

Das Digitalisierungszentrum an der Bayerischen Staatsbibliothek

Marianne Dörr

Mit der Schaffung des neuen Förderbereichs „Retrospektive Digitalisierung von Bibliotheksbeständen“ hat die Deutsche Forschungsgemeinschaft die Gründung von zwei sogenannten Digitalisierungszentren initiiert. Diese sollen Vorreiter- und Beratungsfunktion für die wachsende Zahl von Projekten im Förderbereich haben und Wissen und Erfahrungen in folgenden Bereichen erwerben und aktiv weitervermitteln:

- Technik der Digitalisierung
- Bereitstellungssysteme und Präsentation im WWW
- Standards und „Best practices“
- Verbindung zu vorhandenen Bibliotheksverbundsystemen
- Sicherung der langfristigen Verfügbarkeit der Dokumente.

Im Sommer 1997 begann an der Bayerischen Staatsbibliothek die Einrichtung eines solchen Digitalisierungszentrums (*Münchener Digitalisierungszentrum, MDZ*). Im Pendant, dem *Göttinger Digitalisierungszentrum (GDZ)* an der Staats- und Universitätsbibliothek Göttingen, hatte die Arbeit bereits im April desselben Jahres begonnen. Zum Zeitpunkt der Gründung waren in der BSB schon einige Digitalisierungsaktivitäten im Gange: So wird im VD17-Projekt seit 1996 vom Mikrofilm digitalisiert, die entstehenden Bilder in einer Text-Bild-Datenbank verwaltet und zur Recherche angeboten¹⁾. Bei der Konversion des Katalogs 1953 - 1981 waren 2,2 Millionen Katalogkarten direkt gescannt worden, so entstand der erste deutsche Image-Katalog²⁾. Die Einrichtung eines Digitalisierungszentrums und die Durchführung der kompletten Digitalisierung ganzer Sammlungen erfordern aber grundlegendere Planungen.

Grundlagen: Konzeption des Zentrums

Die BSB hatte bei der Einrichtung des Zentrums ursprünglich intendiert, weitestgehend auf Outsourcing zu setzen und nicht nur Scan-Arbeiten, sondern auch Software-Entwicklung und -Betreuung sowie Hardware-Bereitstellung durch einen Dienstleister vornehmen zu lassen. Dies erwies sich jedoch als wirtschaftlich (noch?) nicht tragfähig. So entwickelte sich folgende Konzeption: Die Bereitstellung erfolgt auf eigener Hard- und Software durch die BSB. Die Digitalisierung i.S. von Scannen wird jedoch konsequent an Dienstleister vergeben. Für unterschiedliche Projekte sind jeweils unterschiedliche Geräte

zur Durchführung notwendig - die Anschaffung eines kompletten Geräteparks erforderte enorme Investitionen, die wirtschaftlich kaum gerechtfertigt werden könnten. Denn eine dauernde und gleichmäßige Auslastung wäre kaum erreichbar; die schnelle technologische Entwicklung in diesem Bereich macht rasche Erneuerung notwendig, um auf dem Stand der Technik zu sein, entsprechend geschultes und flexibles Personal wäre eine weitere Voraussetzung - diese Bedingungen sind in der Praxis kaum erreichbar bzw. nicht finanzierbar. Ein Outsourcing der Scan-Dienstleistung erlaubt dagegen, jeweils den Dienstleister mit der passenden Erfahrung und dem passenden Equipment auszuwählen. Das Marktangebot hat sich hier in den letzten Jahren erweitert. Einholung von Angeboten, Besuche beim Dienstleister, um sich gegenseitig sowohl über die eingesetzten Geräte als auch über den Workflow bzw. den Umgang mit dem Material zu informieren, gehören zu einer guten Zusammenarbeit, die Lerneffekte auf beiden Seiten bewirkt. Outsourcing bedeutet im übrigen nicht zwangsläufig, daß die zu digitalisierenden Bestände auch außer Haus gegeben werden müssen: Bisher wurde bereits in zwei Projekten durch externe Dienstleister mit deren Geräten im Haus gearbeitet. Flexible Lösungen für den Einzelfall, das Einzelprojekt müssen entwickelt und ausgehandelt werden.

Bereitstellung digitalisierter Daten im WWW

Das Ziel einer Digitalisierung ist es, eine verbesserte Zugänglichkeit, neue Nutzungsformen der Materialien zu erreichen. Die Art der Bereitstellung spielt deshalb eine zentrale Rolle.

Der Entscheidung für ein Bereitstellungssystem war in der Bayerischen Staatsbibliothek eine lange Beratungs- und Informationsphase vorausgegangen. Das erste von einem Dienstleister vorgeschlagene System erwies sich als zu unflexibel und den Anforderungen einer digitalen Bibliothek nicht gewachsen, der zweite Vorschlag eines Dokumentenmanagementsystems war in der Anpassung an einzelne Projekte zu aufwendig und damit zu teuer. Auch aufgrund dieser Erfahrungen bildete sich ein Katalog unterschiedlicher, formaler und inhaltlicher Kriterien heraus, auf den die in engere Wahl genommenen Systeme geprüft wurden.

Im formalen (technisch-organisatorischen) Bereich waren folgende Punkte besonders stark gewichtet:

- Standardprodukt, keine Eigenentwicklung
- Anwender in verwandten Bereichen vorhanden
- geringer Installations- und Betreuungsaufwand

- vertretbare Kosten (zu kalkulieren sind hier nicht nur die Anschaffung und Wartung, sondern auch der für die Installation und Projektanpassung notwendige Aufwand)
- Skalierbarkeit
- Zukunftssicherheit (Formate).

Von der inhaltlichen Leistungsfähigkeit her waren folgende Kriterien entscheidend:

- Nachbildung von Dokument-Strukturen für eine komfortable Navigation im Dokument
- strukturierte Suche
- Volltext-Suche
- flexible Anpassung von Darstellung und Suchmasken an die jeweilige Sammlung
- Zusammenstellung von Sammlungen
- Suche über Sammlungen hinweg
- Import- und Exportschnittstellen für strukturierte Daten (möglichst SGML – mit Öffnung zu XML als Format für Migration und Langzeitarchivierung).

Auf der Basis dieser Kriterien wurde die Entscheidung für das WebPublishingTool *DynaText* von Inso getroffen. Diese Software besteht aus unterschiedlichen Komponenten:

- den DynaText Publishing Tools (DynaTag, InSted Stylesheet-Editor)
- dem DynaText CD/LAN Browser und
- der WWW-Komponente DynaWeb.

DynaText basiert auf der Standardized General Markup Language (SGML), die als layout-unabhängige Beschreibungssprache auch für die Zwecke der Langzeitarchivierung empfohlen wird.

Mit DynaText können sogenannte elektronische Bücher erstellt werden. Wenn die Originaldaten bereits in SGML- bzw. XML³⁾-Strukturierung vorliegen, geschieht dies ohne zusätzliche Konvertierungen. Die Indizierung setzt auf der SGML-Struktur auf. Auch in Word erstellte Dateien können importiert werden: Sie werden mittels DynaTag zunächst in SGML oder XML konvertiert und dann publiziert. Graphikformate (Images) sind einbindbar. DynaWeb, der Web-Server des Produkts, generiert aus einem SGML oder XML basierten elektronischen Buch dynamisch HTML und erlaubt so das Angebot bzw. den Zugriff über die Standard-Web-Browser. Bei der Suchfunktionalität steht sowohl eine leistungsfähige Volltext-Suche wie auch die Suche über strukturierte Indizes zur Verfügung. Die strukturierte Suche basiert auf der SGML-Auszeichnung der entsprechenden Dokumentteile. Damit wird eine Suchfunktionalität er-

zeugt, die der Suche in den Feldern einer Datenbank vergleichbar ist. Die Präsentationsoberfläche ist über Stylesheets frei gestaltbar. Mit dieser Architektur ist getreu der SGML-Philosophie eine Trennung von Dokumentenstrukturierung einerseits und der Sicht auf das Dokument andererseits realisiert. DynaText /-Web ist im Verlagsbereich stark vertreten. Es wird weiterhin in verschiedenen Bibliotheksprojekten in den USA eingesetzt (z. B. Berkeley: American Heritage Virtual Archive Project; California Heritage Collection; Digital Scriptorium der Duke University). In Deutschland wurden auch schon Umsetzungen mit Bibliotheks-Katalog-Daten, also sehr stark strukturierten Daten, realisiert.

Für die aktuell laufenden Projekte hat das MDZ zur Erstellung von SGML-ausgezeichneten Dokumentinstanzen auf vorliegende Dokumenttypdefinitionen (DTD) zurückgegriffen: Bei den Projekten, in denen Image-Digitalisierungen im Vordergrund stehen, wird zur Erstellung navigationsfähiger Bücher die e-bind-DTD der UC Berkeley verwendet; bei Volltexten wird mit DTDs der Text Encoding Initiative (TEI) gearbeitet.

DynaText /-Web ist ein eigenes, aber kein hermetisch geschlossenes System. Einzelne Sammlungen bzw. einzelne elektronische Bücher sind über eine herleitbare URL direkt ansteuerbar. Es kann also aus einem anderen System direkt auf sie verlinkt werden. Andererseits können einzelne Sammlungen durch Paßwortschutz nur für bestimmte Nutzergruppen freigegeben werden, was besonders im Hinblick auf ein Angebot noch urheberrechtlich geschützter Materialien wichtig ist. Eine Postscript-Druckausgabe wird unterstützt.

Das MDZ geht nicht davon aus, daß ein System alle Anforderungen, die im Rahmen von Digitalisierungsprojekten auftreten, erfüllen kann. DynaText bietet jedoch eine flexible und gestaltbare Plattform für die textbasierten Dokumente. Die integrierte Volltext-Suche ist, wie auch amerikanische Erfahrungen zeigen, gerade bei der zunehmenden Zahl von volltextdigitalisierten Dokumenten ein entscheidender Vorteil. Außerdem stellt der Einsatz eines SGML-basierten Systems ein wichtiges Argument im Hinblick auf die Langzeitarchivierung der Daten dar.

Eigene Projekte

Wissenserwerb und -vermittlung setzt umfangreiche eigene Erfahrungen voraus. Das MDZ führt deshalb verschiedene Projekte durch und bemüht sich laufend um Kooperationen für neue Aufgaben. Derzeit sind in der Bayerischen Staatsbibliothek folgende Projekte in Arbeit:

Decretum Gratiani

Das Decretum Gratiani, benannt nach seinem Verfasser, dem Mönch und Rechtsgelehrten Gratian, der in der 1. Hälfte des 12. Jahrhunderts in Bologna lebte, bildet den ersten Teil des später im Corpus Iuris Canonici zusammengefaßten römisch-katholischen Kirchenrechts.

Die Edition Emil Friedbergs von 1879 wurde bereits Mitte der 80er Jahre durch die Monumenta Germaniae Historica digital erfaßt und in Form einer Konkordanz aufbereitet. Diese Konversion bezog sich aber nur auf den Text der Edition, nicht auf den kritischen Anmerkungsapparat. Diese ASCII-Daten wurden dem Digitalisierungszentrum von der MGH zur Aufbereitung und Präsentation für das Web überlassen.

Das Projekt hat deshalb ein zweifaches Ziel:

- Verbesserung der Erschließung: Durch die zusätzliche Digitalisierung der Buchseiten als Images wurde die komplette Information inklusive des kritischen Apparats zugänglich.
- Außerdem sollte die Kombination von Volltextrecherche und digitalisiertem Image erprobt werden.

Das Decretum Gratiani ist als Prototyp bereits im WWW verfügbar unter der URL <<http://mdz.bsb.badw-muenchen.de:6336/>>.

Stenographische Berichte des Deutschen Reichstags (1867-1895)

Die Stenographischen Berichte, gedruckt in über 400 Bänden, sind eine der wichtigsten Quellen der neueren deutschen Geschichte. Digitalisiert wurden für den Zeitraum 1867-1895 165 Bände mit ca.100.000 Seiten.

Nach umfangreichen Tests zur Ermittlung der geeigneten Vorlagenvariante (Original, Masterkopie des Papier-Reprints, Mikrofiche) und der optimalen Auflösung (300 oder 600 dpi) wurde - wegen der problematischen Schriftart und -größe - schließlich mit 600 dpi und einer Farbtiefe von 1 Bit digitalisiert. Als Webversionen stehen GIF und PDF (als Arbeitsversion mit den vollen 600 dpi) zur Verfügung.

Aufgrund der großen Materialmenge können die Reichstagsberichte nicht im Volltext angeboten werden. Der Druck in Frakturtype läßt wegen der immer noch zu hohen Fehlerraten auch einen OCR-Einsatz nicht ratsam erscheinen. In den als Images digitalisierten Bänden kann wie im gedruckten Buch navigiert werden. Ein spezifischer Suchzugriff ist auf die einzelnen Sitzungen der einzelnen Jahre möglich. Für eine sachliche Suche wird das Generalregister der Berichte (manuell) in SGML strukturiert erfaßt. Im Register kann dann sachlich gesucht werden, die eingefügten Links führen automatisch zu den

korrespondierenden Image-Seiten der einzelnen Berichtsbände. Die Erfassung ist noch nicht abgeschlossen; das Prinzip kann jedoch an einem Beispielregistereintrag im WWW unter <http://mdz.bsb.badw-muenchen.de:6336/> ausprobiert werden. Die Bände, auf die vom Eintrag aus verwiesen wird, stehen bereits vollständig digitalisiert und zur Navigation aufbereitet zur Verfügung.

Deutsche druckgraphische Buchillustration des 15. Jahrhunderts

In diesem Projekt wurden - unmittelbar vom Original - die Illustrationen von 73 im deutschen Sprachraum gedruckten Inkunabeln der Bayerischen Staatsbibliothek digitalisiert. Es handelt sich in der Mehrzahl um schwarzweiße, aber auch um zahlreiche kolorierte Holzschnitte. Eine digitale Kamera (JenOptik Progres 3012) mit Buchschwinge wurde durch einen Dienstleister für dieses Projekt drei Monate in der „Schatzkammer“ der Bayerischen Staatsbibliothek installiert.

Die Kamera verfügt über ein Auflösungsvermögen von 3.500 x 4.500 Bildpunkten (entspricht bei einer DIN A4-Vorlage ca. 400 dpi), das für die Digitalisierung voll ausgeschöpft wurde. Die Archivversionen der Bilder wurden als TIFF (unkomprimiert) abgelegt. Die Farbtiefe betrug bei farbigen Vorlagen 24 Bit; Schwarzweißvorlagen wurden mit 8 Bit (Graustufen) digitalisiert. Die daraus resultierenden Einzeldateien umfaßten bei Farbbildern 45 MB; bei Graustufenbildern immerhin noch 14 MB. Das Ziel war mit dieser Qualität auch reproduktionsfähige Druckvorlagen ohne erneuten Rückgriff auf die Originalwerke erzeugen zu können. Für die Webversion wurden komprimierte und in der Auflösung auf ca. 100 dpi reduzierte JPEGs erzeugt. Insgesamt wurden 6425 Illustrationen gescannt.

Als Sucheinstieg ist neben den Aufnahmen des Incunabula Short Title Catalogue eine inhalts- und themenbezogene Attributierung der Einzelbilder vorgesehen. Auch hier ist bereits ein Beispiel mit wenigen Bildern im WWW verfügbar (<http://mdz.bsb.badw-muenchen.de:6336/>)

Großes vollständiges Universal-Lexicon aller Wissenschaften und Künste (Zedler) und Grammatisch-kritisches Wörterbuch der hochdeutschen Mundart (Adelung)

Modellhaft untersucht werden sollen die Möglichkeiten der Digitalisierung und Erschließung von Enzyklopädien und Wörterbüchern. Ziel ist es, mit der Digitalisierung der wichtigsten deutschen Enzyklopädie und des maßgeblichen Wörterbuchs des 18. Jahrhunderts in sich komplementäres und für ein breites Spektrum von Fachbereichen wichtiges Quellenmaterial verfügbar zu machen.

Das Zedlersche Universallexikon soll im Image-Modus digitalisiert werden; Tests mit den vom Reprint-Verlag zur Verfügung gestellten Mikrofiches haben begonnen. Außerdem ist für den *Zedler* der testweise Einsatz einer italienischen Strukturerkennungssoftware vereinbart. Für den *Adelung* ist eine Volltexterfassung in Anlehnung an die Wörterbuch-DTD der TEI vorgesehen. Die zugehörigen Images des Originals sollen natürlich ebenfalls angeboten werden. Aus dem Projekt soll auch eine Kosten-Nutzen-Abschätzung zu Methoden der vertieften Erschließung resultieren, wie z. B. einer Standardisierung zur Verbesserung des Suchkomforts (Normierung von Namen und der Orthographie).

Wissenstransfer / Öffentlichkeitsarbeit

Eine wesentliche Aufgabe der Digitalisierungszentren ist der Wissenstransfer und die Öffentlichkeitsarbeit. Die beiden Zentren veranstalten hierzu gemeinsam Kolloquien und Workshops. Im Januar 1998 fand ein erstes großes Kolloquium in Göttingen statt, in dem mit vielen in- und ausländischen Referenten ein breiter Überblick über die Aspekte der Digitalisierung gegeben wurde. Für das zweite Kolloquium in München war eine thematische Zentrierung vorgenommen worden: In zehn Vorträgen wurden unterschiedliche Aspekte der Erschließung digitaler Text- und digitaler Bildsammlungen dargestellt. Dabei bildeten allgemeine Themen den Rahmen. Im textbezogenen Kolloquiumsteil wurde zur Einführung über die Rolle von SGML, XML und DSSSL als neutrale Auszeichnungssprachen und Formatvorlagen referiert; im bildzentrierten Teil wurde beispielsweise auch der „State of the Art“ der automatischen Bilderkennung vorgestellt.

Neben den großen Veranstaltungen sind zunehmend auch kleinere und regional zentrierte Aktionen geplant. Außerdem bilden die persönlichen Beratungen von Interessenten, Antragstellern und anderen Projektnehmern einen wichtigen Teil der Arbeit.

Vernetzung mit anderen Projekten und Arbeitsbereichen

Digitalisierung ist ein neuer Aufgabenbereich, in dem viel experimentiert werden muß. Fragen der Digitalisierungstechnik und technisch-organisatorische Fragen der Projektvorbereitung und -durchführung sowie der Bereitstellung und Präsentation der digitalisierten Dokumente bilden deshalb derzeit noch den Schwerpunkt der Arbeit.

Die technischen Fragestellungen werden vermutlich jedoch zunehmend zurücktreten. Digitalisierung muß - wenn die Attraktion des Neuen nachgelassen hat - ein inhaltlich sinnvoll integriertes Angebot schaffen, mit dem wissen-

schaftlich gearbeitet werden kann. Dies zeigen beispielsweise die extrem hohen Zugriffszahlen auf die Server des *Electronic Text Centers* in Virginia. Das ausgedehnte und gut erschlossene Angebot, das hier für Philologen bereit steht, hat sich weltweit einen Ruf gemacht und wird auch weltweit intensiv genutzt.

Die Bayerische Staatsbibliothek als Sondersammelgebietsbibliothek für Geschichte hat deshalb versucht, inhaltlich die Digitalisierungsprojekte bereits darauf abzustimmen und erweiterbare Projekte anzustoßen. In das Zentrum werden auch Projekte integriert, die verwandte Zielsetzungen haben und die Synergieeffekte erwarten lassen. So hat im Februar mit dem sogenannten „Server Frühe Neuzeit“ ein Projekt begonnen, das im Förderbereich „Virtuelle Fachbibliothek“ angesiedelt ist und den Aufbau eines integrierten Informationsangebots für die Früh-Neuzeit-Forscher intendiert. Auf diesen Schwerpunkt sollen zukünftig verstärkt retrospektiv digitalisierte Materialien hin orientiert werden, damit sich die digitalen Quellensammlungen in ein breiteres und auch auf wissenschaftliche Aktualität ausgerichtetes Angebot einpassen lassen. Natürlich ist das Zentrum auch in die Planungen zum Aufbau einer *Bayerischen Landesbibliothek online* involviert. Das Ziel des Münchener Digitalisierungszentrums ist es, das Wissen und die Erfahrungen im technisch-organisatorischen Feld der Digitalisierung für ein inhaltlich attraktives Online-Angebot zu nutzen, damit eine langfristig tragfähige Perspektive eröffnet werden kann.

Im Hinblick auf größere Zeiträume spielt natürlich die Langzeitarchivierung der digitalen Medien und Publikationen eine zentrale Rolle. Zu Beginn dieses Jahres hat - ebenfalls ins Zentrum integriert - ein Projekt zur Untersuchung der Langzeitarchivierung begonnen. Es wird in Kooperation mit dem Institut für Softwaretechnologie der Universität der Bundeswehr durchgeführt. Die Zielsetzung ist dabei, praktisch anwendbare Verfahren zur langfristigen Verfügbarkeit der digitalen Medien in einer Universal- und Archivbibliothek zu entwickeln. Aber die Archivierung soll auch als vernetzte Aufgabe begriffen werden, die aufgrund der zunehmenden Heterogenität und Vielfalt der zu archivierenden Materialien Arbeits- und Funktionsteilungen zwischen Institutionen und Organisationen erfordern wird. Für diese Aufgabenstellung hat sich international noch kein erprobtes, tragfähiges Konzept herausbilden können, aber gerade in einem Digitalisierungszentrum müssen diese Probleme von Anfang an mitreflektiert und angegangen werden.

Ansprechpartner im Zentrum sind:

*Dr. Markus Brantl, Tel.: (0 89) 2 86 38-23 94,
E-Mail: brantl@bsb.badw-muenchen.de*

Dr. Marianne Dörr, Tel.: (0 89) 2 86 38-26 00,

E-Mail: doerr@vd17.bsb.badw-muenchen.de

Monika Malikioussis, Tel.: (0 89) 2 86 38-23 94

E-Mail: malikioussis@bsb.badw-muenchen.de

Anmerkungen:

- 1) Die VD-17-Datenbank ist verfügbar unter <<http://www.vd17.bsb.badw-muenchen.de>>
- 2) Der BSB-Katalog für die Jahre 1953 - 1981 (IFK) ist verfügbar unter: <<http://193.174.99.237/ifk/ifk.html>>. Im Sommer wird die Endversion, die eine Datenbankintegration in den BSB-OPAC beinhaltet, öffentlich.
- 3) XML steht für Extended Markup Language. Es handelt sich um eine Art reduzierte SGML, die mit (dann spricht man von „well formed“) und ohne Dokumenttypdefinition benutzt werden kann. Netscape und Microsoft haben für die künftigen Browsergenerationen XML-Unterstützung angekündigt. Aus diesem Grund ist ein breites Durchsetzen von XML sehr wahrscheinlich.

