LANGZEITARCHIVIERUNG

Sicherung des Weltkulturerbes am Leibniz-Rechenzentrum

DIE LANGFRISTIGE UND NACHHALTIGE AUFBEWAHRUNG DIGITALER KULTURGÜTER FÜR NACHFOLGENDE GENERATIONEN IST EINE GROSSE HERAUSFORDERUNG UND EINE INVESTITION IN DIE ZUKUNFT.

BSB

Eine Handschrift aus dem Bestand der Bayerischen Staatsbibliothek.

VON BERND REINER

Zum Aufgabengebiet des Leibniz-Rechenzentrums (LRZ) gehört unter anderem die Bereitstellung der Sicherungsumgebung für die Münchner Hochschulen und für die Bayerische Akademie der Wissenschaften. Unter dem Motto „quadratisch, praktisch, sicher“ lässt sich die neue Heimat der Backup- und Archivierungsumgebung des LRZ umschreiben.

Seit dem Umzug des LRZ im Frühjahr 2006 von der Münchener Innenstadt nach Garching steht für die Archiv- und Backupdaten eine Etage des nach modernsten Gesichtspunkten ausgestatteten „Rechnerwürfels“ zur Verfügung.

Das LRZ betreibt drei große robotergesteuerte Archiv- und Backupsysteme unter dem Softwarepaket IBM Tivoli Storage Manager (TSM). Gegenwärtig (Stand Juni 2007) werden über 4.450 Systeme auf Magnetbänder am LRZ gesichert, wobei der tägliche Dateneingang dieser Systeme am LRZ zwischen 6 bis 8 Terabyte liegt. Betrachtet man die Entwicklung des Datenvolumens über die letzten Jahre hinweg, so sind enorme Zuwächse zu verzeichnen. Lag das am LRZ gesicherte Datenvolumen im Januar 2001 noch bei ca. 40 Terabyte, so beträgt das aktuelle Datenvolumen im Archiv- und Backupsystem derzeit ca. 2,3 Petabyte (= 2.300 Terabyte =

2.300.000 Gigabyte). Das entspricht einer gewaltigen Steigerung um den Faktor 58. Die momentane Anzahl der gesicherten Dateien beläuft sich auf 2,5 Milliarden. Für die zukünftige Entwicklung wird mit einer Verdoppelung des Speichervolumens durchschnittlich alle 1,5 Jahre gerechnet.

Bibliotheken im Wandel

Aus traditioneller Sicht sind Bibliotheken und staatliche Archive zuständig für die Langzeitarchivierung von Dokumenten jeglicher Ausprägung. Die klassische Aufbewahrung von Büchern, Zeitschriften und Manuskripten in ihrer natürlichen Form hat sich seit einigen Jahren gewandelt und ausgeweitet. Elektronische Dokumente und Daten in digitaler Form nehmen im Wissenschaftsbetrieb wie auch im gesellschaftlichen Leben insgesamt einen immer höheren Stellenwert ein. Oftmals wird, wie z. B. bei Dissertationen und amtlichen Publikationen, auf ein gedrucktes Exemplar ganz verzichtet. Während die Digitalisierung dem Nutzer den Zugang und den Umgang mit der Information beschleunigt und insgesamt erleichtert, entstehen den Bibliotheken weltweit dadurch, sowohl aus organisatorischer und rechtlicher, als auch aus technischer Sicht, neue Herausforderungen.

Zusätzlich sind die Bibliotheken mit einer jährlich stark steigenden Anzahl von digitalen Objekten konfrontiert. Hierzu zählen nicht nur neue Veröffentlichungen in digitaler Form, sondern auch Digitalisierungsmaßnahmen von Altbeständen der Bibliotheken zur Sicherung des Weltkulturerbes. Diese Vielzahl an Objekten soll nicht nur verwaltet und gespeichert, sondern auch langfristig und nachhaltig zugänglich gemacht werden. Eine solche Aufgabe wird erschwert durch den raschen technologischen Wandel im Bereich der Hard- und Software und durch die natürlichen physikalischen Grenzen der Datenträger. Für die Bibliotheken besteht die Notwendigkeit, neben den konventionellen, traditionellen Archiven elektronische Archivsysteme aufzubauen und zu betreiben.

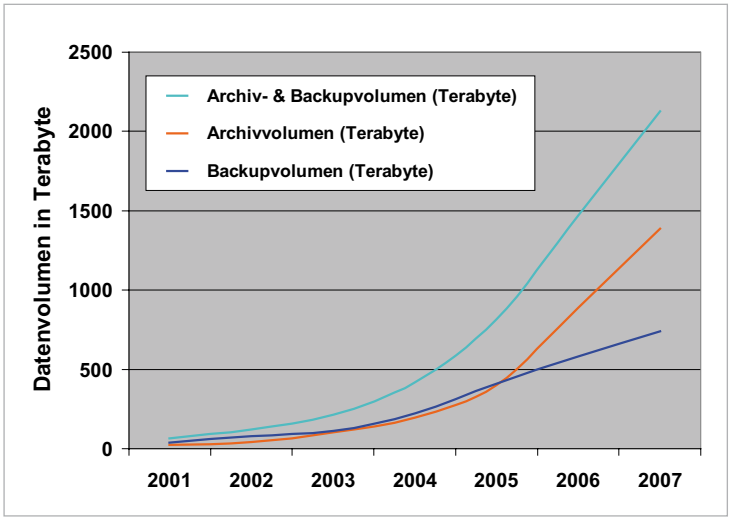
Fruchtbare Kooperation

Für Bibliotheken wie die Bayerische Staatsbibliothek (BSB) sind der Aufbau und ein nachhaltiger Betrieb einer professionellen Archivierungsumgebung nicht zuletzt wegen der fehlenden technischen Infrastruktur nur sehr schwer allein zu stemmen. Hier bietet sich eine Kooperation zwischen der BSB und einem Rechenzentrum wie dem LRZ mit seiner ausgeprägten Archivierungs- und Backupumgebung mit robotergesteuerten Magnetbandsystemen geradezu an.

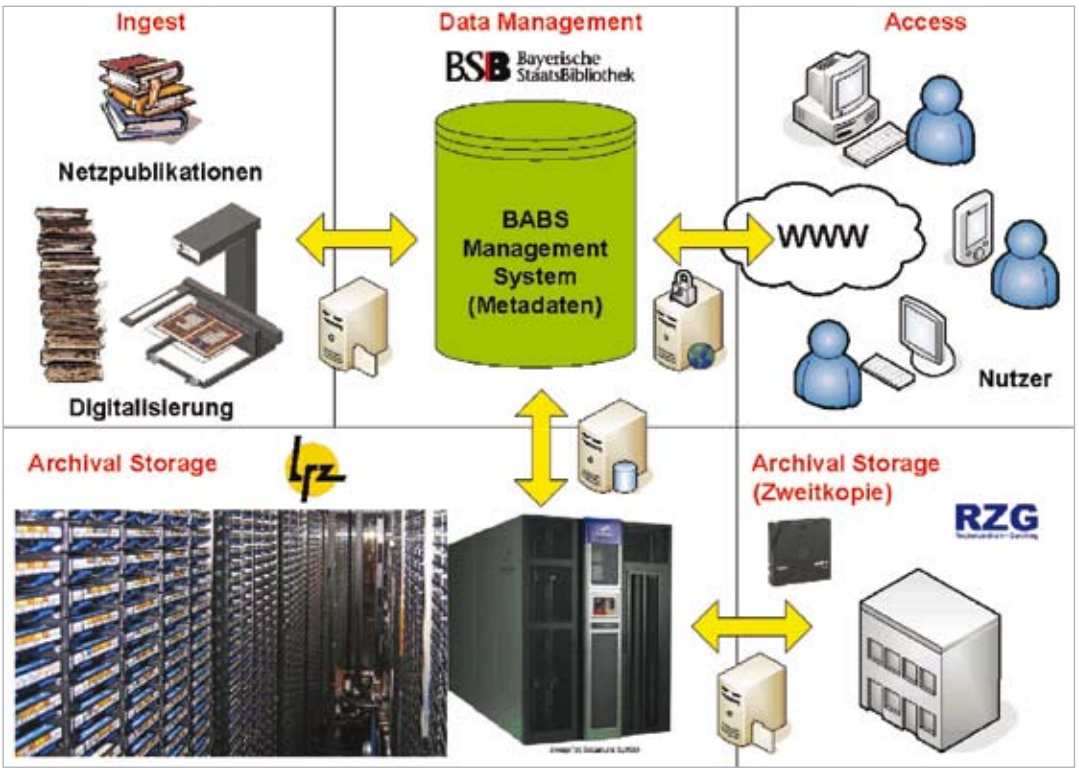
LRZ

Der „Rechnerwürfel" des Leibniz-Rechenzentrums der Bayerischen Akademie der Wisenschaften in Garching.

Die Grundidee hierbei ist, dass jeder Partner genau das tut, was er am besten kann. Während das Rechenzentrum die langfristige Archivierung der digitalen Daten im eigentlichen Sinne (Archival Storage) und die regelmäßig durchzuführenden Erhaltungsmaßnahmen (Preservation Planning) übernimmt, kann sich die Bibliothek auf die bibliothekarischen Aufgaben konzentrieren. Zu diesen Aufgaben zählen die Sammlung, Erschließung (Ingest) und Verwaltung (Data Management) der Dokumente und die Zugriffsteuerung (Access). In einer solchen Kooperation ist es möglich, durch das bereits vorhandene Knowhow im Rechenzentrum und in der Bibliothek ein nachhaltiges Langzeitarchiv aufzubauen. Aus diesem Grund entstand bereits 2004 eine erste Kooperation zwischen dem LRZ und der BSB, um die bisher auf CD archivierten Daten auf die Magnetbänder des LRZ zu migrieren.



Entwicklung des Datenvolumens der Backup- und Archivierungsumgebung am LRZ.



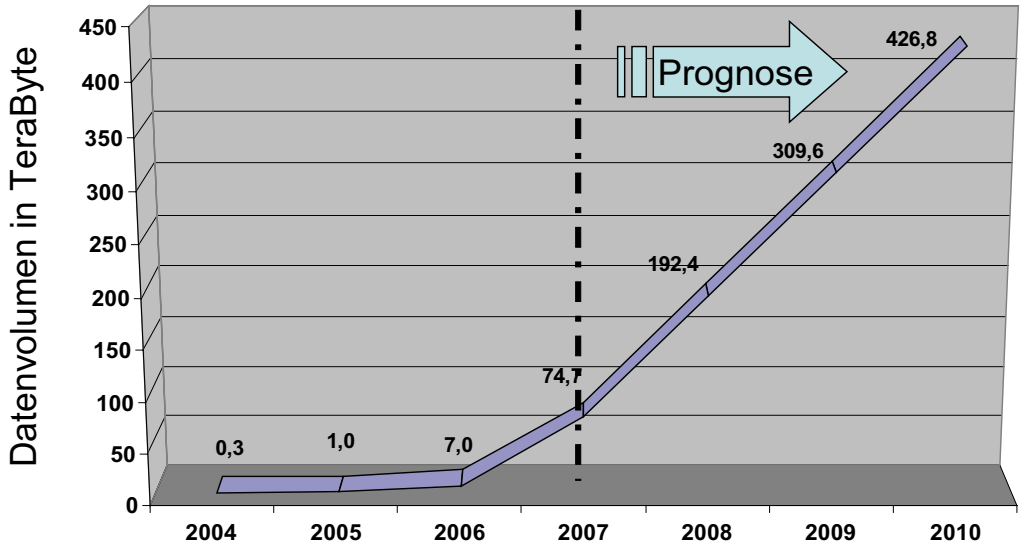
Arbeitsabläufe in der Langzeitarchivierungsumgebung BABS.

Das Projekt BABS

Diese anfängliche Kooperation wurde im Juli 2005 durch ein Projekt im Bereich der Langzeitarchivierung (LZA) ausgeweitet und intensiviert. Das Projekt wurde durch die Deutsche Forschungsgemeinschaft (DFG) gefördert und trägt den Namen BABS (www.babs-muenchen.de). Das Akronym BABS steht für „Bibliothekarisches Archivierungs- und Bereitstellungssystem“. Das Ziel war und ist der exemplarische Aufbau einer organisatorischen und technischen Infrastruktur für die Langzeitarchivierung von Netzpublikationen und sonstigen E-Medien unterschiedlicher Provenienz. Die gewonnene Erfahrung hat zur Etablierung neuer Arbeitsabläufe und Verfahren des Datenmanagements in der BSB und dem LRZ beigetragen. Die realisierte Langzeitarchivierungsumgebung folgt dem allgemeinen Open Archival Information Systems (OAIS) Referenzmodell und deckt somit die Prozesse Ingest, Data Management, Archival Storage und Preservation Planning ab.

Innerhalb der zweijährigen Projektlaufzeit von BABS wurden mehr als 19 Millionen Objekte mit einem Datenvolumen von über 36 Terabyte digitalisiert und im Archivsystem des LRZ gesichert. Die BSB verfügt somit über eines der größten und am schnellsten wachsenden elektronischen Langzeitarchive in Deutschland. Wird die Zweitkopie der Archivdaten berücksichtigt, die aus Sicherheitsgründen beim Archivierungsvorgang zusätzlich erstellt wird, verdoppelt sich die Anzahl der Objekte und das Datenvolumen noch einmal. In der entsprechenden Abbildung (u. l.) ist nur die Entwicklung des Datenvolumens ohne Zweitkopie veranschlagt. Wie sich am Archivvolumen ablesen lässt, kann die Kooperation zwischen dem LRZ und der BSB durchaus als fruchtbar bezeichnet werden.

Entwicklung des Archivvolumens der BSB am LRZ.



Projekt VD16digital

Aufgrund der erfolgreichen Zusammenarbeit im Projekt BABS wurde ein weiteres DFG-gefördertes Projekt mit dem Namen VD16digital mit einer Laufzeit von zwei Jahren ins Leben gerufen. Der Projektstart war im Juli 2007. Das Ziel des Projektes ist die Digitalisierung, Katalogisierung und Langzeitarchivierung der an der BSB vorhandenen, im deutschen Sprachbereich erschienenen Drucke des 16. Jahrhunderts und die sofortige Bereitstellung dieser Digitalisate im Internet. Die Langzeitarchivierung erlaubt und gewährleistet eine vielfältige Nachnutzung der digitalen Master. Die zweijährige Digitalisierungs- und Archivierungsmaßnahme umfasst 36.150 Titel mit über 7,2 Millionen Seiten.

TREVENTUS

Scan-Roboter der Firma Treventus im Einsatz.

Mit VD16digital wird eine neue Dimension im Bereich der Digitalisierung und der Archivierung erreicht. Die bisherige manuelle Buchdigitalisierung (wie bei dem Projekt BABS) wird von zwei Scan-Robotern der Firma Treventus übernommen, die eine Scan-Kapazität von 2.200 Seiten pro Stunde erreichen. Pro Jahr entstehen hier über 100 TByte an neuen Archivdaten, die in den Bandlaufsystemen des LRZ gespeichert werden.

Desasterschutz

Um für den Desasterfall gewappnet zu sein, werden zahlreiche Vorkehrungen getroffen, damit keine Datenverluste auftreten. Dies ist gerade für Langzeitarchive von besonderer Bedeutung. Im Rechnerwürfel wird deshalb für optimierte klimatische Bedingungen (Temperatur und Luftfeuchte) gesorgt, um eine lange Haltbarkeit der Magnetbänder und der Archiv- und Backupsysteme zu erlangen. Außerdem verfügen die Rechnerräume über fortschrittlichste Sicherungsmaßnahmen gegen Brandschäden. Es wird im Brandfall nicht mit Wasser gelöscht, sondern der Raum wird mit dem Edelgas Argon geflutet, um den Brand sehr schnell zu ersticken. Gegenüber der Wasserlöschung sind bei der Argonlöschung keine bzw. nur geringe Schäden an der Infrastruktur zu erwarten. Somit sind die auf den Magnetbändern archivierten Daten mit hoher Wahrscheinlichkeit auch nach einem Brandfall noch zu lesen.

Um eine noch größere Sicherheit zu erlangen, werden am LRZ, wie bereits schon erwähnt, Archivdaten grundsätzlich auf zwei Magnetbänder gespeichert. Die geographische Trennung der Zweitkopie sorgt für weiteren Schutz. Archivkopien werden nicht am LRZ, sondern am Rechenzentrum Garching der Max-Planck-Gesellschaft (RZG) gespeichert. Im Gegenzug werden in dieser Kooperation Archivkopien des RZG am LRZ vorgehalten. Eine redundant ausgelegte Gebäudeinfrastruktur (Strom, Klima usw.) sorgt darüber hinaus für geringe Ausfallzeiten.

Nachhaltigkeit

Langzeitarchivierung („long-term digital preservation“) umfasst Aktivitäten, die dem Erhalt der Verfügbarkeit der Dokumente über die Lebensdauer der Trägermedien und die Grenzen des technologischen Wandels hinaus dienen. Die Aufgabenstellung der Langzeitarchivierung ist daher ein langwieriges Geschäft. Es ist nicht damit getan, Bücher zu digitalisieren und nur mal einfach so zu archivieren. Vielmehr ist für das Erreichen einer nachhaltigen langfristigen Speicherung von Archivdaten und der Aufrechterhaltung der Nutzbarkeit ein enormer stetiger Aufwand zu betreiben. Die Lebensdauer von digitalen Daten ist unter anderem begrenzt durch die Haltbarkeit des

Datenträgers, die Verfügbarkeit der Hardware, des Betriebssystems, des Dateisystems und des Dateiformates. Aus diesem Grunde sind regelmäßige Erhaltungsmaßnahmen unerlässlich. Um ein Bewusstsein für die Problematik zu schaffen, sei hier die notwendige periodische Überführung (Migration) veralteter Dateiformate in aktuelle Formate erwähnt. Es ist nur noch schwer möglich, eine Microsoft Word Datei der ersten Version, die nicht regelmäßig auf neuere Formate migriert wurde, mit aktueller Hardware und Software zu lesen. Um den schnellen Wandel der Dateiformate abzuschwächen, helfen speziell für die Langzeitarchivierung standardisierte Dateiformate, wie z. B. das PDF/A-Format.

LRZ

Blick in ein robotergesteuertes Magnetbandsystem mit tausenden Magnetbändern.

Regelmäßiger Wandel

Um die Lesbarkeit von Archivdaten auf lange Frist zu gewährleisten, ist es notwendig, Daten spätestens vor Erreichen der Haltbarkeitsgrenze der Datenträger auf neue Datenträger zu migrieren. Aus Sicht des LRZ werden diese Grenzbereiche bei weitem nicht erreicht, da die Speichersysteme einer ständigen Erneuerung unterworfen sind. Die Erneuerung wird durch das immense Datenwachstum und durch den schnellen technischen Fortschritt im Bereich der Server, Festplatten, Schreib- und Lesegeräte und der Datenträger regelrecht erzwungen.

Am LRZ werden die erwähnten Systeme etwa alle fünf Jahre durch leistungsfähigere Systeme ersetzt. Allein die Robotersysteme (Libraries) haben mit bis zu zehn Jahren eine höhere Standzeit. Durch diese ständige Erneuerung ergibt es sich von selbst, dass Daten von veralteten Magnetbändern in einem vorgegebenen Rhythmus von fünf Jahren auf neue Magnetbandgenerationen wandern. Das LRZ kann bereits jetzt auf eine mehr als 15-jährige Erfahrung im Bereich der technischen Migration von Archivdaten zurückblicken. Daten wurden mehrfachen Hardware- und Softwaremigrationen unterzogen. Im Zusammenspiel mit der am LRZ vorhandenen Erfahrung in diesem Bereich garantiert modernste Hard- und Software die Sicherheit der Daten.

Neue Herausforderungen

In den vergangenen Jahren haben das LRZ und die BSB gezeigt, dass auf dem Gebiet der Langzeitarchivierung eine Kooperation zwischen Bibliothek und Rechenzentrum ideal ist. Die Tätigkeitsfelder der beiden Kooperationspartner ergänzen sich hervorragend. Das immense Datenwachstum und bisher ungelöste Probleme im weiten Bereich der Langzeitarchivierung bieten auch in Zukunft neue und spannende Herausforderungen. Um allerdings ein nachhaltiges Langzeitarchiv zu betreiben, sind noch einige Hürden zu nehmen. Eine dieser Hürden ist die Realisierung einer dauerhaften Finanzierung.

Der Autor ist wissenschaftlicher Mitarbeiter des Leibniz-Rechenzentrums der Bayerischen Akademie der Wissenschaften und dort zuständig für die Betreuung von Datei- und Speichersystemen, Archivierung und Langzeitarchivierung.

Internet:

www.babs-muenchen.de