Matthias Groß

Digitale Langzeitarchivierung mit Rosetta im Bibliotheksverbund Bayern

Die digitale Langzeitarchivierung stellt eine der größten Herausforderungen für die Bibliotheken im 21. Jahrhundert dar. Der Vortrag beschreibt den im Bibliotheksverbund Bayern verfolgten Ansatz mit dem Digital-Preservation-System Rosetta von Ex Libris, das an der Bayerischen Staatsbibliothek in München eingeführt und von weiteren bayerischen Universitätsbibliotheken im Vorfeld eines künftigen Verbundeinsatzes erprobt wird. Dabei werden insbesondere folgende Aspekte eingehend dargestellt: die Skalierbarkeit des Systems hinsichtlich großer Datenvolumina bei zudem komplex strukturierten Objekten; besondere Herausforderungen des Einsatzes eines digitalen Langzeitarchivs auf Verbundebene; hierzu gehört auch die Anbindung an den Verbundkatalog (Aleph 500); sowie der „light archive"-Ansatz mit Endnutzerpräsentation.

Der Herbst ist die angemessene Jahreszeit, um über die Vergänglichkeit nachzudenken: Die Blätter fallen von den Bäumen – sie fallen manchmal auch aus sauren Büchern –, aber auch das Sammeln und die Aufbewahrung sind in dieser Jahreszeit fest verankert, das Eichhörnchen kann uns hier als Leitfigur dienen. Auch für die digitale Langzeitarchivierung ist der Herbst eine inspirierende Jahreszeit.

Wenn wir uns zunächst der *Haltbarkeit von Datenträgern* zuwenden, so stimmt es nachdenklich, dass eine aus einem Abbruchhaus wieder ans Tageslicht geförderte Zeitung, die zufällig irgendwo hineingerutscht war, nach 55 Jahren immerhin noch zu wesentlichen Teilen lesbar ist, während eine CD-ROM in der freien Natur nur wenige Tage bis Wochen überlebt. Räumt man ein, der Vergleich sei unfair, so zeigt die Langzeitbetrachtung einiger Zeitungsseiten, die ein Sturm hinter den Schneefang eines Nachbarhauses geweht hatte, doch auch hier einen Abbauzeitraum von mehreren Jahren trotz aller Witterungsunbilden. Zudem ist dem digitalen Datenträger der tückische Prozess des Verfalls, der sich bei immer höherer Datendichte schon im mikroskopischen Bereich verheerend auswirkt, äußerlich nicht ohne Weiteres anzusehen. Und selbst wenn die Datenträger lange halten, so sorgt doch der stete Innovationsschub für einen raschen Wandel bei den marktüblichen Medien und Endgeräten. Daher ist man schon vor einigen Jahren dazu übergegangen, für Zwecke der Archivierung die Daten vom physischen Träger zu lösen und auf eine abstrakte Speicherschicht zu übertragen, deren konkrete Ausprägung jeweils in gewissen Abständen ausgetauscht werden muss, die Daten werden also alle paar Jahre auf die dann aktuellen Trägermedien umkopiert. Diesen Vorgang bezeichnet man in der Literatur als *bitstream preservation* und stellt sich dabei die Dateien als Folgen von Nullen und Einsen vor, die es unverändert durch die Zeit zu transportieren gilt. Der aktuelle Lösungsansatz besteht im Wesentlichen aus der periodischen Migration, zur Sicherstellung der Unversehrtheit der Informationen kommen Prüfsummen-

verfahren zum Einsatz. Die aktuelle skalierbare Basistechnologie dafür stellen Bandrobotersysteme, sogenannte *tape libraries*, dar; diese werden außer für die eigentliche Archivierung auch für die kostengünstige Aufbewahrung von Backups eingesetzt. Das Bandarchiv registriert dabei innerhalb seiner Verwaltungsschicht, wie oft auf jedes Band zugegriffen wurde, um durch rechtzeitiges Umkopieren der Inhalte auf neue Bänder das Risiko zu großen mechanischen Verschleißes auszuschalten. Für die *bitstream preservation* liegen bereits umfassende Erfahrungen für die Migration großer Datenbestände vor.

Das gewichtigere Teilproblem der digitalen Langzeitarchivierung ist jedoch: Was bedeuten diese Folgen von Nullen und Einsen, wie sind sie zu interpretieren und darzustellen? Es geht einher mit dem Veralten von Software und deren Standards. Auch bei diesem Problem begegnen wir wieder dem Ansatz der Migration, jedoch nicht unbedingt als zeitlichem Hintereinander von jeweils allein verfügbaren Darstellungen, sondern mit der Möglichkeit, verschiedene Repräsentationen derselben intellektuellen Entität gemeinsam in einem System zu verwalten. Wenn wir den berühmten Stein von Rosetta mit seiner Inschrift in dreierlei Gestalt betrachten, so können wir uns vorstellen, dass man zunächst für diejenigen, die den Hieroglyphen-"Master" nicht mehr unmittelbar lesen konnten, eine erste *Formatmigration* durchgeführt hat, bezeichnenderweise ins Demotische, also eine Volksausgabe im wahrsten Sinne des Wortes; um den Text noch weiteren Nutzerkreisen zu erschließen, hat man außerdem noch ein griechisches Derivat hinzugefügt. Dies illustriert, dass man bei Formatmigrationen die bisherigen Daten nicht unbedingt verwirft, sondern die neue Repräsentation dem Bestand hinzufügen kann. In der Praxis wird man hierbei je nach Zweck und Begleitumständen pragmatische Entscheidungen zu treffen haben, zumal aus Kostengründen. In der Regel wird man aber die originale Datei stets mitführen, nicht zuletzt, um bei aller Sorgfalt erst nachträglich festgestellte Defizite bei der Migration durch den Einsatz verfeinerter Methoden ausgleichen zu können. Im Rahmen dieses kurzen Überblicks kann nicht näher auf die Einzelheiten und Herausforderungen der Format- und Risikoanalyse eingegangen werden.

Die logische Herausforderung, wie man die Bedeutung der Daten durch die Zeit transportiert, wird zudem durch die Dimension *Datenvolumen* verstärkt. Konkret im Fall der Bayerischen Staatsbibliothek ist von August 2010 bis August 2011 das digitale Archiv um 33 Prozent gewachsen, der absolute Zuwachs betrug 80 Terabyte, die Zahl der Einzeldateien wuchs von 476 auf 612 Millionen. Dies bedeutet für die Einführung eines Systems mit dem Anspruch, die digitalen Sammlungen im Sinne der digitalen Langzeitarchivierung zu verwalten, also detaillierte Kenntnisse über die Formate der einzelnen Dateien zu besitzen und auf dieser Grundlage die absehbaren Risiken aus Sicht der Einrichtung zu evaluieren, zunächst, dass es einen jährlichen Zuwachs von hier derzeit konkret 150 Millionen Einzeldateien im laufenden Betrieb verarbeiten können muss und daneben möglichst zügig das bereits vorhandene Material in das System hineinzumigrieren ist. Das ist bereits ohne die Be-

trachtung weiterer Bibliotheken in Bayern und ohne die parallele Anzeige von Objekten für Endnutzer eine große Herausforderung.

Wie sind in Bayern die *organisatorischen Rahmenbedingungen* für die digitale Langzeitarchivierung beschaffen? Das ist ein Aspekt, der neben der Technik ganz besonders wichtig ist: Wer sind die Partner, die die Verantwortung übernehmen, wie sind sie organisatorisch eingebunden, und wie kann auch die Finanzierung nachhaltig gestaltet werden – es handelt sich schließlich, wie die Bezeichnung schon verrät, um eine Daueraufgabe.

Die *Bayerische Staatsbibliothek* hat als zentrale Landes- und Archivbibliothek des Freistaats Bayern den Auftrag, ihre Bestände nicht nur hier, heute und jetzt, sondern auch in Zukunft zur Verfügung zu stellen und das kulturelle Erbe des Landes zu bewahren. Sie hat ein regionales Pflichtexemplarrecht, das bis in das 17. Jahrhundert zurückgeht und sich nun auch auf die Ablieferung amtlicher Druckschriften in elektronischer Form erstreckt, bei einer Novellierung des bayerischen Pflichtstückegesetzes wird es auch elektronische Verlagspublikationen berücksichtigen. Im Münchener Digitalisierungszentrum (MDZ) produziert sie selbst eine große Anzahl von Retrodigitalisaten oder lässt diese durch Partner herstellen. Hieraus speist sich der Löwenanteil des digitalen Archivs, eine zentrale Rolle spielt dabei natürlich die Public Private Partnership mit *Google*. Aber auch die Inhouse-Digitalisierungsprojekte etwa für das 16. Jahrhundert oder für Handschriften sind Massendigitalisierungsprojekte und tragen durch ihre spezifischen Formatfestlegungen wesentlich zum Datenvolumen bei. Das Referat Digitale Bibliothek der Bayerischen Staatsbibliothek führt zudem seit über zehn Jahren Projekte zu wichtigen Aspekten der digitalen Langzeitarchivierung durch und ist in maßgebliche Kooperationen eingebunden.

Teil der Bayerischen Staatsbibliothek ist auch die *Verbundzentrale des Bibliotheksverbunds Bayern (BVB)*, sie hat die Aufgabe, EDV-Infrastrukturen für die wissenschaftlichen Bibliotheken in Bayern bereitzustellen. Im Vorfeld der neuen Herausforderungen der digitalen Langzeitarchivierung konnten hier bereits wichtige Erfahrungen mit dem Betrieb des bayerischen Multimedia-Servers gesammelt werden, der auf dem Produkt *DigiTool* der Firma Ex Libris basiert. So war die Ansiedlung des technischen Betriebs der neu einzuführenden Anwendung für die digitale Langzeitarchivierung bei der Verbundzentrale bereits vorbereitet.

Ein wichtiger Partner ist das *Leibniz-Rechenzentrum (LRZ)* der Bayerischen Akademie der Wissenschaften, das einerseits als Höchstleistungszentrum auf nationaler und internationaler Ebene agiert und zum anderen das Hochschulrechenzentrum für die Münchener Universitäten darstellt. Seit 2008 nimmt es zudem im Zuge der Neustrukturierung der EDV-Betriebszentren in der bayerischen Staatsverwaltung die Aufgabe des Verbundrechenzentrums des BVB wahr. Das LRZ ist langjähriger Partner der Bayerischen Staatsbibliothek für die digitale Langzeitarchivierung, bislang auf der Ebene der *bitstream pre-*

servation, und ermöglicht mit seinen skalierbaren Infrastrukturangeboten überhaupt erst ein Engagement in dieser Dimension.

Die zentrale Klientel für die Verbundzentrale sind die *Hochschulbibliotheken*, und im Kontext der digitalen Langzeitarchivierung sind auch die Hochschulrechenzentren mit zu betrachten. Eine wesentliche Grundlage für das bibliothekarische Handeln in Bayern ist die spezifische Form der Kooperation, die sich in den verschiedenen Verbundgremien über lange Jahre etabliert hat; diese hat 2006 auch Eingang in die Neufassung des Bayerischen Hochschulgesetzes gefunden (Art. 16, Abs. 1). Somit hat der Ansatz, einen breiten fachlich fundierten Konsens herzustellen und diesen flächendeckend umzusetzen, die Würdigung seitens des Gesetzgebers erlangt. Das Interesse der Hochschulbibliotheken an der digitalen Langzeitarchivierung richtet sich zum einen auf Digitalisate unikaler Bestände, sodann in natürlicher Weise auf Hochschulschriften – in einem weiten Sinn verstanden –, sowie für erweiterte Möglichkeiten, innerhalb der eigenen Hochschule als Dienstleister fungieren zu können, dem man traditionell bei der langfristigen Aufbewahrung von Inhalten vertraut. Hier tritt insbesondere die Archivierung von Forschungsprimärdaten verstärkt ins Blickfeld, bis zu einer Konkretisierung hin zu echten Kooperationen ist jedoch noch eine gewisse Wegstrecke zu bestreiten, die Problem- und Zielschärfung einerseits und das Heranwachsen eines Vertrauensverhältnisses andererseits umfasst.

Die *Auswahlkriterien* für ein System für die digitale Langzeitarchivierung für die Bayerische Staatsbibliothek und den Bibliotheksverbund Bayern umfassten insbesondere den bereits betonten Aspekt der Skalierbarkeit, sodann das Sicherheits- und Berechtigungskonzept, da ja Objekte verschiedener Einrichtungen und mit unterschiedlichen Rechten zu verwalten sind, die Beachtung von Standards wie dem OAIS-Referenzmodell und das Vorhandensein offener oder gut dokumentierter Schnittstellen sowie den Grad der Unterstützung von Maßnahmen zum Erhalt der Verfügbarkeit, insbesondere Formatmigrationen. Zudem sollte es kein sogenanntes *dark archive* für die reine Archivierung, sondern ein *light archive* mit direktem Zugriff für die Öffentlichkeit auf die Objekte nach Maßgabe der jeweiligen Nutzungsrechte sein. Es sei an dieser Stelle darauf hingewiesen, dass die Endnutzerderivate zwar nicht das im engeren Sinne zu archivierende Gut darstellen, aber da bei ihrem Verlust eine Neugenerierung aus den Masterdateien mit sehr hohem Aufwand an Rechenzeit verbunden wäre, sind sie doch zwingend mit einzubeziehen.

Die Entscheidung fiel 2009 auf das Produkt *Rosetta Digital Preservation System*, dessen Vorgeschichte 2004 in Neuseeland begann, als man sich dort die Frage nach der Zukunft des nationalen digitalen Erbes gestellt hat, woraus über eine sehr ausführliche Spezifikationsphase die Produktentwicklung mit der Firma Ex Libris resultierte. Der produktive Einsatz von Rosetta in Neuseeland begann 2008.

Das Rosetta-Implementierungsprojekt in Bayern startete operativ im Februar 2010. In der Anfangsphase stand die Analyse und Beschreibung der abzubildenden Workflows im Vordergrund. Man kann annähernd davon ausgehen, dass der Projektaufwand linear in der Zahl der umzusetzenden Workflows ist, und je weniger klar diese zu Beginn definiert sind, um so mehr Iterationen muss man im Projekt durchlaufen. Konkret bedeutet etwa die Frage, ob der OCR-Volltext mit in die Langzeitarchivierung einbezogen werden soll, eine solche Weichenstellung, die sich etwa auch auf die Ausgestaltung der Strukturmetadaten auswirkt.

Schon frühzeitig wurde ein sogenanntes Trainingssystem auf einem einzelnen virtuellen Server eingerichtet, womit die Schulungen des Projektteams, das sich aus dem Referat Digitale Bibliothek / Münchener Digitalisierungszentrum sowie aus der Verbundzentrale des Bibliotheksverbunds Bayern zusammensetzt, durchgeführt und die grundlegende Systemkonfiguration erarbeitet wurden. Sodann wurde eine Mehrserverumgebung als eigentliches Projektsystem installiert, da die gemäß Zielsetzung zu verarbeitenden Datenvolumina nur durch Skalierung auf der Hardwareseite zu bewältigen sind. Sie umfasst derzeit vier Applikationsserver, die als virtuelle Server betrieben werden, sowie zwei Datenbankserver, die als physische Server realisiert sind. Es schlossen sich iterative Funktions- und Lasttests mit den jeweils alle zwei bis drei Monate herauskommenden neuen Software-Releases an; später im Betrieb wird man das System seltener aktualisieren.

Es geht zunächst um die Umsetzung von drei Workflows der Bayerischen Staatsbibliothek:

(1) Massendigitalisierung, hier kommen als Dateiformate TIFF und JPEG2000 für die Masterdateien vor,

(2) digitale Pflichtablieferung von amtlichen Druckschriften, in der Regel im PDF-Format, sowie

(3) Website-Harvesting im Kontext der Sondersammelgebiete für fachlich ausgewählte Ressourcen mit Einwilligung der Rechteinhaber; dies ist zwar von der Anzahl der Objekte (auf der Ebene der Websites betrachtet) der kleinste Workflow, aber mit einer hohen Anzahl von Einzeldateien mit einer ganz besonders hohen Formatvielfalt und daher sehr spannend im Hinblick auf Formatanalyse und -migrationen.

Die Herausforderung der Skalierbarkeit bedeutet für uns ganz konkret, dass das System inklusive eines Puffers für die Bewältigung von Spitzenlasten täglich mindestens das Äquivalent von 1000 digitalen Büchern à 300 Seiten verarbeiten können muss; mit den unterschiedlichen Repräsentationen ergibt das über eine Million Dateien, von denen jede analysiert und geprüft werden muss. Auf diesen Durchsatz hin ist die gesamte Infrastruktur inklusive Netzwerkanbindung und Firewall zu optimieren. Auch das System selbst war an einigen Stellen anzupassen, die Firma Ex Libris hat hier bereits Verbesserun-

gen in das Produkt einfließen lassen. Im Vergleich mit den Erfahrungen anderer Anwender hat sich gezeigt, dass insbesondere die große Zahl von Dateien innerhalb einer intellektuellen Einheit die spezifische Herausforderung unseres Projekts darstellt, nicht das Datenvolumen an sich.

Für den *Verbundeinsatz* ist es wichtig, dass der Betrieb in mehreren Bibliotheken unterstützt wird, also die Mandantenfähigkeit gegeben ist. Wir haben dies bislang durch die Simulation zweier Bibliotheken intern getestet und werden dies in eine Phase mit Tests mit mehreren echten Bibliotheken überführen. Im Gegensatz zu unserem DigiTool-Betriebsmodell, das dem vom Verbundkatalog her bekannten Prinzip der Kooperation folgt, wird man bei der digitalen Langzeitarchivierung schon allein vom Anspruch der Aufgabe her stärker beschränken müssen, welcher Mitarbeiter Daten wirklich verändern kann. Zudem stellt sich für den Verbundeinsatz die Frage der dauerhaften Finanzierung auf landesweiter Ebene.

Bei der Einführung von Rosetta geht es auch um die Integration in die bestehende technische Systeminfrastruktur. Die Endnutzer kommen an die Objekte zunächst über Links aus dem Verbundkatalog sowie ihren lokalen Bibliothekskatalogen, die daraus versorgt werden; dazu wird Aleph 500 in zwei Richtungen mit Rosetta verbunden, zum einen werden die bibliographischen Metadaten aus dem Verbundkatalog B3Kat in Rosetta nachgenutzt, zum anderen haben wir insbesondere für den Workflow der digitalen Pflichtablieferung mit Ex Libris eine Möglichkeit entwickelt, die mit einem Objekt mitgelieferten Metadaten nach Aleph zu transportieren und damit automatisiert ein Rumpfkatalogisat anzulegen, welches später je nach Kapazität und Bedarf hochkatalogisiert werden kann. Diese zweite Richtung steht technisch bereit, ist aber vor dem praktischen Einsatz noch bibliothekarisch-fachlich auszugestalten. Später ist die Volltextindexierung der Objekte mit den geeigneten Mitteln anzugehen.

Da der Betrieb eines digitalen Langzeitarchivs, das seinem Namen gerecht werden will, auch über die Lebensdauer eines Produkts hinaus gesichert durchzuführen ist, kommt dem sogenannten *Exit-Szenario* (man denke an Jona und den Walfisch) ganz wichtig und muss bereits bei der Implementierung eines Systems berücksichtigt werden. Bei Rosetta werden alle wesentlichen Informationen in einer Ordnerstruktur im Filesystem abgelegt, so dass mit einigen Kenntnissen bei Bedarf alles Wesentliche auch ohne das System direkt von dort abgegriffen werden kann.

Jenseits der Täler und Schluchten, die ein Einführungsprojekt mit sich bringt, umfasst der Ausblick auf die nächsten Monate an Gipfeln und Höhepunkten den stufenweisen Übergang des Systems in den Produktionsbetrieb sowie die Aufnahme von Tests mit mehreren bayerischen Universitätsbibliotheken als Baustein für die angestrebte Ausweitung des Betriebs. Daneben wird sich zeigen, wie sich die realen Möglichkeiten der Formatanalyse und -migration

nicht nur in Bayern, sondern in der weltweiten Community, die sich um das Produkt Rosetta herum bildet, ganz konkret darstellen.

Literaturverzeichnis

Brantl, Markus; Ceynowa, Klaus; Groß, Matthias; Reiner, Bernd; Schoger, Astrid: Digitale Langzeitarchivierung in Bayern : Vom Projekt zum nachhaltigen Modell. In: Bibliothek Forschung und Praxis (BFP), Jg. 35 (2011), S. 15-25.