

11 (2008)  
AUSGABE 3

**Sonderdruck**

# **B**.**i**.**t** *online*

Bibliothek  
Information  
Technologie

D 52614  
ISSN 1435-7607

**Zeitschrift für Bibliothek, Information und Technologie  
mit aktueller Internet-Präsenz: [www.b-i-t-online.de](http://www.b-i-t-online.de)**



## **Zur Workflowsteuerung der Massendigitalisierung**

**Der Weg der Bücher und der Digitalisate**

Martin Baumgartner, Michael Beer, Berthold Gillitzer, Rosmarie Leichtl, Gabriele Meßmer,  
Karsten Trzcionka, Thomas Wolf-Klostermann und Wilhelm Hilpert

# Zur Workflowsteuerung der Massendigitalisierung

## Der Weg der Bücher und der Digitalisate

Martin Baumgartner, Michael Beer, Berthold Gillitzer, Rosmarie Leichtl, Gabriele Meßmer, Karsten Trzcionka, Thomas Wolf-Klostermann und Wilhelm Hilpert

■ Im Jahr 2007 hat die Bayerische Staatsbibliothek einen vielbeachteten Vertrag mit der Firma Google geschlossen, auf dessen Basis rund eine Million Titel aus dem urheberrechtlich freien Altbestand der Bayerischen Staatsbibliothek digitalisiert werden. Das Projekt ist auf eine Laufzeit von mehreren Jahren ausgelegt. Zudem ist vereinbart, dass die Bücher innerhalb des Freistaates Bayern eingescannt werden. Der Vertrag, basierend auf einer europaweiten „Bekanntmachung zur Teilnahme am Verhandlungsverfahren“ der Bayerischen Staatsbibliothek in Tenders Electronic Daily, stieß auf breite Resonanz in der Presse und löste eine lebhafte Diskussion innerhalb der bibliothekarischen Öffentlichkeit aus.<sup>1</sup>

Es gehört zum Wesen von Public-Private-Partnership-Abkommen, dass die Vorgaben beider Seiten hinsichtlich Vertraulichkeit formuliert und natürlich auch beachtet werden. Für privatwirtschaftliche Unternehmen ist der Schutz ihres Know-hows, insbesondere ihrer Betriebsabläufe und ihrer Technologie oft wettbewerbsentscheidend. So ist „Confidentiality“ das prägende Wort, das den Beziehungen zwischen Google und seinen Partnerbibliotheken zu Grunde liegt. Insofern wird auch dieser Artikel keine Aussagen zur Scantechnologie von Google, zu deren Qualitätssicherungsverfahren oder deren interner Ablauforganisation machen. Wir können Google im Rahmen dieser Ver-

öffentlichung als einen beliebigen Dienstleister betrachten, in dessen Geschäft wir uns nicht einmischen, der sich allerdings in zwei wesentlichen Punkten von allen anderen, bisher in Deutschland tätigen Scandienstleistern unterscheidet. Statt einer Bezahlung erhält – oder genauer behält – Google eine Kopie des Digitalisates und

tet dies, dass der Bayerischen Staatsbibliothek für die Digitalisierung keinerlei direkte Kosten entstehen und dass die Bayerische Staatsbibliothek eine eigene identische Kopie, die sogenannte „Library Digital Copy“ des durch Google erstellten Digitalisates erhält.

### Die Vorbereitungsphase

Anlass für diesen Artikel ist der Abschluss der Vorbereitungsphase dieses Massendigitalisierungsprojektes und der Start in die Produktivphase. Manchem mag die Vorberei-



Abb. 1: Der Altbestand der Bayerischen Staatsbibliothek mag mit über einer Million Titel noch manchen unentdeckten Schatz für die Wissenschaft enthalten.



statt einiger hundert Seiten pro Tag werden viele hundert Bücher pro Tag eingescannt und nachbearbeitet. Grundsätzlich können die beschriebenen Workflows jedoch als paradigmatisch für das Arbeitsfeld Massendigitalisierung betrachtet werden und kommen – in gegebenenfalls modifizierter Form – so auch in anderen Großprojekten der Bayerischen Staatsbibliothek zum Einsatz, beispielsweise im Rahmen der Digitalisierung der 37.000 deutschen Drucke des 16. Jahrhunderts.

Aus anderem Blickwinkel betrachtet bedeu-

tungszeit von über einem Jahr als sehr lang erscheinen, sie ist aber einem Projekt dieser Größenordnung durchaus adäquat und aus der Sicht aller Beteiligten eher zu kurz als zu lang bemessen. Google musste zunächst die erforderlichen personellen und materiellen Ressourcen für ein Projekt dieser Größenordnung bereitstellen. Von Seiten der Bayerischen Staatsbibliothek mussten die Werkzeuge geschaffen werden, die es erlauben, den Strom der Bücher und Digitalisate zu steuern, zu überwachen und sicher zu archivieren.

<sup>1</sup> Vgl.: Klaus Ceynowa: Massendigitalisierung für die Wissenschaft – Zur Digitalisierungsstrategie der Bayerischen Staatsbibliothek. In: Information – Innovation – Inspiration. 450 Jahre Bayerische Staatsbibliothek. Hrsg. von Rolf Griebel und Klaus Ceynowa. München: Saur 2008. S. 241-252.

### Die logistische Herausforderung

Nicht jeder kann abschätzen, welche Herausforderung es darstellt, über eine Million Bücher in der sehr überschaubaren Projektlaufzeit bereitzustellen und zu bearbeiten. Wurde in der Vergangenheit in Deutschland oft schon dann von Massendigitalisierung gesprochen, wenn im Rahmen eines Projektes einige tausend Werke eingescannt wurden, so muss man sich klar machen, dass solche Mengen im Rahmen dieses Vorhabens in weniger als einer Woche bearbeitet werden. Mit diesem Projekt beginnt für die Digitalisierung in deutschen Bibliotheken eine neue Ära, die der industriellen Massendigitalisierung.

Wenn wir weiter oben sagten, dass wir uns in das Geschäft des „Scandienstleisters“ Google nicht einmischen, so ist dies korrekt bis auf zwei, allerdings sehr wichtige Punkte:

- Die Bayerische Staatsbibliothek hat und behält die alleinige Kontrolle darüber, dass konservatorische und bestandserhalterische Gesichtspunkte bei jedem Arbeitsschritt, auch auf der Ebene des Einzelfalles, Anwendung und Berücksichtigung erfahren.
- Die Bayerische Staatsbibliothek arbeitet sehr eng mit Google zusammen, um die Digitalisate auf einem möglichst hohen qualitativen Niveau anbieten zu können.

Das Niveau wird im Kern dem entsprechen, das auch im Rahmen öffentlich-rechtlich geförderter Projekte erreicht wird. Google arbeitet laufend an der Verbesserung der Qualität seiner Digitalisate und seiner Qualitätssicherungsmaßnahmen. Deshalb ist es für die Bayerische Staatsbibliothek durchaus vorteilhaft, nicht der Gruppe der ersten „Google-Bibliotheken“ angehört zu haben. Dass man jedoch Digitalisate, die im Zuge einer industriellen Massendigitalisierung entstanden sind, nicht in jeder Hinsicht mit der Qualität von Digitalisaten einer fünfzigseitigen Inkunabel vergleichen kann, für deren Bearbeitung man mehrere Tage aufgewendet hat, dürfte sich von selbst verstehen.

Massendigitalisierung hat zur Konsequenz, dass es im Rahmen eines solchen Ablaufes nicht oder nur mit sehr hohem personellen Aufwand, den die Bayerische Staatsbibliothek nicht leisten kann, möglich ist, einzelne Titel, Textcorpora oder Systemgruppen gezielt auszuwählen. Einzig die Logistik bestimmt das Vorgehen und das Tempo bei der Digitalisierung.

Schon vor dem Start des Google-Projektes wurden aus den Magazinen der Bayerischen Staatsbibliothek täglich an die 6.000

Bücher für Ortsleihe, Lesesaalleihe, Fernleihe, Dokumentlieferung und die dienstliche Ausleihe (überwiegend für Digitalisierungsprojekte und sonstige Bestandserhaltungsmaßnahmen) bereitgestellt. Selbstverständlich bedeutet dies, dass an jedem Arbeitstag im Magazin zurückgeführt werden müssen. Serviceorientierung drückt sich gewiss nicht nur, aber eben auch in den Nutzungszahlen einer Bibliothek aus. Hinzu kommen etwa 5.000 Bände, die täglich, zur Optimierung der Magazinbewirtschaftung, innerhalb der Magazine umgezogen werden, sowie über 500 Bände, die täglich erstmalig aus dem Neuzugang in die Magazine eingestellt werden. Durch die zusätzlichen Buchbewegungen (Ausheben, Einstellen, Transporte von und zum Projektteam und der Metadatenoptimierung), die pro Arbeitstag aus dem Massendigitalisierungsprojekt resultieren, erhöht sich das Transportvolumen auf deutlich über 20.000 Bände pro Arbeitstag.

Situation vorbereitet, die ihnen abverlangt, schnell und präzise diese anstehenden Entscheidungen zu treffen.

### Die bibliothekarische Herausforderung

Noch wesentlich größer als die logistische Herausforderung ist die bibliothekarische Herausforderung. Diese beinhaltet, dass die Bücher mit korrekten und eindeutig zuzuordnenden Metadaten an Google übergeben werden müssen. Zusammen mit der Übergabe einer Produktionscharge sind auch laufend die zu dieser Charge gehörenden Metadaten, mit einer eindeutigen Zuordnung zu den entsprechenden Werken, zu transferieren. Die Metadaten des für die Digitalisierung in Frage kommenden urheberrechtsfreien Altbestandes liegen zwar fast komplett in elektronischer Form vor, wurden jedoch zum ganz überwiegenden Teil durch Retrokonversion von

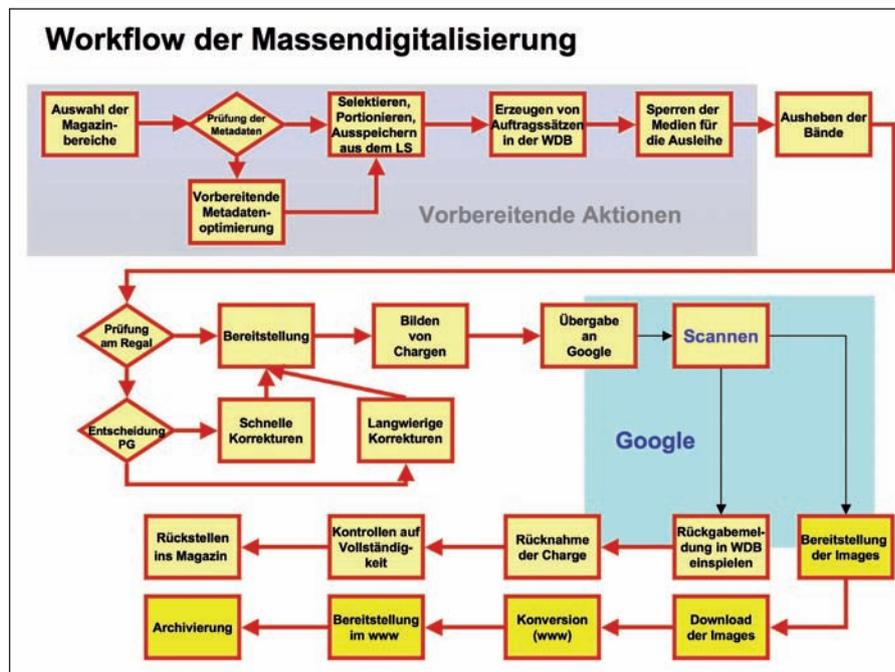


Abb. 2: Der Workflow im Überblick (LS = Lokalsystem; PG = Projektarbeitsgruppe; WDB = Workflowdatenbank)

Nicht vergessen werden darf, dass täglich an mehreren Stellen des logistischen Prozesses zahlreiche Entscheidungen getroffen werden müssen. Die konservatorische Eignung, die Vollständigkeit der Metadaten und deren eindeutige Zuordenbarkeit zum betreffenden Werk, aber auch die Eignung hinsichtlich der Größe oder die Beurteilung der urheberrechtlichen Situation sind Beispiele für Problemstellungen auf der Ebene des einzelnen Bandes, die ad hoc eine Entscheidung verlangen. Die Projektmitarbeiter wurden in Schulungen und vor allem intensiven Einarbeitungsphasen auf diese

handgeschriebenen Katalogen gewonnen. Da die alten Kataloge oft nur unzureichende Informationen für die Übernahme in den EDV-Katalog boten, eine Katalogisierung per Autopsie aber auch in Problemfällen nicht möglich war, sind sie teilweise mit Fehlern behaftet. Für einen Teil dieser Fehler ist es zwingend notwendig, dass sie korrigiert werden, bevor das zugehörige Werk in den Massendigitalisierungsprozess gegeben wird. Die drei häufigsten Fehler sind unvollständige und fehlerhafte Konversionsaufnahmen, eine fehlerhafte Bandstruktur und fehlerhafte Signaturen.

Viele formale Fehler (z. B. eine falsche Bandstruktur) können durch automatisierte Abfragen der Katalogdatenbank identifiziert und als Fehlerlisten ausgegeben werden. Angelernte Hilfskräfte sind nach eingehender Einarbeitung in der Lage, einen Teil dieser Fehlerlisten zu bearbeiten. Dadurch werden diese Fehler für den gesamten Bestand so rechtzeitig bereinigt werden können, dass sie keinen Einfluss auf die Bereitstellung für den Digitalisierungsprozess haben. Der größere Teil der Fehler erfordert aber den Einsatz von Katalogspezialisten, die mit dem Altbestand der Bayerischen Staatsbibliothek sehr gut vertraut sein sollten. Zwar wird nur das unbedingt Notwendige korrigiert, trotzdem wird dieser Teil sicher nicht so schnell erledigt sein, dass die Metadaten jedes Buches, bevor es an der Reihe ist digitalisiert zu werden, schon bearbeitet sind. Auf die Konsequenzen, die aus dieser Einsicht zu ziehen sind, kommen wir später zu sprechen.

### Die Workflowdatenbank

Aber jetzt sind wir entlang des Handlungsstrangs „Fehlerbereinigung der Metadaten“ schon sehr weit in das Projekt eingestiegen. Um zu erklären, wie vorgegangen wird und was es bedeutet, dass ein Buch „an der Reihe ist“, ist ein wenig weiter auszuholen. Als wir das Projekt angingen, wurde uns sehr schnell klar, dass wir ein Instrument brauchen, das uns die Übersicht über den gesamten Workflow behalten lässt und uns auf viele absehbare Fragen Antwort geben kann:

- Ist Buch xy beim Dienstleister?
- Welche Bücher insgesamt sind zurzeit beim Dienstleister?
- Ist Buch xy zurück und vom Dienstleister digitalisiert?
- Haben wir von Werk xy auch das Digitalisat, die Digital Library Copy, erhalten?
- Welche Gründe waren ausschlaggebend, dass Buch xy nicht digitalisiert wurde?
- Wie lange muss Benutzer Mustermann voraussichtlich noch auf Buch xy warten?
- Wie lange muss Benutzerin Musterfrau aus Würzburg noch auf das Digitalisat im Internet warten?
- Wie viele Bücher sind bis heute exakt bearbeitet?
- Hat Buch xy eine ausklappbare Karte oder Graphik, die zunächst nicht eingescannt werden konnte?
- Welche Bücher übergeben wir dem Dienstleister morgen?
- usw.

Rasch wurde deutlich, dass es dieses Instrument in der benötigten und gewünsch-

ten Form nicht bzw. nur in Ansätzen gab. Daher haben wir mit MyBib eine Software, die von der Firma ImageWare für die Auftragsbearbeitung in der bibliothekarischen Dokumentlieferung konzipiert wurde, sowie mit ZEND, der Zentralen Erfassungs- und Nachweisdatenbank der Bayerischen Staatsbibliothek, zwei Softwarewerkzeuge ausgewählt, auf denen wir aufbauen konnten und die beide bereits mit nachhaltigem Erfolg an der Bayerischen Staatsbibliothek im Einsatz sind. ZEND und MyBib wurden vom Münchner Digitalisierungszentrum der Bayerischen Staatsbibliothek und von ImageWare zu einem Werkzeug zur Steuerung von Massendigitalisierungsprojekten fortentwickelt. Beide Teile fügen sich harmonisch zusammen und sind über offene Schnittstellen verbunden. Beide Teile zusammen werden wir in dieser Veröffentlichung als die „Workflowdatenbank der Bayerischen Staatsbibliothek“ (WDB) bezeichnen, wobei der Teil, der auf MyBib aufsetzt, unter dem vorläufigen Arbeitstitel MyBib-WDB geführt wird und vor allem die logistischen Abläufe steuert und dokumentiert. Die für diesen Einsatz in der industriellen Massendigitalisierung erweiterte ZEND nennen wir Google-ZEND. Dieser Teil wird die Abholung der Digitalisate, die Erzeugung von WWW-tauglichen Bildformaten sowie die Bereitstellung und die Langzeitarchivierung steuern. Wie eng die Verzahnung zwischen MyBib-WDB, Google-ZEND und dem Katalogsystem letztendlich ist, zeigt ein einziger wichtiger Vorgang sehr gut, der zum Nachweis des Digitalisates über den Katalog führt. Nach der Bereitstellung des Digitalisates im Repository durch die Google-ZEND erfolgt eine automatische Rückmeldung von Google-ZEND an die MyBib-WDB und diese wiederum veranlasst über das Verbundsystem des Bibliotheksverbundes Bayern den Nachweis im Verbundkatalog wie im lokalen Katalog, durch die Eintragung eines persistenten Links beim entsprechenden Katalogisat. Insgesamt werden auf diese Weise über 30 Arbeitsschritte des Workflows von beiden Softwaretools gesteuert und dokumentiert. Im Folgenden sind die wichtigsten Funktionalitäten nochmals stichpunktartig aufgeführt:

- Gewährleistung der korrekten Zuordnung zwischen zu scannendem Objekt, den Metadaten und dem Digitalisat
- Anlegen von Auftragsätzen für die Digitalisierung und Ausdruck der zugehörigen Auftragszettel
- Erfassung der Gründe, falls ein Werk nicht bearbeitet werden kann
- Steuerung der Nachbearbeitung, z.B. Metadatenkorrektur (falls notwendig)
- Zusammenstellung einer Produktions-

charge und Erfassen aller Werke in einer Charge

- Kontrollierte Übergabe der Bücher einer Charge mit Übergabe der zugehörigen Metadaten an den Scandienstleister
- Rücknahme der Bücher vom Scandienstleister und Überprüfung der Vollständigkeit
- Übernahme der Library Digital Copy vom Scandienstleister
- Bereitstellung im World Wide Web
- Nachweis in lokalen und überregionalen Katalogen durch automatisierte Bildung von URN's und weiteren Informationen (Resolving URL; Hinweis auf Kostenfreiheit usw.) und deren Eintrag in den Katalogen
- Übertragung der Bilddateien, der Strukturdaten und Textdateien fertig bearbeiteter Bände in das Langzeitarchivierungssystem des Leibniz-Rechenzentrums.

Nur durch einen sehr hohen Grad an Automatisierung ist letztendlich die tägliche Bearbeitung von vielen hundert Digitalisaten möglich. Das LRZ betreut die skalierbare Hardware mit ausreichendem Speicherplatz und das Cluster von Servern zur Bildkonversion der Digitalisate der Bayerischen Staatsbibliothek und stellt die benötigten schnellen Datenleitungen zur Verfügung, um die umfangreiche tägliche Produktion verarbeiten zu können.

### Der Workflow im Überblick

Der Altbestand der Bayerischen Staatsbibliothek ist in knapp über 200 verschiedenen Fachsignaturen aufgestellt. Diese Fachsignaturaufstellung wurde bis zum Jahr 1936 beibehalten und strukturiert noch heute alle bis zum Jahr 1936 erschienenen Werke. Es liegt also nahe, entsprechend diesen Fachsignaturen vorzugehen und dabei alle für die Digitalisierung infrage kommenden Werke auszuwählen. Die Reihenfolge, in der die Fachsignaturen abgearbeitet werden, wird viele Monate vor der eigentlichen Digitalisierung festgelegt. Bestimmt wird die Wahl von der prognostizierten Fehlerhäufigkeit bei den Metadaten, der zu erledigenden oder bereits erledigten bestands-erhalterischen Maßnahmen, dem Format<sup>2</sup>, dem Magazinstandort<sup>3</sup> und der mittelfristigen Umzugsplanung für die Magazine. Es werden Bereiche in der Größe von mehre-

2 Die Bayerische Staatsbibliothek hat den Altbestand überwiegend in den drei Formaten Folio (≥ 35 cm), Quarto- (25-35 cm) und Octavo-Format (≤ 25 cm) aufgestellt.

3 Die Bayerische Staatsbibliothek bewirtschaftet zurzeit drei Magazinstandorte: Das Stammhaus in der Ludwigstraße, die Speichermagazine in Garching und ein Ausweichmagazin im Münchner Euroindustrialpark.

ren 10.000 Werken ausgewählt. Die Metadaten, insbesondere aber die Signaturen, werden automatisierten formalen Prüfungen unterzogen. Entsprechende Fehlerlisten werden ausgegeben und wie weiter oben schon angedeutet durch angelernte Hilfskräfte abgearbeitet. Die bearbeiteten Signaturbereiche werden nun im Lokalsystem erneut selektiert, portioniert und gespeichert. Nach Einspeicherung in die MyBib-WDB werden Auftragsätze angelegt. Diesen wird eine eindeutige Nummer zugeordnet, die sogenannte digID, die zeitnah in den Titelsätzen im lokalen Katalog der Bayerischen Staatsbibliothek durch Rückmeldung ergänzt wird. Die digID wird dadurch zum wichtigsten Verbindungsglied zwischen der lokalen Katalogdatenbank und den Daten der MyBib-WDB. Sie wird später ein entscheidendes Element des persistenten Links und bildet damit auch die eindeutige Verbindung zwischen Archivierungssystem und Katalogen. Auf den Auftragszetteln wird sie als Strichcode ausgegeben und begleitet so das Buch auf seinem Weg im Digitalisierungsprozess, bis es zur Bayerischen Staatsbibliothek zurückkommt und wieder seinen Stammplatz im Magazin einnimmt. Die digID lässt auch die Library Digital Copy den „richtigen Platz“ im Archivierungssystem der Bayerischen Staatsbibliothek einnehmen.

Etwa sechs Wochen, bevor wir in der MyBib-WDB den Druck von Auftragszetteln anstoßen, werden die entsprechenden Bände für die Benutzung gesperrt und ausgeliehene Medien dienstlich vorgemerkt. Nur am Rande sei festgestellt, dass wir in wissenschaftlich begründeten Fällen, in der Zeit bis zur Übergabe an den Dienstleister, selbstverständlich eine Ausleihe in die Lesesäle der Bayerischen Staatsbibliothek ermöglichen. Die Digitalisierung selbst, unsere Überprüfung auf Vollständigkeit und das Rückstellen der Bücher nehmen einige Wochen in Anspruch. Bezogen auf den gesamten Digitalisierungsworkflow können die Bände insgesamt etwa drei Monate nicht benutzt werden. Aber nur etwa die Hälfte dieser Zeit ist das Buch absolut unzugänglich. Das ist nur unwesentlich länger, als dies bei jeder nutzungsbedingten Entleihung von vier Wochen Dauer auch der Fall ist. Wir möchten ausdrücklich betonen, dass diese Zeitspanne im Vergleich mit anderen Digitalisierungs- und Bestandserhaltungsvorhaben hervorragend und das Qualitätsmerkmal einer durchdachten Ablaufplanung ist.

Etwa zwei Wochen vor der Übergabe an den Dienstleister werden die Auftragszettel ausgedruckt und anhand dieser Aufträge die Bücher ausgehoben. Das ist ein Vorgang von erheblicher Komplexität, der mit der Vokabel „Ausheben“ völlig unzuläng-

lich umschrieben ist. Die Bücher müssen zunächst hinsichtlich ihrer Eignung für das Scannen (Zustand, Größe, Erscheinungsjahr usw.) beurteilt werden. Es sind drei Fälle zu unterscheiden:

1. Die Auftragszettel werden bei unproblematischen Fällen direkt in das Buch eingelegt, das dann direkt der Digitalisierung zugeführt wird. Die Auftragsverbuchung in der MyBib-WDB wird durch Einscannen der Strichcode-digID mit tragbaren Handscannern mit integrierter, mobiler Speichereinheit vorbereitet. Das Einlesen der auf diese Weise gesammelten digID's von den Speichereinheiten in die Datenbank von MyBib-WDB erfolgt später zentral und bewirkt, dass die einzelnen Aufträge in einen Status überführt werden, der es erlaubt, dass sie in eine Charge zur Übergabe an den Scandienstleister aufgenommen werden können.

2. Bei offensichtlich ungeeigneten oder fehlenden Exemplaren – es gibt natürlich auch falsch eingestellte Bücher – wird der Grund auf dem Auftragszettel festgehalten. Die gekennzeichneten Auftragszettel werden per Einzugsscanner in die MyBib-WDB eingelesen, um die Gründe, die dazu geführt haben, dass das Buch nicht gescannt wurde, dort zu hinterlegen. Sollten sich Rahmenbedingungen ändern oder Folgeprojekte möglich sein, können diese Titel leicht erneut selektiert werden.

3. Fälle, bei denen sich die Ausheber hinsichtlich der Eignung unsicher sind oder für die Unklarheiten im Bezug auf die Metadaten bestehen, bleiben entweder im Regal (wobei auf dem Auftragszettel ein Nachbearbeitungskennzeichen gesetzt wird), oder sie werden dem Leitungsteam der lokalen Projektgruppe zur Begutachtung vorgelegt. Die Projektgruppe muss entscheiden, ob die ihr zugeleiteten Bücher schnell und somit von ihr selbst bearbeitet werden können oder für aufwendige Metadatenkorrekturen an das Sachgebiet Qualitätssicherung der Medienbearbeitung übergeben werden müssen. Ein Teil der Bücher wird auch der Abteilung für Bestandserhaltung oder dem Institut für Buch- und Handschriftenrestaurierung (IBR) der Bayerischen Staatsbibliothek zugeführt werden müssen. Alle diese Entscheidungen müssen effizient und Ziel führend, aber auch mit hoher Verantwortlichkeit für das uns anvertraute Kulturgut, getroffen werden. Die Workflowdatenbank der Bayerischen Staatsbibliothek leistet die hierzu notwendige Unterstützung und Dokumentation, damit der Weg jedes Buches jederzeit nachvollziehbar ist.

Diese Form der regalweisen systematischen Durcharbeitung des Magazins bietet zusätzlich den enormen Vorteil, dass auf diese Weise auch alle Bücher aufgefunden werden, für die aufgrund von fehlerhaften oder

vollständig fehlenden Metadaten kein Auftrag ausgedruckt wurde.

Obwohl ursprünglich nicht unbedingt intendiert, wird uns dieses Vorgehen zwei weitere langjährige Desiderate wie von selbst liefern. Der gesamte urheberrechtsfreie und wertvolle Altbestand der Bibliothek wird einer gründlichen Revision unterzogen und zusätzlich wird eine Schadenserhebung zumindest dahingehend geleistet, dass schwerwiegende Schädigungen, die einer Digitalisierung entgegenstehen, erfasst und behoben werden. Der Unterschied zu bisherigen Schadenserhebungen besteht darin, dass nicht nur ein kleiner Teil des Bestandes untersucht wird und dass anschließend hochgerechnet wird, sondern dass wirklich Band für Band des gesamten Altbestandes in Autopsie untersucht und dies in der MyBib-WDB für jeden Einzelfall entsprechend dokumentiert wird.

Die für Google bereitgestellten Bücher speisen sich also letztendlich aus drei Quellen, dem Strom der Bücher ohne Korrekturbedarf, die direkt zur Digitalisierung gehen, dem Fluss der Bücher, die von der lokalen Projektgruppe vorbereitet werden können, und dem voraussichtlich kleinen Rest der Werke, die von der Medienbearbeitung, einem Buchbinder oder dem Institut für Buch- und Handschriftenrestaurierung zurückkommen. Über die MyBib-WDB wird dann für eine genau definierte Menge von Büchern eine sogenannte Charge gebildet und zum Scannen bereitgestellt.

Was geschieht aber nun mit all den Büchern, deren Metadaten wir nicht schnell genug korrigieren konnten, um sie direkt dem Digitalisierungsprozess zuführen zu können? Wir werden in einigen Jahren, aber in jedem Fall innerhalb der definierten Projektlaufzeit, die gesamten Altbestandsmagazine der Bayerischen Staatsbibliothek ein zweites Mal dem oben geschilderten Prozess unterziehen, also einen zweiten Durchgang durch die Magazine beginnen. Natürlich bleiben bei diesem Durchgang alle Bücher unberücksichtigt, die schon digitalisiert wurden und da dies der weitaus überwiegende Teil sein wird, wird dieser Durchgang sich über eine sehr viel kürzere Zeit hinziehen. Nun wird auch klar, warum wir beim ersten Durchgang relativ großzügig mit den Buchwünschen unserer Nutzer umgehen können. Wenn wir ein Buch ausleihen statt es zu digitalisieren, dann wird es beim zweiten Durchgang oder womöglich einem dritten Durchgang ganz sicher erfasst. Auch Bücher, deren Instandsetzung längere Zeit in Anspruch genommen hat, erhalten auf diese Weise eine zweite Chance auf eine Digitalisierung.

Die Abläufe bei Google werden, wie weiter oben schon ausgeführt, hier nicht beschrie-

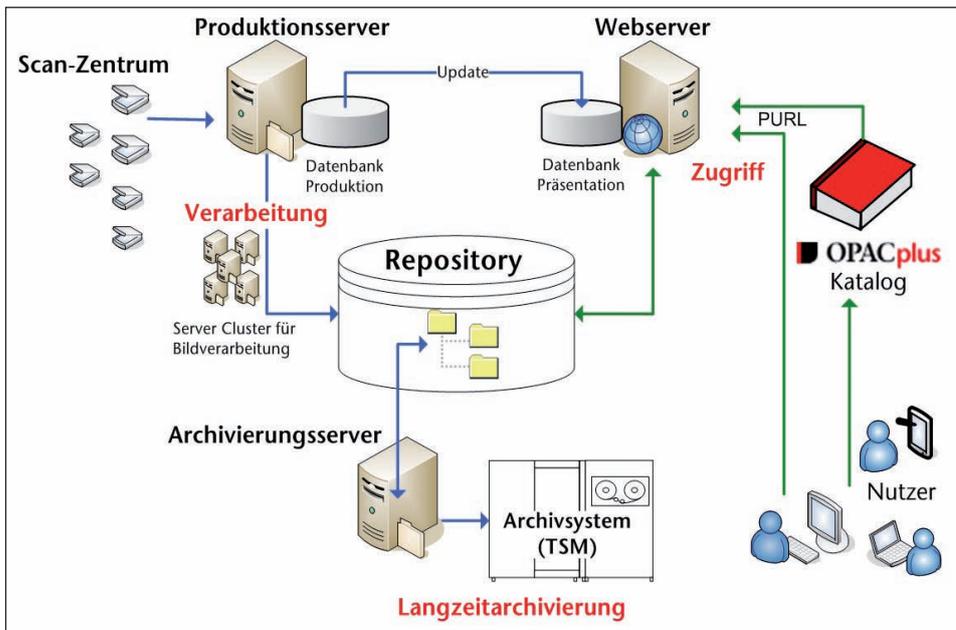


Abb. 3: Die Systemarchitektur des Bereitstellungs- und Archivierungssystems (PURL = Persistent URL; TSM = Tivoli Storage Manager)

ben. Auf jeden Fall werden die Bücher gescannt, Scanqualität und Vollständigkeit werden überprüft und soweit möglich Strukturdaten und Indexdaten für die Volltextrecherche erzeugt. Danach werden die Bücher in kompletten Chargen an die Bayerische Staatsbibliothek zurückgeliefert. Diese chargenweise Rückgabe wie auch die MyBib-WDB erleichtern die numerische Kontrolle, ob auch wirklich alle Bücher wieder zurückgeliefert wurden. Nach der Rücknahme werden die Bücher von erfahrenen Magazinräften, denen auch einzelne Lücken innerhalb der Signaturfolge auffallen würden, in die Magazinregale zurückgestellt. Eine allerletzte Kontrolle auf Vollständigkeit erfolgt über das Einscannen aller beim Zurückstellen gezogenen Auftragszettel per Einzugsscanner in die MyBib-WDB.

### Die Bereitstellung der Digitalisate

Die Bereitstellung und Langzeitarchivierung der Library Digital Copy wird vom Münchener Digitalisierungszentrum (MDZ) in Zusammenarbeit mit dem Leibniz-Rechenzentrum (LRZ) organisiert. Hierfür wurde in den Monaten vor Projektstart eine leistungsfähige technische Infrastruktur aufgebaut und getestet. Ausgangspunkt war die vorhandene und praxiserprobte Zentrale Erfassungs- und Nachweisdatenbank der Bayerischen Staatsbibliothek, welche seit 2004 im MDZ im Einsatz ist. Diese musste für die Zwecke der Massendigitalisierung angepasst und um einige Funktionen erweitert werden. Im Zentrum steht ein skalierbares Repository zur Aufnahme sämtlicher Daten aus dem Projekt. Ein zentraler Pro-

duktionsserver holt die Datenpakete vom Scandienstleister ab und reicht sie zur Konvertierung an ein Server-Cluster weiter. Dort werden die Daten nach dem Standard-Workflow des Münchener Digitalisierungszentrums verarbeitet. Es werden Präsentationsderivate und Strukturdaten erzeugt und je Buch eine durchblätterbare Version für die Nutzung gebildet. Über einen zweiten Produktionsserver wird danach die Archivierung der Dateien im TSM-System (Tivoli Storage Manager) des LRZ angestoßen. Das Archivsystem ist direkt an das Produktionssystem angeschlossen. Am Ende des gesamten Prozesses steht die Bereitstellung des verarbeiteten Digitalisates auf dem Webserver der Bayerischen Staatsbibliothek. Danach erfolgt eine automatische Rückmeldung an die WDB und von dort an das Katalogsystem. Erst jetzt wird das Digitalisat durch Eintrag eines Links in den Katalog auch für den Nutzer sichtbar und benutzbar. Zu guter Letzt sorgt eine Volltextindexierung für eine vollständige Durchsuchbarkeit des Buches.

Das System ist so ausgelegt, dass sämtliche digitalisierten Bücher zeitnah nach dem Scannen verarbeitet und bereitgestellt werden können und kein Rückstau entsteht. Die gesamte Verarbeitungskette, von der Datenabholung über die Imagekonversion bis hin zu Bereitstellung und Archivierung wäre ohne einen hohen Grad an Automatisierung in allen Produktionsschritten nicht realisierbar. Aufgrund der ungewöhnlich großen Datenmenge – an normalen Produktionstagen müssen viele hundert Bücher am Tag verarbeitet und bereitgestellt werden – wäre Handarbeit nicht nur eine Bremse sondern auch eine ständige Fehlerquelle.

Priorität hat die zeitnahe Bereitstellung für unsere Nutzer und die sichere Archivierung aller Daten.

Nach umfangreichen Vorbereitungsarbeiten ist die BSB für die Herausforderungen dieses Großprojektes gut gerüstet. Und so wird schon in wenigen Jahren der urheberrechtsfreie historische Bestand der Bayerischen Staatsbibliothek weitgehend online verfügbar sein. Rund jedes zehnte Buch der Bayerischen Staatsbibliothek kann dann nach wenigen Mausklicks auf den Bildschirmen unserer Nutzer erscheinen rund um die Uhr, weltweit und ohne Wartezeit. Zwar werden es nur virtuelle Kopien sein. Aber es stehen auch alle Bücher nach dem Ende dieses umfangreichen Digitalisierungsprojektes wieder vollzählig und sicher in den Magazinen der BSB und weiterhin jedem Bücherfreund offen, der auf das Rascheln von Papier nicht verzichten mag.

### ■ DIE AUTOREN

**MARTIN BAUMGARTNER**  
Kooperatives Datenmanagement  
Martin.Baumgartner@bsb-muenchen.de

**MICHAEL BEER**  
Ltg. d. Qualitätssicherung Erschließung  
Michael.Beer@bsb-muenchen.de

**DR. BERTHOLD GILLITZER**  
Stellv. Leitung Benutzungsdienste  
Berthold.Gillitzer@bsb-muenchen.de

**DR. WILHELM HILPERT**  
Leitung Benutzungsdienste  
Wilhelm.Hilpert@bsb-muenchen.de

**ROSMARIE LEICHTL**  
Leitung Workflowteam Google  
Rosmarie.Leichtl@bsb-muenchen.de

**GABRIELE MESSMER**  
Koordination Digitale Bibliothek  
Gabriele.Messmer@bsb-muenchen.de

**KARSTEN TRZCIONKA**  
Leitung Dokumentverwaltung  
Karsten.Trzcionka@bsb-muenchen.de

**DR. THOMAS WOLF-KLOSTERMANN**  
Digitale Bibliothek  
Thomas.Wolf-Klostermann@bsb-muenchen.de

Alle Autoren:  
Bayerische Staatsbibliothek  
80328 München



ImageWare Components GmbH  
Am Hofgarten 20  
53113 Bonn  
Germany  
www.imageware.de