Data Curation oder (Retro-)Digitalisierung ist mehr als die Produktion von Daten

Klaus Kempf, Bayerische Staatsbibliothek München[1]

Zusammenfassung:
Nahezu jede Bibliothek, die über Altbestand verfügt, beschäftigt sich heute mit dem Thema Digitalisierung. Dabei wird oftmals übersehen, dass die dauerhafte und möglichst nutzerfreundliche Bereitstellung (und letztlich die Archivierung) der erzeugten Daten organisatorische Anforderungen ganz eigener Art an die bestandshaltende Institution stellt. Anders als bei analogen Materialien ist es mit dem sachgerechten Vorhalten und der Ausgabe des Sammelguts in buchklimatisch möglichst einwandfreien Räumlichkeiten nicht getan. Elektronische Dokumente, egal welcher Herkunft, welchen Umfangs, Typs und Inhalts, verlangen – weitgehend unabhängig von ihrer tatsächlichen Inanspruchnahme – ein aktives und extrem aufwändiges Datenmanagement sowie eine intensive Datenpflege. Dazu kommen der Betrieb und das regelmäßige Update der für das Datenangebot notwendigen Hard- und Software. Dieses Tun wird in der angelsächsischen Bibliothekswelt gemeinhin mit dem Begriff „digital curation" oder „data curation", manchmal auch mit „data stewardship" umschrieben. Organisationsstrukturen und Arbeitsabläufe (Geschäftsgänge) des digitalen Servicekonzepts der Bibliothek müssen darauf ausgelegt sein, die notwendigen personellen und finanziellen Ressourcen müssen dafür eingeplant und bereitgestellt werden. In deutschen Bibliotheken wird diesem wesentlichen Aspekt des digitalen Arbeitens nach wie vor viel zu wenig Aufmerksamkeit geschenkt, dabei ist er für die Qualität des (digitalen) Servicekonzepts und seiner Nutzerakzeptanz von herausragender Bedeutung. In dem Beitrag sollen vor dem Hintergrund der langjährigen Erfahrungen in der Bayerischen Staatsbibliothek (BSB) die Dimension und Komplexität des Problems aufgezeigt sowie mittlerweile praxiserprobte Lösungsansätze vorgestellt werden.

Summary:
Today almost every library with historic collections faces the question of digitisation. However, it is often ignored that a long-term and user-friendly provision of access to (and ultimately the preservation of) the produced data poses special organisational challenges for the respective institution. Unlike with print materials, it is not enough to store and circulate the digital library collection in facilities with the best possible climate. Regardless of their origin, extent, type and content, and largely independent from their actual use, digital documents need an extremely active and complex data management as well as intensive data maintenance. Additionally, the necessary hard- and software for providing the data must be operated and regularly updated. In the Anglo-American library world, these activities are usually called „digital curation" or „data curation", sometimes also „data stewardship". The library must design the organisational structures and workflows (processes) of its digital service concept to meet these requirements, and the necessary personnel and financial resources must be budgeted and made available. German libraries still pay far too little attention

1 Ich danke den Kolleginnen Frau Gabriele Meßmer und Frau Dr. Margarete Wittke für die vielfältigen Anregungen und die Unterstützung bei der Erstellung dieses Beitrags.

to this crucial aspect of digital work, although it is of outstanding importance for the quality of the (digital) service concept and its acceptance by the users. The paper demonstrates the dimension and complexity of this challenge against the background of the long-term experience of the Bavarian State Library and presents field-proven solutions.

Zitierfähiger Link: http://dx.doi.org/10.5282/o-bib/2015H4S268-278
Autorenidentifikation: Kempf, Klaus: GND 1043401806
Schlagwörter: Digitale Bibliothek; digital curation; data curation; data stewardship; digitale Bestandspflege; Digitalisierung

1. Einleitung und Begriffsklärung

Jede Institution, die digitale Daten in einem gewissen Mindestumfang produziert, sieht sich früher oder später mit dem Problem konfrontiert, dass digitale Daten im Unterschied zu konventionellen, analogen, also gedruckten Beständen ein hohes Maß an Betreuung erfordern, will man sie auf Dauer nutzbar erhalten. Der deutsche Begriff „digitale (Langzeit-)Archivierung" bringt diesen Umstand nur ungenügend zum Ausdruck, da hier im üblichen Sprachgebrauch die langfristige und verlustfreie Datensicherung, also der Nachhaltigkeitsaspekt, eindeutig im Vordergrund steht. Dabei wird häufig übersehen, dass die größte Herausforderung im Bereich der digitalen Bestandserhaltung in der möglichst permanenten und umfassenden Online-Bereitstellung der Daten liegt und die Bewältigung dieser Herausforderung die eigentliche Kernaufgabe der Datenhaltung darstellt. Erfordert bereits die Datenhaltung in einem sog. dark archive ein nicht zu unterschätzendes Maß an Datenpflege, so ist der Aufwand um ein Vielfaches höher, wo ein stark nachgefragter Datenbestand rund um die Uhr möglichst nutzerfreundlich, d.h. den aktuellen technischen Standards entsprechend, zugänglich gehalten werden soll. Letzteres ist nur mit einem erheblichen Zeit- und Ressourcenaufwand möglich. Diese komplexe und in ihrem Umfang dynamisch wachsende Aufgabenstellung, die im angloamerikanischen Sprachraum heute mit dem Begriff „digital curation" oder auch „data curation", manchmal auch „data stewardship", umschrieben wird, ist nach Auffassung von Lewis künftig eine, wenn nicht die zentrale Aufgabe wissenschaftlicher Bibliotheken, die über relevante digitale (Alt)Bestände verfügen.[2]

Entsprechend der Definition des Begriffs „data curation" von der Graduate School of Library and Information Science der University of Illinois ergeben sich für eine Einrichtung, die digitale (Forschungs-)Daten anbietet, vier wesentliche Teilaufgabenfelder: Sie muss sicherstellen, dass

1. die Such- und Findbarkeit der Daten, u.a. mittels verbesserter Sichtbarmachung und Visualisierung, gegeben ist (data retrieval).

2 David Lewis sieht bei einem unterstellten dynamischen Wachstum der Open-Access-Bewegung zunehmend weniger Ressourcenbedarf in der Beschaffung bzw. Lizenzierung von elektronischer Literatur, vielmehr dagegen im Bereich der „digitalen Bestandspflege", eben der „data curation", der nunmehr digital vorliegenden Altbestände und/oder der gesammelten Open-Access-Inhalte/-Contents. Vgl. dazu näher: Lewis, David: A strategy for academic libraries in the first quarter of the 21st century. In: College & Research Libraries 68 (2007), H. 5, S. 418–434, hier: S. 426 und Grafik auf S. 427. http://dx.doi.org/10.5860/crl.68.5.418.

2. Sie muss Vorkehrungen zur Qualitätssicherung, u.a. zur Korrektur der Daten einschließlich der dazugehörigen Metadaten treffen (maintain quality).
3. Sie soll zusätzlichen Mehrwert generieren (add value), etwa durch tiefere Erschließung bzw. Anreichern von Daten/Metadaten, Herstellen neuer Kontexte, aber auch durch die Entwicklung und das Angebot neuer, originärer Dienstleistungen.
4. Sie soll die Daten zur möglichst vielfältigen Nachnutzung zur Verfügung stellen (re-use).[3]

2. Entwicklung des digitalen Datenbestands in der Bayerischen Staatsbibliothek

Als ab 1997 die ersten Objekte bzw. Bücher an der Bayerischen Staatsbibliothek (BSB) digitalisiert wurden, war neben dem eigentlichen Scannen auch die Produktion der „digitalen Bücher" für das Internet noch Handarbeit. Die hochauflösend eingescannten Bilder mussten in browsergeeignete Formate konvertiert und es musste pro Buch per Hand eine in Excel erstellte sog. Strukturdatei angefertigt werden, die u.a. die Bilder in der richtigen Reihenfolge aufführte und damit das Blättern im digitalen Buch ermöglichte. Anfangs wurden alle Daten noch aufwendig per FTP ins Netz hochgeladen. Im Schnitt wurden so 300 digitale Bücher im Jahr produziert. Die Masterfiles, also die Originaldateien im TIFF-Format in höchster Auflösung, wurden zunächst noch auf CDs gebrannt und so archiviert.

Dies änderte sich radikal mit der Entwicklung und dem Einsatz des zentralen digitalen Produktionstools, der Zentralen Erfassungs- und Nachweisdatenbank, kurz ZEND genannt.[4] Dieses modular aufgebaute, auf open source-Software basierte, inhouse entwickelte und stetig weiter gepflegte und perfektionierte Tool umfasst grundsätzlich den gesamten Produktionsworkflow von der Aufgabe der Digitalisierungsorder – für Einzelaufträge heute auch über das Internet durch das in die ZEND integrierte Tool ERaTo möglich – mit allen notwendigen Zusatzinformationen bis zum automatischen Datentransfer in das digitale Langzeitarchiv beim Leibniz-Rechenzentrum (LRZ).

Damit wurden vom Jahre 2003 an nach und nach alle Prozessschritte nach dem Scannen vollständig automatisiert, so dass 2006 einige hundert Titel im Jahr und in den Folgejahren u.a. durch den Einsatz von Scanrobotern (mit Beginn des VD 16/2-Projekts im Jahre 2007),[5] aber auch durch die Erweiterung der Rechnerkapazitäten (Nutzung eines Linux-Clusters mit 35 Rechnern ab 2009) die Produktion sprunghaft zugenommen hat und schließlich ein Jahresspitzenwert von rund 12.000 Titeln (2009) erreicht wurde. Mittlerweile hat sich die Produktion im hauseigenen Scanzentrum, in dem vermehrt Stücke aus den Sondersammlungen und auch Handschriften digitalisiert werden, die beide von Natur aus mehr Aufwand verursachen, bei ca. 7.000 Bänden im Jahr eingependelt.

3 Siehe dazu: Cragin, Melissa H. u.a.: An educational program on data curation. Poster for 2007 STS Conference poster session. http://hdl.handle.net/2142/3493 (30.10.2015).

4 Vgl. zur Entwicklung und dem Einsatz der digitalen Produktionsplattform ZEND u.a. Brantl, Markus; Schoger, Astrid: Das Münchner Digitalisierungszentrum zwischen Produktion und Innovation. In: Rolf Griebel; Klaus Ceynowa (Hg.): Information, Innovation, Inspiration. 450 Jahre Bayerische Staatsbibliothek, München: Saur, 2008, S. 253–280.

5 Vgl. dazu näher: Brantl, Markus u.a.: Massendigitalisierung deutscher Drucke des 16. Jahrhunderts – Ein Erfahrungsbericht der Bayerischen Staatsbibliothek. In: Zeitschrift für Bibliothekswesen und Bibliographie 56 (2009), H. 6, S. 327–338. http://dx.doi.org/10.3196/186429500956655.

Bei den Google-Digitalisaten war der höchste Ausstoß in den Jahren 2010 bis 2014 mit ca. 200.000 Titeln pro Jahr zu verzeichnen. Bis heute (Stand: September 2015) wurden insgesamt, also im Wege der Eigenproduktion und durch Google, über 1,1 Millionen Digitalisate mit ca. 300 Millionen Seiten erstellt und im Netz veröffentlicht. Die digitale Gesamtdatenmenge ist um ein Vielfaches größer, da Bilddateien in unterschiedlichen Auflösungsstufen hergestellt und aufbewahrt werden und mittlerweile zudem Volltexte für einen großen Teil der Digitalisate zur Verfügung stehen.

Der Herstellungsprozess eines digitalen Buches endet nur vorläufig damit, dass die Originalbilddateien und eine Kopie der Strukturdatei, die die dazugehörigen Metadaten umfasst sowie seit neuestem auch eine mit OCR erzeugte Volltextversion, in ein digitales Archiv beim LRZ verschoben werden. Da „digitales Archivieren" zumindest im Sprachgebrauch der BSB gleichzusetzen ist mit der Forderung nach einer durchgehenden Bereitstellung bzw. Zugänglichmachung der einmal (teuer) erzeugten Daten, gibt es kein wirkliches Ende des Produktionsprozesses. Beim Umgang mit digitalen Daten gibt es ganz generell kein wirkliches Ende. Bei den Digitalisaten handelt es sich quasi um „lebendes Material", dessen Neu- oder Weiterverarbeitung jederzeit möglich sein muss. Eine umfassende „digitale Bestandspflege" oder „data curation" ist daher das Gebot der Stunde.

3. Auffindbarkeit der Daten (retrieval)

Die Such- und Findbarkeit der digitalen Dokumente wird primär über die einschlägigen (beschreibenden) Metadaten sichergestellt. Die digitalen Objekte werden im Laufe des Produktionsvorgangs, spätestens aber bei ihrer Bereitstellung mit einem Set von Metadaten versehen, die in der BSB großteils in einer komplexen Strukturdatei, die dem eigentlichen digitalen Objekt „beigepackt" ist, enthalten sind. Das reicht von beschreibenden Metadaten, die schon bei der Auftragsvergabe automatisch dem Katalog entnommen werden, wie z.B. die bibliographischen Daten, bis zur URN (= Uniform Research Name) der Digitalisate, die in einem eigenen Produktionsschritt kurz vor der Dokumentbereitstellung bei der sie verwaltenden national zuständigen Instanz, der Deutschen Nationalbibliothek (DNB), online abgerufen wird und die sie dauerhaft unter der gleichen Internetadresse aufrufbar macht (= persistenter Identifikator). Durch die Eintragung der URN in den lokalen Bibliothekskatalog bzw. die damit hergestellte Verknüpfung von Katalognachweis und elektronischer Ressource sind die Digitalisate über den OPAC recherchierbar bzw. dort direkt (im JPG-Format) einsehbar und zugänglich. Alle digitalen Objekte, die zur Nutzung über das Internet freigegeben sind, sind natürlich auch über die Homepage des Münchner Digitalisierungszentrums (MDZ) recherchierbar. Dort sind sie darüber hinaus im Sammlungskontext sicht- und durchsuchbar, wobei eine Recherche durch die Nutzerin oder den Nutzer sinnvoll eingeschränkt bzw. erweitert werden kann. Insgesamt stehen heute über 200 verschiedene Sammlungen im Netz. Ferner werden über die ZEND (vollautomatisch) die hauseigenen Portale bzw. die Virtuellen Fachbibliotheken permanent mit einschlägigen Nachweisen und einem direkten Zugriff auf die neu erzeugten und angebotenen retrodigitalisierten Werke versorgt. Über die ZEND angestoßen und gesteuert erfolgt mittels einer OAI-Schnittstelle last but not least auch die fortlaufende Versorgung von internationalen Nachweissystemen, wie der Europeana oder des WorldCat mit BSB-Digitalisaten und damit deren weltweite Online-Bereitstellung.

4. Erhaltung und Verbesserung der Datenqualität (maintain quality)

Hier muss man im Falle der BSB insbesondere zwischen der Arbeit an den eigen- und den fremderzeugten, d.h. i.d.R. den aus dem Massendigitalisierungsprojekt mit Google stammenden Daten (einschließlich der dazugehörigen Metadaten) unterscheiden. Auch bei den im eigenen Scanzentrum produzierten Bilddateien kommen gelegentlich Fehler vor, die nachträglich korrigiert werden müssen. Die Fehlerhäufigkeit korreliert dabei direkt mit dem Aspekt, ob im betreffenden Projekt, in dem die Daten erzeugt worden waren, eine (durchgehende) Qualitätskontrolle vorgesehen, d.h. finanziert worden ist, oder nicht. Immer dort, wo Daten aus einem Projekt mit Massendigitalisierungscharakter – auch bei hauseigener Produktion – stammen, d.h. nach BSB-interner Lesart Projekte mit einem Umfang von mehr als einer Million digitalisierter Seiten betroffen sind und keine intensive Qualitätskontrolle erfolgte, kann von einer gewissen Fehleranfälligkeit ausgegangen werden.

Im Rahmen des Digitalisierungsprojekts mit Google, Google Books, wurde mittlerweile die Millionengrenze (in Bänden) deutlich überschritten. Google ging nach eigenen Berechnungen zuletzt in dem gesamten Projekt von einer Fehlerquote von 1,37 % aller Seiten aus, d.h. im Umkehrschluss 98,63 % aller Seiten sind in Ordnung. Dies scheint angesichts des beträchtlichen Outputs ein vertretbarer Wert. Wenn man die absoluten Zahlen sieht, ergibt sich für die BSB bei Zugrundelegung von einer Million Bänden mit durchschnittlich 300 Seiten einen Korrekturbedarf von 4.110.000 Seiten. Vor diesem Hintergrund und angesichts der stetig steigenden Nutzernachfrage kann hier nicht mehr im Sinne einer Einzelfallkorrektur vorgegangen werden, sondern es bedarf auch für die Fehlerfindung bzw. -behebung eines systematischen Ansatzes, vor allem aber entsprechender organisatorisch-technischer Vorkehrungen, also der Einrichtung differenzierter Korrekturworkflows. Über die Notwendigkeit, die Fehlerkorrektur aktiv und systematisch anzugehen, bestand im Übrigen zwischen Google und der BSB immer Einigkeit. So wurden (und werden) die Daten aus der Google-Produktion laufend sowohl von Google als auch durch die BSB verbessert. Einige häufiger auftretende Produktionsfehler konnten durch BSB-eigene Bildanalyseverfahren automatisiert ausfindig gemacht werden. Google erhielt entsprechende Rückmeldungen und konnte gezielt Neulieferungen anstoßen. Eine größere Anzahl von Fehlern, insbesondere fehlende oder verdrehte Seiten, die nicht automatisch erkannt werden können, wird i.d.R. von Nutzerinnen und Nutzern gemeldet und entweder in Zusammenarbeit mit Google oder auch durch eigenes Nachscannen im hauseigenen Scanzentrum beseitigt. Bei diesen Workflows greifen manuelle und maschinelle Vorgänge ineinander. Sie beginnen aber immer mit der intellektuellen Analyse des Fehlertyps, d.h. es wird festgestellt,

- ob das Digitalisat schwarze Seiten enthält,
- ob nur einige wenige Seiten fehlen, z.B. ausklappbare Faltkarten,
- ob mehrere Fehlertypen vorkommen.

Anhand des Fehlertyps wird dann entschieden,

- ob ein maschinelles Reprozessieren möglich ist,
- ob manuell Faltkarten (von eigener Hand) nachgescannt und maschinell eingefügt werden müssen
- oder ob – in Ausnahmefällen – ein komplettes Neuscannen nötig ist.

Wenn der Text lesbar ist und keine Seiten fehlen, wird grundsätzlich nicht nachgescannt. Bloße „Schönheitskorrekturen" werden also nicht vorgenommen.

Ein besonderes aufwändiger Korrekturworkflow – bezogen auf die Eigenproduktion – ist der sog. Abschaltworkflow, der in Kraft tritt, wenn ein Digitalisat aus Urheberrechtsgründen gänzlich aus dem Netz genommen werden muss. Letzteres passiert leider hin und wieder mit Bezug auf das Digi20-Projekt. Hier wurden in der Zeit von Juli 2009 bis Herbst 2015 in Absprache mit ausgewählten deutschen Verlagen der Geschichts- und Kulturwissenschaften Neuerscheinungen gescannt und mit einer moving wall von fünf Jahren frei zugänglich ins Netz gestellt.[6] Obwohl seitens der betreffenden Verlage vorab eine definitive und umfassende Rechteklärung zugesichert wurde, kam und kommt es immer wieder zu Fällen, in denen nach erneuter Rechteprüfung durch die Justiziarin der BSB die Entscheidung gegen einen Verbleib des Dokuments im Netz fällt und der äußerst komplexe und aufwändige Korrekturmechanismus in Gang gesetzt werden muss.

Last but not least wird die Qualität der im Netz bereitgestellten Images auch dadurch gehoben, dass seit geraumer Zeit nach jeder Verbesserung älterer Daten und bei Neuproduktionen standardmäßig eine weitere Auflösungsstufe hinzugefügt wird, die auch bei zukünftig höheren Bildschirmauflösungen eine angemessene Qualität der Bildausgabe erlaubt.

Die Maßnahmen zum Erhalt und Verbesserung der digitalen Daten betreffen nicht nur die eigentlichen Objektdaten, sondern auch die dazu gehörigen Metadaten. Letztere werden in der durch ZEND gesteuerten digitalen Produktion an der BSB, wie bereits dargelegt, ganz überwiegend in einer Strukturdatei bei den betreffenden Objekten mitgeführt. Diese Strukturdateien werden kontinuierlich verbessert und den jeweiligen Anforderungen entsprechend angepasst. So werden bei jedem Update der Titeldaten auch die Strukturdateien neu erstellt. Mit der Einrichtung des sog. elektronischen Lesesaals in der BSB und der Einstellung von ausgewählten retrodigitalisierten Werken dort wurde es notwendig, zusätzliche (Rechte-)Informationen in die betreffenden Strukturdateien aufzunehmen. Darüber hinaus verdient in diesem Zusammenhang Erwähnung, dass vor vier Jahren im Zuge der Umstellung der Zeichenkodierung von ASCII auf Unicode eine Million Strukturdateien konvertiert und erneut in das Langzeitarchiv gestellt wurden.

5. Erschließung, Anreicherung, Herstellung neuer Kontexte (add value)

Der überwiegende Teil der im Wege der Retrodigitalisierung erzeugten Daten stehen heute noch ohne weitere Erschließung im Netz. Allerdings gibt es Möglichkeiten, dies für jedes einzelne Buch zu jedem Zeitpunkt zu ändern. Das am häufigsten dafür benutzte Instrument ist ein ZEND-Modul, der Table of Content-, kurz ToC-Editor, der es jederzeit – vergleichsweise einfach und mit vertretbarem Aufwand – ermöglicht, halbautomatisch ein Inhaltsverzeichnis bzw. die hierfür erforderlichen (logischen) Strukturdaten zu erstellen. Dies ist ein Vorteil im Vergleich zu anderen Systemen, bei

6 Projekt-Homepage: http://digi20.digitale-sammlungen.de/ (30.10.2015). – Das Projekt und seine Intentionen bzw. Inhalte sind näher beschrieben in: Schäffler, Hildegard; Seiderer, Birgit: Digitalisierung im urheberrechtsgeschützten Bereich – das Projekt Digi20. In: Zeitschrift für Bibliothekswesen und Bibliographie 58 (2011), H. 6, S. 311–315. http://zs.thulb.uni-jena.de/receive/jportal_jparticle_00247961 (01.12.2015).

denen das Inhaltsverzeichnis direkt beim Scannen erfasst werden muss. Mit dem ToC-Editor können digitalisierte Bestände genau dann, wenn sie in den Fokus des Interesses rücken, mit zusätzlichen Daten angereichert werden. In größerem Umfang erfolgte dies im vorletzten Jahr beispielsweise bei der „Zeitschrift für bayerische Landesgeschichte", die zum Teil schon vor über zehn Jahren digitalisiert wurde.

Zur Erschließung im weiteren Sinne gehört auch die OCR-Volltexterkennung. Bei einigen Projekten wurde sie von vornherein eingeplant (und durch die DFG auch mitfinanziert). Für die Mehrheit der Digitalisate war sie bislang aber noch nicht vorgesehen, vor allem deshalb, weil zum einen die Lizenzen für die Frakturschrifterkennung lange Zeit extrem teuer und qualitativ nicht befriedigend waren sowie zum anderen der Arbeitsprozess zu aufwändig war.[7] Das hat sich in jüngster Zeit geändert. Mit der OCR-Software Abbyy FineReader Recognition ist ein Produkt auf dem Markt, das vertretbare Ergebnisse auch bei Vorlagen in Frakturschrift liefert, und dies zu Lizenzbedingungen, die aus BSB-Sicht akzeptabel erscheinen. Darüber hinaus ist der OCR-Bearbeitungsschritt mittlerweile so in den ZEND-Gesamtworkflow integriert worden, dass er eigentlich, wie oben schon ausgeführt, zum Routine-Workflow gehört. Vor diesem Hintergrund wurde in einem ersten Schritt bzw. Teilprojekt vor ca. zwei Jahren die OCR-Erkennung von ca. 30.000 (Bavarica-)Titeln in Angriff genommen. Dafür wurden Kopien der Originaldateien (rund sechs Millionen Seiten) aus dem Datenarchiv geholt, in den OCR-Prozess verschoben und die bereits vorhandenen digitalen Bücher mit den Volltexten im XML-Format und im PDF-Format angereichert.

OCR-Software wird höchstwahrscheinlich niemals eine hundertprozentige Erkennungssicherheit erreichen. Es wird daher angestrebt, zukünftig auch Volltexte per Hand online korrigieren zu können – im besten Fall gemeinsam mit Forschenden bzw. Forschergruppen. Dabei ist es vorstellbar, dass die Volltextdaten mit Zusatzinformationen angereichert werden, indem etwa Personen- und Ortsnamen mit den Identifikatoren der Gemeinsamen Normdatei (GND) ausgezeichnet werden, so dass sie präziser recherchiert und Verknüpfungen hergestellt werden können.

Die Zusammenstellung in neuen Sammlungskontexten stellt ebenfalls eine Möglichkeit dar, aus der Masse der Digitalisate Gruppen zu bilden, die ganz gezielt weiter erschlossen werden können. Dies geschieht beispielsweise derzeit im Rahmen eines Zeitungsprojekts, bei dem aus der riesigen Menge von Google-Digitalisaten einzelne Zeitungstitel ausgewählt und deren Ausgaben mit Inhaltsverzeichnissen erschlossen werden, die es später ermöglichen, einzelne Ausgaben über eine Kalenderfunktion gezielt anzusteuern und die gegebenenfalls interessierenden Artikel und sonstigen Inhalte in adäquater Form auszudrucken bzw. herunterzuladen.

7 Bei den BSB-eigenen Retrodigitalisaten überwiegt derzeit mit ca. 90 % der Anteil der noch nicht durch Volltext erschlossenen Titel; für rund 10 % oder auch etwas mehr wurden mittlerweile gezielt im Rahmen von Projekten Volltexte erzeugt, z.B. für Bavarica-Titel, aber auch für Dokumente mit Osteuropabezug. Die von Google produzierten Digitalisate wurden dagegen von Anfang an mit OCR-erzeugten Volltexten ausgestattet, die allerdings zunächst eine bedenkliche Qualität aufwiesen. Dank anhaltender Qualitätsbemühungen und der von Google selbst entwickelten Software Aksara scheint nun ein Durchbruch geschafft und damit auch eine Lösung für das Problem der Frakturschrifterkennung gefunden worden zu sein.

Auch über die Vernetzung der bestehenden Angebote untereinander und/oder mit der Gemeinsamen Normdatei (GND) können die Digitalen Sammlungen tiefer erschlossen und damit insgesamt besser zugänglich gemacht werden. Besonders intensiv wurde dies bislang bei dem von der DFG geförderten ältesten Digitalisierungsprojekt der BSB, den „Reichstagsprotokollen", durchexerziert.[8] Hier wurde nicht nur aus den Reichstagshandbüchern die sehr differenziert abfragbare Abgeordnetendatenbank aufgebaut, sondern diese wurde sowohl mit den Redebeiträgen als auch den Artikeln in der Allgemeinen bzw. der Neuen Deutschen Biographie über die einzelnen Abgeordneten verlinkt. Den Schlüssel für die Vernetzung des gesamten Angebots bildet die Identifikationsnummer aus der vormaligen Personennamendatei bzw. heute der GND.

Eine ganze Reihe von Sammlungen verfügt über eigens geschaffene Weboberflächen mit zusätzlichen Retrieval-Möglichkeiten, die nur über eine tiefere Erschließung verwirklicht werden konnten. Ein besonders schönes Beispiel für eine allmähliche Verbesserung der Recherchemöglichkeiten ist das Zedler-Lexikon.[9] Es wurde nach der Digitalisierung der analogen Vorlage zunächst in einer grob strukturierten reinen Blätterversion ins Netz gestellt. Später wurden die Lemmata als Sprungmarken erschlossen, die schließlich in einem dritten Schritt in einem Kooperationsprojekt mit der Herzog-August-Bibliothek Wolfenbüttel systematisiert wurden.

Zu den „added values" zählen auch verschiedene Zusatzdienste, die die BSB, aufsetzend auf ihrem „digitalen Datenschatz", anbietet. Das ist zum einen die sog. Bildähnlichkeitssuche. Sie bietet die Möglichkeit, innerhalb eines Bestandes von mehreren Millionen (derzeit ca. 5,5 Millionen) Images eine Bildähnlichkeitssuche anzustoßen. Mit ihr können ähnliche Bildmotive im Portal aufgrund formaler Merkmale (Farben, Formen, Kontraste) recherchiert werden. Dieses innovative Feature, das gemeinsam mit dem langjährigen Technologiepartner der BSB, dem Fraunhofer Heinrich-Hertz-Institut, Berlin, entwickelt wurde, eröffnet gerade der kunsthistorischen Forschung neue Perspektiven. Das Kriterium ist allein die Ähnlichkeit von Bildmotiven nach rein äußerlichen Merkmalen (Farben, Texturen, markante Formen und Kontraste); inhaltliche Kriterien spielen hierbei im Moment keine Rolle.[10]

Im Rahmen des landeskundlich-kulturwissenschaftlichen Onlineportals bavarikon wurde auch die Bereitstellung von 3-D-Objekten über das Netz entwickelt bzw. als Dienstleistung ins Leben gerufen.[11] Für 3D-Digitalisierungen stehen an der BSB aktuell zwei 3D-Scansysteme zur Verfügung, die in der Lage sind, neben der 3D-Vermessung synchron die Textur (detaillierte Farbwerte) der Objekte

8 Startseite: http://www.reichstagsprotokolle.de/index.html (30.10.2015).

9 Startseite: http://www.zedler-lexikon.de/index.html?c=startseite&l=de (30.10.2015).

10 Vgl. näher zur Bildähnlichkeitssuche und 3D-Digitalisierung: Ceynowa, Klaus; Brantl, Markus: Visuelle Suche und virtuelle Interaktion. Neues aus der Innovationswerkstatt der Bayerischen Staatsbibliothek. In: Bibliotheks-Magazin. Mitteilungen aus den Staatsbibliotheken in Berlin und München 8 (2013), H. 2, S. 15–20. https://www.bsb-muenchen.de/fileadmin/imageswww/pdf-dateien/bibliotheksmagazin/BM2013–2.pdf (30.10.2015) bzw. http://staatsbibliothek-berlin.de/fileadmin/user_upload/zentrale_Seiten/ueber_uns/pdf/Bibliotheksmagazin/BibliotheksMagazin_2013–02.pdf (30.10.2015). - Ceynowa, Klaus; Brantl, Markus: Innovationen aus der Bayerischen Staatsbibliothek. Bildähnlichkeitssuche und 3D-Interaktion. In: Bibliotheksforum Bayern 7 (2013), H. 3, S. 162–165. https://www.bibliotheksforum-bayern.de/fileadmin/archiv/2013–3/PDF-Einzelbeitr%C3%A4ge/BFB_0313_03_Ceynowa_V03.pdf (30.10.2015).

11 Startseite des Portals bavarikon: http://www.bavarikon.de/ (30.10.2015).

zu erfassen: ein Farb-3D-Scanner für die hochauflösende 3D-Digitalisierung von kleinformatigen Objekten und ein Laserscanner für größere Objekte mit einem Messbereich von zwei bis 100 Metern. Dieser Service ist vor allem für Museen von besonderem Interesse. Mittlerweile haben fünf 3D-Digitalisierungskampagnen stattgefunden (Archäologische Staatssammlung, Staatliches Museum Ägyptischer Kunst, Staatliche Antikensammlung und Glyptothek, Staatliche Münzsammlung, Deutsches Medizinhistorisches Museum). Die digitalisierten Objekte, die jeweils aufwändig nachbearbeitet werden müssen, werden ab Herbst 2015 schrittweise online gestellt.

Eine weitere innovative Dienstleistung betrifft die Anfertigung und das Angebot unterschiedlicher Apps. Man greift dabei u.a. auf Datenbestände zurück, die im Kontext des bavarikon-Projekts produziert und in diesem Zusammenhang angeboten werden. Auch in anderen Kontexten und unabhängig von bavarikon werden Apps entwickelt, wie zuletzt die App „30 Klassiker in Erstausgaben aus den Beständen der BSB“.[12]

Die BSB erschließt sich mit ihren vielgestaltigen und sehr gefragten digitalen Daten auch neue Geschäftsfelder. Dazu gehört zum einen der Einsatz eines neuen Viewertyps, des IIIF-Viewers, der u.a. die Annotation durch Benutzerinnen und Benutzer selbst erlaubt.[13] Noch einen Schritt weiter wird man künftig im Sinn einer interaktiven Dienstleistung mit dem Angebot von Virtuellen Forschungsumgebungen gehen, die u.a. ganz im Sinne von big data umfangreiche Bereitstellungen von ausgewählten digitalen Datenbeständen an Forschende zum Inhalt haben.

6. Nachnutzung (re-use)

Bereits mit dem ersten im MDZ erstellten Digitalisat wurde mit dem Ziel, das digitale Duplikat möglichst lange und vielfältig nachnutzen zu können, nach dem Motto verfahren: Tue es einmal, aber richtig. Damit wurde eine zukunftsweisende Entscheidung getroffen: Jedes Digitalisat wird seither in der bestmöglichen Qualität gescannt, d.h. im TIFF-Format und mit mindestens 300 dpi. Dies führt allerdings gleichzeitig zum Entstehen beträchtlicher Datenmengen. Die Originaldateien eines großformatigen Buches mit über 500 Seiten können leicht 40 bis 80 GB umfassen, was der Belegung von 10 bis 20 DVDs gleichkäme. Und obwohl beim LRZ dafür mittlerweile Online-Speicher in einer Größenordnung von mehreren hundert TB zur Verfügung stehen, können die Originaldateien im Speichersystem nicht einfach nur aufbewahrt werden so wie sie eben anfallen. Es ist vielmehr ein ausgefeiltes Datenmanagement zu betreiben, damit es nicht zum Datenüberlauf und damit einhergehend zum Datenverlust oder zur Datenbeschädigung kommt. Das bedeutet, dass die produzierten Daten so schnell wie möglich in das Archiv verschoben und aus dem Arbeitsbereich gelöscht werden müssen. Beim Erreichen der Speicherkapazitätsgrenze muss die Produktion gegebenenfalls automatisch gestoppt werden. Sie darf erst wieder aufgenommen werden, wenn neue Kapazitäten bereit stehen bzw. die Altproduktionsdaten „abgeräumt“ wurden.

12 Hierzu wird in den Zeitschriften „Bibliotheks-Magazin. Mitteilungen aus den Staatsbibliotheken in Berlin und München“ und „Bibliotheksforum Bayern“ jeweils in Heft 1/2016 ein ausführlicher Bericht erscheinen.

13 Zu IIIF (International Image Interoperability Framework) siehe: http://iiif.io/about.html bzw. zu Apps und Demos: http://iiif.io/apps-demos.html (30.10.2015).

Die gute Qualität der Originaldaten ist die Voraussetzung für eine sinnvolle Nachnutzung der Digitalisate. Insbesondere die Anfertigung von Reproduktionen in Druckqualität, die von Nutzerinnen und Nutzern über die Serviceleistung „Dokumentlieferung Altes Buch" (mit dem online-Bestelltool ERaTo) bestellt werden kann, vereinfacht sich durch die Nachnutzung der Digitalisate radikal. Statt ein Buch aus dem Magazin zu holen, eine oder mehrere Seiten zu scannen und das Buch zurückzustellen, können über ein einfaches Webformular die betreffenden Seiten in Originalqualität aus dem Archiv angefordert werden und stehen kurze Zeit später zur Auslieferung an die Nutzerinnen und Nutzer bereit.

Es zeigt sich bereits seit längerem, dass mit den digitalisierten Büchern nicht nur online gearbeitet wird, sondern dass ein großer Bedarf besteht, die Digitalisate in Form von PDF-Dateien herunterzuladen und offline zu nutzen. Die Downloadzahlen sind beeindruckend. In diesem Jahr wurden bislang (Stichtag: 26.05.2015) knapp 400.000 Digitalisate in PDF-Form heruntergeladen, obwohl der Download pro Sitzung beschränkt ist, um die Server nicht zu überlasten. Seit der Einrichtung des PDF-Downloads wurden fast vier Millionen digitalisierte Bücher ausgeliefert.

Alle Retrodigitalisate

Jahr	Anzahl	Ø/Tag	TB	GB/Tag
2015	394695	2723	40.13	276.76
2014	870173	2385	90.70	248.49
2013	669534	1835	68.52	187.73
2012	794900	2178	76.75	210.27
2011	640262	1755	55.89	153.12
2010	346346	949	28.16	77.15
2009	192147	527	12.77	34.99
2008	23850	66	1.88	5.15
Gesamtzeitraum	3931907	---	374.8	---

MDZ-Digitalisate

Jahr	Anzahl	Ø/Tag	TB	GB/Tag
2015	157730	1088	15.24	105.10
2014	364408	999	36.39	99.70
2013	350657	961	33.78	92.55
2012	429823	1178	38.33	105.01
2011	423143	1160	33.87	92.79
2010	274808	753	21.48	58.85
2009	184677	506	12.19	33.40
2008	23850	66	1.88	5.15
Gesamtzeitraum	2209096	---	193.16	---

Google-Digitalisate

Jahr	Anzahl	Ø/Tag	TB	GB/Tag
2015	236965	1635	24.89	171.66
2014	505765	1386	54.31	148.79
2013	318877	874	34.74	95.18
2012	365077	1001	38.42	105.26
2011	217119	595	22.02	60.33
2010	71538	196	6.68	18.30
2009	7470	21	0.58	1.59
Gesamtzeitraum	1722811	---	181.64	---

Abbildung: Statistik der PDF-Downloads von Retrodigitalisaten der BSB, 2008–2015, Stichtag: 26.05.2015

In noch größerem Umfang werden digitalisierte Datenbestände der BSB künftig im Rahmen von „DaFo" (= Datenlieferung an die Forschung) weiterverwertet werden können. Dabei handelt es sich um ein Verfahren, bei dem wissenschaftliche Nutzerinnen und Nutzer selbst online auch hochaufgelöste Bilder zu Forschungszwecken bestellen können. Diese werden (automatisiert) aus dem Archiv geholt und zum Download bereitgestellt. Bis vor kurzem war dies sehr aufwändig nur von Hand möglich, vermittelt durch das Personal der Bibliothek. Auf Dauer war dieser Aufwand nicht vertretbar, da er zu chronischen Personalengpässen führte. Es ist zu erwarten, dass der neue Dienst auch ohne spezielle Werbemaßnahmen in kurzer Zeit sehr intensiv nachgefragt werden wird. Derzeit wird das neue, automatisierte Verfahren noch getestet. Die Bestellmenge wird pro Nutzer/in und Tag vorerst mengenmäßig beschränkt sein. Die Bereitstellung soll, so die Planung, innerhalb von vier Wochen nach Bestellung erfolgen.

7. Fazit

Die Datenmengen entfalten eine Eigendynamik, die alle früheren Überlegungen und Berechnungen zur „digitalen Bestandspflege" und den dafür notwendigen Personalressourcen über den Haufen geworfen haben. Neben der Aufrechterhaltung und Optimierung des laufenden Datenmanagements

ist ein stetig neues Angebot attraktiver Dienste rund um den eigenen „digitalen Datenschatz" unerlässlich. Darüber hinaus ist die permanente Erweiterung der Hardware, die fortlaufende Aktualisierung der benutzten Software, die hinreichende Absicherung gegen Datenverluste durch eine lückenlose Kontrolle des Datenflusses und eine möglichst zyklische Migration der Datenbestände eine absolute Pflicht. Nur die alles umfassende Datenpflege, verstanden als fortlaufender Prozess von der Datenentstehung bis zur Datenarchivierung – im Englischen kurz und bündig mit „data curation" umschrieben – sorgt dafür, dass der so teuer aufgebaute digitale Datenbestand „lebendig" und damit auf Dauer gebrauchsfähig bleibt.

Literaturverzeichnis

- Brantl, Markus u.a.: Massendigitalisierung deutscher Drucke des 16. Jahrhunderts – Ein Erfahrungsbericht der Bayerischen Staatsbibliothek. In: Zeitschrift für Bibliothekswesen und Bibliographie 56 (2009), H. 6, S. 327–338. http://dx.doi.org/10.3196/186429500956655.

- Brantl, Markus; Schoger, Astrid: Das Münchner Digitalisierungszentrum zwischen Produktion und Innovation. In: Rolf Griebel; Klaus Ceynowa (Hg.): Information, Innovation, Inspiration. 450 Jahre Bayerische Staatsbibliothek, München: Saur, 2008, S. 253–280.

- Ceynowa, Klaus; Brantl, Markus: Innovationen aus der Bayerischen Staatsbibliothek. Bildähnlichkeitssuche und 3D-Interaktion. In: Bibliotheksforum Bayern 7 (2013), H. 3, S. 162–165. https://www.bibliotheksforum-bayern.de/fileadmin/archiv/2013–3/PDF-Einzelbeitr%C3%A4ge/BFB_0313_03_Ceynowa_V03.pdf (30.10.2015).

- Ceynowa, Klaus; Brantl, Markus: Visuelle Suche und virtuelle Interaktion. Neues aus der Innovationswerkstatt der Bayerischen Staatsbibliothek. In: Bibliotheks-Magazin. Mitteilungen aus den Staatsbibliotheken in Berlin und München 8 (2013), H. 2, S. 15–20. https://www.bsb-muenchen.de/fileadmin/imageswww/pdf-dateien/bibliotheksmagazin/BM2013–2.pdf (30.10.2015) bzw. http://staatsbibliothek-berlin.de/fileadmin/user_upload/zentrale_Seiten/ueber_uns/pdf/Bibliotheksmagazin/BibliotheksMagazin_2013–02.pdf (30.10.2015).

- Cragin, Melissa H. u.a.: An educational program on data curation. Poster for 2007 STS Conference poster session. http://hdl.handle.net/2142/3493 (30.10.2015).

- Lewis, David: A strategy for academic libraries in the first quarter of the 21st century. In: College & Research Libraries 68 (2007), H. 5, S. 418–434. http://dx.doi.org/10.5860/crl.68.5.418.

- Schäffler, Hildegard; Seiderer, Birgit: Digitalisierung im urheberrechtsgeschützten Bereich – das Projekt Digi20. In: Zeitschrift für Bibliothekswesen und Bibliographie 58 (2011), H. 6, S. 311–315. http://zs.thulb.uni-jena.de/receive/jportal_jparticle_00247961 (01.12.2015).

