Retrospektive Digitalisierung von Bibliotheksbeständen

Berichte der von der Deutschen Forschungsgemeinschaft einberufenen Facharbeitsgruppen 'Inhalt' und 'Technik'

Berlin 1998
DEUTSCHES BIBLIOTHEKSINSTITUT

dbi-materialien ; 166

Schriften der Deutschen Forschungsgemeinschaft

DFG-Projekt „Vorbereitung des Aufbaus einer verteilten digitalen Forschungsbibliothek in bibliothekarischer, fachlicher und technischer Hinsicht"

Projektdurchführung: *Niedersächsische Staats- und Universitätsbibliothek Göttingen*

Projektkoordinator und Herausgeber: *Prof. Dr. Elmar Mittler*

Projektbetreuung und Redaktion: *Dr. Norbert Lossau*

Die Deutsche Bibliothek - CIP-Einheitsaufnahme

Retrospektive Digitalisierung von Bibliotheksbeständen : Berichte der von der Deutschen Forschungsgemeinschaft einberufenen Facharbeitsgruppen „Inhalt" und „Technik" ; [DFG-Projekt Vorbereitung des Aufbaus einer verteilten digitalen Forschungsbibliothek in bibliothekarischer, fachlicher und technischer Hinsicht] / Deutsches Bibliotheksinstitut. [Projektkoordinator und Hrsg.: Elmar Mittler]. - Berlin : Dt. Bibliotheksinst., 1997
(Dbi-Materialien ; 166 : Schriften der Deutschen Forschungsgemeinschaft)
ISBN 3-87068-966-8

Berlin, Dezember 1997

Herstellung und Vertrieb:
Deutsches Bibliotheksinstitut
Alt-Moabit 101 A
10559 Berlin

Druck: Ernst Knoth, 49303 Melle

Gedruckt auf alterungsbeständigem Papier.
Diese Veröffentlichung entstand mit Förderung des Bundes und der Länder.

Vorwort

In ihren gemeinsamen Empfehlungen „Neue Informations-Infrastrukturen für Forschung und Lehre" (Februar 1996) haben der Bibliotheksausschuß und die Kommission für Rechenanlagen der Deutschen Forschungsgemeinschaft als neue Nutzungsmöglichkeiten der elektronischen Kommunikations- und Publikationstechniken die Bereitstellung und Nutzung wissenschaftlicher Forschungsliteratur in digitaler Form über Kommunikationsnetze direkt am PC-Arbeitsplatz des Wissenschaftlers in den Blick gerückt und vorgeschlagen, die Digitalisierung geeigneter Bestände von Bibliotheken durch ein eigenes Förderprogramm zu fördern. Damit sollte von vornherein eine ‚Verteilte Digitale Forschungsbibliothek' aufgebaut werden, um der neuen Nutzungsform rasch eine breite materielle Grundlage zu geben, ihr Leistungspotential zu demonstrieren und sie als Bibliotheksdienst zu etablieren.

Die Forschungsgemeinschaft nahm die Empfehlung auf und konnte dem Bibliotheksausschuß in dessen Frühjahrssitzung 1996 die Bereitstellung von Fördermitteln für ein solches Programm, und zwar erstmalig bereits für 1997, ankündigen. Zeit war nicht zu verlieren, dies umso weniger, als ein Angebot elektronischer Texte im allgemeinen und die Durchführung von Digitalisierungsprojekten im besonderen für deutsche Bibliotheken noch weitgehend Neuland war und auch von Seiten der Nutzer umschriebene Anforderungen zumeist fehlten. Sowohl für die technische Ausführung als auch für die Fragen der inhaltlichen Materialauswahl waren praktikable Konzepte nötig, um binnen kurzem Interesse und qualifizierte Projektanträge anzuregen und das Programm zügig umzusetzen.

Zwei Arbeitsgruppen aus Fachleuten zur Ausarbeitung entsprechender Konzepte zusammenzurufen und zu koordinieren, erklärte sich Herr Prof. Dr. Elmar Mittler, der Direktor der Göttinger Staats- und Universitätsbibliothek, bereit und ging die Aufgabe, durch Mittel der Forschungsgemeinschaft unterstützt, energisch an. Er wirkte ebenfalls in einer dritten, von der Forschungsgemeinschaft eingesetzten, Arbeitsgruppe mit, die Richtlinien für die Antragstellung und Begutachtung in dem neuen Förderprogramm erarbeitete. Die Richtlinien verabschiedete der Bibliotheksausschuß in seiner Herbstsitzung 1996, so daß ein Begutachtungsausschuß eingesetzt werden und das Programm pünktlich starten konnte. Im Frühjahr 1997 lagen auch bereits die Entwürfe der hier im Druck veröffentlichten Berichte der Arbeitsgruppen ‚Inhalt' und ‚Technik' vor und standen Interessenten zur Verfügung, von dem ‚Technik'-Bericht auch ein Extrakt in Form eines von der Geschäftsstelle der Forschungsgemeinschaft erstellten Merkblattes „Praktische Hinweise zur retrospektiven Digitalisierung von Bibliotheksbeständen".

Das Förderprogramm „Retrospektive Digitalisierung von Bibliotheksbeständen" ist von Anfang an mit Projektanträgen in großer Zahl und von guter

Qualität sehr erfolgreich angelaufen. Allen Mitgliedern der ‚Arbeitsgruppe Inhalt' und der ‚Arbeitsgruppe Technik' sowie auch der ‚Arbeitsgruppe Richtlinien' ist für ihre intensive Arbeit, ohne die dieser erfolgreiche Start nicht möglich gewesen wäre, sehr zu danken. Einen besonderen Anteil daran haben als „treibende Kräfte" mit hohem Engagement und mit Expertise Herr Prof. Dr. Elmar Mittler und sein Projektbearbeiter Herr Dr. Norbert Lossau sowie auch bei der Deutschen Forschungsgemeinschaft Herr Dr. Jürgen Bunzel. Ihnen gebührt für die erreichten Ergebnisse ganz besonderer Dank!

Bonn, Dezember 1997

Dr. Peter Rau
Universitäts- und Landesbibliothek Bonn
Vorsitzender des Begutachtungsausschusses
„Retrospektive Digitalisierung von Bibliotheksbeständen"

Inhalt

Empfehlungen zur inhaltlichen Auswahl von Bibliotheksmaterialien für die retrospektive Digitalisierung

Bericht der Facharbeitsgruppe Inhalt zur Vorbereitung des Programms „Retrospektive Digitalisierung von Bibliotheksbeständen" im Förderbereich „Verteilte Digitale Forschungsbibliothek"

Mitglieder der Arbeitsgruppe:

Prof. Dr. Helmut Altrichter, Friedrich-Alexander-Universität Erlangen-Nürnberg, Philosophische Fakultät 1

Dr. Ewald Brahms, Deutsche Forschungsgemeinschaft, Bonn

Prof. Dr. Bernhard Fabian, Westfälische Wilhelms-Universität Münster, Englisches Seminar

Prof. Dr. Horst Gronemeyer, Staats- und Universitätsbibliothek Hamburg

Prof. Klaus-Dieter Lehmann, Die Deutsche Bibliothek, Frankfurt a. M./Leipzig

Dr. Norbert Lossau, Niedersächsische Staats- und Universitätsbibliothek Göttingen (DFG-Projekt „Verteilte Digitale Forschungsbibliothek")

Prof. Dr. Elmar Mittler, Niedersächsische Staats- und Universitätsbibliothek Göttingen

Dr. Ulrich Ott, Deutsches Literaturarchiv/Schiller-Nationalmuseum, Marbach

Prof. Dr. Winfried Scharlau, Westfälische Wilhelms-Universität Münster, Mathematisches Institut,

Prof. Dr. Michael Schlaefer, Akademie der Wissenschaften zu Göttingen, Deutsches Wörterbuch

Dr. Manfred Vorholzer, Generaldirektion der Bayerischen Staatlichen Bibliotheken, München

Arbeitssitzungen am 26. Juli 1996, 19. Dezember 1996 und 20. Februar 1997 (Göttingen)

1
Vorbemerkungen

Die Deutsche Forschungsgemeinschaft hat für das Jahr 1997 ein neues Förderprogramm aufgelegt, das die retrospektive Digitalisierung ausgewählter Bibliotheksbestände umfaßt. Ziel des Programms ist es, durch den Einsatz digitaler Technik die wissenschaftliche Literaturversorgung zu verbessern. Im Vordergrund stehen dabei:

- der Direktzugriff auf für die Forschung und Lehre wichtige Bestände
- der Mehrfachzugriff auf vielgenutzte Literatur
- die digitale Bereitstellung schwer zugänglicher Bestände
- die erweiterte Nutzung bisher nur wenig bekannter Materialien.

Der Schwerpunkt dieses Förderprogramms liegt somit weniger auf dem sog. historisch wertvollen Erbe, sondern vielmehr auf der aus heutiger Sicht forschungsrelevanten Literatur.

Die von Bibliotheken und Fachwissenschaftlern ausgehende Initiative zur Digitalisierung von Dokumenten und anderen Materialien soll auch entsprechende Aktivitäten der Verlage anregen und unterstützen. Sind die Bibliotheken bei der Erstellung und Bereitstellung von digitalen Kernsammlungen zunächst älteren, urheberrechtsfreien Materials erfolgreich, wird dies auch im geisteswissenschaftlichen Bereich die Akzeptanz der digitalen Medien erhöhen und den Markt für derartige Aktivitäten erweitern. In der Folge ist auch mit einem verstärkten Engagement der Verleger im Bereich der aktuellen Forschungs- und Studienliteratur zu rechnen.

Zur Vorbereitung des neuen Förderprogramms hat die DFG neben einer technischen Arbeitsgruppe eine inhaltliche Facharbeitsgruppe zusammengerufen, die aus Fachwissenschaftlern und Bibliothekaren besteht. Die Facharbeitsgruppe Inhalt *(AG Inhalt)* legt in den vorliegenden Empfehlungen einen Kriterienkatalog für die Auswahl zu digitalisierender Materialien vor.

Darüber hinaus trifft sie Aussagen zur Vorgehensweise bei der inhaltlichen Selektion von Bibliotheksbeständen und zu den Rahmenbedingungen für den Aufbau digitaler Sammlungen.

Die AG Inhalt betont die Notwendigkeit und die Möglichkeit der Verbesserung der Literaturversorgung durch die retrospektive Digitalisierung von Bibliotheksbeständen. Sie betont zugleich, wenn nicht vorrangig, die große Chance, die die Digitalisierung für die Verbesserung der Arbeitsgrundlagen in den *Fach*disziplinen bietet, die in erster Linie oder ausschließlich mit Texten arbeiten. Durch die Digitalisierung werden neue Arten des Zugriffs auf Texte ermöglicht, die über die bisher üblichen und möglichen Zugriffsmöglichkeiten

beträchtlich hinausgehen. Größere Textcorpora können bearbeitet und neue Fragestellungen können entwickelt werden; zugleich ergibt sich eine größere Sicherheit der Aussage über empirische Befunde. Um solche Verbesserungen zu erreichen, gilt es, das Spektrum der Digitalisierungsmöglichkeiten vom Image-Scannen bis zur Volldigitalisierung von Texten auszuschöpfen. In vielen Fällen erweist sich die Bilddigitalisierung als wesentlicher erster Schritt, der Ausgangspunkt für die weitergehende Erfassung und Erschließung bedeutet, aber auch einen dauerhaften Wert behält.

2
Die retrospektive Digitalisierung von Bibliotheksmaterialien

Das Angebot an Bibliotheksmaterialien in elektronischer Form hat in den letzten Jahren in beträchtlichem Umfang zugenommen. Die Fragestellung, ob Publikationen nur in elektronischer Form, als Druck und in elektronischer Form oder nur als Druck vorliegen sollen, wird in zunehmendem Maße Thema der bibliothekarischen wie der fachwissenschaftlichen Diskussion. Dabei kann man bei der Literatur aus jüngster Zeit davon ausgehen, daß sie in der Regel bereits bei der Entstehung, spätestens aber für den Druck, in elektronische Form gebracht wird. Damit ist diese Literatur im Prinzip als digitaler Volltext verfügbar.

Die retrospektive Digitalisierung bereits gedruckt vorliegender älterer Materialien zum Aufbau einer Verteilten Digitalen Forschungsbibliothek ist in zwei Stufen realisierbar, die aufeinander aufbauen:

1. *Image-Scannen* (Einlesen der gedruckten Vorlage über geeignete Scanner)

Das Ergebnis des Image-Scannens ist ein in Pixel (Bildpunkte) zerlegtes Bild bzw. Image der Vorlage, das mit dem Computer weiterverarbeitet werden kann.

2. *Volltexterfassung* der gedruckten Vorlage

Die Volltexterfassung gedruckter Vorlagen ist auf zwei Wegen möglich:

a) automatisierte Erfassung durch eine Texterkennungssoftware (OCR)

b) manuelle Erfassung von Texten

Ergebnis des Texterkennungsprozesses ist ein Text im ASCII-Format. Die Volltexterfassung ist eine weiterführende Form der Digitalisierung von gedruckt vorliegenden Texten, die in vielen Fällen auf das Image-Scannen aufbaut.

2.1 Die Volltexterfassung gedruckter Vorlagen

Die AG Inhalt sieht - im Rahmen des Förderprogramms - die Volltextdigitalisierung in vielen Fällen als optimale Form der Bereitstellung eines Textes für die wissenschaftliche Arbeit an. Nur in dieser Form können Textcorpora in kürzester Zeit nach Stichwörtern durchsucht werden, wobei die mögliche Verknüpfung von Stichwörtern zusätzliche Informationen liefern kann. Aufgefundene Textstellen können bei diesen Voraussetzungen auch in Textverarbeitungsprogramme zur weiteren Nutzung exportiert werden.

Besonders sinnvoll erscheint eine Volltexterfassung für bestimmte Textformen oder Textgattungen. Dies sind in der Regel kürzere Texte, die konsultiert werden - im Gegensatz zu Texten, die eine extensive Lektüre erfordern. Solche Texte sind Eintragungen in Wörterbüchern oder Nachschlagewerken (im weitesten Sinne) sowie Nachweise in Bibliographien und Verzeichnissen verschiedenster Art.

In einer Reihe von Fällen kann die Kombination der layoutgetreuen Abbildung einer Textvorlage (Image) mit einem volltextdigitalisierten Text sinnvoll sein. Sie ist deswegen prinzipiell in Betracht zu ziehen, weil sich dadurch nicht nur ein differenzierter Zugriff auf einen Text ergibt, sondern sich auch neue Erkenntnismöglichkeiten erschließen können.

Die Volltexterfassung stößt jedoch gerade bei älteren Druckvorlagen häufig an Grenzen, da automatisierte Texterkennungsprogramme hier versagen. Probleme bereiten u.a. die Schriftqualität der Vorlage, Verschmutzungen, uneinheitlicher Schriftsatz und in neuerer Zeit nur selten verwendete Schriftarten (wie beispielsweise Fraktur). Eine ökonomisch vertretbare Volltexterfassung dieser Texte im Ganzen ist deshalb zur Zeit nur in Ausnahmefällen mit befriedigendem Ergebnis möglich.

Vor diesem Hintergrund empfiehlt die AG Inhalt nachdrücklich Pilotprojekte, in denen exemplarisch die Möglichkeiten der Volltexterfassung von Druckvorlagen anhand einer begrenzten Zahl von geeigneten Objekten untersucht wird, um die Entwicklung von Bearbeitungstools für die Retrodigitalisierung voranzutreiben und das Bereitstellen dieser Materialien zu optimieren.

2.2 Das Image-Scannen gedruckter Vorlagen

Schon mit der ersten Stufe der digitalen Konversion - dem reinen Image-Scannen - erreicht man, daß die Texte direkt am Arbeitsplatz des Forschenden, auch außerhalb der Bibliothek und frei von räumlichen und zeitlichen Beschränkungen, bereitgestellt werden können. Der Wissenschaftler kann darüber hinaus an Bestände gelangen, die in der eigenen Bibliothek nicht vorhanden sind oder die aus konservatorischen Gründen nur eingeschränkt benutzbar sind.

2.3 Verbindung der Image-Digitalisierung mit erschließenden Volltexten

Der wesentliche Nachteil der reinen Imagedigitalisierung von Texten besteht darin, daß der gezielte Zugriff auf einzelne Wörter nicht möglich ist. Dieser Nachteil kann durch ergänzenden Erschließungsaufwand teilweise ausgeglichen werden. So ist die Bereitstellung volltextdigitalisierter Inhaltsverzeichnisse und - soweit vorhanden - auch Register, die über elektronische Verknüpfung („Verlinkung") den gezielten Zugriff auf einzelne Imageseiten ermöglichen, ein so wesentlicher Zugewinn an Nutzungsmöglichkeiten, daß die

AG Inhalt empfiehlt, im DFG-Förderprogramm „Retrospektive Digitalisierung von Bibliotheksbeständen" diese kombinierte Technik zur verpflichtenden Mindestanforderung zu machen.

Dem Benutzer wird durch dieses Angebot das mühselige sequentielle Suchen am Bildschirm erspart. Die auf diese Weise zur Verfügung gestellten Navigationsmöglichkeiten sind eine entscheidende Voraussetzung um zu verhindern, daß „Textfriedhöfe" in digitalisierter Form entstehen, wie dies heute häufig bei schlecht erschlossenen Mikroformensammlungen der Fall ist.

Die hier aufgestellten Forderungen an die Retrodigitalisierung lassen sich durch den Einsatz vorhandener Dokumentenmanagementsysteme und ihre gezielte Weiterentwicklung realisieren (vgl. den Bericht der Facharbeitsgruppe Technik zur „Retrospektiven Digitalisierung von Bibliotheksbeständen").

3
Aufbau der Verteilten Digitalen Forschungsbibliothek

Mit dem Förderprogramm „Retrospektive Digitalisierung von Bibliotheksbeständen" sollten insbesondere Projekte unterstützt werden, die

- weiterführende Perspektiven erkennen lassen
- Ansätze für weitergehende Digitalisierungsvorhaben bieten
- einen Anstoß für die längerfristige Zusammenarbeit mit personellen und institutionellen Vertretern der Wissenschaft geben
- in Kooperation mit Verlagen und sonstigen Inhabern von Rechten die Digitalisierung vom urheberrechtsfreien Material auch auf urheberrechtsrelevante Literatur ausdehnen.

Für den Aufbau der Verteilten Digitalen Forschungsbibliothek sollten insbesondere folgende Aspekte berücksichtigt werden:

1. *Thematisch orientierte Sammlungen von herausragendem Interesse für die Forschung*[1]

Wichtige Aspekte sind dabei insbesondere:

- der Beitrag zum Schutz singulärer und bestandsschutzwürdiger Materialien und
- die Verbesserung des Zugangs zu schwer zugänglichen Materialien.

2. *Materialien von grundlegender fachwissenschaftlicher Bedeutung*

Die Auswahl von Materialien im Bereich von forschungsrelevanter Grundlagenliteratur sollte in Kooperation von Bibliothekaren und Fachwissenschaftlern erfolgen. Die Arbeitsgruppe konnte für ihre Empfehlungen in einem ersten Ansatz Listen relevanter Titel auswerten, die aufgrund einer Rundfrage an über 25 Sondersammelgebiets- und anderen großen Bibliotheken erstellt wurden.

---

1 Digitalisierungsvorhaben dieser Art werden z.B. von den Bibliotheken der AG *Sammlung Deutsche Drucke* mit Themenbereichen wie „Deutsche druckgraphische Buchillustration des 15. Jahrhunderts" (BSB München), „Deutsche Drucke des 17. Jahrhunderts zur Festkultur des Barock" (HAB Wolfenbüttel), „Itineraria und deutsche Nordamericana" (SUB Göttingen), „Flugschriften und Kleinschrifttum 1848" (StuUB Frankfurt), „Musikdrucke des 19. Jahrhunderts" (SB zu Berlin - PK) und „Ausgewählte Zeitschriften und Zeitungen des deutschsprachigen Exils" (DDB, Frankfurt a. M.) vorbereitet.

Dabei zeigte sich ein differenziertes Bild für die Einsatzmöglichkeiten der Retrodigitalisierung entsprechend der jeweiligen Textgattung und bei einzelnen Fachgebieten, das die Grundlage für die folgenden Empfehlungen bildet.

3. Materialien mit besonderer Intensität der Nutzung

Eine weitere Möglichkeit zur Auswahl von Bibliotheksmaterialien ist die Intensität der Nachfrage in der Benutzung. Ergänzend zum systematischen Vorgehen in einzelnen Fachgebieten sollte häufig über die Fernleihe genutzte Forschungsliteratur ausgesucht werden, um gezielt dem Bedarf der Benutzer zu entsprechen. Modellhaft könnte ein solches Projekt z.B. auf der Grundlage der Zentralkataloge in Niedersachsen und Hamburg bzw. des Online-Bestellsystems des GBV durchgeführt werden; auch die auf Bestellungen beruhenden Mikrofichesammlungen z.B. der HAB Wolfenbüttel und der SUB Göttingen könnten dazu herangezogen werden.

4. Kooperation Fachwissenschaftler/Bibliothekare bei der Auswahl

Die Auswahl von Bibliotheksmaterialien aus dem Bereich der forschungsrelevanten Grundlagenliteratur wird schwerpunktmäßig sicher an Sammelschwerpunkt- und Spezialbibliotheken erfolgen. Anzustreben ist dabei die Kooperation von Bibliothekaren, wissenschaftlichen Instituten und Fachwissenschaftlern, wie sie für die Fächer Geschichte (München) und Mathematik (Berlin/Göttingen) bereits konkret geplant ist. Im Rahmen solcher Kooperationen sind dabei nicht nur Listen von zu digitalisierender Literatur zusammenzustellen, zu klären sein wird vielmehr auch die jeweils optimale Methode der Digitalisierung und Erschließung. Unterschiede zwischen den einzelnen Fächern sollten dabei im Sinne einer möglichst vielseitigen Vorgehensweise nutzbar gemacht werden.

Erkenntnisse aus den geschilderten Kooperationen könnten darüber hinaus durch Vertreter der genannten Institutionen in andern Orts geplante Digitalisierungsvorhaben eingebracht werden.

4
Empfehlungen für einzelne Textgattungen und Fächer

4.1 Grundfragen

Die Leitfrage für jedes Projektvorhaben beim Aufbau der Verteilten Digitalen Forschungsbibliothek sollte lauten:

Von welchem Material ist mit welcher Digitalisierungstechnik der bestmögliche Nutzen für die Forschung zu erzielen?

Besondere Vorteile der Digitalisierung sind die räumlich und zeitlich unbegrenzte Bereitstellung von Forschungsmaterialien und die Verbesserung der Arbeitsgrundlagen für die Forschung. Bei der Bestandsauswahl für Digitalisierungsprojekte sollten deshalb zerstreute und singuläre Materialien Berücksichtigung finden. Diese dürften zudem nicht in erster Linie für die regionale, sondern vielmehr für die nationale und internationale Forschung von Interesse sein.

Die Nutzungsmöglichkeiten des digitalisierten Materials der Verteilten Digitalen Forschungsbibliothek sollten auch die interdisziplinäre Forschung anregen. Ein kumulierter Index aller volltextdigitalisierten Buchregister beispielsweise kann bei der Suche nach einzelnen Stichwörtern den Interessenten auf digitalisierte Dokumente aus unterschiedlichsten Fachgebieten führen.

4.2 Textgattungen

Enzyklopädien und Nachschlagewerke

Für diese Literaturart besteht eine hohe Nachfrage in der Nutzung. Aufgrund des Nachschlagecharakters sind Enzyklopädien zudem für eine punktuelle Benutzung am Bildschirm geeignet.

Wünschenswert wäre ein Projekt, das sich mit der digitalisierten Bereitstellung älterer, urheberrechtsfreier Enzyklopädien befaßte. Da sie für viele Forschungszweige noch immer von einem hohen Informationswert sind und da andererseits die Bibliotheken, sofern sie nicht über Reprints verfügen, sie aus Gründen der Bestandssicherung nur mit Einschränkungen den Benutzern zur Verfügung stellen können, wäre der Zugang über den PC des Forschers - sei es im Netz oder auf CD-ROM - ein großer Gewinn.

Ein Projekt, das sich zunächst auf einige wenige bedeutende Enzyklopädien beschränkt (z.B. auf Zedler, Krünitz, Ersch/Gruber) sollte untersuchen,

- ob und mit welchem Aufwand nach heutigem Stand der Technik eine automatisierte Erfassung durch eine Texterkennungssoftware möglich ist

- in welcher Weise die Werke, falls nur ein Image-Scanning möglich ist, bei manueller Erfassung der Lemmata am zweckmäßigsten erschlossen werden können. Dabei gilt es zu klären, ob eine Mischung der Lemmata verschiedener Enzyklopädien und eine Verknüpfung mit neuerer Terminologie (Schlagwortnormdatei) sinnvoll ist.

Um eine Abschätzung der Gesamtkosten zu ermöglichen, wäre ein Vorprojekt mit der Einschränkung auf einen kleinen Abschnitt des Alphabets zweckmäßig.

Biographische Nachschlagewerke

Der Bedarf für eine Digitalisierung wird derzeit nicht gesehen, da der Grad der Erschließung und Zugänglichkeit für diese Gattung bereits heute aufgrund der Mikrofichematerialien, die durch automatisierte Register erschlossen sind, als zufriedenstellend betrachtet werden kann.

Bibliographien, Kataloge, Verzeichnisse

Bei der retrospektiven Digitalisierung von Materialien aus diesem Bereich wird bezüglich der Digitalisierungstechnik von Fall zu Fall entschieden werden müssen. In bestimmten Fällen kann es sinnvoll sein, sachlich zusammengehörige Bibliographien oder Verzeichnisse in Form der Image-Digitalisierung zusammenzufassen und durch zusätzliche *finding aids* so aufzubereiten, daß eine Benutzung dieser Werke, die häufig mühsam in verschiedenen Bibliotheken oder an verschiedenen Stellen einer Bibliothek aufgesucht werden müssen, schnell und problemlos am Arbeitsplatz eines Wissenschaftlers möglich wird.

In anderen Fällen ist eine Image-Digitalisierung von begrenztem Nutzen. Eine Reihe von älteren Bibliographien oder Verzeichnissen können in Zukunft nur dann richtig ausgenutzt werden, wenn sie in eine volldigitalisierte Form überführt und damit auf eine bisher nicht mögliche Weise recherchierbar werden. Ein besonders wichtiges Korpus solcher Verzeichnisse liegt in den Meßkatalogen (seit 1564) vor, die interdisziplinär von hohem Interesse und Erkenntniswert sind, bislang aber nur völlig unzureichend ausgewertet werden konnten. Hier müßte eine Volldigitalisierung sogar mit einem erheblichen editorischen Aufwand verbunden werden.

Die Digitalisierung reiner Bibliographien ohne Verknüpfung mit Volltexten erscheint nicht empfehlenswert.

Sprachwörterbücher von historischem Wert

Wörterbücher des 17.-19. Jahrhunderts stellen wichtige kultur- und sprachgeschichtliche Zeugnisse dar. Sie sind sowohl für philologisch-sprach-

wissenschaftliche als auch für literatur- und geschichtswissenschaftliche Fragestellungen unentbehrliche Quellen und Hilfsmittel.

Unter dem Blickpunkt der Literaturversorgung stellt sich jedoch nicht selten das Problem der Erreichbarkeit und der synoptischen Benutzbarkeit dieser Werke. Daran haben auch die Reprintfassungen vieler älterer Wörterbücher, die in den letzten zwanzig Jahren veröffentlicht wurden, nicht viel verändert, da auch diese Reprints nur bedingt verfügbar sind. Durch die Aufnahme eines Grundbestands älterer lexikographischer Standardwerke in die Verteilte Digitale Forschungsbibliothek könnte die Literaturversorgung erheblich verbessert werden. Darüber hinaus kann durch die elektronische Bereitstellung eine qualitative Verbesserung der Wörterbuchnutzung bzw. Wörterbuchforschung erschlossen werden. Innerhalb einer solchen Zielsetzung sind unterschiedliche digitale Erschließungsstufen mit je spezifischer und insgesamt wachsender Nutzungsmöglichkeit zu unterscheiden.

Die erste Stufe der retrodigitalen Wörterbucherschließung stellt eine Erfassung der Werke durch Image-Scannen dar. Verbunden mit einer Indizierung der Wörterbuchimages auf Stichwortebene erschließt sich dem Wissenschaftler bereits ein komfortables Navigieren in der Makrostruktur der Werke ebenso wie neue Möglichkeiten synoptischer Nutzung verschiedener Artikel in parallelen Datenverarbeitungsfenstern.

Die mit den indizierten Image-Digitalisierungen eröffneten wissenschaftlichen Optionen der Wörterbuchbenutzung würden über eine Volltextdigitalisierung in einem erheblichen Umfang erweitert. Unter einer Volltextdigitalisierung wird hier die Zugangserschließung auf der Ebene der Einzelzeichen, Zeichenfolgen, aber auch der Artikelstrukturen verstanden. Trotz der vergleichsweise höheren Kosten für dieses Verfahren kommt ihm im Rahmen der Verteilten Digitalen Forschungsbibliothek aus wissenschaftlicher Sicht exemplarische Bedeutung zu.

Schwerpunkte der Bedeutung einer solchen Volltextdigitalisierung liegen

- im Bereich der Umschreibung wissenschaftlicher Standards bzw. Minimalanforderungen für Wörterbuchdigitalisierungen,
- in der Entwicklung einer textsortenangemessenen Datenstrukturierung und Datenerschließung sowie
- im Aufbau einer erweiterbaren digitalen Plattform für Erforschung, Planung bzw. Entwicklung lexikographischer Projekte.

Kultur-, Literatur- und Fachzeitschriften, Rezensionsorgane, Feuilletons

Für die Digitalisierung der Zeitschriften spricht neben der starken Nachfrage nach derartigen Quellenmaterialien durch die Forschung der hohe Grad der Erschließung (*Index der deutschsprachigen Zeitschriften*, *Rezensionsindex*), der in einer Reihe von Projekten (Frankfurt, Göttingen, Marbach u.a.) geleistet

wurde bzw. wird. Eine Studie sollte den bereits vorhandenen Grad der Erschließung dieser Zeitschriften und Zeitungen klären und Vorschläge für die jeweils sinnvollen Anforderungen für Erschließung und Digitalisierung unterbreiten. Beispielhaft sei an dieser Stelle auch die Erschließung des *Journal des Luxus und der Moden* an der Herzogin Anna Amalia Bibliothek in Weimar erwähnt, bei der einzelne Artikel ausgewertet und indexiert werden. Die Verknüpfung einer derartigen Erschließung mit den Imageabbildungen der Zeitschrift könnte ein Modellfall für das digitale Angebot dieser Textgattung sein. Ein anderes Beispiel ist das durch die Deutsche Forschungsgemeinschaft finanzierte Projekt der Universität Tübingen zur Erschließung des Feuilletons der „Frankfurter Zeitung" 1918-1933.

Als konkrete Projekte werden insbesondere auch die Digitalisierung der *Göttingischen Gelehrten Anzeigen (GGA)* sowie der *Allgemeinen Deutschen Bibliothek (ADB)* empfohlen.

Minimaler Standard für die Bereitstellung der digitalisierten Zeitschriften sollte die Zugriffsmöglichkeit auf einzelne Artikel über elektronische Inhaltsverzeichnisse sein.

Bei Fachzeitschriften ist insbesondere die Volltextdigitalisierung jener Teile in Erwägung zu ziehen, die für die Fachwissenschaft von zentraler, bleibender Bedeutung sind. Solchen Überlegungen wird aus Sicht der AG Inhalt beispielsweise Rechnung getragen, wenn von Mittelalterhistorikern die Digitalisierung des Rezensionsteils des Zentralorgans der deutschen Mittelalterhistorie (des „Deutschen Archivs") oder von Mathematikern die selektive Digitalisierung der „Fortschritte der Mathematik" vorgeschlagen wird.

Monographische Literatur

Für den Bereich der Monographie besteht fächerübergreifend das Problem der objektivierbaren Auswahl. Besonders in den historisch-philologischen Fächern mit ihrer umfangreichen Produktion an Monographien läßt sich bei der Auswahl nur weniger Titel urheberrechtsfreien Materials der Vorwurf der Beliebigkeit schwer von der Hand weisen. Darüber hinaus ist der Stand der Forschung in vielen älteren Monographien meist nur noch partiell aktuell; ohne systematische Ergänzung durch urheberrechtsrelevantes Material ist es kaum möglich, ein sinnvolles Programm zu erstellen.

Handbücher

Im Bereich der monographischen Literatur im weiteren Sinne ist die Digitalisierung von Handbüchern eine für die Forschung wertvolle Möglichkeit. Dabei sollten einige wenige Fächer exemplarisch ausgewählt werden, deren Werke historisch und interdisziplinär von Interesse sind. Die Theologie beispielsweise erscheint für ein solches Vorhaben besonders geeignet.

Im Rahmen von derartigen Modellprojekten sollten die neuen Möglichkeiten zur Nutzung der digitalisierten Handbücher erprobt werden, wobei insbesondere dem Bereich der Erschließung große Aufmerksamkeit zu schenken ist. Ziel sollte es sein, durch konkrete Projekte in der Zusammenarbeit Fachwissenschaft - Bibliotheken Standards für das Anbieten von forschungsrelevanter Grundlagenliteratur in digitalisierter Form zu setzen, bevor eine Prädiktion auch von Verlagsseite einsetzt.

Eine neue Qualität der Nutzung sollte durch die parallele Bereitstellung verschiedener Auflagen erreicht werden. Der interessierte Forscher sollte den komfortablen Zugriff auf alle Auflagen von seinem Arbeitsplatz aus erhalten, entweder im Online-Zugriff über das Netz oder, falls gewünscht, auch offline als CD-ROM. Die jetzige Praxis sieht hingegen in aller Regel so aus, daß mehrere Forschungsreisen zu unterschiedlichen Bibliotheken erforderlich sind und selbst für den Fall des Bestandes in einer Bibliothek unterschiedlichste Aufstellungsorte aufgesucht werden müssen.

Die parallele Nutzung der älteren und neueren Auflagen läßt für die Forschung interessante, zusätzliche Erkenntnisse erwarten und ermöglicht es zugleich, die Gesamtinformation aller Ausgaben nutzbar zu machen. Für den Bereich der Theologie könnten auf diese Weise z.B. große Sachlexika mit umfangreichen Artikeln wie das *Lexikon für Theologie und Kirche* und *Religion und Geschichte in der Gegenwart* genannt werden.

Im Rahmen solcher Modellprojekte sollte zugleich die Kooperation mit den betroffenen Verlagen modellhaft entwickelt werden, da auch urheberrechtsrelevante Auflagen in das Angebot einzubeziehen wären.

Historische Quellen und literarische Texte

In die Digitalisierung einbezogen werden sollten auch historische Quellen und literarische Texte. Eine hohe Forschungsrelevanz ist in der Regel gegeben, die Nachfrage in nationalen und internationalen Wissenschaftskreisen ist gesichert. Eine entsprechende Erschließung, wie sie für die Handbücher beschrieben wurde, ist allerdings auch für zahlreiche Quellenwerke notwendig.

Die Ausgangslage bezüglich der Bedeutung von Quellen und literarischen Texten ist in den einzelnen Fächern sicherlich unterschiedlich. Analog zur Vorgehensweise bei der Digitalisierung von Handbüchern sollten deshalb zunächst gezielt Modellprojekte durchgeführt werden. Für die hier genannten Textgattungen werden insbesondere die folgenden Fächer für geeignet gehalten:

- Germanistik, Philosophie, (z.B. Erstausgaben von Autoren, für die bisher eine Historisch-Kritische Gesamtausgabe - noch - nicht vorliegt)
- Theologie
- Geschichte

- Rechts- und Staatswissenschaft

Für die Parlamentaria und Gesetzesblätter wird die Durchführung eines Pilotprojektes empfohlen, um u.a. Methoden der jeweils adäquaten Digitalisierungstechnik und der qualitativen Erschließung zu erproben. Berücksichtigt werden sollten dabei in jedem Fall methodische Erkenntnisse aus bereits laufenden Projekten mit partiell ähnlicher Aufgabenstellung.

Auch bei den historischen Parlamentaria von der Paulskirche bis zu den Protokollen des Bundestages wird sich der Aufwand der Digitalisierung wohl nur rechtfertigen lassen, wenn sie mit einer tieferen Erschließung als in der gedruckten Form einhergeht.

Archivalische Quellen

Es sollte an geeigneten Beispielen modellhaft untersucht werden, ob sich neue historisch-kritische Editionstypen entwickeln lassen, in denen die genetischen Apparate durch digital verfügbare Images (Faksimiles) der handschriftlichen Textzeugen entlastet bzw. in ihrer Aussagekraft gesteigert werden können (vgl. die im Verlag Stroemfeld-Roter Stern entwickelte, bisher mit konventionellen Mitteln arbeitende Editionsmethodik). Der Anstoß dazu muß wohl von den Fachwissenschaften ausgehen. Textbestände, bei denen eine häufige Benutzung der Autographen mit der Notwendigkeit des Schutzes der Originale zusammentrifft, sollten ergänzend zu den bestehenden wissenschaftlichen Textausgaben als Images digitalisiert werden (z.B. Teile des Goethe-Nachlasses).

Bildquellen[2]

Bis heute ist die Verbreitung von Bildern im Druck aufwendiger als die von Texten. Deshalb konnte das überlieferte Bildmaterial von den Kulturwissenschaften bisher nur im Ausnahmefall seiner Bedeutung entsprechend konsultiert und ausgewertet werden. Die Digitalisierung ermöglicht nun, Bilder wie Texte zu vermitteln und zu nutzen. So sollten in die Digitalisierung auch Materialien einbezogen werden, deren Informationswert mehr oder minder auf ihren bildlichen Elementen beruht, also beispielsweise Emblembücher, Landkarten, Ornament- und Porträtstiche, höfische Festdokumentationen, Plakate oder Flugblätter. Da diese Materialien aus konservatorischen Gründen in der Regel Nutzungsbeschränkungen unterliegen, würde ihre Bereitstellung in der Verteilten Digitalen Forschungsbibliothek von besonderem Wert sein.

Ethnologische und sonstige kulturgeschichtliche Dokumentationsfotografien könnten, soweit sie für die Forschung den Wert primären Quellenmaterials

---

2 Die AG Inhalt dankt L. Heusinger vom Bildarchiv Foto Marburg für diesen Beitrag.

besitzen und die inhaltliche Erschließung eine ausreichende Differenzierung erreicht, berücksichtigt werden.

5
Pilotprojekte beim Aufbau der Verteilten Digitalen Forschungsbibliothek

Als methodische Vorgehensweise zum Aufbau der Verteilten Digitalen Forschungsbibliothek empfiehlt die AG Inhalt die Förderung von Pilotprojekten in unterschiedlichen Bereichen.

Digitalisierungstechnik

Die Technik zur Digitalisierung sollte insbesondere mit dem Ziel der Entwicklung und Erprobung von Verfahren zur Volltexterfassung älterer, in der Regel für die automatisierte Texterkennung nicht zugänglicher Druckvorlagen untersucht werden.

Bereitstellung forschungsrelevanter Grundlagenliteratur

Die Möglichkeiten zur optimalen Bereitstellung von Literatur aus diesem Bereich könnten exemplarisch an Handbüchern der Theologie untersucht werden. Betont werden sollte in diesem Zusammenhang die Chance für Bibliotheken und Fachwissenschaft zur Setzung von Standards, an denen sich in der Zukunft auch rein kommerzielle Projekte von Verlagen messen lassen müssen.

Bereitstellung von Quellenmaterial

Hierunter fallen Pilotprojekte zur Digitalisierung von Kultur- und Literaturzeitschriften, in denen bereits vorhandene Erschließungsinstrumente mit den digitalisierten Materialien zu verknüpfen sind. Analog kann für Parlamentaria und Gesetzesblätter verfahren werden.

6

Rahmenbedingungen für den Aufbau der Verteilten Digitalen Forschungsbibliothek

Der Aufbau der Verteilten Digitalen Forschungsbibliothek wird nur Erfolg haben, wenn folgende Rahmenbedingungen bei den Projektnehmern beachtet werden:

Bereitstellung der digitalisierten Materialien in einer für die Forschung qualitativ ansprechenden Weise

Auf die notwendige Erschließung digitalisierter Materialien wurde bereits an anderer Stelle hingewiesen. So soll beispielsweise im Bereich der Kulturzeitschriften auf bereits vorhandene Ergebnisse aus Erschließungsprojekten aufgebaut werden, damit der zielgerichtete Zugriff sichergestellt ist. Als Minimalanforderung bei der Digitalisierung von Büchern sollten in der elektronischen Bereitstellung die Inhaltsverzeichnisse und - soweit vorhanden - die Register im Volltext angeboten werden, von denen aus dem Benutzer über Verlinkung der inhaltliche Zugriff auf einzelne Imageseiten möglich ist.

Überregionaler Nachweis der digitalen Sammlungen

Auf bibliographischer Ebene ist der überregionale Nachweis lokaler digitaler Sammlungen zu gewährleisten. Lokale Insellösungen, die von der wissenschaftlichen Gemeinschaft nicht wahrgenommen und deshalb nicht adäquat genutzt werden können, sind unter allen Umständen zu vermeiden. Der Nachweis sollte in den regionalen Verbünden und in EROMM erfolgen.

Online- und Offline-Bereitstellung der digitalisierten Dokumente

Die bestehenden und leider in gewisser Form immer wiederkehrenden Netzprobleme können eine komfortable online-Nutzung digitalisierter Dokumente gravierend erschweren. Der punktuelle Zugriff auf einzelne Imageseiten unter inhaltlichen Gesichtspunkten bietet vor diesem Hintergrund den technischen Vorteil, daß der Nutzer nicht in jedem Fall das gesamte Buch über die Datenleitung holen muß, sondern aufgrund der ihm zur Verfügung gestellten volltextdigitalisierten Register und Inhaltsverzeichnisse einzelne Imageseiten bzw. -kapitel zielgerichtet selektieren kann. Außerdem wird empfohlen, auch offline-Versionen digitalisierter Bücher (CD-ROM) anzubieten und häufiger genutzte Materialien lokal oder regional zu spiegeln.

Verhandlungen zur Klärung urheberrechtlicher Fragen beim Aufbau der Verteilten Digitalen Forschungsbibliothek

Die Lösung der Problematik des Urheberrechts muß im weiteren Verlauf der Digitalisierungsinitiative einen zentralen Platz einnehmen, da eine Beschränkung auf urheberrechtsfreie Materialien auf Dauer sicher nicht ausreichend ist.

Ein Blick auf einzelne Fächerlisten der SSG- und DBV-Bibliotheken hat die Problematik des (durch die AG Inhalt zunächst vorgegebenen) Auswahlkriteriums „Urheberrechtsfreiheit" deutlich gemacht. Vermieden werden sollte u.a. die Aufsplitterung fortlaufender Werke in urheberrechtsfreie (und deshalb digitalisierbare) und urheberrechtsrelevante und deshalb zunächst nicht digitalisierbare Teile. Auch in solchen Fällen sollten Verhandlungen mit Verlagen aufgenommen werden, um das gesamte Werk digitalisieren zu können. Dabei sollen auch kooperative Vorgehensweisen erprobt werden.

Aufbau der Verteilten Digitalen Forschungsbibliothek mit Unterstützung von Service- und Kompetenzzentren

Die Urheberrechtsproblematik aber ebenso auch technische und inhaltliche Aspekte machen die Errichtung von Service- und Kompetenzzentren erforderlich,[3] die auch wichtige Koordinationsaufgaben übernehmen sollen. Bereits bestehende Kompetenz auf dem Gebiet der retrospektiven Digitalisierung soll hier nicht nur in der pilothaften Anwendung geeigneter Systemlösungen kontinuierlich fortentwickelt werden. Ziel ist insbesondere auch die Weitergabe des Wissens und der Erfahrungen an andere Projektträger des neuen Förderprogramms, u.a. im Rahmen von Workshops. Die rechtlichen, technischen und organisatorischen Lösungen für retrospektive digitalisierte Dokumente sollen nicht isoliert von den Anforderungen durch aktuelle digitale Dokumente erfolgen. Gefragt sind systemübergreifende Lösungen. Sinnvoll ist deshalb auch die Zusammenarbeit mit Der Deutschen Bibliothek, die als nationale Archivbibliothek Konventionen und Verfahren zur Langzeitverfügbarkeit digitaler Publikationen entwickelt.

Die Einführung und Einhaltung von Standards kann auf diesem Wege erleichtert werden.

Die personelle und technische Infrastruktur in den beiden Zentren sollte darüber hinaus sicherstellen, daß die Bereitstellung der digitalisierten Materialien überregional, schnell und vor allem auch dauerhaft erfolgt.

---

3 Geplant sind Einrichtungen an der SUB Göttingen und der BSB München, die im Frühjahr/Sommer 1997 ihre Arbeit aufnehmen werden.

Retrospektive Digitalisierung von Bibliotheksbeständen für eine Verteilte Digitale Forschungsbibliothek

Bericht der Arbeitsgruppe Technik zur Vorbereitung des Programms „Retrospektive Digitalisierung von Bibliotheksbeständen" im Förderbereich „Verteilte Digitale Forschungsbibliothek"

Mitglieder der Arbeitsgruppe:

Prof. Dr. Rudolf Bayer, Technische Universität München, Fakultät für Informatik

Dr. Jürgen Bunzel, Deutsche Forschungsgemeinschaft, Bonn

Dr. Marianne Dörr, Bayerische Staatsbibliothek München

Dr. Reinhard Ecker, Beilstein-Institut bzw. ABC Datenservice GmbH, Frankfurt/Main

Dipl.-Math. Heinz-Werner Hoffmann, Hochschulbibliothekszentrum NRW, Köln (als Gast für die AG der Verbundsysteme)

Dr. Norbert Lossau, Niedersächsische Staats- und Universitätsbibliothek Göttingen (DFG-Projekt 'Verteilte Digitale Forschungsbibliothek')

Prof. Dr. Elmar Mittler, Niedersächsische Staats- und Universitätsbibliothek Göttingen

Dipl.-Inf. Christian Mönch, FB Informatik der J.W. Goethe-Universität Frankfurt

Dr. Wilhelm R. Schmidt, Stadt- und Universitätsbibliothek Frankfurt

Dr. Hartmut Weber, Landesarchivdirektion, Stuttgart

Arbeitssitzungen am 14. Mai 1996 (Frankfurt a. M.), 29.-30. Juli 1996 (München), 12.-13. Dezember 1996 (Göttingen)

Die Retrodigitalisierung von Bibliotheksbeständen

Der Bibliotheksausschuß und die Kommission für Rechenanlagen der Deutschen Forschungsgemeinschaft (DFG) haben sich in ihren gemeinsamen Empfehlungen „Neue Informations-Infrastrukturen für Forschung und Lehre" dafür ausgesprochen, die Nutzung der neuen Kommunikations- und Publikationstechniken zur Verbesserung der wissenschaftlichen Arbeitsbedingungen beim Zugriff und bei der Verarbeitung von Literatur, sowie von wissenschaftlichen Daten und Informationen verstärkt zu fördern. Um elektronische Texte direkt am Arbeitsplatz des Wissenschaftlers bereitzustellen soll in einem Kernbereich der Förderung wissenschaftliche Forschungsliteratur aus den Beständen von Bibliotheken digitalisiert und über Kommunikationsnetze zugänglich gemacht werden.

Zur Vorbereitung des neuen Programms der retrospektiven Digitalisierung wurde eine AG Technik ins Leben gerufen. Ihre Aufgabe ist die Bewertung der heute zur Verfügung stehenden technischen Möglichkeiten zur Digitalisierung, Speicherung, Verwaltung und Bereitstellung von digitalen Dokumenten. Die ersten Ergebnisse dieser Untersuchung wurden in dem vorliegenden Bericht zusammengefaßt und sollen potentiellen Antragstellern des neuen Förderprogramms als konkrete Hilfestellung dienen.

Einführung

Das Angebot an Bibliotheksmaterialien in elektronischer Form hat in den letzten Jahren in beträchtlichem Umfang zugenommen. Die Fragestellung, ob Publikationen nur in elektronischer Form, als Druck und in elektronischer Form oder nur als Druck vorliegen sollen, wird in zunehmendem Maße Thema der bibliothekarischen wie der fachwissenschaftlichen Diskussion. Dabei kann man bei der Literatur aus jüngster Zeit davon ausgehen, daß sie in der Regel bereits bei der Entstehung, spätestens aber für den Druck, in elektronische Form gebracht wird. In zunehmendem Umfang wird aber auch verlangt, bereits gedruckt vorliegende Literatur älterer Jahrgänge direkt am (EDV-) Arbeitsplatz verfügbar zu haben. Der räumlich und zeitlich unbegrenzte Zugriff auf solche ansonsten vielleicht nur schwer beschaffbare oder häufig nachgefragte Bibliotheksbestände kann so realisiert werden.

Das neue Förderprogramm hat deshalb seinen Schwerpunkt dezidiert auf die retrospektive Digitalisierung von Bibliotheksbeständen gelegt.

Der Aufbau einer Verteilten Digitalen Forschungsbibliothek (VDF) bedeutet für deutsche Bibliotheken in technischer und organisatorischer Hinsicht das Betreten von Neuland. Ziel ist es, die Ergebnisse der Digitalisierungsprojekte für Forschung und Studium möglichst rasch und umfassend zugänglich zu machen, um die Akzeptanz dieser neuen Bibliotheksdienstleistung zu demon-

strieren und die Dienste in Reaktion auf Benutzerbedarf und Benutzungsanforderungen sukzessive weiter zu verbessern.

Technische Grundlage für die Bereitstellung digitalisierter Bibliotheksbestände werden in erster Linie Dokumentmanagementsysteme (DMS) und Multimedia-Ausstattungen sein, die zukünftig zum standardmäßigen Funktionsumfang lokaler Bibliothekssysteme gehören werden. Beschaffungsmittel für solche Ausstattungen sind im Hochschulsonderprogramm III ausgewiesen.

Ein wichtiges Ziel ist es jedoch, von vornherein auch einen integrierten und einheitlichen Zugriff auf die Gesamtheit der digitalisierten Bestände zu ermöglichen. Dies erfordert die Föderation der unterschiedlichen lokalen Lösungen im Kontext einer verteilten digitalen Bibliothek. Hierfür müssen gemeinsame Konventionen und „good practices" vereinbart werden.

Gerade für kleinere Einrichtungen wird es nicht immer möglich sein, rasch die erforderlichen lokalen Systemausstattungen zu schaffen und aus eigener Kraft das erforderliche Know-How aufzubauen.

Daher kommt insbesondere in der Anfangsphase der Entwicklung sogenannten Service- und Kompetenzzentren eine besondere Bedeutung zu, wie auch Erfahrungen aus bereits laufenden Digitalisierungsinitiativen in den Vereinigten Staaten, Großbritannien, Frankreich oder Australien zeigen.[1] Der Aufbau derartiger Zentren ist an der Staats- und Universitätsbibliothek (SUB) Göttingen und der Bayerischen Staatsbibliothek (BSB) München vorgesehen. Zu den Aufgaben der Kompetenzzentren zählen u.a.:

- Aufbau einer Basis-Infrastruktur zur raschen, überregionalen Bereitstellung der Ergebnisse von Digitalisierungsprojekten im Internet,
- Aufbau prototypischer Systeme für Dokumenten-Management und Präsentation der „Verteilten Digitalen Forschungsbibliothek" im WWW,
- Verknüpfung der „Verteilten Digitalen Forschungsbibliothek" mit den vorhandenen Bibliotheksverbundsystemen,
- Anpassung und Weiterentwicklung vorhandener Systeme,
- Initiativfunktion bei der Vereinbarung von Konventionen, Standards und „good practices",
- Einbindung lokaler Lösungen in das Gesamtsystem einer „Verteilten Digitalen Forschungsbibliothek",

---

1 Vereinigte Staaten: American Memory (1. grosse Digitalisierungsinitiative), Home Page: American Memory from the Library of Congress (http://lcweb2.loc.gov/) und National Digital Library Federation (http://lcweb.loc.gov/loc/ndlf/); Großbritannien, eLib Home page (http://ukoln.bath.ac.uk/elib/); Australian Cooperative Digitisation Project, 1840-45, (http://www.nla.gov.au/ferg/)

- Sicherung der dauerhaften überregionalen Bereitstellung der digitalen Dokumente.

Zudem stehen sie als Ansprechpartner für andere Bibliotheken und Institutionen im Bereich der retrospektiven Digitalisierung von Bibliotheksmaterialien zur Verfügung.

In diesem Zusammenhang ist auch die Bedeutung der kooperativen Zusammenarbeit aller Beteiligten beim Aufbau der VDF hervorzuheben. Der Leitgedanke einer „National Digital Library Initiative", wie er sich in den Vereinigten Staaten im Rahmen der nationalen Digitalisierungsinitiative entwickelt hat, sollte auch für die deutsche Initiative tragend werden.

Unter Beachtung der Komplexität des gesamten Bereiches der Digitalisierung hat sich die AG Technik entschlossen, in dem vorliegenden Bericht gewisse Schwerpunkte zu setzen. Diese betreffen zum einen die Bibliotheksmaterialien, zu denen Aussagen getroffen werden. Es erscheint zum jetzigen Zeitpunkt nicht möglich, auf die ganze Vielfalt dieser Materialien einzugehen (Photos, Karten, Bildvorlagen etc.). Es werden daher in erster Linie die technischen Rahmenbedingungen für eine digitale Konversion von Büchern untersucht.

Zum anderen ist die Erschließung der digitalisierten Dokumente ein umfassender und äußerst vielschichtiger Komplex. Sie erstreckt sich von der reinen Bilderfassung über eine Volltexterfassung bis zur Strukturierung der Texte mit *SGML (Standard Generalized Markup Language*) oder der Umwandlung in das Austauschformat *PDF (Portable Document Format)*. Die speziell auch im angloamerikanischen Bereich angewandte Strukturierung von digitalisierten Dokumenten in SGML richtet sich dabei zunehmend nach den jüngst entwickelten Richtlinien der *TEI (Text Encoding Initiative*), die ein sorgfältig ausdifferenziertes Beschreibungsinstrumentarium für elektronische Texte zur Verfügung stellen. Derart strukturiert werden hier im übrigen nicht nur die Dokumente selbst, sondern auch die sog. 'finding aids', also Katalogeinträge, Register etc.

Im Zusammenhang mit dem Förderprogramm der DFG ist davon auszugehen, daß der Schwerpunkt der Aktivitäten zunächst auf gedruckt vorliegenden Materialien liegen wird.

In einem ersten Schritt werden hier Bilder der gedruckten Vorlagen erzeugt. Erfahrungen aus Projekten im Bibliotheksbereich (vgl. DFG-Projekt zur Digitalisierung der Titelblätter von Beständen der Bibliothek „Öttingen-Wallerstein"), in denen bereits heute Bild-Digitalisierungen bereitgestellt werden, zeigen, daß der Benutzer großes Interesse an solchen Images hat.

Die zweite Stufe der digitalen Konversion, die Volltexterfassung, ist bei älteren Büchern mit Problemen behaftet. Uneinheitlicher Schriftsatz, Vergilbungen und in neuerer Zeit nur selten verwendete Schriftarten (z.B. Fraktur) be-

reiten bei einer automatisierten Texterkennung große Schwierigkeiten. Ist das Erstellen einer digitalen Volltextfassung aus diesen Gründen ökonomisch nicht durchführbar, ist der gezielte Zugriff auf einzelne Wörter im Text nicht möglich. Um so größere Bedeutung kommt daher bei der reinen Bilddigitalisierung einer ergänzenden Erschließung der Texte zu. Über volltextdigitalisierte Inhaltsverzeichnisse und - soweit vorhanden - Register wird dem Benutzer der punktuelle Zugriff auf einzelne Seiten-Bilder ermöglicht.

Langfristiges Ziel wird aber sein, nicht nur diese Materialien zu einem späteren Zeitpunkt als Volltexte zur Verfügung zu stellen sondern möglichst bald, auch in Kooperation mit Verlagen und anderen Inhabern von Rechten, neuere Literatur in eine digitale Forschungsbibliothek aufzunehmen.

Der vorliegende Bericht legt als Grundschema bei der Behandlung technischer Detailfragen die einzelnen Schritte bei der Durchführung eines Digitalisierungsvorhabens zugrunde:

1. Digitales Erfassen
2. Speichern
3. Erschließen und Verwalten
4. Suchen und Zugreifen
5. Bereitstellen und Nutzen
6. Rechteverwaltung

Im folgenden wird ausführlich auf die Themenbereiche 1 bis 5 eingegangen. Mit dem Bereich 6, der Rechteverwaltung, wird man sich zu einem späteren Zeitpunkt eingehend befassen.

1
Digitales Erfassen

1.1 Scanner

Der Scanner ist ein Lesegerät, das über eine geeignete Software (gedruckte) Vorlagen für die Weiterverarbeitung mit einem Computer in maschinenlesbare Form umwandelt.[2]

Er wird als Peripheriegerät an den Computer angeschlossen. Dabei ist es von Vorteil, wenn er über eine SCSI-Schnittstelle als Subsystem angesteuert werden kann. Diese Schnittstelle - zur Zeit SCSI-2 - erlaubt neben dem gleichzeitigen Anschluß mehrerer intelligenter Subsysteme auch die unproblematische Anbindung dieser Systeme an den Computer. Für den Einsatzzweck der Digitalisierung ist zudem die hohe Übertragungsgeschwindigkeit der Daten von Bedeutung.

Die durch den Scanner erzeugten Bilder oder Images werden in Pixel (Bildpunkte) zerlegt. Für die Strukturierung dieser Images gibt es eine Vielzahl unterschiedlicher Formate, auf die an anderer Stelle noch ausführlich eingegangen wird.

Scanner sind in unterschiedlicher Ausprägung mit jeweils spezifischen Funktionalitäten und in allen Preisklassen auf dem Markt: Handscanner, Flachbettscanner, Einzugscanner und Trommelscanner.[3] In jüngster Zeit wurde diese Palette um einen neuen Typ bereichert, den sog. Buch- oder Aufsichtscanner.

Handscanner

Der Handscanner, praktisch aufgrund seiner Größe und, wie ein Laptop, gut zu transportieren, kann beim Scannen mit einer Auflösung von bis zu 400 dpi bereits durchaus respektable Leistungen erbringen und auch für farbige Vorlagen eingesetzt werden. Aufgrund seiner geringen Lesebreite (maximal ca. 11 cm) ist er für die Digitalisierung größerer Textmengen ungeeignet sowie aus Gründen der Bestandserhaltung (direkte Berührung) bedenklich.

---

2 Eine anschauliche Beschreibung der Funktionsweise des Scanners erhält man bei: Wolfgang Limper, *OCR und Archivierung*, München 1993, S.77ff.

3 Zu einer Übersicht verschiedener Scannertypen siehe: Heiner Henninges, *Scannen: Technik und Praxis*, München 1994, S. 61 ff., 102 ff.

Flachbettscanner

Der Flachbettscanner hat von der Form her die größte Ähnlichkeit mit einem kleinen Bürokopierer. Die Vorlage wird auf eine Glasplatte gelegt, ein Schrittmotor bewegt eine Sensoreneinheit (CCD-Zeile) samt Optik zum Abtasten an den aufgelegten Materialien vorbei. Das Scannen von farbigen Vorlagen bereitet keine Probleme, Auflösungen von 600 dpi sind keine Seltenheit mehr. Durch Interpolation können bis zu 2400 dpi erreicht werden. Neben dem gängigen A4-Scanner werden auch A3- und in Sonderfällen A0-Modelle angeboten.

Wie der Kopierer auch hat der Flachbettscanner beim Einsatz für das Scannen von Büchern einen großen Nachteil: da die Vorlagen möglichst dicht auf die Glasplatte aufgelegt werden müssen, ist ein gewisser Druck auf den Buchrücken unvermeidlich. Dieser nicht gerade schonende Umgang mag bei neuerer Literatur noch hingenommen werden; für die geplante Digitalisierung älterer, in der Erhaltung gefährdeter oder besonders schützenswerter Bücher ist dieser Typ des Scanners sicher nicht einsetzbar.

Einzugscanner

Während beim Flachbettscanner die Abtasteinheit an der Vorlage vorbeigeführt wird, ist es beim Einzugscanner die Vorlage, die bewegt wird. Bezüglich Auflösung und Farbscannen kann man sie in etwa mit dem Flachbettscanner vergleichen. Sie können in der Regel Vorlagen im Format A3 verarbeiten, möglich sind Formate bis A0.

Die Stärke des Einzugscanners liegt in der Möglichkeit der raschen Verarbeitung großer Mengen. Können die Vorlagen für den Einzelblatteinzug aufbereitet werden (z.B. durch das Aufschneiden von Zeitschriftenheften), ist dieser Scannertyp für die Massendigitalisierung sicher eine gute Wahl.

Trommelscanner

Der Trommelscanner wird heute in erster Linie bei der professionellen Bildverarbeitung im Reprobereich eingesetzt und kann extrem hohe Auflösungen (bis 4000 dpi) erreichen. Für das Scannen von Büchern ist seine Mechanik, die das Spannen der Vorlage auf eine Trommel erfordert, nicht geeignet.

Buch- oder Aufsichtscanner

Der jüngste unter den oben genannten Scannertypen ist der Buch- oder Aufsichtscanner. Beide Namen sind sprechend und bezeichnen zum einen das Einsatzgebiet dieses Geräts, das Scannen gebundener Bücher, und zum anderen seine Funktionsweise, das Scannen mit einem Lesekopf von oben auf das Buch herab.

Bei der Entwicklung dieses Scannertyps hat sicher die technische Ausrüstung für die Mikroverfilmung Pate gestanden. Deutlich wird dies besonders bei dem von der Firma Zeutschel (Tübingen) angebotenen Buchscanner *Omniscan 3000* mit Buchwippe. Die Standardausstattung bei dieser Ausführung mit Grundgestell, vertikaler Säule, Beleuchtungsvorrichtung und Buch-Aufnahmewippe mit Glasplatte wird Mikroverfilmern bekannt vorkommen. Zu einem Scanner wird dieses System erst durch den an einer vertikalen Säule oberhalb der Auflage befestigten Scan-Kopf, einen CCD-Zeilenscanner. Dieser stammt von Kodak und wurde dort für den *Kodak Imagelink 200*-Buchscanner eingesetzt.

Die Art der Ausstattung zeigt, worauf bei diesem Scanner Wert gelegt wurde: die Möglichkeit des schonenden Umgangs mit dem (alten) Buch. Die Buchwippenfunktion ermöglicht lt. Herstellerangabe das Scannen von Büchern mit einer Dicke bis zu 15 cm.

Von Minolta wird der Scanner *PS3000* angeboten. Anfänglich nur als geschlossenes System zum Anschluß an einen Digitalkopierer oder Drucker verwandt, gibt es ihn seit kurzem auch mit einer Schnittstelle zur Anbindung an den PC.

Ein Probeeinsatz dieser beiden Scanner in der Fotostelle der SUB Göttingen erbrachte - beim Scannen eines Buches (Oktav-Format) von 300 Seiten (=156 Aufnahmen) - eine Stundenleistung von 156 Scans (Minolta), 104 Scans (Zeutschel ohne Buchwippe) und 62,4 Scans (Zeutschel mit Buchwippe).

Ein weiterer Buchscanner wurde im Januar 1997 von der Firma Rank Xerox (XBS, Düsseldorf) auf den Markt gebracht. Funktionalität und Einsatzmöglichkeiten sind prinzipiell der des Minolta-Produkts vergleichbar.

Im Überblick bieten sich die technischen Daten dieser drei Buchscanner wie folgt dar:

Technische Daten	Minolta Buchscanner PS3000	Zeutschel (Kodak) Buchscanner Omniscan 3000 mit Buchwippe	Xerox Di Bookeye
Vorlagen-format	bis DIN A3	bis DIN A2	bis DIN A3 (optional DIN A2)
Vorlagen-stärke	bis 10 cm	bis 15 cm	bis 10 cm
Auflösung	400 dpi	A3 und A4: 400 dpi A2: 300 dpi	300 dpi
Scanmodus	Text, Photo	keine Angabe	Text, Photo
Bildwieder-gabe	bitonal s/w; (rechnerisch auch Grau-stufen)	bitonal s/w; (rechnerisch auch Graustufen)	bitonal s/w; (rechnerisch auch Grau-stufen)
Scange-schwindigkeit	1,27 Sek./A4	5 Sek./ A4, ca. 9 Sek./ A3	2,5 Sek./ A4, 3,2 Sek./ A3 4,0 Sek./ A2
Schnittstelle zum PC	z.Zt. Video-Schnittstelle; ISIS-Schnitt-stelle geplant	SCSI 2-Schnittstelle; ISIS-Schnittstelle geplant	Fujitsu-kom-patible Video-schnittstelle (M3097); ISIS-Schnitt-stelle wird zur Zeit erprobt
Daten-Ausgabe	TIFF-G3/G4	TIFF-G4	TIFF-G4

Kamerascanner

Als Spezialist für alte Dokumente und Handschriften wird von IBM der *Pro/3000* Kamerascanner angeboten. Die Firma weist ausdrücklich auf die spezifische Einsatzmöglichkeit dieses Gerätes hin. So wurde er beispielsweise für die Digitalisierung alter Handschriften in der Vatikan-Bibliothek eingesetzt sowie zur Zeit für die Bestände der Lutherhalle in Wittenberg. Die exzellente Qualität und die präzise Farbwiedergabe gehen allerdings zu Kosten der Scanzeit. Hier werden ca. 8 Minuten pro Scan gerechnet.

In Schweden wurde für den Einsatz im Archivbereich ein Kamerascanner für bitonale, Halbton- und Farbvorlagen entwickelt, dessen Vorteile vom Hersteller neben dem großen Schärfentiefebereich (bis zu 25 cm) insbesondere in der Möglichkeit zum schnellen Ausdruck gesehen werden, der durch die Verbindung mit einem in Deutschland entwickelten Spezialmodul erreicht wird. Die Bilddaten werden dabei mit einer hohen Auflösung unter Umgehung des internen Drucker-Controllers direkt über ein Hochgeschwindigkeitskoaxialkabel an den Drucker (z.B. HP-Laserjet 4v) geleitet.

Die speziellen Funktionalitäten der beiden hier erwähnten Scanner schlagen sich allerdings auch im Preis nieder, der bei beiden Scannern je nach Ausstattung die 100.000 DM-Grenze übersteigen kann.

1.2 Scan- und Bildbearbeitungssoftware

Jeder der zuvor genannten Buchscanner wird über eine eigene Software angesteuert, die neben dem Einlesen der Vorlage auch Funktionalitäten der Bildbearbeitung anbietet. Erwähnt seien beim Einscannen das automatische Entfernen des Schattens von Falz und Rändern, das Scannen im Text- und Fotomodus und eine 'Finger-erase'-Funktion. Standardbildbearbeitungsfunktionen sind Kontrastverbesserung, Drehen, Ausrichten, Skalieren etc.

Weiter Möglichkeiten zur Bearbeitung der Images wie das Schreiben zusätzlicher Informationen in den TIFF-Header des digitalen Masters, bietet standardmäßig keines der eingesetzten Programme. Die SUB Göttingen strebt aus diesem Grund in Kooperation mit einem Systemintegrator, der Firma Satz-Rechen-Zentrum (SRZ) in Berlin, die Entwicklung einer Scan- und Bildbearbeitungssoftware an, die alle Erfordernisse der Imageerstellung und -bearbeitung, wie sie in dem vorliegenden Bericht definiert werden, erfüllen.

1.3 Erstellen der Images

Die Umwandlung gedruckter Vorlagen in digitale Dokumente ist grundsätzlich auf zwei Wegen vorstellbar:

1. Die Digitalisierung direkt vom Buch
2. Die Verfilmung des Buches mit anschließender Digitalisierung des Mikrofilms[4]

Ein Blick auf laufende Digitalisierungsvorhaben zeigt, daß beide Verfahren gängig sind. Die Library of Congress hat in ihren Ausschreibungen für exter-

---

4 Für die Digitalisierung vom Mikrofilm ist der Abschlußbericht der AG „Digitalisierung" des DFG Unterausschusses „Bestandserhaltung" zu beachten. Zudem sollten die Anforderungen an die Verfilmung zur Langzeitarchivierung, wie sie in dem Projekt VD 17 erarbeitet wurden, auch im Programm zur „Retrodigitalisierung" beachtet werden.

ne Scan-Dienstleister detaillierte Konditionen für beide Vorgehensweisen formuliert.

Im Rahmen der nationalen Digitalisierungsinitiative in Australien zu Materialien aus der Zeit von 1840-1845 wird grundsätzlich der Weg über die Mikroverfilmung gegangen.

Vorhandene oder eigens für den Zweck der Digitalisierung erstellte Mikrofilme lassen sich vergleichsweise kostengünstig mit Hilfe spezieller Mikrofilmscanner digitalisieren. Die Filmdigitalisierung wird als Serviceleistung angeboten. Die Digitalisierung vom Mikrofilm führt zu besonders guten Ergebnissen und läßt sich besonders wirtschaftlich durchführen, wenn bei der Erstellung der Mikroformen und bei der Filmdigitalisierung selbst die entsprechenden Hinweise der Arbeitsgruppe „Digitalisierung" des Unterausschusses Bestandserhaltung der Deutschen Forschungsgemeinschaft beachtet werden.[5] So sollen als Mikroform Rollfilme 35mm möglichst mit Bildmarken (Blips) verwendet werden, die weitgehend automatisch digitalisiert werden können. Die Filme sollen mindestens eine den DIN-Normen entsprechende Qualität hinsichtlich der Filmdichte und der Wiedergabeschärfe (Lesbarkeit) aufweisen. Die einheitliche Ausrichtung und Positionierung der Vorlagen (Bücher) und ein einheitlicher Verkleinerungsfaktor über einen kompletten Film hinweg fördern einen weitgehend automatischen und damit rationellen Digitalisierungsvorgang. Schließlich erleichtert eine gute Strukturierung des Mikrofilms mit einer durchdachten Filmorganisation und Aufnahmedokumentation die mit der Digitalisierung zu verbindende formale und inhaltliche Aufbereitung der digitalisierten Images.

Da ordnungsgemäß verarbeitete Mikrofilme auf Polyesterunterlage als alterungsbeständige Informationsträger gelten, soll immer dann über die Zwischenstufe des Mikrofilms digitalisiert werden, wenn damit zugleich Sicherungs-, Schutz oder Erhaltungszwecke für Objekte verfolgt werden, die in ihrer Erhaltung gefährdet oder bereits beschädigt sind. Darüber hinaus kann es sich als wirtschaftlicher erweisen, insbesondere Bücher und andere Vorlagen, die nicht mit Flachbett- oder Einzugscannern rationell verarbeitet werden können, über die Zwischenstufe des Mikrofilms zu digitalisieren, da beim heutigen Preisgefüge bei solchen Objekten die Filmdigitalisierungskosten zuzüglich der Verfilmungskosten vielfach unter den Kosten für die unmittelbare Digitalisierung liegen. Der zusätzlich entstandene hochwertige Mikrofilm steht auch in diesen Fällen als relativ anspruchslos zu lagernder analoger Langzeitspeicher zur Verfügung, der unter anderem beliebig oft zur Digitalisierung und ggf. zusätzlich für den Zweck der Fernleihe herangezogen werden kann.

---

5 Marianne Dörr und Hartmut Weber: „Digitalisierung als Mittel der Bestandserhaltung? Abschlußbericht einer Arbeitsgruppe der Deutschen Forschungsgemeinschaft", in: ZfBB 44 (1997) 1, S. 55-78

Bei der Erstellung des Mikrofilms wird zukünftig auch verstärkt die Entwicklung der COM (Computer Output on Microfilm)-Techniken zu berücksichtigen sein. Diese sieht zunächst eine qualitativ hochwertige Digitalisierung, dann die Konversion der digitalen Vorlage auf Mikrofilm vor.[6]

Prinzipiell sollte jedes Buch, nicht zuletzt aus konservatorischen und ökonomischen Gründen, nur einmal gescannt oder verfilmt werden. Die Qualität der erstellten Images muß demnach so beschaffen sein, daß eine etwaige Weiterverarbeitung wie Komprimierung und Konvertierung, aber auch die Bearbeitung mit einer Texterkennungssoftware, von diesen „Erst-" bzw. „Einmal"-Scans vorgenommen werden kann. Unterschiedliche Versionen sind deshalb von einer Vorlage zu erstellen.

1.3.1 Auflösung beim Scannen

Die Entscheidung über die zu wählende Auflösung sollte grundsätzlich im Zusammenhang mit der geplanten Verwendung der Scans und der Art der zu digitalisierenden Vorlage gesehen werden. Die Arbeitsgruppe „Digitalisierung" hat in ihrem Abschlußbericht in Anlehnung an amerikanische Veröffentlichungen vorgeschlagen, beim Digitalisieren vom Original oder vom Mikrofilm die Auflösung von der Schriftzeichengröße der Vorlagen abhängig zu machen.[7] Sie orientiert sich dabei an dem für die Beurteilung der Wiedergabequalität graphischer Zeichen international gebräuchlichen Quality Index (QI) und schlägt vor, für die Präsentation von Images unter Berücksichtigung der Speicheranforderungen eine mittlere Qualität (QI=5) festzulegen. In Verbindung mit normalem Schriftgut und gängigen Druckwerken sollen demnach beim bitonalen Digitalisieren Auflösungen von mindestens 300 dpi angestrebt werden.

Technisch möglich und in großen amerikanischen Digitalisierungsprojekten als Standard angestrebt wird für s/w-Vorlagen eine Auflösung von 600 dpi.[8]

---

6 Digital to Microfilm Conversion: A Demonstration Project 1994-1996; Final Report to the National Endowment for the Humanities, PS-20781-94, Anne R. Kenney, Cornell University Library, Department of Preservation and Conservation, Ithaca, NY 14853, URL: (http://www.library.cornell.edu/preservation/pub.htm)

7 a.a.O, S. 64f.; höhere Auflösungen sind anzustreben, wenn die digitale Form die alleinige Überlieferungsform ist, s. S. 73f.

8 Vgl. die Empfehlungen von Anne R. Kenney und Stephen Chapman in *Digital imaging for libraries and archives,* Ithaca, NY: Dept. of Preservation, Cornell University Library, 1996. Umgesetzt wird diese Empfehlung u.a. in den Projekten *JSTOR* (großes Zeitschriftendigitalisierungsprojekt der A.-W- Mellon Foundation, http://www.jstor.org/), *Making of Amerika* MOA (unter Beteiligung der Cornell University, University of Michigan, http://moa.cit.cornell.edu/MOA/moa-main_page.html) und *Open Book* (Yale University, http://www.library.yale.edu/preservation/pobweb.htm).

Diese Auflösung stellt sicher, daß das Digitalisat als Grundlage für andere Ausgabeformen von hoher Qualität (hochqualitativer Ausdruck, COM) dienen kann.

Beim Digitalisieren mit Graustufen sollten Auflösungen zwischen 250 und 300 dpi gewählt werden, Farbvorlagen benötigen eine vergleichbare Qualität.

Wird zu einem späteren Zeitpunkt die Behandlung der digitalisierten Dokumente mit einer Texterkennungssoftware nicht ausgeschlossen, wird eine Auflösung von mindestens 400 dpi empfohlen. Tests, unter anderem an dem renommierten *Electronic Text Center* an der University of Virginia, haben hier eindeutig ergeben, daß gerade kleine Schriftgrößen bei einer Bearbeitung mit OCR-Software im Falle von 400 dpi deutlich besser erkannt werden als bei 300 dpi.[9]

Beim Digitalisieren von Fotografien sind je nach Detailreichtum geringere Auflösungen ausreichend oder höhere Auflösungen (bis 600 dpi) erforderlich. Wichtiger ist dabei allerdings die Digitalisierung mit Graustufen. Bei gerasterten Abbildungen in Büchern darf die Auflösung beim Digitalisieren die Rasterauflösung nicht überschreiten.

1.3.2 Farbtiefe

Beim Scannen direkt vom Buch (bitonal s/w) wird in der Regel mit einer Farbtiefe von 1 bit per Pixel gearbeitet werden. Handschriften, Zeichnungen mit Bleistift oder Farbstift, (auch Bleistiftanmerkungen in Verbindung mit gedruckten Texten), Schreibmaschinenschrift mit Gewebefarbbändern, farbige Illustrationen und Zeichnungen, Darstellungen mit verschiedenen Grauabstufungen und Fotografien in schwarz-weiß oder Farbe sollen je nach Vorlage mit 16 oder 256 Graustufen digitalisiert werden. Entsprechendes gilt für die Digitalisierung vom Mikrofilm.[10] Sollen Grautöne (Handschriften usw.) vom üblichen panchromatischen AHU-Mikrofilm wiedergegeben werden, der den Kontrast von vornherein steigert, genügt in der Regel eine Digitalisierung mit 16 Graustufen (4 Bit). Wird von einem Halbton-Mikrofilm mit feiner Grauabstufung digitalisiert, sollen 256 Graustufen (8 Bit) dargestellt werden. Allgemein gilt, daß beim Digitalisieren mit Graustufen die Auflösung bei gleicher Wiedergabequalität reduziert werden kann.

1.3.3 Dateiformate der Images

Die Bandbreite der möglichen Dateiformate für Images ist beeindruckend. Leistungsfähige Viewer-Software mit Lesemöglichkeiten für mindestens 20

---

9 Electronic Text Center - University of Virginia (http://www.lib.virginia.edu/etext/ETC.html)

10 Dörr/Weber, S. 64

unterschiedliche Formate ist inzwischen Standard. Hinzu kommen die verschiedenen Versionen ein- und desselben Formats, die, ähnlich wie bei Softwareupgrades, von einigen Firmen für ihre Produkte in gewissen Abständen auf den Markt gebracht werden.

Eine klare Unterscheidung ist zwischen dem beim Einscannen mit hohem Qualitätsanspruch erstellten Image und den zum späteren Zeitpunkt über das Internet zur Verfügung gestellten Bildern zu treffen. Das Scan-Image übernimmt im Rahmen der Retrodigitalisierung die Funktion eines „digitalen Masters", der auf geeigneten Speichermedien zur langfristigen Verwendung abgelegt wird und im Zuge einer Pflegeroutine in regelmäßigen Abständen auf Lesbarkeit und Kompatibilität zu überprüfen ist. Unter dem Gesichtspunkt der Langfristarchivierung des digitalen Masters ist bei der Auswahl eines Dateiformats unbedingt darauf zu achten, daß auf Standards zurückgegriffen wird, die im Rahmen späterer Konvertierungsvorhaben ohne nennenswerte Probleme der neuen Systemumgebung angepaßt werden können.

Das Image, welches der Benutzer auf Anforderung am Bildschirm sieht, wird durch Konvertierungsläufe vom digitalen Master erstellt und kann niedrigeren Qualitätsanforderungen genügen als die Archivierungsversion.

Eine weitere Version kann für das Herunterladen ganzer Image-Dokumente erstellt werden. Diese Download-Version ist für den Benutzer, der den online-Text ständig verfügbar haben möchte, von großer Bedeutung. Vor dem Hintergrund bekannter Netzleitungsprobleme bezüglich des Datendurchsatzes ist es ihm auf diesem Wege möglich, den gewünschten Text auf dem eigenen Arbeitsplatzrechner lokal gespeichert zu halten.

1.3.3.1 Digitaler Master

Die Anforderungen, die an den digitalen Master gestellt werden, sind aus der Art der Digitalisierungsvorlagen abzuleiten. Das Hauptaugenmerk der AG Technik war hier auf Textmaterialien, in erster Linie also auf bitonale (s/w) Vorlagen gerichtet. Eine verbindliche Empfehlung für *ein* Dateiformat des digitalen Masters abzugeben, hält die AG Technik zum jetzigen Zeitpunkt nicht für angebracht, da sich auf diesem Gebiet ein möglicher Wechsel der bisherigen Standards andeutet.

Das TIFF-Format[11]

Für bitonale Vorlagen hat sich in der Praxis das von der Firma Aldus entwikkelte TIFF-Rasterformat zu einer Art quasi-Standard herauskristallisiert. Reizvoll für viele Anwender ist dabei wohl besonders die Möglichkeit, der einzel-

---

11 The Unofficial TIFF Home Page (http://rushmore.jpl.nasa.gov/~ndr/tiff/#shouldi)

nen Imagedatei Informationen beizugeben, die in das 'Image File Directory' der Datei geschrieben werden. Diese Informationen sind, wie auch der Name des Formats sagt, nach Kategorien gegliedert. In der zur Zeit aktuellen Version 6.0 (in der Spezifikation von Juni 1992),[12] gibt es über 90 Kategorien, in denen Informationen zum Image untergebracht werden können (zur Auflösung, Farbtiefe, Größe etc.). Einige Felder sehen dabei auch die Aufnahme von Informationen im ASCII-Format vor. (In *Anlage 1* befindet sich eine Übersicht über die Kategorien, die bei der Imageerstellung belegt werden sollten.) Die Library of Congress empfiehlt aus diesem Grunde TIFF als Format für die Archivierung bitonaler Images von Handschriften und gedruckten Vorlagen.

Da sich die Verwendung des unkomprimierten TIFFs aufgrund der zu bewältigenden Speichermengen für die Archivierung großer Textmengen nicht eignet (1 s/w A4-Seite unkomprimiertes TIFF bei 400 dpi Auflösung = ca. 2 Mb), wird die Verwendung der verlustfreien (Fax)-Komprimierung Gruppe 4 (Standard der ehemaligen *CCITT*, heute *ITU*) empfohlen. Die Größe einer Imagedatei bei dieser Komprimierung liegt dann zwischen 100 und 150 Kb.

Das PNG-Format

In der jüngsten Zeit ist ein neues Dateiformat für Rasterimages dabei, die Welt des World Wide Web zu erobern. *Portable Network Graphics (PNG*, sprich: PING) wurde von einer Gruppe von Graphik- und Programmierungsspezialisten unter der Leitung des WWW Consortium (W3C) - Mitglieds Chris Lilley entwickelt.[13] Hintergrund der Entwicklung ist der Erwerb des Patentrechts für das gängige LZW-Komprimierungsverfahren durch die Unisys Corp., die in der Folge Lizenzgebühren von den Anbietern forderte, die ihre Images im kommerziellen Bereich einsetzten. Die so lizenzierte Komprimierungsform wird beispielsweise bei dem Grafikaustauschformat GIF eingesetzt und ebenfalls bei der Komprimierung von TIFF-Dateien, wenn es sich um Farbimages handelt.

Die Beachtung von PNG empfiehlt sich insbesondere vor dem Hintergrund einer Quasi-Standardsetzung dieses Format für den Datentransfer im Internet durch die jüngsten offiziellen Empfehlungen der Internet Engineering Task

---

12 TIFF Revision 6.0
(http://icib.igd.fhg.de/icib/it/defacto/company/aldus/read.html#ExtraSamples)

13 Den Hinweis auf PNG und Informationen zu der Bedeutung dieses neuen Formats verdankt die AG Technik R. Bayer. Nähere Informationen über dieses Format erhält man unter den folgenden Adressen: PNG (Portable Network Graphics) Home Page (http://www.wco.com/~png/), Specification (http://www.boutell.com/boutell/png/). Eine nützliche Zusammenfassung gibt James Felici in der Zeitschrift *Publish* „International Report" January 1997 (http://www.publish.com/0197/international/).

Force (IETF) und des World Wide Web Consortiums (W3C). Neben dieser offiziellen Empfehlung und der Tatsache, daß PNG vollständig in den Bereich 'Public Domain' fällt, gibt es auch technische Gründe, die für eine Verwendung von PNG als Dateiformat für den digitalen Master sprechen. So bietet PNG bei Farbvorlagen eine Farbtiefe von bis zu 48 Bits und für Graustufen 16 Bits an (zum Vergleich: TIFF bietet 24 Bits bei Farbe und 8 Bits bei Graustufen). Man sollte in diesem Zusammenhang jedoch darauf hinweisen, daß die bisher angebotene Farbtiefe im Normalfall sicher ausreicht. Im Bereich der Komprimierung scheint die bei PNG eingesetzte DEFLATE-Komprimierung für bitonale Vorlagen effektiver zu sein als Fax Gruppe 4 bei TIFF. Die Komprimierung für Farbimages kann darüber hinaus in der Zukunft zu Lizenzproblemen führen, weil TIFF hier das bereits erwähnte LZW-Verfahren anwendet.

Für TIFF als digitalen Master, jedenfalls bei der Digitalisierung von bitonalen Vorlagen, spricht hingegen weiterhin die oben beschriebene Möglichkeit der umfangreichen Informationsmitgabe in die Imagedatei selbst, was in diesem Umfang und in der strukturierten Form bei PNG nicht möglich ist.

Aus Sicht der Arbeitsgruppe kommen beide genannten Formate für Digitalisierungsvorhaben in Frage, wobei TIFF bei abgeschlossenen und derzeit laufenden Digitalisierungsvorhaben mit Abstand am häufigsten eingesetzt wird.

1.3.3.2 Benutzungsversion für den Online-Zugriff

Für die Bereitstellung der Images über das Internet sollte vom digitalen Master mindestens eine Benutzungsversion erstellt werden. Bei der Auswahl eines geeigneten Dateiformats für diese Version sollten insbesondere die Frage der Unterstützung durch gängige Web-Browser und die Größe der Datei in bezug auf den Datentransfer und eine rasche Performanz des Bildschirmaufbaus berücksichtigt werden.

Da die Anzeige von TIFF-Dateien zur Zeit von gängigen Web-Browsern noch nicht unterstützt wird, sollte für die Bereitstellung über das Internet ein anderes Dateiformat gewählt werden.

Dafür kamen in der Vergangenheit vor allem zwei Komprimierungsformate in Betracht:

GIF

Das *Graphics Interchange Format (GIF)* der Firma Compuserve, das den LZW-Kompressionsalgorithmus benutzt und in zwei Spezifikationen, GIF 87a und GIF 89a, vorliegt. Im Mikrocomputerbereich kommt besonders seine hardwareübergreifende Verwendungsmöglichkeit als Austauschformat zum Tragen. Die Komprimierungsverfahren nach dem LZW-Algorithmus, die wie-

derkehrende Binärfolgen erkennen und ersetzen, zählen heute zum Standard der Textkomprimierung.

Da GIF lediglich eine Farbtiefe von 1 bis 8 Bit (s/w bis 256 Farben) erlaubt, ist seine Verwendung nur für bitonale und Halbtonvorlagen sinnvoll.

JPEG

Das *JPEG*-Format (nach der *Joint Photographic Experts Group*), ein Standard für die Komprimierung digitaler Standbilder. Das Verfahren der Datenreduktion erlaubt eine äußerst effektive Datenkomprimierung, die, je nach Verwendungszweck, individuell bestimmbar ist. Theoretisch geht die Skala für diese Komprimierung von 0 bis 99, realistisch ist wohl ein Komprimierungsfaktor bis etwa 40. Weitergehende Komprimierungen würden die Qualität des angebotenen Bildes zu stark beeinträchtigen. Aufgrund der Datenreduktion handelt es sich bei JPEG um eine Komprimierungsverfahren, das mit Informationsverlust arbeitet, anders als die GIF oder TIFF G4 eingesetzten Verfahren. Diese Tatsache sollte bei allen Konvertierungsaktivitäten mit diesem Format immer beachtet werden.

JPEG wird im amerikanischen Digitalisierungsprogramm bevorzugt für die Komprimierung von Farb- und Graustufenbildern eingesetzt.

PNG

Stärker noch als im Fall des digitalen Masters ist PNG als Alternative für die Benutzungsversion zu erwähnen. Über die Farbtiefe von GIF im Bereich bi- und halbtonaler Vorlagen hinaus deckt es zusätzlich den gesamten Bereich der Farbvorlagen ab, wobei es - anders als JPEG mit 24 Bit - bis zu 48 Bit Echtfarben unterstützt. Das Komprimierungsverfahren ist nicht nur - im Gegensatz zu LZW bei GIF - lizenzfrei sondern auch effektiver (um 10 - 30 %).

Aufgrund der offiziellen Empfehlungen des W3C und der IETF wird PNG in neueren Versionen von Web-Browsern standardmäßig unterstützt werden, Plug-Ins werden bereits heute angeboten (z.B. *PNG Live* für Netscape). Die nahe Zukunft wird zeigen, ob PNG sich auch in der Praxis als Standard im Grafikformatbereich für das Web durchsetzen wird.

Die AG Technik sieht eine verbindliche Empfehlung für eines der beschriebenen Dateiformate als Benutzungsversion zum jetzigen Zeitpunkt als nicht sinnvoll an. Die jeweilige Auswahl wird von Fall zu Fall durch die Art der Vorlagen (Text/Strich, Halbton, Farbe) mitbestimmt werden. Im Laufe ihrer weiteren Tätigkeit wird die AG Technik im übrigen neue Kompressionsverfahren, wie beispielsweise die *Cartesian Perceptual Compression (CPC)* oder den er-

weiterten Wavelet-Standard, ein Verfahren aus dem Bereich der Videokompression[14], beobachten und gegebenenfalls ihre bisherigen Hinweise ergänzen.

1.3.3.3 Downloadversion

Das einmalige Herunterladen digitalisierter Dokumente auf den eigenen Rechner wird, insbesondere vor dem Hintergrund von Netzleitungsproblemen bezüglich des Datendurchsatzes, eine der wesentlichen Nutzungsmöglichkeiten der digitalen Forschungsbibliothek werden. Dabei kann der einmal lokal abgespeicherte Text als Grundlage für die Bildschirm- und die Druckausgabe dienen.

Um diese Downloadfunktion für den Benutzer komfortabel zu gestalten, erscheint eine Übertragung im HTML-Format für längere, strukturierte Texte nicht ausreichend. Stattdessen bieten sich solche Formate an,die speziell für die Beschreibung des Layouts ganzer Dokumente entwickelt wurden. Hier sind in erster Linie *PostScript* und das *Portable Document Format (PDF)* zu nennen, beide aus dem Hause Adobe.

PostScript

PostScript wurde Mitte der 80er Jahre als Seitenbeschreibungssprache zur Ansteuerung von Druckern konzipiert mit dem Ziel, formatübergreifend ein einheitliches Layout zu gewährleisten. Mittlerweile wird PostScript auch als Format für die elektronische Distribution von Texten verwendet. Aufgrund der spezifischen und kostenintensiven Anforderungen an die Hardware im Druckausgabebereich, die nur unzureichend über den Einsatz von Viewersoftware (z.B. *Ghostscript view)* umgangen werden können, hat dieses Format sich im breiten Nutzerkreis nicht etablieren können.

PDF

Durchzusetzen scheint sich hingegen das für den Dokumentenaustausch konzipierte *Portable Document Format (PDF)*, das mit der *Acrobat*-Software der Firma Adobe erzeugt, verwaltet und angesehen werden kann.

Zur Klarstellung sei darauf verwiesen, daß hier zunächst nicht an die PDF-Formatierung von Volltexten gedacht ist. Hierfür wäre, wie auch für die Strukturierung mit SGML, eine vorhergehende Texterkennung erforderlich. Vielmehr werden bei der Erstellung einer Downloadversion die Bitmap-Images in das PDF eingebunden.

---

14 Peter Maaß, Martin Böhm, Hartmut Schachtzabel: „Effiziente mathematische Methoden in der Bildverarbeitung", *Informations- und Kommunikationstechnologien im Land Brandenburg* (02/1996), S. 129-133.

Das Layout für die Ausgabe der PDF-Dokumente ist plattform- und applikationsunabhängig festgelegt und für Bildschirm und Drucker gleichermaßen dargestellt. Im Gegensatz zu PostScript-Dokumenten kann der Ausdruck von PDF-Dateien unproblematisch auf jedem Laserdrucker erfolgen.

Die PDF-Dateien können leicht mit Hilfe der entsprechenden „Capture"-Software von Adobe erstellt werden. Für die Ansicht dieser PDF-Dokumente ist mit dem *Acrobat Reader* ein spezieller Viewer erforderlich, der frei erhältlich ist und auf dem jeweiligen Dokumentenserver einer Bibliothek zum Herunterladen angeboten werden könnte. In naher Zukunft dürfte zudem mit der standardmäßigen Plug-In-Einbindung dieses Viewers in gängige Netz-Browser zu rechnen sein.

1.4 Volltexterfassung

Der erste Schritt bei der digitalen Konversion von gedruckt vorliegenden Texten ist das Image-Scannen, dessen Ergebnis ein in Pixel (Bildpunkte) zerlegtes Bild bzw. Image der Vorlage ist, das mit dem Computer weiterverarbeitet werden kann.

Ein weitergehender Schritt ist die Volltexterfassung der nun als Images vorliegenden Dokumente. Die Suche nach einzelnen Wörtern oder die Übernahme von Textteilen zur eigenen Weiterbearbeitung ist erst nach diesem Arbeitsschritt möglich.

Die Volltexterfassung ist auf zwei Wegen realisierbar:

1. Automatisierte Erfassung durch Texterkennungsprogramme (OCR)
2. Manuelle Erfassung von Texten

1.4.1 Automatisierte Erfassung durch Texterkennungsprogramme (OCR)

Für die automatisierte Erkennung von Pixelgrafiken als Texte gibt es eine Vielzahl von Texterkennungsprogrammen, sog. OCR- bzw. ICR-Software (*OCR=Optical Character Recognition, ICR=Intelligent Character Recognition*).

Diese Programme verwenden unterschiedliche Ansätze zur Erkennung von Zeichen.

Neben dem 'Mustervergleich' (Pattern Matching), bei dem ein gescannter Text Pixel für Pixel der Grafik mit den im jeweiligen Programm gespeicherten Mustern vergleicht, kommt zunehmend die 'Merkmalanalyse' (Feature Recognition) zum Einsatz. Dabei werden typische Merkmale eines einzelnen Zeichens erfaßt.

Beim Scannen sauberer Vorlagen mit leicht lesbaren Schriften in guter Druckqualität lassen sich Trefferquoten von über 99% erzielen. Diese auf den ersten Blick imponierende Trefferzahl ist allerdings mit Vorsicht zu genießen,

bedeuten 99% Tefferquote doch immerhin noch 20 Zeichenfehler auf einer Manuskriptseite von 2000 Zeichen.

Die retrospektive Digitalisierung von Bibliotheksbeständen wird zunächst schwerpunktmäßig die ältere Literatur erfassen (vgl. die Erläuterungen in der *Einführun*"). Gerade bei älteren Druckvorlagen kommt man jedoch nicht einmal in die Nähe solcher Trefferwerte. Unterschiedliche Tests mit Texten aus dem 19. Jahrhundert haben ergeben, daß lediglich Trefferquoten von 60-70% zu erwarten sind, was wiederum bei 2000 Zeichen etwa 600-800 falsche Zeichen bedeuten würde, ein vollkommen unbrauchbares Ergebnis.

Zu klären ist in diesem Zusammenhang der Einsatzzweck der volltextdigitalisierten Bücher. Werden sie lediglich für die Volltextsuche im Hintergrund bereitgehalten - d.h., der Benutzer sieht nur das Image, kann aber in der ASCII-Version nach einzelnen Zeichenfolgen suchen - können niedrigere Trefferquoten eher toleriert werden, als wenn die ASCII-Version selbst am Bildschirm für die Benutzung freigegeben werden soll.

Uneinheitlicher Schriftsatz, Verschmutzungen, magelhafte Schriftqualität und in neuerer Zeit eher selten verwendete Schriftarten wie beispielsweise Fraktur stellen jedes OCR-Programm zunächst vor große Probleme. Natürlich besteht gerade für professionelle Programme die Möglichkeit des Trainierens von Schriften. Dies setzt aber hohen Personalaufwand voraus, ein Kostenfaktor, der von den Projektnehmern im Rahmen der retrospektiven Digitalisierung in aller Regel nicht selbst getragen werden kann. Die AG Technik empfiehlt deshalb im Grundsatz, automatisierte Texterkennungsverfahren nur dann einzusetzen, wenn keine nennenswerten Korrekturarbeiten zu erwarten sind.

Im übrigen bleibt abzuwarten, wie die Entwicklung von Tools zur Texterkennung voranschreitet. Eine Qualitätsverbesserung für oben genannte Problemfälle verspricht ein Verfahren, daß in jüngster Zeit unter Zuhilfenahme mathematischer Methoden und Gleichungen entwickelt wurde.[15] Potentielle Anwender seien in diesem Zusammenhang auf eine interessante Studie hingewiesen, die im Rahmen eines vom Land Brandenburg mit Lottomitteln geförderten Pilotprojektes „Anforderungen an einen Computerarbeitsplatz für die vergleichende Textanalyse von mittelalterlichen deutschen Rechtshandschriften und -büchern" eine detaillierte Untersuchung über den Einsatz von OCR-Software im geisteswissenschaftlichen Bereich angestellt hat.[16]

---

15 a.a.O.

16 Kai Schirmer, Friedrich Scheele, Werner Peters in: *Neue Anwendungen der Informations- und Kommunikationstechnologien*. Informations- und Kommunikationstechnologien im Land Brandenburg. (2a/1994), S. 115-133.. Vgl. auch Wolfgang Limper, *OCR und Archivierung*, München 1993.

1.4.2 Manuelle Erfassung von Texten

Die manuelle Eingabe von Texten wird erst dann in Frage kommen, wenn Druckvorlagen für eine automatisierte Textererkennung nicht geeignet sind. Erste Umfragen unter Dinstleistungsanbietern in diesem Bereich haben ergeben, daß die Erfassung von 1000 Zeichen zwischen 1,50 DM (bei einfacher Erfassung) und 6,- DM (bei doppelter Erfassung) liegen. Die Erfassung wird gewöhnlich in sog. Niedriglohnländern durchgeführt. Inwieweit gerade ältere Vorlagen, z.B. in Frakturschrift, in diesen Ländern zu den o.g. Konditionen tatsächlich erfolgreich erfaßt werden können, werden entsprechende Testläufe zeigen müssen.

Die manuelle Erfassung von Texten wird aufgrund der hohen Kosten in der Regel nicht für ganze Werke erfolgen können. Empfohlen wird aber die Erfassung einzelner Strukturelemente eines Buches wie des Inhaltsverzeichnisses und des Registers, um im Zuge der Erschließung über entsprechende 'Verlinkung' einen gezielten inhaltlichen Zugriff auf einzelne Imageseiten zu ermöglichen.

1.5 Strukturbeschreibung von Dokumenten

Die Strukturbeschreibung von Texten setzt im Ablauf der Digitalisierung auf der im Volltext erfaßten Vorlage auf. Die Problematik der Volltexterfassung älterer Textmaterialien wurde im vorhergehenden Abschnitt ausführlich erläutert. Vor diesem Hintergrund wird die Frage der Strukturbeschreibung in den vorliegenden „Technischen Hinweisen" nur überblicksartig angesprochen. Eine ausführliche Diskussion soll zu einem späteren Zeitpunkt erfolgen, wenn die Volltextdigitalisierung im Rahmen der digitalen Konversion auch qualitativ gesehen einen nennenswerten Platz einnimmt.

Unter Strukturbeschreibung von Dokumenten versteht man die formatunabhängige Kennzeichnung bzw. Markierung von distinktiven strukturellen Elementen eines Textes, wie Überschrift, Absatz etc. Beschrieben wird somit die logische Struktur eines Dokumentes, weniger sein Layout. Verschiedene Beschreibungssprachen werden mittlerweile eingesetzt, am bekanntesten dürfte wohl die *Hypertext Markup Language (HTML)* sein, die sich zum Standard für den Einsatz im World Wide Web entwickelt hat und neben der Strukturbeschreibung auch die Möglichkeit zu Querverweisen innerhalb und außerhalb eines Textes bietet. HTML baut wiederum auf der *Standard Generalized Markup Language (SGML)* auf, die auch als ISO-Norm (8879) zur logischen Beschreibung von Texten definiert wurde.

Im Unterschied zu HTML verfügt SGML über ein wesentlich differenzierteres Beschreibungsvokabular. SGML-Dokumente bestehen in der Regel aus drei Teilen:

1. Syntaxvereinbarung (SGML-Deklaration)

2. Dokumenttypdefinition (unter der Abkürzung *DTD* bekannt)
3. Dokumentausprägung (d.h., das Dokument selbst).

Angewandt wird SGML heute in mehreren Bereichen, z.B. im Verlagswesen zur Erstellung einer ausgabenunabhängigen Struktur von Dokumenten (Publikation in gedruckter Form, als CD-ROM, im Internet).

Im Bereich der Digitalisierung erfährt SGML besonders im amerikanischen Raum eine starke Verbreitung. Im Rahmen des *National Digital Library Program* und seines Vorgängers, *American Memory*, wurde eine Vielzahl von Dokumenten unter Zuhilfenahme von SGML strukturiert. Die Library of Congress hat zu diesem Zweck die *American Memory DTD* für digitalisierte historische Dokumente definiert und setzt sie in unterschiedlichen Projekten ein.

Seit einigen Jahren gibt es Bestrebungen, in Kooperation von Informatikern und Geisteswissenschaftlern Richtlinien für die elektronische Auszeichnung und den Austausch von Texten zu erarbeiten, die sog. *Text Encoding Initiative (TEI)*. Als Ergebnis liegen - seit 1994 als Buch, CD und Internet-Fassung - eine Reihe von SGML-konformen DTDs vor, die ein differenziertes Beschreibungsinstrumentarium für die Wiedergabe verschiedener Textsorten (Lyrik, Drama, Prosa u.a.) zur Verfügung stellen.

2
Speichern

2.1 Speicherung digitalisierter Ressourcen für die Benutzung

Das Speichersystem als Teil der digitalen Forschungsbibliothek bedeutet in erster Linie die Bereitstellung von Massenspeicher für die digitalisierten Ressourcen, die über das lokale und weltweite Datennnetz abgerufen werden können. In der Architektur der digitalen Bibliothek ist hier die zentrale Stelle, an der die im internen Produktionsprozeß fertigggestellten digitalisierten Dokumente für die Online-Benutzung vorgehalten werden. Verläßlichkeit und gute Performanz-Zeiten bezüglich des Datentransfers sind Voraussetzung für eine komfortable Benutzung und damit für eine breite Akzeptanz der digitalen Forschungsbibliothek.

Gespeichert wird in diesem System die für den Online-Zugriff erstellte Benutzungsversion (s. Ziffer 1.2.3.2) eines digitalen Dokumentes. Nimmt man hier die durchschnittliche Größe einer Imagedatei für eine Seite mit ca.100 KB an, ergibt sich bei einem Buch von 300 Seiten ein Speicherbedarf von 30 MB. Eine Sammlung von 1000 Büchern nimmt damit bereits einen Speicherplatz von 30 GB ein.

Die bekannten Speichersysteme stellen magnetische, optische und magnetooptische Speichermedien zur Verfügung. Mit Blick auf ihre Verwendung in der digitalen Bibliothek sind verschiedene technische Faktoren wie z.B. Speicherkapazität und Transferzeiten ebenso zu betrachten wie die entstehenden Kosten.

2.1.1 Festplattensysteme

Festplattensysteme lassen sich heute, unter der Voraussetzung einer entsprechend proportionierten Server- und Controllerkonstellation auf Speicherkapazitäten bis in den Terabyte-Bereich aufrüsten. Charakteristisch für diese Systeme sind die schnellen Zugriffs- und Transferzeiten. Diese Eigenschaften werden in der verteilten digitalen Forschungsbibliothek eine bedeutende Rolle spielen, da durch entsprechenden Erschließungsaufwand im Bereich von Inhaltsverzeichnis und Register gerade der schnelle punktuelle Zugriff auf einzelne Seiten ermöglicht werden soll.

Zum jetzigen Zeitpunkt noch ein beträchtlicher ökonomischer Faktor sind die Anschaffungskosten für magnetische Speichermedien, die sich auf etwa 400,- DM pro Gigabyte belaufen.Hier wird jedoch in der Zukunft eine deutliche Kostenreduzierung zu erwarten sein.

Magnetische Speichermedien für den schnellen Zugriff stehen auch in Form von RAID-Array-Systemen (*RAID=Redundant Array of Inexpensive Disks*) zur

Verfügung. Dies sind Festplattensysteme, die softwaregesteuert verschiedene Stufen der Datensicherung gewährleisten und dabei unter anderem mit Verfahren der Festplattenspiegelung und der logischen Aufteilung der Daten auf einzelne Platten arbeiten. Das 1987 an der University of Berkeley definierte „Fünf-Ebenen-Modell" wurde mittlerweile um die Ebenen 6 und 7 erweitert. Die RAID 7 Architektur ermöglicht den Zugriff mehrerer Hosts auf ein Array-System.

Der Sicherheitsanspruch in den einzelnen Projekten ist hier sicher individuell zu definieren.

2.1.2 Optische Plattenspeichersysteme

Optische und magneto-optische Speichermedien werden heute in erster Linie aus Kostengründen gerne als Massenspeicher eingesetzt. Insbesondere auf die CD-R (R=Recordable) und WORM (Write Once Read Multiple) soll im folgenden kurz eingegangen werden.

Die CD-R und die WORM sind beschreibbare Speichermedien, wobei in beiden Fällen der Vorgang des Beschreibens in mehreren Arbeitsgängen geschehen kann.

Die CD-R hat eine Speicherkapazität von 650-780 MB und ist heute als sog. Rohling für ca. 10 - 15 DM erhältlich. Zum Beschreiben erforderlich sind ein CD-R-Laufwerk (auch CD-Brenner genannt) und die dazugehörige Schreibsoftware.

Die Speicherkapazität der WORM hängt unter anderem von ihrer Größe ab. 12" erreichen zur Zeit ca. 12 GB, 14" bis ca. 18 GB. Die Kapazität für das 5,25"-Format liegt bei ca. 0,6 GB (demnächst über 1 GB), ihr Preis beträgt 35 - 40 DM. Das Beschreiben erfolgt in ähnlicher Weise wie bei der CD-R. Die Einführung neuer Techniken wie dem Mehrschichtenspeicher und dem Kurzwellenlaser wird zu einer Steigerung der Speicherkapazität um den Faktor 5 bis 15 führen.

Vor dem Hintergrund des Bestrebens nach dem Einsatz von Standards im Bereich der digitalen Bibliothek sollte unter den optischen Speichermedien wohl die CD-R empfohlen werden, da der ISO-Standard 9660 hier im Gegensatz zum proprietären WORM-Format eine weitgehende Lesbarkeit der CD-R auf allen gängigen CD-ROM Laufwerken garantiert.

Im System als Massenspeicher eingesetzt werden können die beschriebenen CD-R beispielsweise über CD-ROM-Jukeboxen, deren Kapazität durch entsprechende Konfiguration ebenfalls im Terabytebereich liegt.

Sprechen die im Verhältnis zu Festplattensystemen niedrigen Kosten eines optischen Speichersystems für den Einsatz desselben in einer digitalen Bibliothek, dürfen auf der anderen Seite die langsameren Zugriffs- und Transferraten des optischen Systems nicht vernachlässigt werden. Berücksichtigt

man zudem bei der optischen Speicherung außer den Anschaffungskosten auch den Bereich der Wartung, ist hier bei den Jukeboxen aufgrund der empfindlichen Mechanik mit nicht unerheblichen Kosten zu rechnen.

Vorstellbar ist deshalb ab einer bestimmten Größenordnung die Kombination von Festplatten- bzw. RAID-Array-Systemen für den raschen, häufigen Zugriff auf digitalisierte Dokumente mit CD-ROM-Jukeboxen, auf denen weniger frequentierte Daten vorgehalten werden. Der Einsatz einer hierarchischen Speichermanagement-Software kann die Verwaltung dieses Kombinationssystems unterstützen.

Die Konfiguration eines entsprechenden Massenspeichersystems für die Online-Bereitstellung von Dokumenten über das Internet und die effiziente Integration in die Gesamtarchitektur der digitalen Forschungsbibliothek wird - kostenmäßig bedingt - nicht durch jeden Projektnehmer erfolgen können. Hier werden die beiden Service- und Kompetenzzentren in Göttingen und München entsprechende Pilot- und auch Dienstleisterfunktionen übernehmen können.

2.2 Speicherung zum Zwecke der Langzeitsicherung

Die Erfahrungen aus laufenden und abgeschlossenen Digitalisierungsprojekten zeigen in puncto Langfristarchivierung der digitalisierten Dokumente eine klare Tendenz. Die Daten des digitalen Masters werden auf optische Speichermedien, zumeist auf CD-R geschrieben und, physikalisch getrennt von der Benutzungsversion, gelagert. Zu erwägen ist dabei die Erstellung eines Doppelsatzes von jeder Speichereinheit und aus Sicherheitsgründen die Lagerung an unterschiedlichen Orten.

Wie bei jedem größeren EDV-Einsatz mit wertvollen Daten ist die Festlegung einer Pflegeroutine für diese Daten unabdingbar. Unter Beobachtung der technischen Entwicklung im Bereich der Computer- und Speichersysteme sowie der Speichermedien muß sichergestellt werden, daß die digitalisierten Dokumente jeder Zeit lesbar zur Verfügung gestellt werden können. Die rasante Innovation im EDV-Bereich hat dabei zwei Seiten: zum einen werden technische Weiterentwicklungen der Hard- und Software in der Zukunft sicher Verbesserungen für die verteilte digitale Forschungsbibliothek bringen. So ist für die Speichermedien in den nächsten Jahren zu erwarten, daß immer größere Datenmengen auf immer kleineren - und vermutlich auch billigeren - Datenträgern untergebracht werden können (vgl. z.B. die Entwicklung des Schichtenspeichers im Bereich der optischen Speichermedien).

Zum anderen aber schafft die schnelle Weiterentwicklung von Hard- und Software Probleme für eine diachronische Kompatibilität der technischen Komponenten einer digitalen Bibliothek. Kurze Innovationszyklen machen ständige Investitionen im Hard- und Softwarebereich erforderlich, um den

jeweiligen technischen Anforderungen der Zeit zu entsprechen. Präzise Aussagen über die Folgekosten dieser ständigen Migration lassen sich heute jedoch noch nicht treffen.

Sorgfältig zu überlegen ist daher immer auch der Weg der retrospektiven Digitalisierung über den Mikrofilm. Wird über die Zwischenstufe eines alterungsbeständigen Mikrofilms guter Qualität digitalisiert, steht in diesem analogen Medium ein Langzeitspeicher zur Verfügung, von dem auch immer wieder digitalisiert werden kann. Die Problematik einer planmäßigen Migration der Bilddaten stellt sich in diesem Fall nicht. Der umgekehrte Weg, digitale Daten auf Mikrofilm zur Langzeitsicherung auszugeben, ist bisher noch nicht gangbar.[17] Die Ausgabe ist zwar technisch möglich (Computer Output on Microfilm -COM), die Wiedergabequalität ist jedoch unbefriedigend und läßt erneute Digitalisierung mit hinreichender Qualität nicht zu.

17 Dörr/Weber, a.a.O., S. 72f.; zur Sicherungs- und Migrationsproblematik s. S. 68, S. 75.

3
Erschließen und Verwalten

Der komfortable und effektive Zugriff auf die digitalisierten Bücher, Zeitschriften und andere Dokumente ist Voraussetzung für den Erfolg der Verteilten Digitalen Forschungsbibliothek. Unter allen Umständen zu vermeiden gilt es, daß „Textfriedhöfe" in digitalisierter Form entstehen, wie sie bereits heute im Bereich schlecht erschlossener Mikroformsammlungen anzutreffen sind.

Die Erschließung wird deshalb auf drei Ebenen erfolgen:

1) Die traditionelle formale und inhaltliche Erschließung, wie sie von Bibliotheken in konventioneller und heute überwiegend in elektronischer Form betrieben wird, zielt auf den systematischen und strukturierten Nachweis von Büchern, Zeitschriften und anderen Bibliotheksmaterialien in lokalen und überregionalen Katalogen. Dem Benutzer wird ein zumeist differenzierter Sucheinstieg über bibliographische Beschreibungsattribute wie Autor, Titel, Erscheinungsjahr etc. sowie über natürlichsprachige und klassifikatorische Inhaltsdeskriptoren angeboten. Im Rahmen der retrospektiven Digitalisierung von Bibliotheksbeständen werden diese Erschließungsmethoden weiterhin eingesetzt werden, insbesondere auch zum Zwecke des überregionalen Nachweises der digitalisierten Dokumente in den Verbundkatalogen. Nicht selten sind die betreffenden Bestände bereits elektronisch erfaßt.
2) Die Titelaufnahmen müssen zur Beschreibung der spezifischen Attribute der digitalisierten Ressourcen (Online-Ressource, CD-R, etc.) durch zusätzliche Informationen ergänzt werden. Neben Adressinformationen zum lokalen und überlokalen Zugriff auf die Online-Ressource sind dies vor allem technische Daten zur Beschreibung des digitalisierten Masters bzw. des Digitalisierungsverfahrens (Auflösung beim Scannen, Farbtiefe, Dateiformat, etc.). Diese technischen Angaben sind vor allem für Nachweissysteme von Bedeutung, die bereits vorliegende Digitalisierungen zur Vermeidung von Doppelarbeit erfassen. Die Arbeitsgruppe empfiehlt, wie bei Mikroformen auch, Nachweise von Digitalisierungen in die internationale Datenbank EROMM aufzunehmen.
3) Zielten die unter 1) und 2) beschriebenen Erschließungsverfahren auf das Dokument als Einheit, können durch zusätzlichen Erschließungsaufwand mit Mitteln komplexer Dokumentenverwaltungsprogramme Strukturen des einzelnen Dokumentes in digitaler Form für den effektiven, zielgerichteten Zugriff zur Verfügung gestellt werden. Als Minimalanforderung wird hier von der AG Technik die Bereitstellung von Inhaltsverzeichnissen und - soweit vorhanden - von Registern festgehalten.

Die so bei der Erschließung gewonnenen Nachweis- und Strukturinformationen zu dem einzelnen digitalisierten Dokument werden zusammenfassend unter dem Begriff 'Metadaten' subsummiert.

3.1 Bibliographische und technische Metadaten[18]

Die formale und inhaltliche Beschreibung digitaler Dokumente sollte schon aus Gründen der Konsistenzsicherung primär in dem Erschließungskontext erfolgen, in dem auch die primäre Erfassung von Metadaten bezüglich anderer elektronischer und konventioneller Dokumente angesiedelt ist. Primärer Erschließungskontext ist mithin in der Regel das für die jeweilige Bibliothek maßgebliche regionale Verbundsystem bzw. für Zeitschriften die Zeitschriften-datenbank (ZDB).

Zusätzlich müssen die Metadaten in geeigneter Weise lokal repliziert werden, um einen primären Sucheinstieg über das lokale Bibliothekssystem vor allem auch für den Fall von Instabilitäten im Weitverkehrsnetz zu gewährleisten.

Dabei wird in einem ersten Schritt die formale-inhaltliche Beschreibung des digitalisierten Dokumentes im jeweiligen Verbundsystem erfolgen. In vielen Fällen können dabei die bibliographischen Daten der für das betreffende Papierdokument vorliegenden Aufnahme wiederverwendet werden. Das Datenmodell der Verbünde muß zu diesem Zweck um spezifische Informationsbereiche wie Dateiformat, Adresse des digitalisierten Dokumentes für den Online-Zugriff etc.erweitert werden. Zu berücksichtigen sind dabei die Erfordernisse der bibliographischen Beschreibung in EROMM. Wie für Mikroformen sollte auch für die Digitalisierung ein international abgestimmmtes Datensegment im Hinblick auf EROMM definiert werden.da vorgesehen ist, die digitalen Master auf europäischer Ebene in dieser Datenbank nachzuweisen.

Nach Gesprächen zwischen Vertretern der SUB Göttingen und des Gemeinsamen Bibliotheksverbundes (GBV)[19] zeichnet sich folgendes Datenmodell als geeignet ab:

- Zusätzlich zum Datensatz der Digitalisierungsvorlage wird ein eigener Datensatz für den digitalen Master angelegt. In diesem Datensatz erfolgt die Beschreibung des digitalen Masters bezüglich der physikalischen Form, des Datums der Digitalisierung, Ort und Host des digitalen Masters sowie einiger weiterer Angaben auf der bibliographischen Ebene.

---

18 Die Beschreibung der Aufgaben des Verbundsystems erfolgte mit freundlicher Unterstützung von Dr. Gradmann, Direktor der Verbundzentrale des GBV.

19 Von seiten der SUB nahmen an diesen Gesprächen Herr Becker, Frau Cremer, Dr. Lossau und Dr. Sperber teil, Vertreter des GBV war Dr. Gradmann.

- Auch die technische Beschreibung des digitalen Masters wird auf dieser Ebene vorgenommen. Geklärt wird zur Zeit noch, inwieweit Codes für diese technischen Angaben eingesetzt werden können.
- Die Belegung der erforderlichen Kategorien wird sicher von Verbund zu Verbund unterschiedlich sein, auf eine Vereinheitlichung im Sinne des abstrakten Beschreibungsmodells sollte jedoch allein schon mit Blick auf die Austauschbarkeit der Daten im nationalen und internationalen Umfeld großer Wert gelegt werden.
- Die Beachtung und Einbeziehung der Metadatendiskussion aus der jüngsten Vergangenheit - Dublin Core, Warwick Framework[20] - wird in der Zukunft eine wesentliche Rolle spielen.

Als Format der in die Datenbank des lokalen Verwaltungssystems zu ladenden bibliographischen Metadaten empfiehlt die AG Technik das maschinelle Austauschformat für Bibliotheken MAB. Dazu ist es erforderlich, die Annahmen, die hinter dem oben beschriebenen Datenmodell stehen, in MAB umsetzen zu können. Dies bedeutet insbesondere die Erweiterung der lokalen MAB-Ebene. Ein entsprechender Vorschlag wird von den Service- und Kompetenzzentren an den MAB-Ausschuß mit der Bitte um vorrangige Behandlung herangetragen werden.

Mittelfristig wünschenswert ist die Konvertierung von MAB in SGML, das als plattform- und systemunabhängiges Format insbesondere in den Vereinigten Staaten auch für Metadaten eingesetzt wird, um so mit Blick auf die Langfristarchivierung eine formatunabhängige Version für die Zukunft bereitstellen zu können. Das *Electronic Text Center* an der University of Virginia verfährt bei der Strukturierung ihrer bibliographischen Metadaten nach den Richtlinien der bereits erwähnten Text Encoding Initiative (TEI).

Sinnvoll scheint in diesem Zusammenhang ein an dem besagten *Text Center* praktiziertes Verfahren zu sein, nach dem bei der Digitalisierung die bibliographischen Metadaten auch in die einzelnen Imagedateien des digitalen Masters selbst geschrieben werden. Images können so jederzeit problemlos einer bibliographischen Einheit zugeordnet werden. Bei Verwendung des TIFF-Formats für den digitalen Master kann für diesen Zweck das 'Comment-Field' verwendet werden, das Schreiben der Metadaten kann softwaregesteuert im Batchbetrieb erfolgen.

Die Aufgabe der Verwaltung der bibliographischen Metadaten soll in den entstehenden digitalen Sammlungen von einem Dokumentenverwaltungssystem, basierend auf einer relationalen Datenbank, übernommen werden

---

20 Vgl. hierzu die Resource Page zum Dublin Core Metadata Element Set: (http://www.oclc.org:5046/ research/dublin_core/) und die Zusammenstellung in *Organizing the Global Digital Library II: Metadata Meetings* (Library of Congress) (http://lcweb.loc.gov/catdir/ogdl2/metadata.html).

(s. Ziffer 3.3). Durch den jeweiligen Anwender muß dabei für den Aufbau der Datenbankstruktur die Festlegung von Kategorien bzw. Beschreibungsinhalten vorgenommen werden, die aus dem Bibliothekskatalog in das DMS übernommen werden sollen. Gleichzeitig ist eine Entscheidung darüber zu treffen, welche dieser Datenfelder für die Suche indexiert werden sollen.

3.2 Strukturelle Metadaten

Während die bibliographischen Metadaten die Grundlage für den Zugriff auf das Dokument als Ganzes bieten, dienen die strukturellen Metadaten (Inhaltsverzeichnis, Register) als Grundlage für eine komfortable Benutzung des Dokumentes selbst. Die Erschließung in diesem Bereich ist dezidiert benutzerorientiert und setzt auf die durch den Autor für die Druckvorlage bereits geleistete inhaltliche Erschließungstätigkeit auf. Dabei gibt das Inhaltsverzeichnis einen ersten Überblick über Inhalt und Gliederung des Buches, das (oder die) Register greifen sinn- und bedeutungtragende Begriffe auf. Sie erlauben den inhaltlich orientierten punktuellen Zugriff auf einzelne Seiten.

Erst durch die Bereitstellung derartiger elektronischer „Navigationshilfen" wird auch der eigentliche Sinn des digitalisierten Buches erreicht. Der Benutzer kann jederzeit auf ein gewünschtes Werk von seinem elektronischen Arbeitsplatz aus zugreifen, orientiert sich inhaltlich über das Inhaltsverzeichnis und gegebenenfalls Register und greift dann gezielt auf einzelne Seiten zu. Das sequentielle Lesen längerer Texte am Bildschirm wird hingegen in der Regel nicht erwünscht sein.

3.2.1 Erstellen von elektronischen Inhaltsverzeichnissen und Registern

Eine der Grundforderungen der DFG an die retrospektive Digitalisierung ist das Erstellen von Daten, die systemübergreifend genutzt werden können. Bei der Volltexterfassung von Inhaltsverzeichnissen und Registern sind deshalb in einem ersten Schritt sog. Rohdaten (in der Regel im ASCII-Format) bereitzustellen, die dann, in einem weiteren Schritt, in spezifische Verwaltungsformate überführt werden können. In jedem Fall werden die Rohdaten auch in ihrem ursprünglichen Format für etwaige spätere Konvertierungsverfahren dauerhaft archiviert.

Zweck der Bereitstellung von Inhaltsverzeichnissen und Registern ist es, den gezielten Zugriff auf Teile des digitalisierten Dokumentes zu ermöglichen. Dafür ist eine Verknüpfung der im Volltext erfaßten Daten mit den entsprechenden Seitenzahlen des Dokumentes erforderlich, um den Begriffen aus Inhaltsverzeichnis und Register eine 'Seitenadresse' zuzuordnen. Wie diese Verknüpfung realisiert wird, bleibt dem jeweiligen DMS überlassen. Hilfreich für eine automatisierte Verknüpfung ist das Vorliegen einer Seitenbeschrei-

bung für die einzelnen Imagedateien in maschinenlesbarer und strukturierter Form.

3.2.1.1 Kumulierte Register - dokumentübergreifend

Ein Mehrwert für den Forscher kann sich dann ergeben, wenn die einzelnen Register aller digitalisierten Bücher - 1. innerhalb einer Sammlung und 2. sammlungsübergreifend - kumuliert werden und von dem derart kumulierten Index der Zugriff auf einen bestimmten Begriff (bzw. die Imageseite, auf der sich ein einzelner Begriff befindet) ermöglicht wird, der in mehreren Büchern vorkommt (z.B. *Hamburg* in einem geographischen, einem historischen und einem literarischen Werk). Dem interessierten Forscher wird so der erweiterte Blick auf solche Sach- und Literaturgattungen ermöglicht, die über den eigentlichen Rahmen seines Fachgebietes hinausgehen.

Technische Vorgaben und Festlegungen für die Implementierung kumulierter Register sind noch zu erarbeiten.

3.3 Verwaltung der digitalisierten Dokumente und ihrer Metadaten

Für die Verwaltung der digitalisierten Dokumente und der dazugehörigen Metadaten kommen drei grundlegende Systemarchitekturen in Betracht:

1. Die bibliographischen Daten werden zentral in einem Katalog (z.B. lokaler OPAC oder Bibliotheksverbundkatalog) gehalten, die entsprechenden Dokumentdateien (inkl. elektronischem Inhaltsverzeichnis und Register) werden in einem hierarchisch gegliederten Dateiensystem auf einem Dokumentenserver für den Online-Zugriff bereitgestellt. Die Struktur der digitalisierten Sammlung, bzw. die interne Struktur der digitalisierten Dokumente kann dabei durch die Hierarchie des Dateisystems abgebildet werden. Das von der Universitätsbibliothek Bielefeld im Rahmen eines DFG-Projektes entwickelte System BIEBLIS ist an diesem Ansatz orientiert.
2. Ein Dokumenten-Management-System (DMS) kommt zum Einsatz. Hier sind zwei Optionen denkbar:

   2.1. In der relationalen Datenbank des DMS werden alle Arten von Metadaten gespeichert, die Dokumentdateien selbst werden jedoch außerhalb des DMS auf einem Dokumentenserver abgelegt. Der Zugriff auf die Dokumente erfolgt über die Metadaten im DMS.

   2.2. Auch hier werden die Metadaten im DMS gehalten. Zusätzlich werden jedoch auch die für den Online-Zugriff bereitgestellten Dokumentdateien in das DMS importiert.

Für die letzte Option spricht aus der Sicht der AG Technik unter anderem der schnellere Zugriff auf die digitalisierten Dokumente bzw. die einzelnen Seiten

sowie ihre unproblematischere Adressierung bei der Bereitstellung im Netz. Auf die Frage der Adressierung wird in Kapitel 4 noch ausführlich eingegangen.

Hauptaufgabe des DMS in einer digitalen Bibliothek ist die Verwaltung der zuvor definierten Metadaten in einer relationalen Datenbank und ihre Zusammenführung mit den entsprechend zu strukturierenden Imagedateien eines Dokumentes. Damit ermöglicht das DMS den komfortablen und gezielten Zugriff auf das Dokument als Einheit und auf Teile des Dokumentes.

Im Verlauf der Vorbereitung des neuen Förderprogramms zur „Retrospektiven Digitalisierung" wurden auch einige vorhandene Systemlösungen im Bereich Dokumenten-Management-System untersucht. Speziell für den Einsatz in einer 'digitalen Bibliothek' werden Systeme von IBM (*Digital Library*)[21] und Rank Xerox (*XDOD/DocuWeb*)[22] angeboten. Bei diesen Produkten ist bereits eine Zusammenstellung verschiedener Softwarekomponenten (Scan- und Bildbearbeitungssoftware, Datenbank, Web-Interface u.a.) erfolgt, wobei im Falle der *Digital Library* von IBM ein höheres Maß an Grundkonfiguration der einzelnen Komponenten erforderlich ist. Andere Produkte wie das Volltextdatenbanksystem *MYRIAD* der Firma TransAction (München)[23] oder das Dokumenten-Management-System *SAROS* der Firma FileNet[24] müssen erst zu einem System der 'digitalen Bibliothek' weiterentwickelt werden. Dabei kann in unterschiedlichem Umfang auf leistungsfähige Systeme, basierend auf relationalen SQL-Datenbanken, aufgebaut werden. Der Weg zur Weiterentwicklung solcher Standardprodukte geht in der Regel über sog. Systemintegratoren.

Spezifische Funktionalitäten wie beispielsweise der automatisierte Import von bibliographischen Metadaten aus dem Bibliotheksverbundkatalog in das DMS oder die Schaffung gezielter Zugriffsmöglichkeiten auf das imagedigitalisierte Buch über Register werden von keinem der genannten Systeme in der Standardversion angeboten.

Der pilothafte Einsatz eines solchen Systems und die gezielte Weiterentwicklung werden zu den Aufgaben der beiden Service- und Kompetenzzentren in Göttingen und München gehören.

---

21 IBM *Digital Library* (http://www.software.ibm.com/is/dig-lib/ibmdl1a.htm)

22 Xerox Products for Digital Libraries (http://www.xerox.fr/ats/ad/digilib/xrxprod.html), Forschungszentrum in Europa, Grenoble RXRC: Site Map (http://www.xerox.fr/sys/sitemap.html)

23 http://www.tasmuc.de

24 http://www.filenet.com/

4
Suchen und Zugreifen

Die Suche nach den digitalisierten Dokumenten und der Zugriff auf dieselben soll grundsätzlich über das Internet erfolgen. In diesem Zusammenhang ist die Frage der Benennung und Adressierung der einzelnen Dokumente von grundlegender Bedeutung. Der folgende Abschnitt gibt dazu in einem einführenden Überblick die ersten Erkenntnisse wieder, die im Rahmen des von der DFG geförderten Projektes zur Digitalisierung von Dissertationen am Fachbereich Informatik der Universität Frankfurt/Main gewonnen wurden.

4.1 Die Adressierung elektronischer Dokumente für den Online-Zugriff

4.1.1 Benennung elektronischer Ressourcen

Um digitale Ressourcen nutzen zu können, müssen sie identifiziert werden. Dies geschieht anhand zugeordneter Namen, die gemäß eines definierten Benennungsschemas gebildet werden. Namen müssen in ihrem Geltungsbereich eindeutig sein, um eine Ressource identifizieren zu können. Darüber hinaus sollten Namen persistent sein. Ein persistenter Name wird nur einmal während der Existenz des Benennungsschemas vergeben und ist selbst dann noch mit einer Ressource verknüpft, wenn diese nicht mehr existiert oder nicht mehr zugreifbar ist.

Um die Benennung einer unbegrenzten Zahl von Ressourcen zu erlauben, sollte ein Benennungsschema skalierbar sein. Ein skalierbares Schema ermöglicht die Bildung unendlich vieler eindeutiger Namen. Eine Möglichkeit, ein skalierbares Namensschema zu schaffen, besteht in einer hierarchischen Gliederung der Namen (z.B. analog zu Internet Domain-Namen oder durch variabel lange Namen).

Man unterscheidet unter anderem zwischen ortsgebundenen und ortstransparenten Namen. Ortsgebundene Namen identifizieren einen Ort und damit indirekt die Ressource, die sich an diesem Ort befindet (z.B. IP-Addressen). Ortstransparente Namen bezeichnen eine Ressource selbst (z.B. Internet Domain-Namen). Um auf die so bezeichnete Ressource zugreifen zu können, muß der ortstransparente Name in eine Ortsangabe übersetzt werden.

Ortsgebundene Namen sind im allgemeinen einfacher zu handhaben, da sie keinen zusätzlichen Resolutionsschritt zum Zugriff auf die Ressourcen fordern. Darüber hinaus ist eine genaue Ortsangabe in der Regel eindeutig. Der Nachteil von ortsgebundenen Namen liegt in ihrer mangelnden Persistenz. Ändert eine Ressource ihren Aufenthaltsort, oder wird sie durch eine andere ersetzt, ist der ortsgebundene Name nicht länger gültig.

Ortstransparente Namen besitzen den Vorteil der Persistenz. Das heißt, ein ortstransparenter Name bezeichnet immer ein und dieselbe Ressource, unabhängig von ihrem gegenwärtigen Aufenthaltsort. Allerdings erfordert die Verwendung ortstransparenter Namen zusätzlichen Aufwand. Ortstransparente Namen müssen zum Zugriff auf die Ressource durch einen „Name Server" aufgelöst werden, d.h., auf eine Ortsangabe abgebildet werden. Ortstansparente Namen sind nicht a priori eindeutig. Es bedarf einer Vereinbarung oder Standardisierung, um ihre Eindeutigkeit, und damit ihre Persistenz zu gewährleisten.

4.1.2 Bennungsschemata im Internet

Die *Internet Engeneering Task Force* (IETF) entwickelt Standards zur Benennung von Ressourcen, die über das Internet zugreifbar sind. Die von der IETF entwickelten Benennungsschemata werden unter dem Begriff Uniform Resource Identifier (URI) zusammengefaßt. Zur Zeit sind zwei Benennungsschemata identifiziert: Uniform Resource Locator (URL)[25] und Uniform Resource Name (URN)[26].

4.1.2.1 Uniform Resource Locator

Ein URL dient der Angabe des Ortes einer Ressource. URLs sind folgendermaßen aufgebaut:

<Schemakennzeichen>:<schemaspezifischer Teil>

Das bekannteste URL-Schema ist sicherlich das http-Schema zur Angabe des Ortes einer Ressource, auf die über das Hypertext Transport Protokoll (HTTP)[27] zugegriffen werden kann. Bei URLs ist der schemaspezifische Teil in eine Host-Kennung, einen lokalen Pfad und einen Suchausdruck unterteilt, der folgende Form hat:

http://<Host-Kennung>/<lokaler Pfad>?<Suchausdruck>

Die Host-Kennung wird als Internet Domain-Name oder als IP-Addresse des Rechners, auf dem sich die Ressource befindet, angegeben. Der optionale

---

25 T. Berners-Lee, L. Masinter, M. McCahill, *Uniform Resource Locators (URL)*, Network Working Group, RFC 1738, <URL:ftp://ftp.nic.de/pub/doc/rfc/rfc-1700-1799/rfc1738.txt>

26 K. Sollins. L. Masinter, *Functional Requirements for Uniform Ressource Names*, Network Working Group, RFC 1737, <URL:ftp://ftp.nic.de/pub/doc/rfc/rfc-1700-1799/rfc1737.txt>

27 T. Berners-Lee, R. Fielding, H. Frystyk, *Hypertext Transfer Protocol - HTTP/1.0*, Network Working Group, RFC 1945, <URL:ftp://ftp.nic.de/pub/doc/rfc/rfc-1900-1999/rfc1945.txt>

lokale Pfad identifiziert eine Ressource innerhalb des Rechners. Der optionale Suchausdruck kann von der Ressource ausgewertet werden, um spezifische Teile der Ressource auszuwählen. Ein Beispiel für einen URL ist:

http://www.diglib.de/dms?docid=123-4567-89.abcdef

Mit dem Beispiel-URL wird das Dokument mit der lokalen Kennung 123-4567-89.abcdef aus dem DMS das unter http://www.diglib.de/dms zu erreichen ist, ausgewählt. Die Auswertung des Suchausdrucks setzt natürlich voraus, daß es sich bei der mit der Host-Kennung und dem lokalen Pfad identifizierten Ressource um eine aktive Ressource, z.B. ein CGI-Skript, handelt.

4.1.2.2 Uniform Resource Names

Der größte Nachteil eines URL ist seine mangelnde Persistenz, bereits nach wenigen Monaten sind viele URLs nicht mehr gültig. Die Gründe dafür sind vielfältig und können sowohl technischer als auch administrativer Art sein. Mögliche Ursachen sind z.B. die Umstrukturierung eines Rechenzentrums, das Outsourcing von Diensten oder die Migration zu neuen Standards wie IPv6.

Die IETF versucht seit einiger Zeit, ein ortstransparentes Benennungsschema für Internet-Ressourcen zu entwickeln, die Uniform Resource Names (URNs). Einige der an URNs gestellten Anforderungen sind:

- Globale Gültigkeit: Ein URN hat überall dieselbe Bedeutung, er kennzeichnet eine Internet-Ressource.
- Globale Eindeutigkeit: Jeder URN ist genau einer Internet-Ressource zugeordnet.
- Persistenz: Die Zuordnung eines URN zu einer Internet-Ressource ist zeitlich unbeschränkt gültig.
- Skalierbarkeit: Der Namensraum der URNs muß beliebig erweiterbar sein, um eine Benennung aller Ressourcen zu ermöglichen.

Die funktionalen Anforderungen an URNs sind detailliert in Sollins/Masinter beschrieben[28]. Die Standardisierungsbemühungen dauern an, zur Zeit ist kein Standard für ein URN-Schema verfügbar. Aller Voraussicht nach wird sich der Standardisierungsprozeß noch über ein bis zwei Jahre hinziehen.

---

28 Sollins/Masinter, a.a.O.

4.1.3 Benennung von Dokumenten innerhalb der Verteilten Digitalen Forschungsbibliothek

Die in der Verteilten Digitalen Forschungsbibliothek (VDF) abgelegten Dokumente sollen nicht nur in Dokument-Management-Systemen (DMS) gespeichert werden, sondern auch in existierenden Katalogen (lokale Kataloge sowie Verbundkataloge) nachgewiesen werden. Dadurch soll unter anderem der direkte Zugriff auf ein digitales Dokument aus dem OPAC und aus dem Verbundkatalog heraus ermöglicht werden. Analog zum Standortnachweis gedruckter Werke muß eine Signatur für digitale Dokumente erzeugt werden, die den „Standort" des digitalen Dokuments wiedergibt.

Das für diese Signatur gewählte Benennungsschema muß drei Forderungen erfüllen:

- Es muß die Eindeutigkeit der Namen unter allen DMS der VDF bzw. weltweit garantieren.
- Die Namen müssen persistent sein.
- Es sollte mittelfristig die Benennung von Teilen eines Dokuments (Seiten, Kapitel, etc.) erlauben, um eine sammlungsübergreifende Referenzierung auf der Basis eines kumulierten Index zu ermöglichen.

Die Merkmale der URNs lassen sie als geeignet für eine Benennung von Dokumenten erscheinen. Da noch kein Standard für URNs vorliegt, ist es nicht möglich, ein URN-Schema zur Verwendung anzugeben, denn dies wäre notwendig proprietär und inkompatibel zu zukünftigen Standards. Also muß ein anderes Verfahren zur eindeutigen, persistenten Benennung gewählt werden, das kompatibel zu existierenden Programmen und Protokollen ist (z. B. WWW-Browser und HTTP).

Um die Eindeutigkeit der Namen zu gewährleisten, bietet sich die Verwendung des URL des speichernden DMS sowie einer DMS-internen Dokumentenkennung an. Mit diesem Vorschlag macht man sich die Eigenschaft der Eindeutigkeit der URLs zunutze. Diese ist durch die Internet Naming Authority garantiert, die die Vergabe von Rechnernamen und Adressen überwacht. Mit Hilfe der DMS-internen Dokumentenkennung, die innerhalb des DMS natürlich eindeutig sein muß, kann dann ein Dokument VDF-weit eindeutig benannt werden.

Die Erstellung eines kumulierten Index erfordert die Möglichkeit, Inhalte (z. B. einzelne Seiten) eines Dokuments zu referenzieren. Da die internen Strukturen der verwendeten DMS nicht bekannt sind, können Dokumentteile nicht direkt, z. B. über einen Dateinamen, identifiziert werden. Daher muß eine Möglichkeit zu einem einheitlichen Zugriff auf Teile von Dokumenten mit Hilfe einer einheitlichen Schnittstelle geschaffen werden. Hierzu werden Dokumentnamen in der VDF durch eine Kombination von einer ortsgebundenen Benennung der einzelnen DMS mit einer standardisierten Benennung von

Dokumenten gebildet. Die Identifikation des DMS erfolgt über die Host-Kennung und den lokalen Pfad der URL des Dokuments. Die Identifikation des Dokuments (oder später der Dokumentteile) erfolgt über den Suchausdruck der URL. Dieser enthält neben der DMS-internen Kennung des Dokuments Informationen über den gewünschten Teil des Dokuments. Dokumentnamen innerhalb der VDF werden nach folgendem Muster gebildet:

http://<Host-Kennung des DMS>/<lokaler Pfad>?<Info 1>& ... &<Info n>

Dabei haben die Informations-Felder des Suchausdrucks eine der folgenden beiden Formen:

1. <Schlüssel>
2. <Schlüssel>=<Wert>

Welche der beiden Formen gewählt wird, ist vom angegeben Schlüssel abhängig. In *Anlage 2* findet sich ein Entwurf für mögliche Schlüssel und anzugebende Werte. Zum Schlüssel 'docid', der die Angabe der lokalen Dokumentkennung erlaubt, muß z.B. ein Wert angegeben werden. Der Schlüssel „index" beispielsweise, der den Index eines Dokuments referenziert, wird ohne Wert angegeben.

Während die Namen der Schlüssel vorgegeben sind, können die Werte weitgehend frei belegt werden. Sie müssen lediglich der für URLs spezifizierten Syntax für Suchausdrücke genügen und dürfen nicht die Zeichen „&" und „=„ enthalten (*Anlage 3* führt die für Werte erlaubten Zeichen im einzelnen auf).

Jeder gültige Dokumentname muß den Schlüssel „docid" mit einem zugewiesenen Wert enthalten. Der Wert entspricht der internen Dokumentkennung des durch den Namen referenzierten Dokuments im DMS. Zusätzlich kann genau einer der folgenden Schlüssel angegeben werden: „page", „section", „figure", „table", „index", „references", „title".

Die Dokumentkennung kann frei gewählt werden, solange sie den Einschränkungen für Werte genügt. Dokumentkennungen sollten darüber hinaus skalierbar sein, z.B. durch eine hierarchische Gliederung, um zukünftige Erweiterungen eines DMS nicht zu behindern.

Einige Beispiele für gültige Dokumentnamen in der VDF wären:

http://www.diglib.de/dms?docid=123-4567-89.abcdef
http://www.diglib.de/dms?docid=1.b&section=3.2
http://www.diglib.de/dms?docid=1.b&index
http://www.diglib.de/dms?docid=1.b&title

Der HTTP-Server des DMS muß in der Lage sein, Schlüssel und deren Werte aus dem Suchausdruck zu extrahieren und eine HTML-Seite zu generieren, auf der die gewünschten Informationen zugänglich sind. Dies kann z.B. durch den Einsatz eines CGI-Skripts in Verbindung mit einem Standard HTTP-Server oder durch die Installation eines speziellen Servers geschehen.

4.1.4 Persistenzerhaltung durch Persistent Uniform Resource Locator

Das vorgeschlagene Benennungsschema kombiniert die ortsgebundene Benennung der DMS mit der ortstransparenten Benennung der in ihnen gespeicherten Dokumente. Dadurch werden Dokumentnamen unempfindlich gegen Umstrukturierungen innerhalb eines DMS, solange sich der Zugangspunkt zum DMS nicht verändert. Eine Änderung der Addresse eines DMS würde aber nach wie vor dazu führen, daß die Namen der in dem DMS gespeicherten Dokumente ungültig werden. Um trotz einer eventuell unumgänglichen Standortänderung des DMS die Gültigkeit der Namen zu garantieren, kann das Konzept des *Persistent Uniform Resource Locator* (PURL) 4 verwendet werden.

Ein PURL ist ein URL, der unbegrenzt gültig ist. PURLs referenzieren im allgemeinen die entsprechenden Internet-Ressourcen nicht direkt, sondern einen PURL-Server, der PURLs in gültige URLs umwandelt (also eine Abbildung innerhalb des Namensraums der URLs durchführt). Änderungen des Standorts eines DMS können durch Aktualisierung der Einträge im PURL-Server transparent erfolgen. Man erhält so persistente Namen für Dokumente. Da PURLs mit Hilfe eines in HTTP definierten Redirection-Mechanismus realisiert sind, sind sie konform zu existierenden Internet-Standards. Dies hat den Vorteil, daß jedes standardkonforme Programm einen PURL genau wie einen URL verarbeiten kann.

Der Einsatz von PURLs erfordert natürlich zusätzlichen Aufwand auf Seiten des Betreibers eines DMS. Spätestens wenn ein DMS seinen Standort ändert, muß ein PURL-Server an der alten Adresse eingerichtet werden. Daher sollte von vorneherein die Möglichkeit zum Einsatz von PURLs berücksichtigt werden.

4.1.5 Migration zu Uniform Resource Names

Mittelfristig ist eine Migration zu URNs einzuplanen, da diese eine Internetkonforme, ortstransparente Benennung ermöglichen. Das heißt, daß Standardwerkzeuge eingesetzt werden könnnen, um Ressourcen, die durch URNs identifiziert werden zu bearbeiten. Jede Realisierung einer digitalen Bibliothek sollte zumindest eine Erweiterung um URNs vorsehen. Dies schließt die Möglichkeit ein, URNs in die Standortfelder der Kataloge einzutragen.

4.2 Zugang zur digitalen Sammlung

Die AG Technik empfiehlt als Zugang zur digitalen Sammlung einer Bibliothek drei Möglichkeiten:

1. Direkter Einstieg über die Homepage der anbietenden Bibliothek
2. Einstieg über eine Suchanfrage an den lokalen und regionalen Bibliothekskatalog
3. Einstieg über eine Homepage der „Verteilten Digitalen Forschungsbibliothek".

4.2.1 Direkter Einstieg über die Homepage der anbietenden Bibliothek

Die erste Möglichkeit des Zugriffs läuft über die Homepage einer Bibliothek. Der Benutzer möchte in diesem Fall direkt auf (eine) bestimmte digitale Sammlung zugreifen.

Der Begriff der digitalen Sammlung ist dabei als Gliederungskriterium von zentraler Bedeutung. Wird eine Bibliothek im Rahmen der retrospektiven Digitalisierung in inhaltlich und thematisch unterschiedlichen Bereichen tätig, empfiehlt sich die Gliederung der digitalen Bibliothek in mehrere Sammlungen.

Für die technische Umsetzung bedeutet dies die Markierung der Zugehörigkeit eines Dokumentes zu einer bestimmten Sammlung. Im Bibliothekskatalog wird deshalb in eine Kategorie der Name der Sammlung eingetragen, die Kategorie sollte sinnvollerweise für die Suche indexiert werden. Diese Kategorie wird dann zusammen mit anderen Beschreibungsfeldern in die Datenbank des DMS überführt und dort ebenfalls als Suchindex aufbereitet. Damit ist auch die Grundlage für den sammlungsspezifischen Zugriff geschaffen.

Der Zugriff über die Homepage geschieht somit durch das Anklicken einer Option „digitale Bibliothek" oder „einzelne Sammlung". In beiden Fällen wird der Benutzer auf das Suchformular des User-Interfaces für das DMS herabgeführt und kann dort, je nach Umfang und Komfort des dort implementierten Recherchemoduls parametrisch nach Kategorien wie Autor, Titel etc. suchen, mit Boole'schen Operatoren verknüpfen oder natürlichsprachig sowie über Listen suchen. Die sammlungsübergreifende Recherche greift dabei auf den Gesamtindex des Beschreibungsfeldes „Name der Sammlung" zu, die sammlungsspezifische Recherche nur auf einen Teilbereich.

4.2.2 Einstieg über eine Suchanfrage an den lokalen und regionalen Bibliothekskatalog

Viele Bibliotheken bieten bereits heute einen Internetzugriff auf ihren lokalen Online-Katalog bzw. Verbundkatalog über gängige Web-Browser. Ziel in einer Verteilten Digitalen Forschungsbibliothek muß es mittelfristig sein, alle digitalen Sammlungen auf diesem Weg ansprechen zu können. Der Benutzer

kann dann auf die ihm vertraute Oberfläche des Bibliothekskatalogs zugreifen und über die dort angebotenen Suchindizes ein - mehr oder weniger breites - Rechercheangebot nutzen.

In den Katalogeinträgen zu einzelnen Treffern erhält der Benutzer die Information, daß zu einem gesuchten Buch eine 'Online-Version' angeboten wird. Zugleich ist eine Option vorhanden, diese digitale Fassung des Dokumentes anzusehen. Diese Funktionalität ist bereits heute in einigen Verbünden vorhanden.

Auf die technische Umsetzung dieser Anforderung aus dem Bibliothekskatalog an das DMS wurde auch unter 4.1 bereits eingegangen. Vorstellbar ist der Einsatz von CGI-Skripten, die die Internet-Adresse des DMS und als Anforderungsargument die interne Verwaltungsnummer des digitalen Dokumentes innerhalb des DMS enthält.

Anders als die Verwaltung der Dokumente über eine reine Verzeichnisstruktur auf einem Server hat die Verwaltung mit Hilfe einer Datenbank für den Zugriff auf das Dokument einen entscheidenden Vorteil: es gibt hier nur eine http-Sammeladresse für die Datenbank, die einzelnen Dokumente werden über ihre Verwaltungs-Identnummer angesprochen. Sollte sich diese im Zuge einer Umstrukturierung einmal ändern, bedeutet es keine Mühe, über eine Konkordanztabelle im DMS auf die veränderte Identnummer zuzugreifen.

Die Verknüpfung der digitalisierten Sammlungen mit lokalen OPACs fällt in den Verantwortungsbereich der jeweiligen Bibliotheken. Technische Lösungen für die Verknüpfung der lokalen digitalen Bibliotheken mit regionalen und überregionalen Verbundsystemen gehören zum definierten Aufgabenbereich der einzurichtenden Kompetenzzentren. Der überregionale Zugriff sollte Bestandteil der Fortführung des Projekts „Verbund deutscher Bibliotheks- und Fachinformationssysteme“ (DBV/OSI) werden.

4.2.3 Zugriff auf verschiedene lokale Systeme der Verteilten Digitalen Forschungsbibliothek

Der Name der *Verteilten* Digitalen Forschungsbibliothek impliziert das Vorhandensein einer Reihe lokal verteilter digitaler Sammlungen an verschiedenen Bibliotheken und Institutionen. Für die Akzeptanz dieses Angebots von entscheidender Bedeutung wird die Vernetzung dieser Sammlungen sein, um dem interessierten Benutzer unter einer Oberfläche Zugriff auf alle diese Sammlungen zu ermöglichen, wobei über die Funktionalität von überregionalen Katalogen hinausgehend auch der navigatorische Zugriff über (hierarchisch) strukturierte Listen, sowie ein Retrieval über sämtliche Metadaten, insbesondere also Begriffe aus Inhaltsverzeichnissen und Registern ermöglicht werden sollte.

Die Frage der Realisierung einer Vernetzung lokaler, verteilter Datenbanken kann zur Zeit noch nicht abschließend beantwortet werden. Vorstellbar ist der Einsatz von Suchmaschinen ebenso wie der Aufbau einer 'Super-Datenbank', in die die erforderlichen Metadaten aller Sammlungen eingespielt werden. Im Bibliotheksbereich bekannt geworden ist in der jüngsten Zeit das Beispiel des *Karlsruher Virtuellen Katalogs* (KVK), eines Meta-Suchinterfaces für Bibliothekskataloge im World Wide Web, das Suchanfragen an mehrere Kataloge parallel weitergibt und dem Benutzer so ein eigenes Ansteuern verschiedener Ressourcen erspart. Die Funktionsweise des KVK ist jedoch nur bei einer kleinen, beschränkten Anzahl anzusprechender Kataloge gesichert.

Um die Option des Einsatzes einer Suchmaschine in der VDF vorzubereiten, hält die AG Technik es für erforderlich, einige Rahmenbedingungen für den Einsatz von Datenbanken im jeweils gewählten Dokumentenverwaltungssystem verbindlich festzulegen. Dabei geht die AG Technik davon aus, daß nicht jeder Antragsteller die Verwaltung und Pflege der Metadaten selbst übernehmen muß. In diesem Fall ist er jedoch verpflichtet, die Daten zur Sicherung des überregionalen Nachweises an andere, geeignete Institutionen abzugeben, die sich ihrerseits zur Erfüllung der Anforderungen bereiterklären. Eine solche Aufgabe können beispielsweise die Service- und Kompetenzzentren in Göttingen und München übernehmen.

Verwaltet der Projektnehmer selbst die Metadaten zu seiner digitalen Sammlung, hat er dafür Sorge zu tragen, daß der strukturierte Zugriff auf sein Verwaltungssystem gesichert ist.

Bei der Auswahl und Konfiguration des DMS sind daher die folgenden Punkte unbedingt zu berücksichtigen:

1. *Einheitliche Strukturierung und Indexierung der Datenbank des eingesetzten DMS für die Suchansprache.*

Ein noch näher zu bestimmendes Profil von Beschreibungsfeldern soll in jedem lokalen System analog geführt werden. Koordinierende Funktion zur Festlegung dieses Grundprofils werden die geplanten Service- und Kompetenzzentren übernehmen. Über dieses 'Core-Set' an Beschreibungsfeldern hinaus liegt die Tiefe der inhaltlichen und sachlichen Erschließung in der Verantwortung der lokalen Systemanbieter.

2. *Offenlegung von Schnittstellen des lokalen DMS*

Geprüft wird in diesem Zusammenhang die Anbindung der lokalen Systeme über eine Z39.50-Schnittstelle mit einheitlicher Implementationsumgebung (Preferred Record-Syntax u.a.).

Die Service- und Kompetenzzentren werden die Entwicklungen im Bereich der Suchmaschinen für den Einsatz in der Verteilten Digitalen Forschungsbibliothek sorgfältig beobachten. Geprüft wird zur Zeit beispielsweise die Weiterentwicklung und der Einsatz des von der Firma Rank Xerox in ihrem

europäischen Forschungszentrum in Grenoble entwickelten Prototypen des *Constraint based Knowledge Broker.*

Um kurzfristig eine sammlungsübergreifende Suche in den Metadaten der im Verlaufe der nächsten Zeit digitalisierten Dokumente zu ermöglichen, sind von den Projektnehmern die bibliographischen und strukturellen Metadaten (Inhaltsverzeichnis, Register) in jeweils getrennten HTML-Dateien bereitzustellen, Dadurch können bereits im WWW eingesetzte Suchmaschinen für das sammlungsübergreifende Retrieval genutzt werden.

Letztendlich wird die konstruktive Zusammenarbeit der lokalen Anbieter untereinander und mit den Servicezentren die Voraussetzung für das erfolgreiche Durchsetzen eines überregionalen digitalen Dokumentenangebotes sein.

5
Bereitstellen und Nutzen

Ein entscheidender Punkt für das Erreichen einer breiten Akzeptanz von Büchern in digitalisierter Form wird die Art und Weise sein, in der sie dem Benutzer angeboten werden. Bei der Durchführung eines Digitalisierungsprojektes sollte deshalb diesem sensiblen Bereich eine besondere Aufmerksamkeit geschenkt werden. Technisch gesehen handelt es sich hierbei um das Design des Web-User-Interfaces, das die Verbindung zwischen DMS und World Wide Web darstellt. Die dort enthaltenen Funktionalitäten müssen eine komfortable Nutzung des Dokumentes erlauben.

Es ist dabei nicht zu erwarten, daß industriell-kommerzielle Softwareprodukte im Bereich der 'digitalen Bibliothek' sämtliche Anforderungen von Benutzern von Beginn an befriedigen können. Vielmehr werden Datenanbieter (z.B. Bibliotheken) und Firmen in kooperativer Zusammenarbeit an einer Optimierung des Designs arbeiten müssen.

In Kapitel 4 wurde der Sucheinstieg in eine digitale Bibliothekssammlung beschrieben. Im folgenden werden Möglichkeiten der Bereitstellung bzw. Nutzung einzelner digitaler Dokumente erläutert.

Der Benutzer wird nach einer Recherche und der Anzeige entsprechender Treffermengen auf einen bestimmten Titel geführt. Konkret bedeutet dies die Entscheidung für den ersten Einstieg in die digitale Repräsentation der gedruckten Vorlage. Vorstellbar wäre hier z.B. ein Thumbnail-Image des Titelblattes mit einer html-Seite der Metadaten zum digitalen Dokument oder, falls vorhanden, das elektronische Inhaltsverzeichnis des Buches.

Unbedingt erforderlich sind auf jeder Seite, auf der man sich innerhalb des Dokumentes bewegt, eine Reihe von Navigationshilfen, z.B. grafische Repräsentationen im Rahmen einer Kopfzeile. Der Umfang dieser Navigationshilfen für das digitale Buch kann sicher individuell definiert werden, der folgende Überblick nennt einige hierfür in Frage kommende Optionen:

a) Metadaten:

Info: hier kann der Benutzer die Informationen aus den im DMS gespeicherten Beschreibungsfeldern zu 'seinem' digitalen Dokument einsehen

b) Navigation im digitalen Dokument:

Register: der Benutzer erhält den Zugriff auf das elektronische Register des Dokumentes, in dem er sich befindet

Seite: zum Ansteuern einer beliebigen Seitenzahl

Anfang: Springen an den Anfang eines Dokumentes

Ende: Springen an das Ende eines Dokumentes

Vor: Eine Seite vorgehen

Zurück: Eine Seite zurückgehen

Table-of-content: der Benutzer wird wieder auf das elektronische Inhaltsverzeichnis geführt

Hilfe: über das Hilfemenü sollte eine detailierte Beschreibung mit Fallbeispielen zur Navigation und für die Suche in der *Digital Library* zugänglich sein.

c) Ausgabe des digitalen Dokumentes

Download: Zusätzlich zur Funktion >Save as< im Web-Browser wird hier im User-Interface eine Option zum Download des digitalisierten Dokumentes angeboten. Die Frage der Zugriffsrechte ist dabei vom System von Fall zu Fall zu klären.

Print:

1. der zentrale Ausdruck von Dokumenten (z.B. in der Bibliothek)

Der Benutzer erhält hier die Möglichkeit, Textpassagen aus einem Buch, die ihn interessieren, zusammenzustellen und als Druckauftrag an die Bibliothek weiterzugeben. Die Bibliothek ist zu diesem Zweck gehalten, an geeigneter Stelle Kapazitäten für einen qualitativen Ausdruck vorzuhalten. Alternativ bietet sich hierfür die Inanspruchnahme eines externen Dienstleisters an.

2. der dezentrale Ausdruck am Arbeitsplatz des Benutzers

Wünschenswert ist eine für den Benutzer einfach gehaltene Möglichkeit zum Ausdruck auf dem eigenen Drucker. Dies kann kurzfristig durch die Bereitstellung einer PDF-Version zum Download in Verbindung mit dem *Acrobat Reader* realisiert werden.

Ausgabe auf Offline-Medien (CD-R):

Heutige Probleme bei der Datenfernübertragung lassen es sinnvoll erscheinen, größere Datenmengen, z.B. ein Buch oder längere Abschnitte aus demselben, dem Benutzer offline zur Verfügung zu stellen. Aufgrund der niedrigen Herstellungs-kosten und des weit verbreiteten Einsatzes der CD-R auch am Arbeitsplatz des Wissenschaftlers und Studenten spricht vieles für die Wahl dieses optischen Datenträgers.

Der Benutzer muß einen entsprechenden Hinweis auf dieses Angebot erhalten, verbunden mit einem geeigneten Bestellformular.

Der Zugriff auf die Daten der CD-R sollte im Interesse des Nutzers langfristig gesehen mit vergleichbaren Navigationsmöglichkeiten gegeben sein, wie sie für das einzelne Dokument im User-Interface zum Web bereits beschrieben

wurden. Download- und Printfunktion für den Arbeitsplatz des Benutzers sind ebenfalls zu empfehlen.

Eine Minimallösung könnte das Schreiben einer PDF-Version zusammen mit dem *Acrobat Reader* auf die CD-R sein.

Datenspiegelung

Eine weitere Möglichkeit zur Umgehung von Engpässen bei der Datenfernübertragung ist die Spiegelung häufig frequentierter Dokumente bzw. Sammlungen auf dem lokalen Server einer Bibliothek. Hierzu ist die Bereitschaft zur Kooperation der einzelnen Anbieter erforderlich. Bei der technischen Umsetzung können die Service- und Kompetenzzentren Hilfestellung geben.

Zusammenfassung

Die AG Technik hat in ihrem hier vorgelegten Bericht die verschiedenen technischen Komponenten der Architektur einer 'Digitalen Bibliothek' vorgestellt. Aspekte wie das digitale Erfassen, Erschließen und Verwalten, Speichern, Suchen und Zugreifen sowie das Bereitstellen und Nutzen wurden dabei unter verschiedenen Gesichtspunkten beschrieben.

Neben der Deskription der technischen Komponenten hat die AG Technik darüber hinaus Empfehlungen für einzelne Komponenten im Hinblick auf das neue Förderungsprogramm der Deutschen Forschungsgemeinschaft abgegeben. Diese Empfehlungen tragen teilweise hinweisenden, informatorischen Charakter, teilweise sind sie aber auch als Verpflichtung für die Antragsteller formuliert. Die Deutsche Forschungsgemeinschaft wird diese Empfehlungen, soweit sie formelle Bewilligungsbedingungen für DFG-geförderte Projekte werden sollen, in dem Merkblatt „Technische Hinweise zur Durchführung von Projekten zur retrospektiven Digitalisierung" nochmals gesondert veröffentlichen.

Die Abgabe verpflichtender Empfehlungen ist auf einem Gebiet wie der EDV mit ihren kurzen Innovationszyklen immer ein Risiko, das es zu bedenken gilt. Diese Empfehlungen beziehen sich daher in der Regel nicht punktuell auf einzelne technische Details (Dateiformate, Speichermedien, DMS etc.). Vielmehr definieren sie dort, wo Rahmenbedingungen gesetzt werden, die für den Erfolg des Aufbaus einer Verteilten Digitalen Forschungsbibliothek grundlegende Voraussetzungen schaffen. Wie diese Rahmenbedingungen dann im einzelnen technisch umgesetzt werden, liegt letztendlich im Ermessen der Projektnehmer. Voraussetzung für die Projektförderung ist aber, daß sie umgesetzt werden.

Grundsätzlich läßt sich sicher festhalten, daß bei der retrospektiven Digitalisierung dort auf Standards zurückgegriffen wird, wo die Technik sie bereits heute zur Verfügung stellt. Dies betrifft z.B. die Auswahl des Dateiformats für den digitalen Master (TIFF bzw. PNG) oder die Verwendung von digitalen Speichermedien für die Langfristsicherung (CD-R). Prinzipiell sollte jeder Antragsteller sich mit Blick auf die zu digitalisierenden Vorlagen auch überlegen, bei der Digitalisierung den Weg über den Mikrofilm zu gehen, gerade unter Berücksichtigung der Kriterien des Bestandsschutzes und der Langfristarchivierung.

Zusammenfassend seien die Punkte noch einmal hervorgehoben, die von der AG Technik als bindende Verpflichtung für jeden Bewilligungsempfänger im Förderprogramm „Retrospektive Digitalisierung" vorgeschlagen wurden:

- Überregionaler Nachweis der digitalisierten Dokumente in den Bibliotheksverbundkatalogen

- Bereitstellung der digitalisierten Dokumente für den Online-Zugriff im Internet
- Wahl eines Dateiformats für die Benutzungsversion, das mit gängigen Netz-Browsern gelesen werden kann
- Einheitliche Strukturierung und Indexierung der Datenbank des eingesetzten DMS für ein Grundprofil an Beschreibungsfeldern
- Offenlegung von Schnittstellen des lokalen DMS
- Kooperation mit den Service- und Kompetenzzentren bei der Erarbeitung und Einhaltung von technischen Standards, insbesondere auch bei der Einbindung der eigenen digitalen Sammlung in eine Verteilte Digitale Forschungsbibliothek
- Beachtung der Erfordernisse der Bestandserhaltung durch Nutzung bestandsschonender Verfahren bei der Digitalisierung (oder Verfilmung)
- Sicherung der langfristigen Verfügbarkeit der digitalen Ressourcen.

Die folgenen Aufgaben sollten durch die beiden Kompetenzzentren für retrospektive Digitalisierung in gesonderten Arbeitsgruppen behandelt werden:

- Vorgaben für die Digitalisierung und Erschließung von Bildvorlagen
- Organisatorische und technische Konzeptionen für hochqualitative Dokumentausdrucke in regionale verteilten Druckzentren, sowie Download und Druckausgabe am Benutzerarbeitsplatz
- Technische Vorgaben zur Implementierung kumulativer Register und die Realisierung einer dedizierten Suchmaschine für die „Verteilte Digitale Forschungsbibliothek"
- Formatvorgaben für Rohdaten struktureller Metadaten zu Digitalisierten Dokumenten; Spezifikation und Pflichtenheft für eine Standard-Softwarelösung.

Literaturempfehlungen (Auswahl)

- **Elektronische Zeitschriften und Bibliographien mit Berichten u.a. zum Themenkomplex Digitalisierung**

D-Lib Magazine (http://www.ukoln.ac.uk/dlib/)

The Public-Access Computer Systems Review (http://info.lib.uh.edu/pacsrev.html)

RLG DigiNews (http://www.rlg.org/preserv/diginews/index.html#contents)

Scholarly Electronic Publishing Bibliography (http://info.lib.uh.edu/sepb/sepb.html)

- **Retrospektive Digitalisierung - übergreifende Aspekte**

Conway, Paul, *Preservation in the Digital World*, Commission on Preservation and Access - Commission Publications (3/96, 24pp.)

Fleischhauer, Carl [Technical Coordinator, National Digital Library Program, Library of Congress], *Digital Historical Collections: Types, Elements, And Construction*, 21. August 1996 (http://lcweb2.loc.gov/ammem/elements.html)

Kenney, Anne R., Chapman, Stephen, *Digital imaging for libraries and archives,* Ithaca, NY: Dept. of Preservation, Cornell University Library, 1996

Preserving Digital Information: Final Report and Recommendations in the final report of the Task Force on Archiving Digital Information, co-sponsored by RLG and the Commission on Preservation and Access, Mai 1996 (http://www.rlg.org/ ArchTF/)

Steps in the Digitization Process (Library of Congress, Januar 1996), (http://lcweb2.loc.gov/ammem/award/docs/stepsdig.html)

Vereinigte Staaten - National Digital Library Program: Technical Papers [Library of Congress and Ameritech]: „Further Technical Background" (http://lcweb2.loc.gov/ammem/award/further.html) und „Selected topics from NDLP internal documentation"
(http://lcweb2.loc.gov/ammem/award/docs/select.html)

- **Einzelne Digitalisierungsprojekte**

Cornell

Kenney, Anne R., Personius, Lynne K., *Joint Study in Digital Preservation: Report: Phase I,* January 1990-December 1991 (http://palimpsest.stanford.edu/cpa/reports/joint/index.html)

Cornell Digital Library: MOA [Making of America] Project (http://moa.cit.cornell.edu/MOA/moa-main_page.html)

TULIP

Final Report, 1996 (http://www.elsevier.nl:80/homepage/about/resproj/trmenu.htm)

Yale (Project Open Book)

(http://www.library.yale.edu/preservation/pobweb.htm)

Waters, Donald, Weaver, Shari, *The Organizational Phase of Project Open Book,* September 1992 (http://palimpsest.stanford.edu/cpa/reports/openbook.html)

Waters, Donald, Weaver, Shari, *The Setup Phase of Project Open Book,* June 1994 (http://palimpsest.stanford.edu/cpa/reports/conway.html)

Conway, Paul, *Conversion of Microfilm to Digital Imagery: A Demonstration Project,* (Performance Report on the Production Conversion Phase of Project Open Book),Yale University Library, August 1996

Imaging

Kenney, Anne R. and Chapman, Stephen, *Tutorial Digital Resolution Requirements for Replacing Text-Based Material: Methods for Benchmarking Image Quality,* Commission on Preservation and Access Commission Publications (4/95, 22 pp.)

Fleischhauer, Carl, *Digital Formats for Content Reproductions,* 20. August 1996 (http://lcweb2.loc.gov/ammem/formats.html#V)

Quality Review of Document Images: Internal training guide, 1996 (http://lcweb2.loc.gov/ammem/award/docs/docimqr.html). [This set of instructions is for staff involved in checking images of text received from contractors. The section „Imaging Guidelines for the National Digital Library Program" describes quality problems that have been encountered in practice at the Library of Congress]

Die Adressierung elektronischer Dokumente für den Online-Zugriff

Identifiers for Digital Resources, 1996
(http://lcweb2.loc.gov/ammem/award/docs/ identifiers.html)

The Relationship between URNs, Handles, and PURLs, 1996
(http://lcweb2.loc.gov/ammem/award/docs/PURL-handle.html)

Anlagen

Anlage 1

Schreiben von Informationen in den Header der TIFF-Datei (aktualisierte Version September 1997)[1]

Dieses Dokument beschreibt den Inhalt des TIFF-Headers von gescannten Image-Dateien.

Als Beispielfall wird die folgende Monographie herangezogen:

Chappell, Edward: Voyage of His Majesty's ship Rosamond to Newfoundland and the southern coast of Labrador, London 1818, PPN 135661005.

Die **TIFF-Tags** in der linken Spalte der nachfolgenden Tabelle, die zusätzlich zu den standardmäßigen Einträgen ausgefüllt werden sollen, sind durch Fett-Druck markiert.

Auswahl von Kategorien für die VDF aus der Spezifikation für TIFF 6.0, Juni 1992	Kat.-Nr.	Inhalt
New Subfile Type	254	Angaben standardmäßig durch die TIFF-Spezifikation vorgesehen (=Standard)
ImageWidth: Actual Pixel count	256	Standard
Image Length: Actual Pixel Count	257	Standard
BitsPerSample	258	Standard
Compression	259	Standard
PhotometricInterpretation	262	Standard
Document Name (ASCII-Feld)	**269**	Übernahme der Katalognummer des digitalisierten Dokumentes in die Datenbank des Dokumenten-Management-Systems Beispiel.: [PicaProduktionsNummer mit führendem Katalogschlüssel] *"PPN135661005"*
ImageDescription (ASCII-Feld)	**270**	Enthält bibliographische Informationen in diversen Unterfeldern, außerdem im letzten Subfeld die Verzeichnisstruktur zur Ablage der Image-Datei **ohne** jeweiligen Dateinamen Beispiel: oben erwähnte Monographie zur Neuseeland-Reise von E.Chappell:

1 Nähere Informationen zu diesem Thema sind erhältlich beim Göttinger Digitalisierungszentrum *GDZ*, Dr. Norbert Lossau, Tel.: 0551/ 39-5217, Fax: 0551 /39-5222, E-Mail: lossau@mail.sub.uni-goettingen.de.

Auswahl von Kategorien für die VDF aus der Spezifikation für TIFF 6.0, Juni 1992	Kat.-Nr.	Inhalt
		|<DOC_TYPE>MONOGR|<HAUPTTITEL>Voyage of His Majesty's ship Rosamond to Newfoundland and the southern coast of Labrador | <AUTOREN/HERAUSGEBER>Chappell,Edward |<JAHR>1818|<ERSCHEINUNGSORT>London |<VERZ_STRCT>ChapVoya_13566100
		Beispiel: Zeitschrift, Haupttext von Heft 5 in Band 30
		"|<DOC_TYPE>ZSCHR|<HAUPTTITEL>Zentralblatt für Mathematik.| <JAHR>1945|<ERSCHEINUNGSORT>Berlin |<VERZ_STRCT>ZentMath_12345678_B0030_H005|"
		|
StripOffsets	273	Standard
Samples Per Pixel	277	Standard
RowsPerStrip	278	Standard
StripByteCounts	279	Standard
XResolution: dots per inch	282	Standard
Yresolution: dots per inch	283	Standard
PlanarConfiguration (how the components of each pixel are stored)	284	Standard
PageName (ASCII-Feld)	**285**	Das Feld enthält die folgenden Angabe: **"physikalische Seitenzahl"** Die Angabe wird von der Scansoftware berechnet und bei ordnungsgemäßem Scanverlauf einfach hochgezählt. Die physikalische Seitenzahl wird beidseitig von dem Delimiter-Zeichen "|" (*stroke*) begrenzt: Beispiel: Seite 172 → Eintrag "|00000172|" Sie entspricht der Position einer Seite in der Abfolge aller Seiten. Die Abfolge beginnt mit der ersten bedruckten Seite und endet mit der

Auswahl von Kategorien für die VDF aus der Spezifikation für TIFF 6.0, Juni 1992	Kat.-Nr.	Inhalt
		letzten bedruckten Seite. Eingeschobene leere, unpaginierte etc. Seiten werden mitgezählt. Änderungen im Format der Seitenzahl, doppelte Seitensequenzen (z.B. römisch im Vorwort und arabisch im Hauptteil) werden bei dieser Zählung nicht berücksichtigt (Berücksichtigung erfolgt in der Excel-Begleitdatei). Die *physikalische Seitenzahl* wird direkt bei der Bildung des Dateinamens umgesetzt. Beispiel.: *00000172.tif*
ResolutionUnit	296	Standard
PageNumber	297	Wird nur bei Multiple Page TIFF benötigt
(Scan)Software	**305**	Bsp.: *SRZ Proscan incl. VersionsNR*
Date Time: date and time scanned	306	Bsp.: *19971020*
Artist (ASCII-Feld)	**315**	Bsp.: *Niedersächsische Staats- und Universitätsbibliothek Göttingen, Germany*

Anlage 2

Suchausdruck der URL: Entwurf für mögliche Schlüssel und Werte (Mönch)

Folgende Tabelle zeigt die definierten Schlüssel, gibt an, ob sie einen Wert besitzen, und beschreibt kurz ihre Bedeutung:

Schlüssel	Wert	Funktion
docid	ja	DMS-interne Dokumentkennung Beispiel: a.b.c-1
page	ja	Seitenangabe Beispiele: 1, III, I.3
title	nein	Titelseite des Dokuments (Default)
section	ja	Abschnitt oder Kapitel des Dokuments Beispiele: 3, 2.5.1, II.25
table	ja	Tabelle innerhalb des Dokuments Beispiele: tab 1, 4.3
figure	ja	Abbildung innerhalb des Dokuments Beispiel: abb 2
contents	nein	Inhaltsverzeichnis des Dokuments
references	nein	Literaturverzeichnis des Dokuments
index	nein	Index des Dokuments

Die Angabe der Schlüssel section, contents, table, figure, index und references ist nur dann sinnvoll, wenn das Dokument entsprechend strukturiert wurde.

Anlage 3

Suchausdruck der URL:
Erlaubte Zeichen für Schlüssel und Werte (Mönch)

Schlüssel und Werte eines Namens dürfen die folgenden Zeichen enthalten (sie entsprechen den in [2] für "search"-Folgen definierten Zeichen ohne "&" und "="):

a b c d e f g h i j k l m n o p q r s t u v w x y z

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

0 1 2 3 4 5 6 7 8 9 $ - _ . + ! * ' ( ) , ; : @

Alle anderen Zeichen müssen durch die Escape-Sequenz %XY dargestellt werden, wobei X und Y hexadezimale Ziffern sind, also jeweils eines der folgenden Zeichen:

0 1 2 3 4 5 6 7 8 9 A B C D E F a b c d e f

Anlage 4

Kosten für die Erfassung eines Standardbuches (Ecker)

In dieser Ausarbeitung wird eine Aufstellung der Kosten für das Verfilmen, Scannen, Speichern, Indexieren und teilweise inhaltliche Erschließen (Texterfassen) eines typischen Buches gegeben. Außerdem wird gezeigt, welche Aufgaben sich für eine Vergabe an einen professionellen Dienstleister eignen und welche Aufgaben von der Bibliothek selbst durchgeführt werden sollten.

Die aufgeführten Zeitangaben und Kosten sind aus der Praxis abgeleitete Werte für durchschnittliches Material. In Sonderfällen können erhebliche Abweichungen auftreten.

1 Spezifikationen

1.1 Durchzuführende Arbeiten

Die durchzuführenden Arbeiten werden wie folgt festgelegt:

Scannen:	400 dpi, Schwarz/Weiß-Modus ohne Graustufen
Speichern:	TIFF G4-Format auf CD-R (incl. Index)
Filmen:	35 mm Silberhalogenid-Archivrollfilm unter Berücksichtigung der einschlägigen DIN Normen
Indexieren:	Bibliographische Kerndaten
Inhaltserschließen:	Texterfassung von Inhaltsverzeichnis, Kapitelüberschriften, Register, ggf. auch Volltext

Die Bücher müssen pfleglich behandelt werden; sie dürfen nicht geschnitten oder beschädigt werden. Es werden daher spezielle Buchscanner und Kameras mit Buchwippe eingesetzt.

Die Scan-Images sollen den Zustand der jeweiligen Seite möglichst originalgetreu wiedergeben. Daher werden bei Bedarf die Imageseiten in einer Nachbearbeitung geradegerückt und von erfassungsbedingt enthaltenen dunklen von Schatten oder Strichen (Kanten, Falz) gesäubert. Bei doppelseitiger Aufnahme werden die Seiten getrennt und sequentiell sortiert.

Eine interaktive Entfernung von Grauschleier, der sich aus der Aufnahme von vergilbten Seiten ergeben kann, wird nicht vorgesehen. Wohl jedoch eine teilweise Säuberung durch automatisierte Prozeduren, soweit diese vorhanden sind.

Die Indexierung erfolgt durch maschinelle Übernahme von elektronischen Katalogdaten.

Die Texterfassung von Inhaltsverzeichnis, Kapitelüberschriften, Register erfolgt mit einer Genauigkeit von >99,985% (0,015% Fehlertoleranz); der Volltext wird mit 99,7% Genauigkeit erfaßt.

Zur Sicherung der erforderlichen Qualität werden alle Arbeitsergebnisse einzeln überprüft. Nach Abschluß der Erfassung erfolgt eine Endkontrolle.

Der durchschnittliche Speicherbedarf je gescannte Seite wird auf 150 KByte geschätzt. Bei stark vergilbten Buchseiten kann sich ein wesentlich höherer Speicherbedarf ergeben. In diesem Fall wären unter Umständen zusätzliche Reinigungsschritte vorzusehen.

1.2 Spezifikation des Standardbuches

Dieses Standardbuch wird folgendermaßen spezifiziert:

Spezifikation des Standardbuches	
Format:	Quart, mindestens A5, höchsten A4
Umfang:	300 Seiten Text zu jeweils 4000 Zeichen
Inhaltsverzeichnis:	3 Seiten zu jeweils 2000 Zeichen
Register:	6 Seiten zu jeweils 2000 Zeichen

2 Alternativen für die Vorgehensweise

Für das Scannen und Verfilmen bieten sich folgende Vorgehensweisen an:

a) Getrenntes Scannen und Verfilmen vom Original
b) Verfilmen des Originals und Scannen des Films
c) Scannen des Originals und Erzeugung einer Filmkopie nach dem COM-Verfahren

Die diesen Vorgehensweisen entsprechenden Arbeitsschritte werden im folgenden Abschnitt beschrieben.

3 Arbeitsschritte

3.1 Auswahl der zu erfassenden Bücher

Arbeitsinhalt	Die zu erfassenden Bücher werden (z.B. nach Inhalt, Erhaltungszustand, urheberrechtlichen Gesichtspunkten, Bedarf etc.) ausgewählt und bereitgestellt.
Ausführende Stelle	Die Arbeiten sollten von Mitarbeitern der Bibliothek ausgeführt werden
Qualifikation	Diplom-Bibliothekar
Arbeitsaufwand	unterschiedlich (z.B. 5 Minuten)

3.2 Bibliothekarische Vorbereitung

Arbeitsinhalt	Das zu erfassende Buch wird unter bibliothekarischen Gesichtspunkten für das Scannen/Verfilmen vorbereitet
Details	• Identitätsprüfung (Titel, Autor, ...) und Abgleichung mit den Katalogdaten • Prüfen, ob das Buch aus inhaltlichen und technischen Gründen gescannt werden kann • Festlegung der Indexierung • Festlegung der zu erfassenden Seiten. • Festlegung von Sonderpaginierungen • Festlegung der Bearbeitung von Überformaten oder verkleinerten Seiten (Miniprints) • Festlegung der Erfassung von Seiten mit Abbildungen in Farbe , SW-Halbton oder Rasterung • Bearbeitung von Errata/Corrigenda • Hinweis auf beigelegte Sonderseiten • Bibliotheksseitige Dokumentation
Ausführende Stelle	Die Arbeiten sollten von Mitarbeitern der Bibliothek ausgeführt werden
Qualifikation	Studentische Aushilfe unter fachlicher Anleitung durch einen Diplom-Bibliothekar
Arbeitsaufwand	15-30 Minuten (realistischer Durchschnittswert: 20 Minuten)

3.3 Technische Vorbereitung

Arbeitsinhalt	Das zu erfassende Buch wird unter technischen Gesichtspunkten für das Scannen/Verfilmen vorbereitet
Details	• Erfassung der Buchstruktur • Ggf. Anfertigung von einzelnen Zwischenkopien • Festlegen der technischen Erfassungsparameter (z.B. Schwellwert, Maskenhintergrund, ..)
Ausführende Stelle	Die Arbeiten sollten von Mitarbeitern der Erfassungsstelle ausgeführt werden
Qualifikation	Operator
Arbeitsaufwand	Kann sehr stark differieren, besonders wenn Zwischenkopien erstellt werden müssen. Realistischer Erfahrungswert bei durchschnittlich 8 Minuten.

3.4 Indexierung

Arbeitsinhalt	Eingabe von Begriffen (z.B. Minimalsatz bibliographischer Daten) nach Vorgabe von Abschnitt 3.2 um das Buch suchbar zu machen
Details	• Maschinelle Verknüpfung mit Katalogdaten • Manuelle Eingabe
Ausführende Stelle	Die Arbeiten sollten von Mitarbeitern der Erfassungsstelle ausgeführt werden
Qualifikation	technischer Mitarbeiter
Arbeitsaufwand	5 Minuten

3.5 Erfassung

a) Getrenntes Scannen und Verfilmen vom Original

Arbeitsinhalt	Das Buch wird gescannt und getrennt verfilmt
Ausführende Stelle	Die Arbeiten sollten von Mitarbeitern der Erfassungsstelle ausgeführt werden
Qualifikation	technischer Mitarbeiter
Arbeitsaufwand	60 Minuten

b) Verfilmen des Originals und Scannen des Films

Arbeitsinhalt	Das Buch wird verfilmt; der resultierende Film wird in einem zweiten Schritt eingescannt
Ausführende Stelle	Die Arbeiten sollten von Mitarbeitern der Erfassungsstelle ausgeführt werden
Qualifikation	technischer Mitarbeiter
Arbeitsaufwand	30 Minuten

c) Scannen des Originals und Erzeugung einer Filmkopie nach dem COM-Verfahren

Bei dem COM-Verfahren (Computer Output on Microfilm) werden digitale Daten per Laserbelichtung auf einen Mikrofilm „geschrieben". Das Ergebnis entspricht einem aus der Originalvorlage photographisch erstellten Film.

Dieses billige und qualitativ sehr hochwertige Verfahren hat sich in den letzten Jahren in vielen Anwendungsbereichen auf Basis von 16-mm-Film durchgesetzt. COM-Belichter für den 35-mm-Film sind noch neu und wegen der daher noch hohen Gerätekosten nur dort wirtschaftlich, wo es um die Erzielung höchster Qualitäten geht. Es wird jedoch erwartet, daß dieses Verfahren in wenigen Jahren gerade im Bibliotheksbereich verstärkt angewandt werden wird.

Arbeitsinhalt	Das Buch wird gescannt; die resultierende digitale Datei wird in einem zweiten Schritt zusätzlich auf 35-mm-Film ausgegeben
Ausführende Stelle	Die Arbeiten sollten von Mitarbeitern der Erfassungsstelle ausgeführt werden
Qualifikation	technischer Mitarbeiter
Arbeitsaufwand	50 Minuten

3.6 Nachbearbeitung

a) Getrenntes Scannen und Verfilmen vom Original

Im Normalfall entbehrlich.

b) Verfilmen des Originals und Scannen des Films

Der Nachteil dieser Vorgehensweise liegt vor allem darin, daß beim Verfilmen von gebundenen Vorlagen dunkle Streifen am Rand und (bei doppelseitiger

Aufnahme) am Falz entstehen. Außerdem läßt sich eine Verkantung der Seiten kaum vermeiden.

Die beim Einscannen des Films erhaltenen digitalen Dateien entsprechen damit jedoch nicht mehr den üblichen Anforderungen. Es sind folgende Arbeiten erforderlich: Trennen von Doppelseiten; sequentielles Sortieren der resultierenden Einzelseiten; Entfernen von Schatten und Falzrändern; lotrechtes Ausrichten der Satzspiegel. Diese Arbeiten können interaktiv oder automatisiert erfolgen.

Automatisierte Verfahren erkennen Randstreifen sowie die Lage des Satzspiegels mittels regressionsanalytischer Verfahren. Bei manuellen Verfahren werden vom Bearbeiter die zu bearbeitenden oder zu löschenden Teile der Imageseite mittels Rahmenmarkierung am Bildschirm festgelegt.

Arbeitsinhalt	Trennen von Doppelseiten, Ausrichten, Entfernen von Schatten und Strichen
Ausführende Stelle	Die Arbeiten sollten von Mitarbeitern der Erfassungsstelle ausgeführt werden
Qualifikation	technischer Mitarbeiter
Arbeitsaufwand	Automatisiert: 10-20 Minuten Manuell: 60 Minuten

c) *Scannen des Originals und Erzeugung einer Filmkopie nach dem COM-Verfahren*

Im Normalfall entbehrlich.

3.7 Text-Erfassung

Die Texterfassung kann maschinell (Optical Character Recognition; OCR) oder manuell erfolgen. Für Textteile (Inhaltsverzeichnisse, Kapitelüberschriften, Register), die zur späteren Recherche genutzt werden sollen, ist eine besonders sorgfältige manuelle Doppelerfassung mit interaktiver Verifikation erforderlich. Häufig sind auch Kombinationen von OCR und manueller Eingabe zweckmäßig. Die Erfassung erfolgt üblicherweise auf Grundlage der Imagedatei. Die Kosten für die OCR-Erfassung sind zumeist nur geringfügig niedriger als die manuelle Eingabe in Ländern Südostasiens.

Historische Schrifttypen bilden oftmals ein besonderes Problem, weil maschinelles Lesen wenig Erfolg verspricht und immer weniger Menschen diese Schrift hundertprozentig beherrschen.

Arbeitsinhalt	Eingabe von Inhalten des Buches
Details	• Inhaltsverzeichnis • Kapitelüberschriften • Register • ggf. Fußnoten • ggf. Volltext
Ausführende Stelle	Die Arbeiten können von Mitarbeitern der Erfassungsstelle oder auch der Bibliothek ausgeführt werden
Qualifikation	Datenerfasser (manuelle Eingabe) Technische Mitarbeiter (OCR)
Arbeitsaufwand	Inhaltsverzeichnis: 60 Minuten manuell Inhaltsverzeichnis: 10 Minuten OCR incl. Kontrolle (Kapitelüberschriften: 60 Minuten) manuell (Kapitelüberschriften: 10 Minuten) OCR incl Kontrolle Register: 120 Minuten manuell Register: 20 Minuten OCR incl Kontrolle Volltext: 100 Stunden manuell Volltext: 20 Stunden OCR

3.8 Permanente Speicherung

Arbeitsinhalt	Speicherung auf CD-R (Kapazität z.Zt. 650 MByte = ca. 15 Bücher)
Details	• Images • Index-Daten • Inhaltsverzeichnisse, Kapitelüberschriften • Register • Volltext
Ausführende Stelle	Die Arbeiten sollten von Mitarbeitern der Erfassungsstelle ausgeführt werden
Qualifikation	Technische Mitarbeiter
Arbeitsaufwand	4 Minuten (wenn die CD in einem Arbeitsgang komplett beschrieben wird = ca. 15 Bücher = 1 Stunde)

3.9 Endkontrolle, Abnahme

Arbeitsinhalt	Abschließende stichprobenweise Überprüfung aller Arbeiten
Details	• Vollständigkeit • Erfassungsqualität (Filmen, Scannen, Texterfassung) • Richtigkeit der Seitennummerierung und Reihenfolge der Seiten • Strukturierung • Gesamteindruck
Ausführende Stelle	Die Arbeiten sollten von Mitarbeitern der Bibliothek ausgeführt werden
Qualifikation	Studentische Aushilfe unter fachlicher Anleitung durch einen Diplom-Bibliothekar
Arbeitsaufwand	10 Minuten

3.10 Schlußdokumentation

Arbeitsinhalt	Abschließende Dokumentation über die durchgeführten Arbeiten
Details	Strukturierte Eingabe in Datenbank, Katalog etc.
Ausführende Stelle	Die Arbeiten sollten von Mitarbeitern der Bibliothek ausgeführt werden
Qualifikation	Studentische Aushilfe unter fachlicher Anleitung durch einen Diplom-Bibliothekar
Arbeitsaufwand	5 Minuten

4 Zusammenfassung: Durchschnittlicher Aufwand für das Erfassen des Standardbuches

Arbeitsschritt	Bibliothek	Dienstleister	Total
Auswahl der zu erfassenden Bücher	5 Minuten		5 Minuten
Bibliothekarische Vorbereitung	20 Minuten		20 Minuten
Technische Vorbereitung		8 Minuten	8 Minuten
Indexierung		5 Minuten	5 Minuten
Erfassung		30-60 Minuten	30-60 Minuten
Nachbearbeitung		10-60 Minuten	10-60 Minuten
Text-Erfassung Inhaltsverz./ Kapitelüberschr.	10-60 Minuten	10-60 Minuten	10-60 Minuten
Text-Erfassung Register	20-120 Minuten	20-120 Minuten	20-120 Minuten
Permanente Speicherung		4 Minuten	4 Minuten
Endkontrolle, Abnahme	10 Minuten		10 Minuten
Schlußdokumentation	5 Minuten		5 Minuten
Total			127-357 Minuten
Zusätzliche Volltexterfassung	20-100 Stunden	20-100 Stunden	20-100 Stunden

Ohne die Erfassung des gesamten Volltextes erfordert das Verfilmen, Scannen, Indexieren, Texterfassung von Inhaltsverzeichnis/Kapitelüberschriften und Registern und Abspeichern auf CD-R einen Arbeitsaufwand von - je nach gewählter Vorgehensweise - 2 bis 6 Stunden. Die Bandbreite ergibt sich vor allem aus der unterschiedlichen Art der Texterfassung.

Für ein Buch aus dem 19. Jahrhundert, dessen Schriftzeichen sich nicht für OCR-Erkennung eignen, resultiert ein Zeitaufwand von rund 5 bis 6 Stunden. Diese Arbeiten können folgendermaßen zwischen der Bibliothek und dem Dienstleister (der natürlich auch eine Bibliothek mit entsprechender Ausstattung und Erfahrung sein kann) verteilen:

Arbeitsstelle	Zeit (Minuten)
Bibliothek	40
Dienstleister	57 - 137
Bibliothek oder Dienstleister	180

5 Zusammenfassung:
Durchschnittliche Fremdkosten für das Erfassen des Standardbuches

Die nachstehenden Kostenangaben beruhen auf Angeboten der letzten Monate.

Arbeitsschritt	Einheit	Preis je Einheit		Total je Band	
		von	bis	von	bis
Technische Vorbereitung, Indexierung	1 Band	0	8,00	0	8,00
Verfilmen	100 Seiten	10,50	20,00	31,50	60,00
Scannen vom Film, einfache Nachbearbeitung	100 Seiten	15,00	20,00	45,00	60,00
Nachbearbeitung	100 Seiten	10,00	10,00	30,00	30,00
Texterfassung Inhaltsverz./Kapitelüberschr.	1000 Zeichen	3,50	18,00	21,00	108,00
Texterfassung Register	1000 Zeichen	3,50	18,00	42,00	216,00
Speicherung: Erste CD-R (4000 Seiten)	1 CD	35,00	130,00	2,70	10,00
Zusätzliche Kopie (4000 Seiten)	1 CD	28,00	80,00	2,20	6,20
Total		105,50	304,00	174,40	498,20
Scannen vom Buch	100 Seiten	30,00	92,00	90,00	276,00
Zusätzliche Volltexterfassug	1000 Zeichen	1,50	9,00	2700,00	10800,00

Nicht enthalten sind: Transport, Versicherung (ca. 2% der Kosten).

Einmalige Kosten: Software zur lokalen Kontrolle (einfache Dokumentenverwaltung, Viewing, Print): Zwischen 1000 und 5000 DM.

Anpassung der Erfassungssoftware: nach Aufwand

Rechnerkonfiguration: PC 133 MHz, 32 MB RAM, 2 GB SCSI-HD, hoch auflösender Bildschirm ca. 20", Arbeitsplatz-Laserdrucker (z.B. HP Laserjet 5L), zusammen ca. 5000 - 9000 DM.

Kosten des Hosts: Umwandlung in Internet-Format
Eintrag in das Datenbank-Managementsystem
Speicherung

In diesem Modell entfallen auf die Bibliothek selbst pro Buch weitere ca. 40 Minuten Arbeitsaufwand, der überwiegend von studentischen Aushilfen unter qualifizierter Anleitung erbracht werden kann.

Interessant ist der folgende Vergleich der Komplettangebote:

Billigster Gesamtanbieter

Arbeitsschritt	Einheit	Preis je Einheit	Total je Band
Technische Vorbereitung, Indexierung (im Verfilmen+Scannen enthalten)	1 Band	0	0
Verfilmen	100 Seiten	20,00	60,00
Scannen vom Film, einfache Nachbearbeitung	100 Seiten	20,00	60,00
Nachbearbeitung (entfällt)	100 Seiten	0	0
Texterfassung Inhaltsverz./Kapitelüberschr.	1000 Zeichen	5,00	30,00
Texterfassung Register	1000 Zeichen	5,00	60,00
Speicherung: Erste CD-R (4000 Seiten)	1 CD	35,00	2,70
Zusätzliche Kopie (4000 Seiten)	1 CD	28,00	2,20
Total		113,00	214,90
Scannen vom Buch	100 Seiten	38,00	114,00

Teuerster Gesamtanbieter

Arbeitsschritt	Einheit	Preis je Einheit	Total je Band
Technische Vorbereitung, Indexierung (im Verfilmen+Scannen enthalten)	1 Band	0	0
Verfilmen	100 Seiten	10,50	31,50
Scannen vom Film, sehr einfache Nachbearbeitung	100 Seiten	15,00	45,00
Nachbearbeitung (entfällt)	100 Seiten	0	0
Texterfassung Inhaltsverz./Kapitelüberschr.	1000 Zeichen	18,00	108,00
Texterfassung Register	1000 Zeichen	18,00	216,00
Speicherung: Erste CD-R (4000 Seiten)	1 CD	130,00	10,00
Zusätzliche Kopie (4000 Seiten)	1 CD	80,00	6,20
Total		271,50	416,70
Scannen vom Buch	100 Seiten	k.A.	k.A.

Dabei ist anzumerken, daß der billigere Anbieter die bessere Qualität (auf Basis seiner Muster) zusichert.

Auffällig sind die großen Abweichungen bei den Kosten einzelner Positionen. Es sollte daher geprüft werden, inwiefern ein Splitten des Auftrages auf mehrere Anbieter möglich ist.