Halbzeit – ohne Pause

Stand und Erkenntnisse der industriellen Massendigitalisierung an der Bayerischen Staatsbibliothek

Martin Baumgartner und Wilhelm Hilpert

Die Bayerische Staatsbibliothek ist ihrem erklärten Ziel, den urheberrechtsfreien Altbestand der Bibliothek – etwa eine Million Bände – im World Wide Web für Wissenschaft und Forschung bereitzustellen, in den letzten Jahren insbesondere durch die Zusammenarbeit mit dem Internetkonzern Google ein sehr großes Stück näher gekommen. Mehr als fünfhunderttausend digitalisierte Titel bzw. 150 Millionen digitalisierte Seiten sind mittlerweile direkt im Netz verfügbar und über den OPAC der Bayerischen Staatsbibliothek recherchierbar. Auch wenn über das Digitalisierungsprojekt in Zusammenarbeit mit dem Internetkonzern Google insgesamt 90 % dieser Digitalisate erzeugt wurden, sollen andere umfangreiche Massendigitalisierungsprojekte der Bayerischen Staatsbibliothek, wie das Inkunabelprojekt, das Blockbuchprojekt oder das VD 16-Projekt keineswegs verschwiegen werden. Allein mit diesen drei genannten und zahlreichen anderen Digitalisierungsprojekten ist die Bayerische Staatsbibliothek die erste Adresse in Sachen Digitalisierung im deutschen Bibliothekswesen. Wer heute einen Titel im Altbestand der Bayerischen Staatsbibliothek über eine OPAC-Recherche findet, kann mit einer Wahrscheinlichkeit von 50 % davon ausgehen, dass er den Volltext als Image – und damit in der Regel die benötigte Information – unverzüglich von nahezu jedem Ort weltweit einsehen kann. Mit diesem Service steht die Bayerische Staatsbibliothek im deutschsprachigen Raum einmalig da. Nicht jeden hat dies glücklich gemacht, jedoch unsere Nutzer, Wissenschaftler und Forscher durchaus. Da der Altbestand der Bayerischen Staatsbibliothek den abendländischen Kulturkosmos bis weit in das 19. Jahrhundert hinein mit all seinen vielen Sprachen in umfassender Weise abbildet, profitiert neben der deutschen Forschergemeinde auch die europäische bzw. internationale Forschung in hohem Maße.

Der Claim der Bayerischen Staatsbibliothek – „Information in erster Linie" – erweist sich durch dieses Angebot einmal mehr als eine Aussage, die in eindrucksvoller Weise mit Inhalt und Bedeutung gefüllt ist.

Bevor auf verschiedene Aspekte des Projektes mit Google wie die Auswirkungen der verringerten Verfügbarkeit der Originale während der Projektlaufzeit oder die konservatorischen Auswirkungen näher eingegangen wird, muss zunächst der Begriff der *industriellen* Massendigitalisierung von dem der gängigen Massendigitalisierung abgegrenzt und präzisiert werden.

Industrielle Massendigitalisierung

Die industrielle Massendigitalisierung zielt primär darauf ab, den Informationsgehalt, der im Text eines gedruckten Werkes steckt, vollständig und korrekt in digitaler Form wiederzugeben. Es ist nicht Ziel, künstleri-

sche Details in allen Einzelheiten abzubilden oder höchsten ästhetischen Ansprüchen zu genügen.

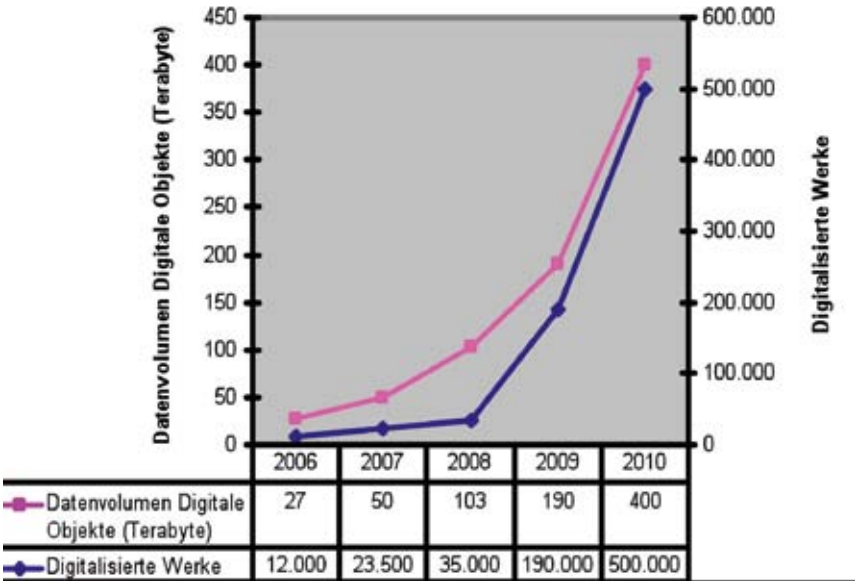
Die industrielle Massendigitalisierung kann etwas konkreter als ein Prozess verstanden werden, der sich von der sonstigen Massendigitalisierung in sechs wesentlichen Punkten unterscheidet:

1. Es findet keinerlei inhaltliche Auswahl des Digitalisierungsgutes statt. Es wird alles digitalisiert, was aus konservatorischer Sicht dafür geeignet ist.
2. Der Prozess darf während der Betriebszeiten nur durch höhere Gewalt zum Stillstand kommen, da sich die Standzeitkosten für Ausrüstung und Personal sehr schnell zu erheblichen Beträgen aufsummieren würden. Software- wie Hardwarekomponenten müssen daher hohen Ansprüchen bezüglich ihrer Ausfallsicherheit genügen und die Reaktionszeiten bei Ausfällen sind kurz zu halten.
3. Die Umsatzzahlen in der industriellen Massendigitalisierung sind gegenüber dem, was gemeinhin als Massendigitalisierungsprojekt bezeichnet wird, mindestens um einen Faktor 10, meist jedoch bis zum Faktor 100 höher.
4. Der Scanvorgang und die Nachbearbeitung sind zwei Prozessschritte, die räumlich und zeitlich klar voneinander geschieden ablaufen.
5. Alle Prozesse der Nachbearbeitung und Bereitstellung sind in einem sehr hohen Grad automatisiert. Veränderungen in Geschäftsgängen sind daher nur nach sorgfältiger Planung und langfristiger Vorarbeit möglich.
6. Eine unmittelbare und jede einzelne Seite betreffende manuelle Qualitätskontrolle der Digitalisate gibt es nicht. Es gibt jedoch ausgefeilte automatisierte Verfahren der Qualitätskontrolle und stichprobenartige Kontrollen durch Mitarbeiter aller beteiligten Einrichtungen.

In einem wichtigen Punkt unterscheidet sich die industrielle Massendigitalisierung jedoch nicht von allen anderen Digitalisierungsvorhaben und -projekten. Der Schutz des originären, materiellen Kulturgutes Buch hat oberste Priorität. Alle Prozessschritte des Workflows werden daran ausgerichtet und gemessen, bis hin zu der Entscheidung, ein Werk nicht zu digitalisieren.

Verfügbarkeit der Bestände während des Digitalisierungsprozesses

Natürlich stellt sich die Frage, ob dieser Service der Bayerischen Staatsbibliothek auch seine dunklen Seiten hat und etwa dazu führt, dass die Bestände während dieses industriellen Digitalisierungsworkflows für lange Zeit der regulären Nutzung entzogen werden?



	2006	2007	2008	2009	2010
Datenvolumen Digitale Objekte (Terabyte)	27	50	103	190	400
Digitalisierte Werke	12.000	23.500	35.000	190.000	500.000

Abbildung 1: Man sieht, dass seit dem Jahr 2008 mit Beginn der Digitalisierung durch Google die Zahl der Digitalisate der Bayerischen Staatsbibliothek sprunghaft angestiegen ist.

Der Altbestand der Bayerischen Staatsbibliothek ist seit dem 19. Jahrhundert grobsystematisch in etwa 250 Fächern aufgestellt. Die Bayerische Staatsbibliothek geht bei der industriellen Massendigitalisierung so vor, dass jeweils große Teile eines Faches (Signatur) für den Digitalisierungsworkflow als Ganzes abgearbeitet werden. Dazu werden zu Beginn des Workflows die aus dem Katalog gewonnenen Metadaten überprüft und soweit notwendig korrigiert. Dies ist eine Aktivität, die dem Service der gesamten Bayerischen Staatsbibliothek zugute kommt und die auch ohne Massendigitalisierung hätte gemacht werden müssen, wenn auch nicht in dieser extrem kurzen Zeit. Anschließend werden alle einschlägigen Werke der vorbereiteten Signaturengruppe in die Workflowdatenbank (WDB) geladen. Im Ausleihclient werden sie einem bestimmten Medientyp zugeordnet, der in der Exemplaranzeige des OPAC dazu führt, dass die Nutzer folgende Mitteilung zum Status eines Werkes erhalten: „Temp. gesperrt, Status an Info erfragbar". Je nach Bestandsmenge einer Signaturengruppe kann diese Sperrung drei bis neun Monate andauern, wobei die durchschnittliche Sperrungszeit bei fünf Monaten liegt. In allen dringenden und wissenschaftlich begründeten Fällen sind wir jedoch mit Hilfe unserer Workflowdatenbank in der Lage, jeden gewünschten Titel zu lokalisieren und aus dem Workflow herauszunehmen, um ihn einem Nutzer zur Verfügung zu stellen. Nur wenn sich der Titel in einer abgeschlossenen Charge befindet und an Google übergeben wurde, ist uns dieses Vorgehen nicht mehr möglich. In solch einem Fall ist die Wartezeit für den Nutzer aber schon deutlich kürzer als bei einer vierwöchigen Ausleihe durch einen Dritten.

Bis heute konnten wir keinerlei Anstieg von Beschwerden wegen der vorübergehenden

hend erschwerten Verfügbarkeit registrieren. Wir führen dies darauf zurück, dass wir unsere Nutzer hinreichend informieren und den Mitarbeitern an den Informationsstellen mit der WDB ein Instrument zur Verfügung stellen, das es ermöglicht, vertiefte Informationen zum Status eines jeden einzelnen Werkes abzurufen und diese an die Nutzer weiterzugeben.

Konservatorische Aspekte und Benutzungsaspekte der industriellen Massendigitalisierung

Digitalisierungsvorhaben nehmen in Bibliotheken in Deutschland wie auch weltweit einen sehr hohen Stellenwert ein. Durch die Digitalisierung ihrer Bestände gelingt es den Bibliotheken, ihre wichtigsten Ziele, die bisher unvereinbar erschienen, in idealer Weise miteinander in Einklang zu bringen. Die Steigerung der Verfügbarkeit und Benutzbarkeit von Dokumenten und Informationen ist plötzlich kein Widerspruch mehr zu deren Bewahrung und Erhaltung für nachfolgende Generationen. Ein digitalisiertes und frei im World Wide Web verfügbares Werk kann zu nahezu jeder Zeit und fast von jedem Ort dieser Welt aus eingesehen werden und dies beliebig oft, ohne dass ihm dadurch Schaden zugefügt würde oder es sich weiter abnutzt. Gleichzeitig ist aber sein Informationsgehalt genauso gut geschützt wie z. B. durch die seit den 1970er Jahren durchgeführten Sicherheitsverfilmungen. Dies ist aber noch längst nicht alles, was Digitalisate zu bieten haben. Über Suchmaschinen und die Texterkennung wird ihre Nutzbarkeit in Bereiche gehoben, die für gedruckte Materialien und Sicherheitsverfilmungen niemals vorstellbar waren.

Die Bayerische Staatsbibliothek besitzt selbst eine vollwertige Kopie eines jeden Digitalisates (Library Copy) und speichert diese auf ihren eigenen Geräten, die am Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften gehostet werden. Die Volltextsuche und die gesicherte Langzeitarchivierung für diese Daten zu realisieren, sind wichtige strategische Ziele der Bayerischen Staatsbibliothek für die nächsten Jahre.

Aber auch die bestandserhaltenden und konservatorischen Auswirkungen einer Digitalisierung gehen weit über die reine Informationssicherung, wie sie oben beschrieben wurde, hinaus. Eine spannende Frage ist in diesem Zusammenhang natürlich, wie sich die Nutzerinteressen und die Ausleihen bei einem großen digitalisierten Bestand verändern. Bisher konnte darüber nur spekuliert werden, und selbstverständlich war auch innerhalb der Bayerischen Staatsbibliothek im Vorfeld der industriellen Massendigitalisierung über die möglichen Auswirkungen der Digitalisierung auf die Nutzung des physischen Altbestandes vielfach diskutiert worden. Dabei standen sich zwei ziemlich konträre Ansichten gegenüber. Die einen glaubten, dass die Nutzung selbstverständlich zurückgehen werde, da ja nun die Inhalte und somit die gesamte Information sofort online verfügbar sei. Eine andere Gruppe dachte, dass durch die Digitalisate das Interesse am Altbestand insgesamt eher ansteigen werde und die Nutzung somit zunehmen oder zumindest auf gleichem Niveau bleiben werde.

Die heute an der Bayerischen Staatsbibliothek vorliegenden Zahlen zeigen, dass die Nutzung der Originale tatsächlich in ganz erheblicher Weise zurückgeht und dies sogar entgegen dem Trend ständig steigender Ausleihzahlen.

In der folgenden Tabelle sind zunächst die Ausleihen in die Lesesäle der Bayerischen Staatsbibliothek für das Bestandssegment mit den Erscheinungsjahren 1701–1840 und als Bezugssystem die insgesamt ansteigenden Ausleihzahlen für die Jahre 2008 bis 2010 wiedergegeben:

1 In der Tabelle sind nur die nichtdienstlichen Ausleihen in die Lesesäle wiedergegeben, die über das Bibliothekssystem abgewickelt wurden. Insbesondere in den Lesesaal für Handschriften und alte Drucke werden auch Bücher über konventionelle Leihscheine ausgeliehen.

Jahr	Ausleihen des Altbestandes mit Erscheinungsjahren 1701–1840[1]	Änderung in Prozent mit 2008 als Bezugsjahr	Ausleihen des Gesamtbestandes	Änderung in Prozent mit 2008 als Bezugsjahr
2008	11.343	100	1.748.000	100
2009	9.532	84,0	1.911.000	109,3
2010	6.763	59,6	2.022.000	115,7

Obwohl zu Beginn des Jahres 2010 nur 15 % des Altbestandes in digitalisierter Form angeboten wurde und erst zum Ende des Jahres knapp 50 % erreicht waren, hat sich dies gravierend auf die Nutzung der Bestände ausgewirkt. Insgesamt ging die Zahl der Ausleihen im Vergleich mit dem Referenzjahr 2008, dem letzten Jahr, in dem noch keine Digitalisate aus der industriellen Massendigitalisierung angeboten wurden, um 40 % zurück. Wenn wir diesen Wert vorsichtig extrapolieren, dann sind nach Abschluss des Projektes in diesem Bestandssegment höchstens noch ein Zehntel der ursprünglichen Ausleihen zu erwarten. Selbst wenn sich die Nutzung statt um 90 % nur um 70 % oder 80 % verringern würde, wäre ein unschätzbarer Beitrag zur Erhaltung dieser Bestände geleistet und das, ohne dass jemand auf benötigte Informationen verzichten muss. Der Rückgang der Benutzungszahlen fällt womöglich auch daher so überraschend deutlich aus, weil uns Gespräche mit Wissenschaftlern gezeigt haben, dass sie in vielen Fällen bereit sind, einige Monate auf ein kostenfreies Digitalisat zu warten, wenn sie sich dafür eine Reise oder einen teuren Digitalisierungsauftrag ersparen können. Auch hierbei sind die Auskünfte aus der WDB von erheblicher Bedeutung.
Erwartungsgemäß fällt für das Bestandssegment mit den Erscheinungsjahren 1501 bis 1700, wie in der folgenden Tabelle zu sehen ist, der Rückgang der Nutzung bei weitem nicht so deutlich aus. Dies verwundert nicht, denn je weiter man sich zeitlich den Anfängen der Buchherstellung nähert, um so mehr Information steckt für die Forschung nicht nur im Inhalt eines Buches, sondern auch in der Art, wie es erstellt wurde, und in den Materialien, aus denen es hergestellt wurde.
Dennoch ist auch in diesem Zeitsegment ein klarer Rückgang der physischen Nutzung und damit der Beanspruchung des Bestandes um 50 % bis 60 % nach Abschluss des Projektes zu erwarten.
Durch diese Zahlen wird sehr deutlich, dass die Digitalisierung nicht nur einen Beitrag zur Informationssicherung und Verbreitung des wichtigsten Kulturgutes Buch leistet, sondern auch einen bedeutenden Beitrag zur Erhaltung und Bewahrung dieses Gutes.
Dem Altbestand der Bayerischen Staatsbibliothek wird aber durch die industrielle Massendigitalisierung nicht nur indirekt, sondern auch ganz direkt Gutes getan. Als wir die Zusammenarbeit mit Google geplant und die Abläufe des Projektes festgelegt haben, war uns völlig klar, dass uns auch Herausforderungen und Hürden erwarten, die wir nicht bis ins letzte Detail planen und vorhersehen können. Galt die primäre Sorge zunächst der Vollständigkeit und Korrektheit der Metadaten sowie der Entwicklung der Workflows inklusive der Steuerungselemente, so zeigte sich sehr schnell, dass der Erhaltungszustand einiger Werke ebenfalls eine erhebliche Herausforderung darstellt. Mehr als 1 % aller Bücher – 200 bis 250 Bände monatlich – werden im Vorfeld der Digitalisierung an das Institut für Buch- und Handschriftenrestaurierung (IBR) gegeben, dort untersucht und in einen Zustand gebracht, der eine schonende Digitalisierung erlaubt und zugleich sicherstellt, dass die Digitalisierung ohne eine Verschlechterung des Erhaltungszustandes möglich ist.
Alle Werke werden zudem im Zuge des Projektes entstaubt, einer grundlegenden Revision unterzogen und neu aufgestellt. Zahlreiche Werke werden in säurefreie Kartons verpackt. Durch diese Maßnahmen wird sich der Altbestand der Bayerischen Staatsbibliothek nach Abschluss des industriellen Massendigitalisierungsprojektes in einem weitaus besseren Erhaltungszustand befinden als zu Beginn der Maßnahme. Bei allen Werken, die aufgrund ihres physischen Zustandes nicht digitalisiert werden, wird in der Workflowdatenbank dokumentiert, welche Schäden vorliegen, um diese Informationen für zukünftige Digitalisierungs- oder Bestandserhaltungsprojekte abrufbar zu haben.

Jahr	Ausleihen des Altbestandes mit Erscheinungsjahren 1501–1700	Änderung in Prozent mit 2008 als Bezugsjahr	Ausleihen des Gesamtbestandes	Änderung in Prozent mit 2008 als Bezugsjahr
2008	5.830	100	1.748.000	100
2009	5.159	88,5	1.911.000	109,3
2010	4.453	76,4	2.022.000	115,7

Qualität der Digitalisate

Die Intention, weitestgehende Fehlerfreiheit zu erreichen, ist bei beiden Projektpartnern sehr hoch, und die Anstrengungen, die in dieser Richtung unternommen werden, sind aufwendig. Dennoch gibt es Grenzen. Wie oben dargelegt, ist es eines der Kennzeichen der industriellen Massendigitalisierung, dass die Qualität der Digitalisate nicht Seite für Seite überprüft wird. Zwischen 200.000 bis 400.000 Seiten täglich einzeln zu begutachten, katapultiert die Kosten der Digitalisierung in Bereiche, die sogar einen Internetkonzern dazu veranlassen, über Kosten-Nutzen-Relationen nachzudenken. Man ist als Projektbeteiligter daher gezwungen, sich mit dem Begriff der „Fehlerrate" anzufreunden und das Augenmerk primär nicht auf den Einzelfehler zu richten, sondern auf systematische Fehler, die es durch laufende Optimierung und Vereinheitlichung der Prozesse zu vermeiden gilt. Fehler werden dabei auf verschiedenen Ebenen gesucht und ausgemerzt. Es beginnt bei Fehlern an einzelnen Scanstationen bis hin zu Fehlerhäufungen in den Scanzentren und letztendlich Fehlern im Gesamtsystem, die überwiegend bei der Nachbearbeitung auftreten. Google hatte bis zum Projektstart mit der Bayerischen Staatsbibliothek aus Vorprojekten nur begrenzte Erfahrungen hinsichtlich der Probleme, die speziell bei der Digitalisierung von alten Büchern auftreten können. Im ersten Jahr der Zusammenarbeit von Google und der Bayerischen Staatsbibliothek gelang es dann gemeinsam, viele Fehler zu kategorisieren und bis zu ihren Ursachen hin zurückzuverfolgen, um sie letztendlich zu eliminieren.
Aber auch der typische Einzelfehler, die vergessene oder schlecht gescannte Seite, wurde keineswegs vernachlässigt. Ein wichtiger Schritt zur Qualitätsverbesserung ist die seit Ende 2010 für die Projektpartner bestehende Möglichkeit, einzelne Seiten eines bereits öffentlich gemachten Digitalisates nachzuscannen und durch Google in das Digitalisat einfügen zu lassen. Dadurch lassen sich fehlerhafte oder fehlende Seiten nun ergänzen. Natürlich könnten wir jederzeit auch

Änderungen an unserer Library Copy vornehmen, nur würden diese Änderungen, bei Erhalt einer verbesserten Version des Digitalisates durch Google, wieder überschrieben. Wenn wir auf die verbesserten Versionen verzichten würden, wären wir von allen Fortschritten, die Google in der Nachbereitung laufend erzielt, abgekoppelt.
Da auch die Bayerische Staatsbibliothek es nicht leisten kann, 150 Millionen bis 200 Millionen Seiten auf Fehler durchsuchen zu lassen, überlassen wir das Auffinden von Einzelfehlern gezwungenermaßen unseren Nutzern. Dieses an der Nachfrage orientierte Vorgehen ist aus unterschiedlichen Katalogkonversionsprojekten an der Bayerischen Staatsbibliothek bestens erprobt und hat sich über Jahre hinweg bewährt.

Präsentation in Katalogen und Portalen

Es gibt eine Vielzahl von Möglichkeiten, die Digitalisate der Bayerischen Staatsbibliothek zu nutzen. Neben der Einbindung in den Online-Katalog der Bayerischen Staatsbibliothek werden die Digitalisate den Nutzern auch in regionalen Katalogen wie dem B3-Kat (Bayern, Berlin, Brandenburg) sowie weltweit in Google Book Search und WorldCat angeboten. Auch bringt die Bayerische Staatsbibliothek ihre Digitalisate in eine große Zahl von öffentlich-rechtlich geförderten Portalen wie Europeana, dem Zentralen Verzeichnis Digitalisierter Drucke (ZVDD), dem Consortium of European Research Libraries (CERL) und künftig auch der Deutschen Digitalen Bibliothek (DDB) ein. In der Europeana stammen zurzeit 80 % aller Bücher aus Deutschland von der Bayerischen Staatsbibliothek.

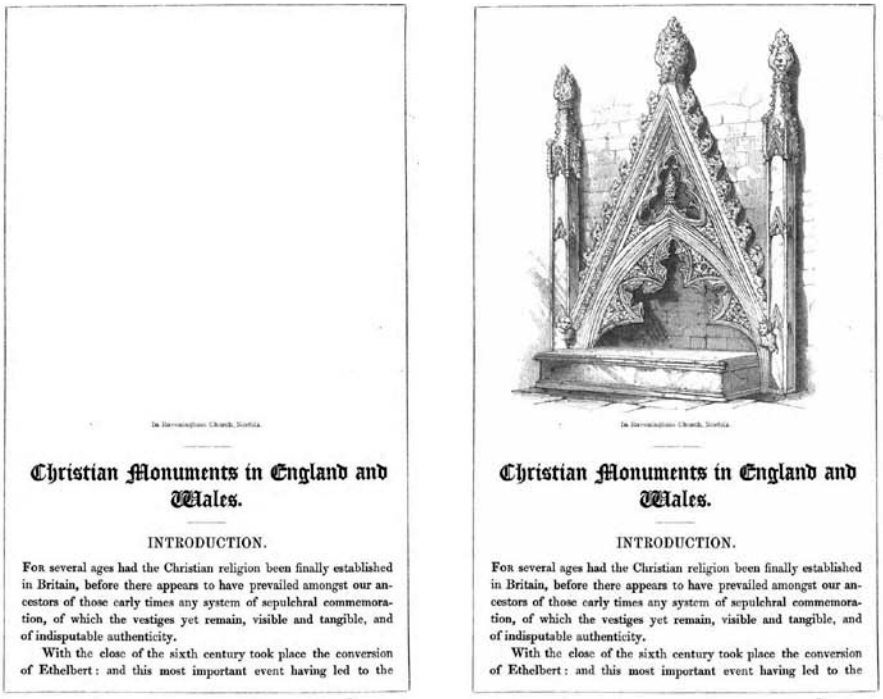


Abbildung 2: Inzwischen behobener Fehler: Statt eines Bildes ist ein weißes Rechteck zu sehen (Vorher-Nachher-Vergleich)

Ausblick

Die Zusammenarbeit der Bayerischen Staatsbibliothek mit dem Internetgiganten Google hat sich für Wissenschaft und Forschung als äußerst fruchtbar erwiesen. Der Weg hin zu einer Million digitalisierter Werke ist zur Hälfte zurückgelegt und das Ziel kommt täglich einen großen Schritt näher. Die Zusammenarbeit ist mittlerweile so erprobt und vertrauensvoll, dass die Bayerische Staatsbibliothek auch Teile des Reservebestandes und die urheberrechtsfreien Lesesaalbestände an Google geben wird. Weiterhin arbeiten wir an einem Geschäftsgang, um Werke mit Erscheinungsjahr bis 1941, von denen die Urheberrechtsfreiheit durch die Todesdaten des oder der Autoren belegt ist, mit einzubeziehen. Diese Titel können nur selektiv aus den Magazinen ausgehoben werden, und es muss durch Einzelfallprüfungen sichergestellt sein, dass wir auch nicht in einem einzigen Fall das Urheberrecht verletzen. Ein Projekt, das die gesamte Bayerische Staatsbibliothek vor eine neue, gewaltige Herausforderung stellt und das dafür sorgen wird, dass wir die angestrebten eine Million digitalisierter Titel ganz erheblich überschreiten werden.

AUTOREN

Martin Baumgartner
Bayerische Staatsbibliothek
Abt. Bestandsaufbau und Erschließung
Kooperatives Datenmanagement
Ludwigstr. 16
80539 München
martin.baumgartner@bsb-muenchen.de

Dr. Wilhelm Hilpert
Bayerische Staatsbibliothek
Leiter der Abteilung Benutzungsdienste
Ludwigstr. 16
80539 München
hilpert@bsb-muenchen.de