Historisches Kulturgut – neue Aufgaben

Konversion des kulturellen Erbes für die Forschung

Volltextbeschaffung und -bereitstellung als Aufgabe der Bibliotheken

Thomas Stäcker, Herzog-August-Bibliothek Wolfenbüttel

Zusammenfassung:
Mit der Transformation des gedruckten Buch zum elektronischen Text verändern sich zentrale Rahmenbedingungen der Bibliothek. Die theoretischen Grundlagen des „Buches" müssen unter dem Gesichtspunkt des Digitalen neu durchdacht und auf ihre praktischen Konsequenzen hin geprüft werden. Vor allem die Transitivität, spezifische Schriftlichkeit und Prozessierbarkeit elektronischer Texte sind Eigenschaften, die Konsequenzen für eine ganze Reihe bibliothekarischer Kernaufgaben haben. Mit Blick auf das kulturelle Erbe, das in Bibliotheken verwahrt wird, stellt sich die Aufgabe, auf diesen Paradigmenwechsel angemessen zu reagieren und Sorge dafür zu tragen, dass das schriftliche und gedruckte Kulturgut auch in einer adäquaten maschinenlesbaren Form zur Verfügung steht. Nach gut zehn Jahren erfolgreicher Imagedigitalisierung muss daher jetzt, nach der Entwicklung entsprechender Techniken, der nächste Schritt zur Herstellung, Aufbereitung und Bereitstellung von Volltext getan werden, um neuen, sich aus der digitalen Wende ergebenden Forschungsanforderungen und Forschungsfragen, die sich z.B. mit Begriffen wie Stilometrie, Clusteranalyse, Topic Modeling etc. verbinden, zu genügen. Den Bibliotheken wächst vor diesem Hintergrund in der Transformation des schriftlichen Kulturgutes und Bereitstellung von Volltexten eine neue Aufgabe zu. Sie können darin einen wichtigen Beitrag zum Aufbau einer Infrastruktur für eine digital arbeitende Geistes- und Kultwissenschaft bzw. die Digital Humanities leisten.

Summary:
The transformation of the book into an electronic text has led to constitutive changes in the functional frameworks of the library. Basic concepts of the notion of the 'book' have to be reconsidered and to be evaluated in view of the practical consequences. Above all, the transitivity of such texts, the specific mode of writing employed, and their ability to be processed are aspects that have considerable impact on several core tasks of the library. In view of the cultural heritage that is preserved in libraries this paradigmatic shift requires an appropriate response, and libraries have the task of converting artifacts of written and printed cultural heritage into machine-readable form and to provide access to it. After ten years of successful image digitization and the development of suitable techniques, the aim is now to produce, enhance and provide access to full text in order to fulfill new research requirements and deal with the issues involved. This involves methods such as stylometry, cluster analysis or topic modeling. In taking on these new tasks libraries can make an important contribution to building up a research infrastructure for the digital humanities.

Zitierfähiger Link (DOI): 10.5282/o-bib/2014H1S220-237

Autorenidentifikation: Stäcker, Thomas: GND 141905573

1. Einleitung

Die Bibliothek der Zukunft ist nicht nur eine Bibliothek der Bücher, sondern auch eine Bibliothek der Texte. Dieser Satz ist auf den ersten Blick merkwürdig und scheint etwas zu fordern, was doch selbstverständlich ist: Bibliotheken kümmern sich um die Bereitstellung von Texten. Doch bei allem Ungefähren, das Texte umgibt[1], trennt Texte und Bücher im üblichen Verständnis ein signifikantes Merkmal. Während Bücher integral mit einem Träger verbunden sind, existieren Texte auch ohne diesen Träger. Texte können von Träger zu Träger wandern und sind gegen ihr Substrat neutral: „l'œuvre se tient dans la main. Le texte se tient dans le language". [2] In diesem Sinne kann man den Text auch vom Dokument unterscheiden, das sich in einem materiellen Substrat wie Papier manifestiert. [3] Ebenso ist der Text anders als das Werk – auch wenn es in der Definition von FRBR dem hier gebrauchten Textbegriff nahekommt, denn das Werk ist zwar auch eine geistige Abstraktion, aber es ist in der Regel innig mit dem Begriff seines Schöpfers verbunden, Texte schließen diese Konnotation nicht notwendig ein. Text wird zwar im Singular gelegentlich wie Werk gebraucht, im Plural jedoch nicht. [4] Zudem ist der Text anders als das Werk mit der Sprache bzw. Schrift verbunden. Der Text ist damit einerseits weiter gefasst und weniger bestimmt bzw. „fixiert" als das Werk, andererseits aber mit der Beschränkung auf die Sprache und Schrift auch enger.[5] Natürlich hängt die Bestimmung von Text davon ab, in welchem Kontext er verwendet wird[6], und obige Charakterisierung ist sicher nicht für alle Fälle brauchbar. Doch für die intendierte bibliothekarische Adaption des Begriffes „Text" ist dies unerheblich, denn es geht vor allem darum, einen Begriff zu entwickeln, der es erlaubt, Eigenschaften wie Schriftlichkeit, Prozessierbarkeit und Transitivität[7] sprachlich angemessen zu verankern und mit Blick auf die Aufgabe der Konversion des gedruckten Erbes fruchtbar werden zu lassen. „Text" in diesem Sinne wird im Kontext der Digitalisierung benutzt, d.h. es geht darum, die spezifischen Eigenschaften des Textes im digitalen Medium und die Auswirkungen, die das digitale Medium auf den Text hat, mit in Rechnung zu stellen. Dabei soll die von Cerquiglini formulierte Einsicht als Leitfaden dienen, dass der digitale Text „malgré la parenté inertie de vocabulaire, sa technique ne ressemble en rien à l'imprimerie"[8]. Mit anderen Worten, hier wird eine grundsätzliche Differenz zum gedruckten oder handgeschriebenen Buch

1 Vgl. z. B. Eggert, Paul: Text-encoding, Theories of the Text, and the 'Work-Site'. In: Literary and Linguistic Computing 20 (2005), H. 4, S. 425-435; Landow, George P.: Hypertext 3.0. Critical Theory and New Media in an Era of Globalization. Baltimor: Johns Hopkins Univ. Press, 2006; Pedauque, Roger T.: Le Document à la lumière du numérique: forme, texte, médium: comprendre le rôle du document numérique dans l'émergence d'une nouvelle modernité. Caen: C & F Éd., 2006 und jüngst Caton, Paul: On the term 'text' in digital Humanities. In: Literary and Linguistic Computing 28 (2013), S. 209-220.

2 Barthes, Roland: De l'œuvre au Texte. In: Revue d'esthétique 3 (1971), S. 226.

3 Pedauque (wie Anm. 1): „Une première definition du document pourrait être représentée par l'équation: Document traditionell = support + inscription", S. 36.

4 Vgl. Caton, Paul (wie Anm. 1).

5 Die metaphorische Weiterung des Textbegriffes in einem außerschriftlichen Sinne kommt hier nicht in Betracht.

6 Pedauque hat hierfür die Analogie zum Begriff der linguistisch verstandenen Pragmatik bemüht (wie Anm. 1): „Le document comme medium: cette dimension enfin pose la question du statut du document dans les relations sociales…", S. 32.

7 Vgl. Barthes (wie Anm. 2): „Il s'ensuit que le Texte ne peut s'arrêter (par exemple à un rayon de bibliothèque); son mouvement constitutif est la traverse (il peut notamment traverser l'œuvre, plusieurs œuvres)", S. 227.

8 Cerquiglini, Bernhard: Éloge de la Variante. Histoire critique de la philologie. Paris: Éd. du Seuil, 1989, S. 29.

behauptet – modisch formuliert: ein Pradigmenwechsel – und gegenüber Analogiebildungen und Adaptionen aus der Ära des Druckes zu Skepsis geraten, auch wenn sprachliche Kontinuitäten Übergänge vermitteln helfen (vgl. das Beispiel des englischen volume, das mit seiner lateinischen Wurzel volumen (=Rolle) wenig gemein hat).

Es ist schon viel geschrieben worden über die Eigenschaften elektronischer oder digitaler Texte bzw. Dokumente. Ich möchte nur drei herausgreifen, die mir für die Behandlung von elektronischen bzw. E-Texten konstitutiv scheinen: Transitivität, Schriftlichkeit und Prozessierbarkeit.

Transitivität bedeutet, dass Texte sich leicht von ihrem Träger lösen lassen. Es heißt nicht, dass der Text gänzlich trägerlos wäre, aber der Träger ist nicht mehr integral mit dem Text verbunden, so wie wir es heute noch von einem Buch kennen. Die Funktion des Trägers bzw. in einem kommunikationstheoretischen Sinne des Mediums, ist dabei eine je unterschiedliche. Schon in der Antike gab es transitorische Texte, z.B. Aufzeichnungen in Wachstafeln. Die Wiederbeschreibbarkeitsfunktion dieser Tafeln bis hin zur Kreidetafel diente vor allem dazu, Texte für zeitlich begrenzte Zwecke sichtbar werden zu lassen, sei es zu pädagogischen oder auch kalkulatorischen Zwecken. Der Text war nach seiner zeitlich begrenzten Präsentation unwichtig und konnte „gelöscht" werden. Der Träger, so wird aus diesem Zusammenhang deutlich, hatte zwei wichtige Funktionen. Einerseits diente er zur Präsentation bzw. Manifestation des Textes – ohne Träger bzw. Medium wäre der Text nicht sichtbar –, andererseits bestimmte der Träger die Funktion des Textes. Gebrauchstexte bedurften keines Trägers, der hohe Anforderungen an seine dauerhafte Verfügbarkeit stellte. Sakrale Texte hatten im Gegenteil hohe Anforderungen an ihre dauerhafte Aufbewahrung und damit an ihren Träger (Stein, Bronzetafeln, Pergament etc.). Mit der innigen Verbindung ging einher, dass der Träger für die Integrität und Originalität des Textes sorgte, der so „festgestellt" und zum Werk wurde, denn die Transitivität der Texte bedeutete im Prozess der Überlieferung auch immer die Gefahr der Verfälschung.[9] Mit dem Aufkommen elektronischer bzw. digitaler Texte[10] rückte die prinzipielle Transitivität des Textes wieder stärker in den Vordergrund. Symptomatisch für die Anfangszeit war, dass man dem temporär auf dem Bildschirm erscheinenden Text keinen oder nur geringem Wert beimaß. Die flüchtige Verbindung des Textes mit seinem Träger (CD-ROM, Festplatte, etc.) schien ihn funktional weder für die dauerhafte Aufbewahrung zu qualifizieren, noch auch seine Originalität sichern zu können. Doch diese aus dem Druckzeitalter stammenden Vorurteile beginnen sich in dem Maß aufzulösen, wie der E-Text zur Normalität und die Transitivität nicht nur als Mangel, sondern auch als Fortschritt begriffen wird. Denn mit dem nahezu Verschwinden des Trägers in elektrischen Ladezuständen des RAM eines Computers verschwinden auch materielle Hindernisse, die die Verbreitung des Textes begrenzt haben. Die in der Geschichte des Textes so wesentliche Möglichkeit des Verbreitens durch Kopieren hat eine Stufe der Mühelosigkeit erreicht, die von allen Einschränkungen wie des aufwändigen handschriftlichen Kopierens oder des Setzens

9 Die Verfälschung, wie wir sie heute verstehen, nämlich als Änderung der schriftlichen Form des Textes, wurde in der Antike und im Mittelalter allerdings nur bei sakralen Texten als Gefahr gesehen. Der schriftlich sichere und festgestellte Text ist vor allem Resultat des Buchdrucks, vgl. Eisenstein, Elisabeth: The printing press as an agent of change. Vol. I and II. Cambridge: Cambridge University Press, 1997: "Of all the new features introduced by the duplicative powers of print, preservation is possibly the most important. ", S. 113.

10 Zuvor im sprachlichen Bereich auch durch Radio und Fernsehen.

für die Druckerpresse absehen lassen. Mit der leichten Kopierbarkeit, die die Transitivität des Textes hervortreten lässt, ändern sich aber auch bibliothekarische Konzepte. Das Sammeln bzw. Erwerbung, die Benutzung, die Ausleihe, das Magazin, der Katalog, der Lesesaal, die Bestandserhaltung etc. wandeln ihre ursprüngliche Bedeutung oder werden obsolet[11]. So benötigen z.B. elektronische Texte kein Magazin im klassischen Sinne und müssen auch nicht ausgeliehen, sondern können kopiert werden, zumindest wenn es die rechtlichen Rahmenbedingungen zulassen.

Mit der Eigenschaft der *Prozessierbarkeit*[12] kommt ein Moment zum Tragen, das für den in der Bibliothek nötigen Wandel ausschlaggebend ist. E-Texte können u.a. durchsucht, indexiert, kombiniert, gewandelt, verlinkt, automatisch analysiert, mit Text Mining Techniken bearbeitet werden, so dass sie ihre ursprüngliche Funktion, die vor allem in ihrer Lesbarkeit bestand, überschreiten und die Bibliothek zwingen, Vorsorge zu treffen, damit Texte auch prozessiert werden können. Dabei beruhen diese Funktionen wesentlich auf der Sprachlich- bzw. der *Schriftlichkeit* der Texte, die sich, was hier theoretisch nicht näher ausgeführt werden kann[13], gegenüber der natürlichen Sprache emanzipiert und unter den Bedingungen des Digitalen Züge einer eigenen Algorithmik annimmt. Fachlich kann die Bibliothek bei der Verarbeitung der Schrift gemäß ihren Intentionen (Erschließung, Benutzbarkeit) auf Methoden zurückgreifen, die vor allem in der Linguistik, aber nicht nur ihr entwickelt wurden und die letztlich dazu führen sollten, die Bibliothekswissenschaft um eine digitale Textwissenschaft zu bereichern. Die Bibliothek, die wie Lipsius bemerkt, zusammen mit der Schrift erfunden wurde[14], muss sich darauf einstellen, und nicht nur das, sie muss ein treibender Faktor in der Konversion des kulturellen schriftlichen Erbes werden, denn es ist absehbar, dass eine effiziente wissenschaftliche Benutzung des in Bibliotheken befindlichen Quellenmaterials nur unter der Voraussetzung möglich sein wird, dass es in aufbereiteter digitaler bzw. re-kodierter Form vorliegt. Gefragt sind vor allem strategische Lösungen, die auf internationalem und nationalem Level arbeitsteilige Prozesse in Gang setzen, um der kultur- und geisteswissenschaftlichen Forschung eine moderne Infrastruktur oder, um einen neuen Begriff zu verwenden: eine virtuelle Forschungsinfrastruktur zur Verfügung zu stellen. Unter nationalen Gesichtspunkten ist die Digitalisierung aller mittelalterlichen Handschriften und Inkunabeln auf deutschem Gebiet (Regionalprinzip) und aller im deutschsprachigen Raum oder in Deutsch gedruckten, heute auf dem Gebiet der Bundesrepublik befindlichen Werke, soweit sie nicht dem Urheberrecht unterliegen (gemischtes Regional- und Sprachprinzip), das Ziel.

Digitalisierung ist dabei ein unscharfer Begriff. Zu unterscheiden sind zunächst die Digitalisierung als Reproduktion von Quellen und deren Konversion in eine maschinenlesbare Form (Retrodigitalisierung)[15], wobei mehr oder weniger konstitutive Komponenten differenziert werden

11 Stäcker, Thomas: Vom Buch zum Text. Sammeln, Erschließen und Benutzen im digitalen Zeitalter. In: Christine Haug; Rolf Thiele (Hg.): Buch – Bibliothek – Region. Wolfgang Schmitz zum 65. Geburtstag. Wiesbaden: Harrassowitz, 2014, S. 353-364.

12 Dino Buzetti: Digital Editions and Text Processing. In: Marilyn Deegan (Hg.): Text editing, print and the digital world. Farnham, Surrey: Ashgate, 2009, S. 45-61.

13 Zum Konzept einer sich von der Sprache emanzipierenden Schrift s. insbesondere Derrida, Jacques: De la grammatologie. Paris: Éd. de Minuit, 1967.

14 Justus Lipsius: Syntagma de bibliothecis. Antwerpen: Moretus, 1602, S. 9. http://diglib.hab.de/drucke/qun-59-9-1/start.htm?image=00011 (17.9.2014).

15 Im Archivkontext redet man auch von der Digitalisierung der Findmittel, also der Metadaten. Diesen eingeschränkten

können, wie die Erfassung von Metadaten, die Imagedigitalisierung, die Erfassung von Strukturdaten, der Volltext (*plain text*) oder der Volltext mit Markup; sodann die Digitalisierung als Praxis digitalen Publizierens (*born digital*). Nachstehend soll der Stand und die Perspektiven der Retro-Digitalisierung aufgezeigt und das Potential deutlich gemacht werden, das sich aus der Transformation einer Bibliothek der Bücher zu einer Bibliothek der Texte ergibt. Die Neuerwerbung von E-Texten im Sinne von *born digitals* tritt organisch hinzu, wird hier aber nicht weiter thematisiert.

2. Stand der Imagedigitalisierung

Die Imagedigitaliserung ist die unmittelbare Voraussetzung für die Volltextdigitalisierung. Daher ist es sinnvoll, zunächst einen Blick auf den Stand der Imagedigitalisierung zu werfen. Wenn man auf die Entwicklung der Digitalisierung in Deutschland zurückblickt, kann man beklagen, dass der Fortschritt so langsam erfolgte und es die einschlägigen Fördereinrichtungen, aber auch die größeren Gedächtnisorganisationen versäumt haben, die Digitalisierung koordiniert voranzubringen. Zahlreiche, sicher verdienstvolle Einzelunternehmungen führten nicht selten zu unnötigen digitalen Dubletten, der zentrale Nachweis blieb zunächst in dem Versäumnis der Verbundsysteme stecken, sich miteinander ins Benehmen zu setzen, um einen gemeinsamen Datenpool aufzubauen und anzubieten.[16] Doch auch die DDB noch auch die Europeana brachten bislang den gewünschten zentralen Zugriff auf Digitalisate. Erfreuliche Spezialportale wie ZVDD[17] für vor allem alte Drucke fanden nicht genügend Rückhalt. Am ehesten kann noch der KVK[18] für sich beanspruchen, einen zentralen Zugriff auch auf digitalisierte Bücher in Deutschland zu bieten.

Auf der europäischen Ebene ist grundsätzlich zu beklagen, dass in den letzten Jahren nur wenig Mittel in die Herstellung von *content* geflossen sind, mit dem widersinnigen Ergebnis, dass zwar die Entwicklung digitaler Werkzeuge gut vorangeschritten ist, es aber zu wenig frei nutzbare Daten in öffentlicher Hand gibt, mit denen diese Tools signifikante Ergebnisse erzielen könnten; die digitalen Geisteswissenschaften bzw. Digital Humanities wurden so um Jahre zurückgeworfen. Angesichts der in EU Rahmenprogrammen verausgabten Summen ist es unverständlich, dass das Potential von Big Data Entwicklungen im Kulturerbebereich verkannt und an den infrastrukturellen Bedürfnissen vorbei substantielle Mittel wo nicht vergeudet, so doch in vielen Bereichen falsch oder erst zu spät an den richtigen Stellen platziert wurden, denn die in der Theorie plausible Vorstellung, dass die europäischen Länder selbst auf nationaler Ebene für den digitalen *content* sorgen müssten, hat sich in der Praxis als illusorisch erwiesen. Gleichwohl verlief die Entwicklung für Deutschland in den letzten Jahren vergleichsweise günstiger als in anderen Ländern, nicht, weil die zuständigen staatlichen Stellen auf Länderebene sich zeitig engagiert hätten – der Aufbruch kam, wenn überhaupt, nur halbherzig, spät und unkoordiniert[19] –, sondern weil die DFG nach einer anfänglichen

Gebrauch mache ich mir hier nicht zu Eigen. Digitalisierung schließt nach meinem Verständnis immer auch die Reproduktion des Originals ein.

16 Vgl. die mehr als deutliche Kritik durch den Wissenschaftsrat: http://www.wissenschaftsrat.de/download/archiv/10463-11.pdf (16.9.2014).

17 http://www.zvdd.de/startseite/ (16.9.2014).

18 http://www.ubka.uni-karlsruhe.de/kvk.html (16.9.2014).

19 Systematische Informationen zu den Digitalisierungsprogrammen der Länder sind nur schwer zu ermitteln, die

Experimentierphase[20] konsequent in die Förderung der Digitalisierung eingestiegen ist und die zuständigen Ausschüsse, insbesondere von Literaturversorgungs- und Informationssysteme (LIS), erkannt haben, wie wichtig die Bereitstellung von digitalen Quellen für die Forschung war und ist und dass digitale Forschung nur auf der Basis digitalisierter Quellen und Daten stattfinden kann. Entscheidend für die sich nun abzeichnende stringentere Kulturgutdigitalisierung waren zwei Faktoren, zum einen die so genannten DFG Praxisregeln[21], die für ein hohes Maß an Standardisierung sorgten und von Anfang an vorausschauend auch mögliche Texterkennungsverfahren bei der Qualität der Imagedigitalisierung im Blick hatten, zum anderen die intensivere Kooperation von großen und mittelgroßen Bibliotheken in verschiedenen so genannten Masterplänen, in denen es schrittweise gelang, frühere Alleingänge zu integrieren und das Ziel der Gesamtdigitalisierung als gemeinsame bibliothekarische Aufgabe ins Bewusstsein zu rücken, ein Prozess, der sich auch für die jetzt anstehende Volltextdigitalisierung als gute Voraussetzung empfiehlt.

Seit gut 10 Jahren digitalisieren Bibliotheken in relativ systematischer Form Ihre Bestände. Den Auftakt machten die so genannten „Massendigitalisierungsprojekte“[22]. Mit dem Googleprojekt der BSB trat ein Projekt hinzu, das seinerzeit ungeahnte neue Größenordnungen eröffnete. Da das Google-Projekt nicht der Systematik der späteren Masterpläne und einem mit anderen Bibliotheken koordinierten Prozess folgen konnte, hat es in organisatorischem Sinne zu einer Zweiteilung des Vorgehens bei der Digitalisierung geführt. Hier die Bestände der BSB, die komplett von Google digitalisiert wurden und werden, dort die Bestände der übrigen Bibliotheken. Alle späteren, Drucke betreffenden Masterpläne bauten auf dieser Konstruktion auf.

Eine Reihe von Initiativen widmete sich gezielt den jeweiligen Materialien. Ein Pilotprojekt zur Entwicklung eines Masterplans zur Digitalisierung der rund 60.000 mittelalterlichen Handschriften in Deutschland wurde vor kurzem begonnen.[23] Digitalisierte Inkunablen werden zentral im ISTC

folgende Übersicht gibt daher nur einen Eindruck zum status quo wieder. In Baden-Württemberg werden Projekte von der Stiftung Kulturgut und dem Land gefördert, das in einem unlängst aufgelegten Programm zur Digitalisierung wertvoller Bestände mehr als € 1,5 Mio zur Verfügung stellte; in Bayern gibt es kein zentrales Programm, sondern nur einzelne Projektförderungen vor allem für den Bereich Bavarica, die über das Portal „bavarikon“ zugänglich sind; Berlin hat Haushaltsmittel von immerhin knapp € 1 Mio pro Jahr für ein „Digitalisierungskonzept für das Land Berlin“ eingestellt, allerdings fließt nur ein Bruchteil in die content –Produktion; Brandenburg fördert im Umfang von etwa € 150.000 Digitalisierungsprojekte im Jahr; Bremen und Hamburg haben kein Digitalisierungsprogramm; Hessen fördert über LOEWE (Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz) auch einzelne Digitalisierungsprojekte; Niedersachsen finanziert zwar ein zentrales Kulturerbe-Portal, aber nicht die Herstellung von content, Nordrhein-Westfalen fördert ein zentrales Portal und verfolgt ein zentrales Archivierungsmodell, „Digitales Archiv NRW“, ein Programm zur content-Herstellung fehlt jedoch; Rheinland-Pfalz besitzt kein Digitalisierungsprogramm, unterstützt aber das Portal „dilibri“; im Saarland gibt es nur Vorüberlegungen; Sachsen hat die SLUB Dresden mit der Koordinierung von Digitalisierungsprojekten beauftragt. Einzelne Vorhaben sind angelaufen; Schleswig-Holstein besitzt kein Programm und fördert allenfalls Einzelprojekte im Kontext der Bestandserhaltung; Thüringen hat kein Programm, fördert aber Einzelprojekte im Volumen von ca. € 300.000.

20 S. das Förderprogramm Verteilte Digitale Forschungsbibliothek der DFG: http://webdoc.sub.gwdg.de/ebook/aw/2004/retro/RappIdeeVDF.pdf (15.9.2014).

21 http://www.dfg.de/formulare/12_151/12_151_de.pdf (14.9.2014).

22 Opitz, Andrea; Stäcker, Thomas: Workshop der Massendigitalisierungsprojekte der Deutschen Forschungsgemeinschaft an der Herzog August Bibliothek Wolfenbüttel. In: Zentralblatt für Bibliothekswesen und Bibliographie 56 (2009), S. 363-373.

23 http://www.bsb-muenchen.de/die-bayerische-staatsbibliothek/projekte/digitalisierung/pilotphase-handschriftendigitalisierung/ (15.9.2014)

und im deutschen Census von der BSB sowie vom Gesamtkatalog der Wiegendrucke in Berlin nachgewiesen.[24] Ziel ist, die rund 27.000 existierenden Inkunabelausgaben digital zur Verfügung zu stellen. Die Digitalisierungsprojekte des 16. und 17. Jahrhunderts bauen auf der Voraussetzung der nahezu fünzigjährigen Katalogisierung von VD 16 und VD 17 auf.[25] Ohne diese jahrzentelangen Vorarbeiten wäre an ein koordiniertes Digitalisierungsprogramm nicht zu denken gewesen und erst im Ausgang von diesen Datenbanken konnte man Teilmengen definieren, die sich Bibliotheken zur Digitalisierung zuweisen lassen konnten. Auf diese Weise wurden im VD 16 bis heute ca. 50.000 von ca. 110.000, im VD 17 ca. 100.000 von ca. 280.000 Drucken digitalisiert. Das noch laufende VD 18 spielt insofern eine Sonderrolle, als die Digitalisierung nicht konsekutiv auf die Katalogisierung folgt, sondern Hand in Hand mit ihr geht. Von den zu erwartenden 600.000 Drucken sind bereits ca. 100.000 digitalisiert und über die Datenbank verlinkt.[26] In den nächsten etwa 10 Jahren sollte die Digitalisierung systematisch auf eine Quote von 80 % gebracht[27] und der Rest nach Bedarf in Einzelprojekten digitalisiert werden können. Voraussetzung ist, dass die Finanzierung der Digitalisierung fortgesetzt und dass auch die Länder ihre Verantwortung in diesem Prozess z.B. bei der Kofinanzierung und der Bereitstellung der erforderlichen Infrastruktur erkennen. Weitgehend unberücksichtigt ist das 19. Jahrhundert. Angesichts der zu erwartenden Mengen und unklaren Katalogisierungs- bzw. Metadatenlage wäre hier zunächst eine kleinere Machbarkeitsstudie und Durchführung einiger Pilotprojekte anzustreben.

Die Grundlagen für die jetzt anstehende Volltexterfassung sind vor diesem Hintergrund relativ gut, zumal die Bibliotheken anders als kommerzielle Anbieter (vgl. z.B. Early English Books Online (EEBO)[28]) ihre Volltextdigitalisate frei anbieten und damit Synergien nutzen können, die kommerziellen oder teilkommerziellen Anbietern nicht zur Verfügung stehen. Die notwendige Metadatenerfassung ist in den letzten Jahren weitgehend29 erfolgt bzw., was die Digitalisierung anlangt, auf gutem Weg, so dass es vernünftig ist, jetzt systematisch in die Volltextdigitalisierung einzusteigen.

3. Volltextdigitalisierung

Volltextdigitalisierung ist aus bibliothekarischer Hinsicht nicht nur ein technisches, sondern vor allem ein administratives Problem. Sie stellt, recht betrachtet, einen speziellen Fall der Erwerbung dar. Bibliotheken verhalten sich nach wie vor zögerlich, was die Erwerbung von Volltexten bzw. E-Texten anlangt. Die aus der Tradition der Bibliothek als sammelnder Einrichtung heraus unverständliche

24 http://www.bsb-muenchen.de/die-bayerische-staatsbibliothek/projekte/erschliessung/inkunabel-census-fuer-deutschland-istc/ (15.9.2014) und GW www.gesamtkatalogderwiegendrucke.de (15.9.2014).

25 Die nationalbibliographische Verzeichnung des VD 16 begann 1969, das des VD 17 1982 (Vorbereitungsphase).

26 http://vd18.de/ (15.9.2014).

27 Hier ist davon auszugehen, dass weitgehend alle relevanten Forschungsfragen adressiert werden können.

28 http://eebo.chadwyck.com/home (15.9.2014).

29 Eine tatsächlich vollständige Erfassung des gedruckten deutschen Kulturerbes bleibt allerdings ein dauerhaftes Ziel. So ist bekannt, dass z.B. im VD 17 Titel mit Druckorten im östlichen Europa des alten deutschen Sprachgebiets nur unzureichend erfasst sind. Dies betrifft gerade das regional begrenzte Kleinschriftum. Andere Regionen, wie der Südwesten Deutschlands, sind ebenfalls nur ungenügend erfasst, weil große Bibliotheken wie Heidelberg, Stuttgart oder Tübingen ihre Bestände nicht oder nur teilweise eingebracht haben. Das jetzt laufende Digitalisierungsprogramm hat aber dazu geführt, dass andere Bestände wie z.B. Erlangen, Hamburg, Jena und Rostock nachträglich ergänzt und schmerzliche Lücken geschlossen werden konnten.

Zurückhaltung bei der Erwerbung von aktuellen E-Texten – meist werden nur Lizenzen erworben[30] –, setzt sich bei der Beschaffung von Volltexten von Werken des kulturellen Erbes fort. Erwerbung von Volltexten oder E-Texten[31] kann auf zwei Wegen erfolgen: entweder durch Herstellung oder durch Kopieren, Herstellung wiederum durch Transformation oder Verfassen (so genannte *digital born* oder digitale, nicht digitalisierte Dokumente).

Kopieren ist ein Erwerbungsprozess, der in der Handschriftenzeit üblich war und im Druckzeitalter aus dem Blick geraten ist[32], weil der Drucker die benötigte Anzahl von Kopien herstellte und händisches Kopieren nach Diktat in Skriptorien entfallen konnte.[33] Mit dem digitalen Dokument bzw. E-Text erhält das Kopieren neue Bedeutung. Beschaffung eines Textes bedeutet im einfachsten Fall, ihn herunterzuladen bzw. eine lokale Kopie herzustellen. Voraussetzung dafür ist, dass es ein Kopierrecht gibt. Ob ein solches besteht, hängt mit dem Herstellungsprozess zusammen und ist für den Erwerber nicht immer leicht zu entscheiden. Von Vorteil erweist sich hier, wenn Texte von ihren Herstellern bzw. den Rechteinhabern markiert sind, z.B. durch eine *Creative Commons Lizenz*.[34] Die Zurückhaltung von Bibliotheken beim Erwerb von Volltexten resultiert auch aus der Form der Texte. Volltexte sind anders als Druckwerke bislang nicht normiert oder genügen modernen Anforderungen, die unter der Maßgabe der *Prozessierbarkeit* von Texten stehen, nicht oder nur unzureichend. Typisch sind im Internet derzeit Texte im PDF Format. Ein PDF ist layoutbasiert und simuliert darin Dokumente des Druckzeitalters. Für die digitale Nachnutzung bzw. Prozessierbarkeit sind aber Dokumente unverzichtbar, deren Strukturen explizit ausgezeichnet wurden. Diese strukturellen Elemente wie Fußnoten, Register, Inhaltsverzeichnis, Haupttext, Paginierungen etc. lassen sich in Texten, die im PDF Format vorliegen, technisch nicht zuverlässig identifizieren. Eine Fußnote ist eine in kleiner Type gedruckte Ziffer, die nur intellektuell durch den Leser zu verstehen ist. Computer vermögen diese Information nicht oder nur unzuverlässig zu verarbeiten. Daher ist PDF als Format prinzipiell ungeeignet, auch für die Langzeitarchivierung, selbst wenn es kommerzielle Anbieter geschafft haben, PDF/A als Langzeitarchivstandard zu „verkaufen".

Die Frage, die sich an dieser Stelle erhebt, ist, wie genau Texte beschaffen sein müssen und sollen, um in ihrer Eigenschaft als E-Texte adäquat genutzt bzw. prozessiert werden zu können. Als sicher kann gelten, dass Texte mit Markup ausgezeichnet sein und einem definierten Zeichencode folgen müssen. Für das Markup hat sich XML durchgesetzt[35], für den Zeichencode Unicode. Allerdings hat sich trotz bereits entwickelnder Standards wie der *Text Encoding Initiative* (TEI)[36] noch kein Publikationsstandard durchgesetzt, der diese Strukturen in verallgemeinerter Form festhielte. Die TEI

30 Es sind durchaus nicht nur rechtliche Probleme, die hier wirksam sind. Einige Einrichtungen scheuen die erforderlichen Umbauten in der Infrastruktur, andere sehen sich mit Verweis auf die DNB nicht in der Pflicht zu archivieren, wieder andere nutzen E-Books nur als ephemeres Angebot.

31 Volltext und E-Text werden meist daraufhin unterschieden, dass erster hergestellt wird, letzterer bereits besteht.

32 Wenn es auch immer wieder Fälle gab, wo vergriffene Werke auf der Grundlage von Bibliotheksexemplaren nachgedruckt wurden. Es gibt sogar darauf spezialisierte Dienstleister.

33 Stein, Peter: Schriftkultur. Eine Geschichte des Lesens und Schreibens. Darmstadt: Wiss. Buchges., 2006.

34 http://de.creativecommons.org/ (16.9.2014).

35 Es ist üblich, aber nicht notwendig, dass die Serialisierung in XML erfolgen muss. Kodierungen in z.B. JSON sind ebenso möglich und kommen gerade in Zusammenhängen in Gebrauch, wo es um lightweigh-Anwendungen geht.

36 http://www.tei-c.org (16.9.2014).

kommen zwar regelmäßig in Editionen zum Einsatz, was sich aus ihrer Herkunft als Markupsprache für Transkriptionen erklärt, aber einerseits kennen sie selbst wiederum vielfältige Varianten, andererseits sind sie bei moderneren Publikationsformen derzeit (noch) eher selten.[37].

Die Herausforderung für die Bibliothek besteht darin, die unterschiedlichen Kopien in ein Zielformat zu überführen, um standardisierte Prozesse der Recherche, von Verfahren der *Digital Humanities* (DH) (s. u.) und zur Langzeitarchivierung ausführen zu können. Das heißt nicht, dass man das Originalformat nicht archivieren könnte und sollte, nur kann die Bibliothek angesichts der zahllosen, oft proprietären oder unzureichend standardisierten Dokumente im Internet nicht individuell für eine vollständige Konversion aller Strukturmerkmale sorgen, zumal die Dokumente selbst (s. PDF oder z.T. auch HTML) diese Information nicht ohne weiteres bieten. So gibt es zu Recht Bemühungen, Basis- oder Standardformate z.B. auf der Basis der TEI[38] zu formulieren, und Bibliotheken tun im eigenen Interesse gut daran, diesen Prozess zu unterstützen und auch Hersteller wie auch Verlage zu ermuntern, Texte in strukturell ausgezeichneten Formaten anzubieten. Sofern Bibliotheken selbst in der Lage sind, Texte über den Herstellungsprozess zu erwerben – entweder indem sie als Repository oder gleichsam Verlag Texte publizieren oder indem sie vorhandene gedruckte Texte in digitale transformieren – lassen sich Standardisierungsfragen leichter adressieren.

Aus dieser knappen Skizze wird deutlich, dass Bibliotheken es nicht mehr nur mit den klassischen bibliographischen Metadaten zu tun haben, deren Formalisierung und Standardisierung weitgehend abgeschlossen ist, sondern dass es nun um die Formalisierung und Standardisierung der strukturellen Metadaten der Texte selbst geht. Damit rückt die Bibliothek enger an den Herstellungsprozess und auch die Forschung heran und kann gerade darin, dass sie den „Aufbereitungsprozess" der digitalisierten Dokumente begleitet, zum Partner der Forschung werden, indem sie für die erforderliche digitale Infrastruktur sorgt.

Aus dem Bereich der digitalen Erwerbung soll im Folgenden allein der Frage der *Transformation* nachgegangen werden, also der Volltextgewinnung auf der Basis bestehender gedruckter Werke[39] und die erforderlichen Arbeitsschritte und neuen Nutzungsszenarien skizziert werden.

37 Vgl. aber die DH Zeitschrift Digital Humanities Quarterly, die ihre Texte in TEI verfasst, sie allerdings auch mit proprietären Tags anreichert: http://www.digitalhumanities.org/dhq/ (16.9.2014).

38 S. das Basisformat des Deutschen Textarchivs http://www.deutschestextarchiv.de/doku/basisformat (16.9.2014).

39 Die Verwendung des kommunikationswissenschaftlichen Begriffes der „analogen Medien", der sich aus dem Sender-Empfänger-Modell der Signalübermittlung ableitet, wird hier und im Folgenden vermieden, weil er dem Sachverhalt einen falschen Akzent verleiht. Die theoretische Frage, ob ein Text „analog", also ohne dekodierende Hilfsmittel oder binär mit dekodierenden Hilfsmitteln übermittelt wird, ist für die hier leitenden Fragestellungen unerheblich, auch wenn die Thematik in der Langzeitarchivierung in manchen Kreisen weiterhin präsent ist. Ich gehe davon aus, dass die Art der Signalübertragung für das Verständnis unerheblich ist, d. h. eine indianische Knotenschrift kann nicht minder unverständlich sein als eine nicht definierte Serie von Bits. Es scheint mir daher sinnvoller, die Dekodierung semantisch zu denken, d.h. der Computer wird in diesem Sinne als ein den binären Code entschlüsselndes Werkzeug ebenso verschwinden wie die Lesebrille beim Prozess des Lesens. Beide sind zwar unerlässlich für die Lektüre, aber phänomenologisch und epistemologisch irrelevant.

4. Transformation des gedruckten kulturellen Erbes in eine maschinenlesbare Form

Auch wenn es bereits interessante Ansätze zur automatisierten Handschriftenerkennung gibt, haben die dafür entwickelten Werkzeuge noch keine Marktreife erreicht und sollen hier außer Betracht bleiben. Die Transformation gedruckter Werke erfolgt gegenwärtig entweder durch Abschreiben oder Texterkennungssoftware (OCR). Für beides gibt es gute Gründe und Anwendungsszenarien. Gegen das Abschreiben sprechen meist Kostengründe, doch gibt es viele Fälle, in denen auf eine qualitativ hochwertigere Form nicht verzichtet werden kann. Aber auch bei OCR ist die Qualitätsfrage nicht zu vernachlässigen, denn angesichts der in den meisten Fällen schlechten Resultate von Standard OCR-Software bei den Drucken der Handpressenzeit muss festgelegt werden, welche Qualität für welche Zielgruppe und Fragestellung ausreichend ist. Dabei gilt, dass spezialisierte OCR auch höhere Preise zur Folge hat.

Unabhängig davon, welche Qualität mit welcher Technik erreicht wird, muss man grundsätzlich mit der Tatsache leben, dass eine hundertprozentige Genauigkeit bei der Wandlung in eine maschinenlesbare Form für die Dokumente, die nicht Gegenstand einer Edition sind, unmittelbar nicht zu erreichen sein wird, d.h. für eine überwältigende Mehrzahl der Dokumente müssen Prozesse zu ihrer laufenden Verbesserung eingeplant werden. Das ist für die Organisation der Erwerbung von Digitalisaten wichtig und stellt eine neue Situation dar. Bisher konnten die erworbenen bzw. hergestellten Werke (nach RAK so genannte Sekundärausgaben) als Imagedigitalisate bzw. Masterfiles abgelegt und mussten nicht weiter bearbeitet werden. Jetzt sind laufende Prozesse zu ihrer Verbesserung nötig.

Um eine flächendeckende Maschinenlesbarkeit zu erreichen, ist die Konversion mit OCR perspektivisch sicher die interessantere Variante. Gleichwohl gibt es noch hohe Hürden zu überwinden. Neben extrinsischen Mängeln wie Scanfehlern weisen Drucke der frühen Neuzeit eine Reihe von Schwierigkeiten auf, die von der OCR zu bewältigen sind. Dazu zählen Mischschriften von Fraktur und Antiqua (aktuelle Software ist meist auf das eine oder andere spezialisiert, der Regelfall ist aber die Mischung), Kursive oder andere Schriften (vor allem griechisch und hebräisch). Häufig trifft man auf Phänomene des Widerdrucks, der Papierverschmutzung oder Bräunung, von Marginalien etc.[40] Derzeit gebräuchliche Softwareprodukte für ältere Drucke sind Abbyy Finereader, tesseract, OCRopus und B.I.T. Tomasi. Wie und unter welchen Voraussetzungen die jeweiligen *tools* günstig sind, ist wiederum vom Anwendungsfall und ökonomischen Betrachtungen abhängig. Unter bibliothekarischen Gesichtspunkten stellt sich die Frage, wie unter der formulierten Bedingung eines strukturierten Volltextes nachnutzbare E-Texte erzeugt werden können. Die Nachnutzbarkeit richtet sich nach den Anforderungen, wie sie derzeit vor allem in den Digital Humanities[41] (DH) formuliert werden.

40 Die Prozesse und Phänomene sind mittlerweile gut untersucht, vgl. die Ergebnisse des IMPACT Projektes http://www.impact-project.eu/ (15.9.2014); s.a. die Projektpublikation der Staatsbibliothek zu Berlin: Volltext via OCR - Möglichkeiten und Grenzen. http://staatsbibliothek-berlin.de/fileadmin/user_upload/zentrale_Seiten/historische_drucke/pdf/SBB_OCR_STUDIE_WEBVERSION_Final.pdf (15.9.2014).

41 Eine Übersicht über die aktuellen Theorien zu diesem Bereich liefert der von Melissa Terras u.a. herausgegebene Reader Defining Digital Humanities. A Reader. Farnham: Ashgate, 2013.

Wie schon erwähnt, ist der vielleicht schwierigste Gedanke bei der Transformation von Volltexten ihre Vorläufigkeit, d.h. nicht nur expansive, sondern auch innere Transitivität. Das heißt nicht nur, dass sie nicht auf Anhieb in einen hundertprozentig richtigen maschinenlesbaren Text transformiert werden können und so der permanenten, gleichsam sich infinitesimal auf das Ziel der Exaktheit zu bewegendenden Verbesserung unterliegen. Es bedeutet auch, dass sich elektronische Dokumente anders als Texte auf Papier weiter anreicheren und strukturell re-kodieren[42] lassen, so dass man auch aus prinzipiellen Gründen nie von einem abgeschlossenen Text sprechen kann. Der Text in dieser Form wandelt sich zu einer „strukturierten Informationsressource"[43]. Angesichts dessen werden Standards umso wichtiger, weil nur sie den Zusammenhalt des Textes wo nicht auf der Ebene der Schrift, so aber auf der Ebene der Struktur sicherstellen können, so dass sich z.B. unterschiedlich angereicherte Transformationen desselben Textes ineinander überführen lassen. Ergebnis fast aller OCR Programme ist auch eine Datei, die die Ergebnisse des OCR Prozesses beinhaltet, das erkannte Layout, ggf. mit basalen strukturellen Informationen wie Illustration oder Tabelle, die identifizierten Wörter (*token*) mit ihren Koordinaten im zugrundeliegenden Digitalisat, der Konfidenzwert des erkannten Wortes oder Buchstabens, ob das Wort korrigiert wurde, ob es in einem Wörterbuch gefunden wurde, etc.[44] Die meisten Programme liefern auch schon standardisierte Formate aus. Derzeit werden vor allem zwei unterstützt, zum einen hOCR[45], zum anderen ALTO[46]. hOCR ist in XHTML formuliert, ALTO in XML. ALTO genießt den Vorzug, ein von der *Library of Congress* unterstützter Standard zu sein, hat aber auch einige Nachteile. So fehlt z.B. die Möglichkeit die Koordinaten einzelner Buchstaben zu markieren. Gleichwohl besitzt man mit dem Vorliegen eines Standards wie ALTO, der auch ein Attribut hat, das über den Korrekturstatus des Dokumentes Auskunft gibt, Mittel, um den wesentlich dynamischen Charakter elektronischer Dokumente im allgemeinen und OCR prozessierter Texte im besonderen abzubilden. Zum weiteren gibt es für ALTO bereits Verfahren, um automatisiert Strukturinformationen zu extrahieren[47]

Die Bibliothek muss dafür Sorge tragen, dass OCR Prozesse mit verbesserter Software, nicht nur desselben Herstellers, zyklisch wiederholt und auch andere Korrekturprozesse, etwa durch Croudsourcing, in das Masterfile integriert werden. Texte geraten so auf der Basis einer einheitlichen Struktur in einen ununterbrochenen Fluss, deren Stabilität einzig im Masterimage des digitalen Faksimilies liegen (Referenzkoordinaten) oder aber in der intellektuellen editorischen Fixierung eines Textes (Text im Sinne von Werk), wobei letzterer durchaus unter Kennzeichnung der Varianzen (Versionierung) fortgeschrieben werden kann und damit Werkspezifika des Druckzeitalters

42 Zum Begriff der Re-Kodierung bzw. Transmedialisierung siehe Sahle, Patrick: Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels. Teil 2: Befunde, Theorie und Methodik. Norderstedt: BoD, 2013, S. 157 ff.

43 Sahle (wie Anm. 42), S. 161.

44 Vgl. z. B. Abbyy: http://www.abbyy-developers.eu/en:tech:features:xml.

45 https://docs.google.com/document/d/1QQnlQtvdAC_8n92-LhwPcjtAUFwBlzE8EWnKAxIgVf0/mobilebasic?pli=1&viewopt=127 (10.9.2014).

46 http://www.loc.gov/standards/alto/ (14.9.2014).

47 Vgl. den aus dem IMPACT-Projekt hervorgegangenen Functional Extension Parser der Universität Innsbruck: http://www.digitisation.eu/fileadmin/user_upload/Deliverables/IMPACT_D-EE_4.3_Functional_Extension_Parser.pdf (14.9.2014).

auflöst48. So ist grundsätzlich sichergestellt, dass eine aus wissenschaftlicher Sicht unverzichtbare verlässliche Zitation möglich ist, entweder auf Basis der Referenzkoordinaten des Images oder aber der jeweiligen Editionsversion. Ausgehend von den identifizierten *token*[49] treten weitere basale Anreicherungselemente für Volltexte dieser Art hinzu. Dazu zählen vor allem die Identifikation von Entities (v.a. Personen- und Ortsnamen)[50] und deren Verknüpfung mit der *Gemeinsamen Normdatei* (GND) oder des *Getty Thesaurus for Geographic Names* TGN, die Einbringung von Lemmata[51], also den grammatischen Grundformen von Wörtern und so genannte Part-of-Speech Tags (POS), die die grammatische Funktion von Wörtern festlegen, z. B. Substantiv, Verb, Präposition etc.[52] Ebenso hinzutreten können auf dieser Ebene die logischen Strukturen wie Abschnitt, Paragraph und Satz. Letztere sind für spätere Analyseprozesse besonders wichtig, denn, will man nicht von der meist wenig aussagekräftigen Seitenlogik des Druckes ausgehen, bilden sie den Container bzw. Rahmen für Boolesche Suchen oder Kookkurrenzen. Der Satz bildet als distinkte grammatische Struktur zudem auch in besonderer Weise eine inhaltliche Einheit, so dass seine Kodierung (*tagging*) interessante Nachnutzungsoptionen eröffnet.

5. Nutzungsszenarien

Man erkennt, dass die Re-Kodierung gedruckter Texte nicht nur eine simple Transformation vom gedruckten ins elektronische Medium ist, sondern dass aus der elektronischen Form neue Anforderungen an die Bibliothek erwachsen. Auch wenn die bei der Transformation des kulturellen schriftlichen Erbes vor allem angesprochenen Geisteswissenschaften in der Fläche nur allmählich und zögerlich neue Methoden adaptieren, ist doch der Trend deutlich erkennbar. Starke Nachfrage nach E-Texten kommt aus dem Bereich der Linguistik und insbesondere den an Bedeutung gewinnenden Digital Humanities. Texte werden Gegenstand quantitativer Analysen und Visualisierungen, die zu neuen Erkenntnissen führen. So erlauben die bekannten *Voyant* Tools[53] nicht nur die Ermittlung von Worthäufgkeiten und -verteilung, sondern auch deren Visualisierung in Graphen oder der bekannten Wordwolke.

48 Der Autorbegriff am Werk wird darin problematisch, selbst wenn Texte Autoren haben. Wikipedia ist ein schönes Beispiel eines multiauktorialen Werkes, an dem als solchem keine Autorschaft reklamiert werden kann.

49 Dieser Begriff aus der Linguistik hat sich hier eingebürgert und bedeutet eine typisierte Zeichenkette, üblicherweise ein Wort, aber nicht nur; so sind auch Zahlen, Abkürzungen oder Satzzeichen token.

50 Dies kann automatisch bzw. semiautomatisch durch NER (named entity recognition)-Software erfolgen.

51 Zum unterstützenden Einsatz kommen hier so genannte Lemmatizer.

52 Gebräuchlich sind in der Linguistik z.B. STTS (Stuttgart-Tübingen Tagset), das man in der Bibliothek für diese Zwecke übernehmen könnte.

53 http://voyant-tools.org/ (16.9.2014).

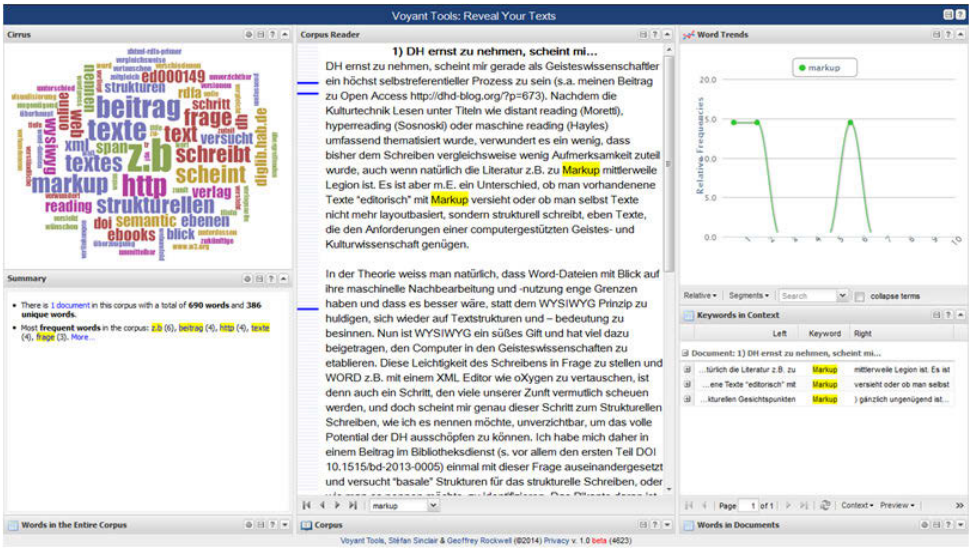

Abb. 1: Voyant Tools

Dieserart Visualisierungstechniken erlauben einen schnellen Überblick über den Inhalt der Werke. So kann man aus der Wortwolke zu Gleditschs botanischen Abhandlungen mit einem Blick wesentliche Dimensionen des Inhalts erfassen.

Abb 2: Gledisch: Botanische Abhandlungen (1789) Quelle: Gleditsch, Johann Gottlieb: Vermischte botanische Abhandlungen. Bd. 1. Berlin, 1789. In: Deutsches Textarchiv, http://www.deutschestextarchiv.de/gleditsch_abhandlungen01_1789 (15.9.2014).

Es entstehen neue ‚Lesemethoden', die man mit Moretti *distant reading* nennen könnte[54] und die nicht nur neue Browsingmöglichkeiten durch eine ansonsten unüberschaubare Textmenge eröffnen, sondern auch die Betrachtung auf die „Textobjekte" wesentlich verändern. Mit Hilfe von *ThemeRiver* Visualisierungstechnik kann man z.B. „thematische Veränderungen in großen Dokumentkollektionen über einen Zeitraum hinweg verfolgen. Die veränderte Breite visualisiert thematische Veränderungen" oder aber man visualisiert bestimmte Themen durch Gebirgslandschaften,[55] so dass z. B. gesellschaftliche Prozesse und Themensetzungen „augenfällig" werden. Das Feld des Text Mining hat zahlreiche Verfahren entwickelt, um aus dem E-Text neue Einsichten zu gewinnen. [56] Ein Beispiel aus „Wer wird Millionär" könnte z.B. mit einer Kookurrenzanalyse beantwortet werden: „Mit wem stand Edmund Hillary 1953 auf dem Gipfel des Mount Everest?", unter den Kookurrenten findet sich die richtige Antwort: Tenzing Norgay.[57] Mittels Clusteranalyse ist es möglich, Dokumente auf die behandelten Themen hin zu untersuchen. Die Themen werden dabei in so genannten Dendrogrammen visualisiert. Gerade die Möglichkeit des Clustering bietet interessante Perspektiven zur Unterstützung der bibliothekarischen Sacherschließung, wie sie erst auf der Basis von Volltexten möglich wird. Gleiches gilt für das Topic Modelling, für das seit einiger Zeit auch freie *tools* zur Verfügung stehen[58]. Mit stilometrischen Verfahren[59] sind Autorenzuweisungen möglich. Dabei steht weniger die Entlarvung von Plagiatoren im Vordergrund als z. B. die Untersuchung von Schriften auf Einflüsse oder auch Einsichten in Arbeitsprozesse.[60] Mittels *Sentiment Analysis* lassen sich Stimmungen einfangen oder Vorurteile identifizieren. GIS Software greift auf Geodaten in Texten zu, Netzwerkanalysen auf Personen, die miteinander in Kontakt stehen. Ein Sonderfall ist die digitale Edition.[61] Sie ist nicht nur ein *tool*, sondern eine neue Publikationsform, die in den Geistes- und Kulturwissenschaften der gedruckten Edition bereits den Rang abgelaufen hat. Sie greift auf digitalisierte Werke zu und baut auf ihnen auf. Gut erfasste Volltexte bilden oft die Grundlage für digitale Editionen. Differenzanalysen der verschiedenen Ausgaben erlauben Kollationsprozesse in einem Umfang, der bisher für unmöglich gehalten wurde.

6. Organisatorische Konsequenzen und neue Dienstleistungen

Für die Bibliothek ergeben sich aus diesen neuen Anforderungen eine Reihe organisatorischer Konsequenzen und neuer Dienstleistungen. Die Aufbereitung von Texten für die Wissenschaft verlangt nach neuen *workflows* und umfasst je nach den finanziellen und technischen Möglichkeiten bzw. dem spezifischen Auftrag der Bibliothek die OCR Transformation, die Lemmatisierung,

54 Moretti, Franco: Conjectures on World Literature. In: New Left Review 1, Jan-Feb 2000. http://newleftreview.org/II/1/franco-moretti-conjectures-on-world-literature (16.9.2014).

55 http://wissensexploration.de/textmining-visualisierung.php (16.9.2014).

56 Heyer, Gerhard u.a.: Text Mining: Wissensrohstoff Text: Konzepte, Algorithmen, Ergebnisse. Herdecke: W3L, 2006.

57 Beispiel nach Heyer (wie Anm. 56), S. 144 f.

58 S. z.B. http://nlp.stanford.edu/software/tmt/tmt-0.4/ (16.9.2014).

59 Wichtig war hier die Arbeit von Burrows, J.: ‚Delta'. A Measure of Stylistic Difference and a Guide to a Likely Authorship. In: Literary and Linguistic Computing 17 (2002), H. 3, S. 267-287.

60 Reynolds, Noel B. u.a.: Who wrote Bacon? Assessing the respective roles of Francis Bacon and his secretaries in the production of his English works. In: Literary and Linguistic Computing 27 (2012), H. 4, S. 409-425. doi: 10.1093/llc/fqs020. http://llc.oxfordjournals.org/content/27/4/409.short?rss=1 (16.9.2014).

61 S. Sahle (wie Anm. 42).

das POS-Tagging (s.o.), die Metadaten- und Strukturdatenerfassung bis hin zur Edition von Werken des kulturellen Erbes. Hinzu kommen qualitätssichernde Aspekte und Anwendung von Instrumenten zur Verbesserung von technisch bedingt fehlerhaften Texten. Die dieserart aufbereiteten Texte des in eine maschinenlesbare Form übersetzen Kulturerbes sind wie Bücher zu archivieren und bilden Sammlungen, die die Linguisten Textkorpora nennen, die die Bibliothek aber anders als die Linguistik nicht nach Sprachen aufbaut, sondern nach den jeweiligen Sammlungsprofilen. Dennoch sind innerhalb des Sammlungsprofils auf der Basis der vergebenen Metadaten differenzierte Klassifikationen, auch nach der Sprache, möglich. Es liegt aber schon hier auf der Hand, dass für spezifische Fragestellungen der Bestand der jeweiligen Bibliothek nicht ausreicht. Daher ist es essentiell, dass Textcorpora frei zur Aggregation mit anderen Textkorpora verbunden werden können, so dass entweder die Bibliothek, die die Forscher betreut, oder der Forscher selbst Texte herunterladen und unter freien Lizenzen nutzen darf. Das umfasst den Download einzelner Texte, von Teilen von Texten (Kapitel, Register, Entitäten, etc.), von zusammengehörigen Texten („Werke"), von ganzen Sammlungen und von Sammlungsteilen (Kriterien: Zeit, Ort, Sachbetreff etc.). Es entstünde so perspektivisch eine Allmende von Texten, die über die jeweiligen Bibliotheksportale bereitgestellt werden. Die Provenienz der Texte müsste schon aus Gründen der Möglichkeit zur Qualitätsbewertung genannt werden wie auch dokumentiert werden muss, von welcher Qualität (Genauigkeit) sie sind und welchen Bearbeitungsstatus sie haben.

Für die Organisation der Bibliothek bedeutet es, dass die Volltextgewinnung in die bestehenden Geschäftsgänge inkorporiert werden muss, z.B. in die integrierte Medienbearbeitung (Erwerbung, Katalogisierung). Das heißt einerseits, dass die Texte wirklich als *file* erworben, meint hergestellt oder kopiert, werden[62], und für deren Langzeitarchivierung gesorgt wird, andererseits dass Erwerbung hier u.U. auch bedeutet, dass kontinuierlich „Neuauflagen", meint verbesserte Versionen, bearbeitet werden müssen, die durch optimierte OCR Prozesse oder manuelle Korrekturen, z.B. per Crowdsourcing oder Kopienvergleich mit anderen Texten, entstehen, wobei eine besondere Herausforderung darin zu sehen ist, verschiedene Varianten ein- und desselben Textes zu einem „besten" Text zu verbinden. Hier ist noch viel Forschungsarbeit notwendig, um einerseits ein pragmatisches Konzept der Versionierung zu etablieren, andererseits aber auch die Zitiersicherheit von solchen Texten sicherzustellen, d. h. durch *persistent identifier* bestimmte Textbereiche und -teile (punktuell und als *range*) zuverlässig ansprechen zu können. Gleiches gilt für die Arbeit am Katalog. Die FRBR bieten anders als die RAK mit dem stabilen Begriff des Werkes zwar gute Grundlagen für die Beschreibung von OCR bedingten Varianten, gleichwohl müssen auch hier im Detail ggf. noch Anpassungen vorgenommen werden, um den Qualitäts- und Annotationsstatus von Dokumenten zu erfassen.

Eine besondere Herausforderung jenseits der reinen Texttransformation besteht in der Aufbereitung, die man als Gegenstand der „Sacherschließung" definieren könnte. Volltexte sollten mit Strukturinformationen angereichert (z.B. mit TEI-Markup), Entitäten markiert, mit einschlägigen Identifiern (GND, TGN, etc.) versehen und ggf. lemmatisiert werden. Verfahren wie *Clustering* und *Topic Modeling* erlauben zudem die traditionelle manuell-intellektuelle Sacherschließung zu erweitern und

62 Oehlmann, Doina: Lizenzen oder Texte, Nutzung oder Hosting? In: Zeitschrift für Bibliothekswesen und Bibliographie 59 (2012), S. 231-235.

Dokumente zumindest semiautomatisch auf der Basis des Volltextes zu erschließen. Auf dieser Grundlage können Volltextkataloge und Suchmaschinen einen direkten Zugriff auf den Text bieten, nicht mehr nur über die Metadaten, sondern in einer Kombination verschiedener Elemente: Volltext, Strukturinformation und Metadaten ließen sich so kombinieren, dass es am Ende möglich sein wird, alle Texte zu extrahieren, die in den Jahren 1620 bis 1640 an der Universität Helmstedt gedruckt wurden und Literatur zitieren, die von Aristoteles handelt und bei denen man bei dieser Gelegenheit noch feststellt, mit welchen Personen diese Arbeiten verbunden sind.

Eine besondere Herausforderung stellt sich für die Benutzung. Sie muss nicht nur dafür sorgen, dass dem Forscher und der Forscherin E-Texte zur Verfügung stehen, sondern auch die Werkzeuge vermitteln, mit denen Volltexte bearbeitet werden können. Ein schönes Beispiel dieser Art der Informationsvermittlung findet man in Angeboten wie in dem Workshop für T*opic Modeling for Humanities Scholars* an der *University of California, Los Angeles* (UCLA).[63]

7. Fazit

Die Bibliothek der Zukunft ist eine Bibliothek der Texte. Angesichts der veränderten Anforderungen an die Nutzung der Bibliothek ist die systematische Transformation des kulturellen Erbes ein Schritt, in der die Volltexterfassung als ein seit Jahren formuliertes Desiderat einen zentralen Platz einnimmt. Mit dem Fortschritt der Imagedigitalsierung und der OCR Technik sind die Voraussetzungen für einen konzertierten Aktionsplan geschaffen. Die Volltextkonversion umfasst nicht nur die Herstellung des *plain text*, sondern auch die Erfassung von Meta- und Strukturdaten, das *tagging* bzw. Annotation der Texte und die Bereitstellung unter freien Lizenzen. Die Bibliothek der Texte ist sichtbar eine andere als die der Bücher. Durch Schrift bestimmte Texte, die durch Transitivität und Prozessierbarkeit gekennzeichnet sind, erfordern nicht nur ein neues theoretisches Fundament, das es erlaubt, sie mit Blick auf ihre Form und Manifestation als E-Texte zu verstehen und zu nutzen, sondern auch nach veränderten Geschäftsgängen und Regelwerken, erweiterten Infrastrukturen und Kompetenzen in den *Digital Humanities*. Wie in der konzertierten Katalogisierung und Imagedigitalisierung des schriftlichen kulturellen Erbes, können Bibliotheken nur in gemeinsamer Anstrengung auch diese Aufgabe bewältigen. Mit der jetzt erfolgten Ausschreibung der DFG[64] „Koordinierte Förderinitiative zur Weiterentwicklung von Verfahren für die Optical-Character-Recognition (OCR)“ bietet sich die Chance, einen wichtigen Schritt in diese Richtung zu tun.

Literaturverzeichnis:

- Barthes, Roland: De l’œuvre au Texte. In: Revue d’esthétique 3 (1971), S. 226.

- Burrows, J.: ‚Delta‘. A Measure of Stylistic Difference and a Guide to a Likely Authorship. In: Literary and Linguistic Computing 17 (2002) H. 3, S. 267-287.

63 http://guides.library.ucla.edu/topic_model (10.9.2014).
64 http://www.dfg.de/foerderung/info_wissenschaft/info_wissenschaft_14_25/ (16.9.2014).

- Caton, Paul: On the term 'text' in digital Humanities. In: Literary and Linguistic Computing, 28 (2013), S. 209-220.
- Cerquiglini, Bernhard: Éloge de la Variante. Histoire critique de la philologie. Paris: Éd. du Seuil, 1989.
- Derrida, Jacques: De la grammatologie. Paris: Éd. de Minuit, 1967.
- Dino Buzetti: Digital Editions and Text Processing. In: Marilyn Deegan (Hg.): Text editing, print and the digital world. Surrey: Ashgate, 2009, S. 45-61.
- Eggert, Paul: Text-encoding, Theories of the Text, and the 'Work-Site'. In: Literary and Linguistic Computing 20 (2005) H. 4, S. 425-435.
- Eisenstein, Elisabeth: The printing press as an agent of change. Vol. I and II. Cambridge: Cambridge University Press, 1997.
- Heyer, Gerhard u.a.: Text Mining: Wissensrohstoff Text: Konzepte, Algorithmen, Ergebnisse. Herdecke: w3L, 2006.
- Justus Lipsius: Sytagma de bibliothecis. Antwerpen: Moretus, 1602. http://diglib.hab.de/drucke/qun-59-9-1/start.htm?image=00011 (17.9.2014)
- Landow, George P.: Hypertext 3.0. Critical Theory and New Media in an Era of Globalization. Baltimore: Johns Hopkins Univ. Press, 2006.
- Moretti, Franco: Conjectures on World Literature. In: New Left Review Nr. 1, Jan-Feb 2000. http://newleftreview.org/II/1/franco-moretti-conjectures-on-world-literature (15.9.2014).
- Oehlmann, Doina: Lizenzen oder Texte, Nutzung oder Hosting? In: Zeitschrift für Bibliothekswesen und Bibliographie 59 (2012), S. 231-235.
- Opitz, Andrea; Stäcker, Thomas: Workshop der Massendigitalisierungsprojekte der Deutschen Forschungsgemeinschaft an der Herzog August Bibliothek Wolfenbüttel. In: ZfBB 56 (2009), S. 363-373.
- Pedauque, Roger T.: Le Document à la lumière du numérique : forme, texte, médium : comprendre le rôle du document numérique dans l'émergence d'une nouvelle modernité. Caen: C & F, 2006.

– Reynolds, Noel B. u.a.: Who wrote Bacon? Assessing the respective roles of Francis Bacon and his secretaries in the production of his English works. In: Literary and Linguistic Computing 27 (2012), H. 4, S. 409-425. doi: 10.1093/llc/fqs020. http://llc.oxfordjournals.org/content/27/4/409.short?rss=1 (16.9.2014).

– Sahle, Patrick: Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels. Teil 2: Befunde, Theorie und Methodik. Norderstedt: BoD, 2013.

– Stäcker, Thomas: Vom Buch zum Text. Sammeln, Erschließen und Benutzen im digitalen Zeitalter. In: Christine Haug; Rolf Thiele (Hg.): Buch – Bibliothek – Region. Wolfgang Schmitz zum 65. Geburtstag. Wiesbaden: Harrassowitz, 2014, S. 353-364.

– Stein, Peter: Schriftkultur. Eine Geschichte des Lesens und Schreibens. Darmstadt: Wiss. Buchges, 2006.

– Terras, Melissa, u.a. (Hg.): Defining Digital Humanities. A Reader. Farnham: Ashgate, 2013.