auf diese Weise vorbereitet und einer Qualitätskontrolle unterzogen, erfolgt die Übergabe an die Archivkomponente. Dort wird auf Basis der Metadaten der Index für die Recherche aufgebaut. Die Archivpakete (AIP – Archival Information Package) werden auf Speicherplatten und parallel auf Band gespeichert. Die abgebende Behörde erhält anschließend eine Archivierungsbestätigung. Nach einem definierten Zeitraum verbleiben die Pakete lediglich auf dem Band, dem kostengünstigeren Speichermedium. Von dort lassen sie sich allerdings bei Bedarf jederzeit auf schnelle Disk-Medien zurückholen.

Ausblick

Der Prozess der Eingangsbearbeitung kann erheblich vereinfacht werden, wenn die Daten bereits in einem Archivierungsformat, vorzugsweise PDF/A vorliegen. Das Bundesarchiv ist in der Lage, Daten aus heterogenen Ausgangsformaten zu verarbeiten. Das Digitale Archiv hat im Oktober 2008 in der ersten Ausbaustufe (40 Terabyte) seinen regulären Betrieb aufgenommen. Das Speichersystem ist bis zu 400 Terabyte skalierbar. Bis 2010 wird das Bundesarchiv in Zusammenarbeit mit den Firmen HP und SER das System stufenweise erweitern. So ist gewährleistet, dass auf zukünftige Änderungen reagiert werden kann. ◄