Literarische und wissenschaftliche Volltexte in elektronischer Form

Im folgenden Beitrag soll von längeren Texten oder umfangreichen Textsammlungen die Rede sein, die in maschinenlesbarer Form vorliegen, ja, das Thema soll im wesentlichen sogar auf die Digitalisierung und Verfügbarmachung bereits gedruckter Texte eingegrenzt werden.[1]

Die Vorteile einer Digitalisierung von Texten für die Verfügbarmachung liegen auf der Hand. Selbst ein gescannter Text, der nur in einem Graphikformat[2] vorliegt, kann von jedem beliebigen Rechner im Netz abgerufen, gelesen und ausgedruckt werden. Dem an seinem häuslichen Schreibtisch arbeitenden Wissenschaftler, Studenten oder interessierten Laien steht so gleichsam auf Tastendruck eine große virtuelle Bibliothek zur Verfügung. In diesem Sinne ist auch das neue Förderprogramm der DFG „Retrospektive Digitalisierung von Bibliotheksbeständen“ [3] zu sehen, das bevorzugt Titel in elektronische Form umwandeln will, die für Forschung und Lehre von besonderer Bedeutung oder schwer zugänglich sind. Technisch ist hier zunächst an die Erstellung von Graphikdateien und die Umwandlung urheberrechtsfreier Titel gedacht, selbst wenn die Kooperation mit Verlagen für neuere Werke nicht ausgeschlossen wird. Neben dem Aspekt der Verbreitung wird aber in jüngster Zeit verstärkt die elektronische Speicherung als Mittel der Bestandssicherung diskutiert.[4] Wesentlich mehr haben allerdings virtuelle Bibliotheken zu bieten, die dem Benutzer nicht nur die Lektüre des Textes, sondern auch eine differenzierte Suche im Text ermöglichen. Hier steht im allgemeinen der Text gar nicht im Vordergrund, sondern vielmehr die Verarbeitung des Textes durch die zugrundeliegende Datenbank, die der Anbieter des Textes entweder mitliefert oder die sich der Forscher selbst erstellen muß. Im allgemeinen dienen diese Texte der schnellen Recherche nach gesuchten Textpassagen, Zitaten oder dem Vorkommen bestimmter Worte und ersetzen die vorher gebräuchlichen Konkordanzen oder Wortlisten, deren Erstellung viel Zeit und Mühe erforderte. Wer einmal versucht hat, das angeblich von Goethe stammende Zitat: „Verflucht, zur rechten Zeit fällt einem nie was ein, und was man Gutes denkt, kommt meist erst hinterdrein“ zu verifizieren, weiß den Wert einer Volltext-Datenbank mit Goethes Werken zu schätzen.[5] Besonders Editoren, die in vergangenen Epochen mitunter tage-, ja wochenlang vermeintlichen oder echten Zitaten und Übernahmen nachjagten, können nun innerhalb weniger Minuten feststellen, daß manche mittelalterlichen Texte (besonders von Theologen, aber auch Juristen) primär aus der Aneinanderreihung von Zitaten anderer Autoren bestehen. Für lexikographische Arbeiten, Stiluntersuchungen, Wörterbücher und linguistische Untersuchungen bieten die elektronischen Versionen unschätzbare Vorteile.

Volltexte in Datenbankform mit entsprechend aufbereiteten Texten (häufig in SGML codiert) werden in vielen Fällen von Verlagen angeboten, während wissenschaftliche Einrichtungen eher dazu tendieren, die Texte als reinen ASCII-Text anzubieten, wobei er vom Nutzer selbst für die Recherche mit einem Volltextdatenbanksystem (WordCruncher[6], TACT[7]) aufbereitet werden muß. Zu beachten ist bei allen diesen Texten die Qualität.[8] Verlagspublikationen geben hier im allgemeinen ihre Vorlagen getreu wieder, besonders wenn die Texte nicht gescannt, sondern von Schreibkräften eingegeben worden sind. Darüber hinaus wird bei der Digitalisierung häufig - um Urheberrecht und damit Tantiemenzahlungen zu umgehen - auf veraltete Textausgaben zurückgegriffen. Der Forscher sieht sich deswegen in der merkwürdigen Lage, daß er für das Auffinden einer Textpassage eine textkritisch veraltete, aber elektronische Ausgabe benutzt, dann aber in der „gedruckten“ kritischen Edition nachschlagen muß, die allein zitierfähig ist.[9] Keineswegs sicher ist es, daß man einen gesuchten Text auch tatsächlich in digitalisierter Form findet, und die Nachweissituation muß als sehr problematisch angesehen werden. So ist es mir z.B. zwar gelungen, eine digitalisierte Fassung von Kafkas Prozeß zu finden, die sich allerdings als eine russische Übersetzung herausstellte und wenige Tage später nicht mehr aufrufbar war.[10]

1 Oßwald, A.: Elektronische Volltexte im Internet. In: ABI-Technik 15,4 (1995) S.415-427. Siehe auch: http://www.ub2.lu.se/UB2proj/LIS_collection/osswald-bt95/bt95-Contents.html

2 Beispiele bei: URL: http://www-fes.gmd.de/library/book1706.html Punkt 1: Elektronische Reprints im Faksimile Format (nicht codierte Information).

3 Koch, M.: Neues DFG-Förderprogramm „Retrospektive Digitalisierung von Bibliotheksbeständen“. ZfBB 44 (1997) S.108-109. Vergl. auch: http://www.dbi-berlin.de/bib_wes/d_lib/d_lib_00.htm

4 Dörr, M.; Weber, H.: Digitalisierung als Mittel der Bestandserhaltung? ZfBB 44 (1997) S.53-75. Vergl. aber auch: Jochum, U.; Wagner, G.: Cyberscience oder vom Nutzen und Nachteil der neuen Informationstechnologie für die Wissenschaft. ZfBB 43 (1996) 579-93.

5 Der Leser möge nur einmal versuchen, dieses Zitat in Goethes Werken zu finden und dann die hierfür benötigte Zeit mit den wenigen Minuten vergleichen, die eine Recherche in der in der Universitätsbibliothek angebotenen Volltextdatenbank erfordert.

6 URL: http://www.wordcruncher.com/software.html.

7 URL: http://www.epas.utoronto.ca:8080/cch/TACT/tact0.html.

8 Man vergißt leicht, daß beim Scannen eine an sich gering erscheinende Fehlerquote von unter einem Prozent im allgemeinen bereits einen Fehler für je zwei (!)Zeilen (120 Buchstaben) bedeutet.

9 Jochum, a.a.O. S.581.

10 Verzeichnet in: http://www.modcult.brown.edu/people/Scholes/modlist/GtoK.html unter Kafka.

Textcorpora auf CD-ROM

Hierbei sind vor allem die umfangreichen CD-ROM-Ausgaben des Verlages Chadwyck-Healey[11] zu nennen, der nicht nur die 221 Bände der Patrologia Latina des Abbès J. P. Migne aus dem 19. Jahrhundert anbietet (PLD), sondern auch die Sophienausgabe der Werke Goethes oder fast die gesamte englische Dichtung der Frühzeit. Die Literatur der klassischen Antike kann man ebenso auf CD erhalten wie die Werke Kants oder Thomas von Aquins.[12] In jedem Fall ist es sinnvoll, die CD-ROM-Angebote der Universitätsbibliotheken und deren Erschließung zu studieren.[13] Dringend notwendig wäre es allerdings, daß die Bibliotheken auch alle in diesen Textcorpora enthaltenen Einzelwerke in ihren Katalogen nachweisen und auf die elektronische Ausgabe verweisen würden.[14] Denn zur Zeit stehen wir vor der fast absurden Situation, daß zwar eine kleine Broschüre von nur fünf Seiten in den Katalogen nachgewiesen wird, nur weil sie gedruckt vorliegt, eine mehrere hundert Seiten lange Abhandlung eines mittelalterlichen Autors aber dort nicht erscheint, nur weil sie nicht als eigene Ausgabe, sondern als Teil der PLD vorhanden ist. Hier müssen sich die Bibliothekare unbedingt von ihrem Regelwerk verabschieden, das eine solche Verzeichnung nicht vorsieht und im übrigen auch die Belange der EDV nur unzureichend berücksichtigt.

11 http://www.chadwyck.co.uk. Vergleiche auch den ausführlichen Prospekt dieses Verlages: „Electronic Publications 1996“, der auf 40 Seiten ca. 90 Volltext- und Faktendatenbanken anbietet. Zu beachten ist, daß Chadwyck-Healey seit einiger Zeit die zunächst nur auf CD-ROM angebotenen Datenbanken gegen eine entsprechende Jahresgebühr (abhängig von der Zahl der gleichzeitigen Zugriffe) auch im Internet zur Verfügung stellt.

12 Die Universitätsbibliothek der Humboldt-Universität bietet zur Zeit Volltextdatenbanken der Werke Kants, Goethes, Thomas von Aquins und die umfangreiche, leider nur unzureichend erschlossene elektronische Version von Mignes Patrologia Latina (Patrologia Latina Database) an. Hier wird man häufig doch wieder auf die gedruckten Erschließungsmittel (etwa: Eligius Dekkers, Clavis patrum latinorum, Steenbrugge 1961) zurückgreifen müssen. Diese Datenbank ist im übrigen ein Musterbeispiel für die Veröffentlichung von veralteten Editionen in elektronischer Form. Die von Migne dargebotenen Texte sind im allgemeinen völlig unveränderte, nur um weitere Druckfehler bereicherte Nachdrucke von Editionen des 16.-18. Jahrhunderts, und selbst die Zuschreibungen an bestimmte Autoren sind in etlichen Fällen falsch. Um einen dort gefundenen Text zitieren zu können, muß in jedem einzelnen Fall überprüft werden, ob nicht eine neuere, zitierfähige Edition vorliegt. Die Ausgabe der Autoren der klassischen Antike (Datenbank Phi - Packard Humanities Institute) kann über das Institut für Geschichtswissenschaften (Tel. 2093-1635) benutzt werden.

13 Die Universitätsbibliothek der Humboldt-Universität bietet ihre CD-ROMs über das Bibliotheksmenü der Fa. H+H (Göttingen) an. Die Volltextdatenbanken sind als solche explizit gekennzeichnet und finden sich unter den jeweiligen Wissenschaftsbereichen, also in der Regel in dem Menüpunkt Geisteswissenschaften. Da unter der Windows-Oberfläche laufende CD-ROMs zur Zeit noch gesondert aufgerufen werden müssen, ist es leider notwendig, auch noch unter dem Menüpunkt CD-ROMs / Windows zu recherchieren. In der Informationsvermittlungsstelle der Universitätsbibliothek wird laufend ein Verzeichnis aller an der Universität vorhandenen CD-ROMs geführt, das zur Zeit telefonisch abgefragt werden kann. Eine Publizierung im WWW ist in Vorbereitung.

Textcorpora und Texte im Internet

Bei der Suche im Internet kann man eine Fülle von Texten finden, die in mehr oder weniger guter Qualität die elektronische Fassung mehr oder weniger berühmter Werke aus allen Bereichen der Wissenschaften darstellen. Einen befriedigenden Katalog dieser virtuellen Bibliothek gibt es leider nicht[15], und die vorhandenen Übersichten oder alphabetisch geordneten Listen[16] sind bisweilen etwas gewöhnungsbedürftig oder sortieren die Autoren nach Vornamen.[17] Eine weitere Möglichkeit besteht natürlich immer darin, den Namen des Autors und/oder des gesuchten Werkes über eine der großen Suchmaschinen zu recherchieren.
Von besonderem Interesse sind allerdings die größeren Projekte, von denen die wichtigsten hier erwähnt werden sollen:

- Projekt Gutenberg: Vom ursprünglich amerikanischen Projekt[18], dessen Ziel es war, die Erstellung und Verbreitung von englischsprachigen Texten zu erreichen, ist nun auch ein deutscher Ableger vorhanden, der überwiegend literarische Texte (meist geringeren Umfangs) im Netz anbietet.[19]
- Oxford Text Archive: Das Oxford Text Archive[20] bietet über 1500 elektronische Texte, die aber nicht direkt gelesen werden können, sondern (insofern sie öffentlich erhältlich sind) über FTP abgerufen werden müssen. Es enthält Werke der verschiedensten Sprachen (darunter eine große Zahl von griechischen, lateinischen und englischsprachigen Autoren sowie auch etliche Hauptwerke der deutschen Literatur), wobei die Grundlage häufig die maßgeblichen textkritischen Ausgaben sind. Die Erstellung der E-Texte erfolgt durch verschiedenste Forscher und Forschergruppen. Leider sind nicht alle Texte frei erhältlich, manche sind nur für den Gebrauch vor Ort zugelassen, und für andere ist die vorherige Erlaubnis des Herstellers nötig. Die öffentlich erhältlichen Texte

14 Ein ähnliches Problem besteht bei den großen Textausgaben auf Microfiche, wo ebenfalls die Verzeichnung der in den jeweiligen Ausgaben enthaltenen Einzeltexte ein bisher ungelöstes Problem darstellt, das wohl nur in der Kooperation zwischen verschiedenen Bibliotheken gelöst werden kann.

15 Vergl.: Alex: A catalog of electronic texts on the Internet. URL: http://www.lib.ncsu.edu/stacks/alex-index.html

16 Oßwald, A.: Reprints im Internet - Links zu im Vortrag erwähnten Quellen (URL: http://www-fes.gmd.de/library/book1706.html) zählt eine Reihe solcher Listen bzw. Quellen auf. Vergl. auch: An index of web sites on modernist artists: URL: http://www.modcult.brown.edu/people/Scholes/modlist/AtoF.html bzw. GtoK.html bzw. LtoR.html bzw. StoZ.html. Es ist mir leider nicht gelungen, den Index zu diesen sehr nützlichen, offenbar von Robert Scholes gestalteten Seiten ausfindig zu machen.

17 etwa: Virginia Tech Etexts: URL: gopher://gopher.vt.edu:10010/10/33

18 URL http://www.promo.net/pg/index.html. Hier findet sich auch ein Menüpunkt: „Other Etext Archives“, der sehr nützliche Informationen bringt.

19 http://gutenberg.informatik.uni-hamburg.de/gutenb/home.html

20 URL: http://weewannie.library.ubc.ca/ota/welcome.html

liegen im SGML-Format vor und sind nach den Vorgaben der TEI[21] ausgezeichnet.

- The Electronic Text Center and On-Line Archive of Electronic Texts: University of Virginia.[22] Diese Sammlung umfaßt über 10.000 Texte und bietet neben den nicht frei verfügbaren großen CD-ROM-Textcorpora eine große Zahl von literarischen und historischen Texten aus dem 18. und 19. Jahrhundert an, die überwiegend aus dem anglo-amerikanischen Bereich stammen. Etliche französische, deutsche und lateinische Titel finden sich ebenfalls. Alle Texte sind im SGML-Format erfaßt und werden für die Nutzung nach HTML konvertiert.
- Projekt WebDOC.[23] Innerhalb dieses internationalen Gemeinschaftsprojektes, das unter Federführung der Pica-Stiftung in Leiden durchgeführt wird, verzeichnen zur Zeit (Stand 10.1.97) 14 deutsche und niederländische Bibliotheken, die amerikanische Research Libraries Group RLG und internationale Verlage die von ihnen erworbenen bzw. herausgegebenen elektronischen Dokumente. Über eine gemeinsame Datenbank sind diese Publikationen abrufbar, sofern die entsprechenden Berechtigungen vorliegen. Die Berechtigungen sind in drei Klassen eingeteilt: Frei verfügbar, nur für Angehörige einer bestimmten Universität verfügbar, nur gegen Gebühr verfügbar. Jede der beteiligten Bibliotheken verfügt über einen eigenen Dokumentenserver. Das Projekt WebDOC ist zur Zeit noch im Aufbau, es finden Verhandlungen mit einer größeren Zahl potentieller Teilnehmer statt, um weitere Dokumente und vor allem Zeitschriften anbieten zu können. Auch die Humboldt-Universität wird sich überlegen müssen, ob eine Beteiligung sinnvoll ist, denn jetzt erscheint schon bei dem Versuch, nur auf den Katalog zugreifen zu wollen, die Meldung: „You are not authorized“.

Hier konnten natürlich nur einige wichtige Einrichtungen, Projekte und Organisationen dargestellt werden, die elektronische Volltexte anbieten. Sie sollen Appetit auf mehr machen und dazu verleiten, sich auf die Suche zu machen, wenn man einen Text in einer elektronischen Fassung benötigt. Viel Arbeit wartet noch auf die Bibliothekare, die es in internationaler Kooperation unbedingt in die Wege leiten müßten, daß alle diese Texte über eine Suchoberfläche recherchierbar gemacht werden. Sicherlich wird man es nicht erreichen, daß die Metadaten aller Texte in einer Datenbank vorgehalten werden können. Dies ist aber auch nicht notwendig, wenn man sich auf eine einheitliche Struktur der Metadaten und Standardschnittstellen einigt. In einem solchen Fall könnte nach Art des Karlsruher Virtuellen Katalogs in den unterschiedlichsten Datenbanken recherchiert werden, ohne daß der Benutzer merkt, daß es sich um verschiedene Quellen handelt. Trotz der vielfältigen Möglichkeiten, heute schon Texte in elektronischer Form bekommen und damit auch maschinell analysieren zu können, sei eine ketzerische Bemerkung zum Schluß gestattet: Wer ein Buch einfach nur lesen will, ist vielleicht mit einem Gang in die Bibliothek und der Ausleihe der Papierversion noch am besten bedient.

Norbert Martin
Universitätsbibliothek

21 URL: http://etext.virginia.edu/TEI.html

22 URL: http://www.lib.virginia.edu/etext/ETC.html. Vergl. auch: Seaman, David: The Electronic Text Center and On-Line Archive of Electronic Texts. In: Elektronisches Publizieren und Bibliotheken. Hrsg. v. K. W. Neubauer. Frankfurt/M. 1996 (Zeitschrift für Bibliothekswesen und Bibliograhie. Sonderheft 65), S. 55-57.

23 URLs: http://www.gwdg.de/~sub/0_digbib.htm; http://www.pica.nl/docs/en/webdoc/webproj.html.