

Korpusbildung

GERHARD LAUER

Die Weltbibliothek und ihre Korpora. Bibliothekarische Infrastruktur als Bedingung der Möglichkeit zur Korpusbildung

Die Rede von der Weltbibliothek ist keine Metapher mehr. Immer größere Bestände sind digital erschlossen, wenn nicht digital erzeugt, und lassen sich zu Bibliotheken zusammenführen, die in ihrer Größe und Erschließungstiefe noch vor kurzem undenkbar waren. Doch der Mehrwert liegt nicht in der Weltbibliothek als solcher, sondern in der Möglichkeit, aus Bibliotheken hochauflösende Korpora zusammenstellen zu können. Forschung auch und gerade an der kulturellen Überlieferung wird zunehmend computerbasierte Forschung sein, die auf jeweils neu erstellten Korpora basiert. Bibliotheken haben die Chance, solche Forschung wesentlich mit zu ermöglichen.

»World library« is no longer merely a metaphor. Ever larger numbers of publications are now digitally catalogued, if not digitally born, and can be combined to create libraries, the size and cataloguing detail of which would previously have been unthinkable. However the true value of the world library does not lie in its mere existence, but in the possibility to put together high resolution corpora from different libraries. Research, especially into cultural heritage, will increasingly be carried out using computers and based on newly created corpora. Libraries have the opportunity to play a major role in facilitating such research.

Der Buchhistoriker und Bibliothekar Robert Darnton hat jüngst in der *New York Review of Books* ein leidenschaftliches Plädoyer für eine Weltbibliothek gehalten, genauer für eine digitale Weltbibliothek.¹ Die sei als eine öffentliche Einrichtung umso nötiger, als die Kommerzialisierung des Wissens durch wenige Wissenschaftsverlage und Internetfirmen die Erkenntnismöglichkeiten immer mehr einschränke. Mit einer durchschnittlichen Preissteigerung für eine Fachzeitschrift in der Chemie von 33 Dollar in den 1970er-Jahren auf mehr als 4.000 Dollar heute, mit der Oligopolisierung des Wissens unter Google, Reed Elsevier, Wiley-Blackwell und Springer wird mit öffentlichen Geldern das Wissen privatisiert. Demokratisierung des Wissens ist das Gebot der Stunde, nun einer langen Stunde, wie wir nur zu gut wissen. Denn alle Versuche, das Urheberrecht in Deutschland wissenschaftsfreundlicher zu machen oder in den USA durch den Fair Access to Science and Technology Research Act (FASTR) den freien Zugang zu den öffentlich finanzierten Publikationen und Daten zu erreichen, treffen auf Widerstand, selbst wenn der volkswirtschaftliche Schaden noch so hoch sein mag.²

Dass andere Wege möglich sind, zeigen Beispiele wie das arXiv.org-Repository, die Public Library of Science (PLOS) oder die OpenEdition Books jeden Tag aufs Neue. Für Darnton ist auch der erfolgreiche Start der Digital Public Library of America 2013 ein weiteres

Exemplum für die Möglichkeit eines weltweiten Zugangs zum Wissen, das gerade das digitale Zeitalter in einem ganz anderen Maßstab erschließt. Der Start der vollständigen Version der Deutschen Digitalen Bibliothek am 31. März 2014 im Rahmen der Europeana-Initiative fügt sich hier ein. Immer ist die Idee eine ähnliche, das kulturelle Wissen möglichst frei zugänglich zu machen, das in vielen Sprachen und ohne Barrieren. Das alles ist nicht nur für Robert Darnton Grund, die Formulierung von der Weltbibliothek nicht nur als eine Metapher oder Utopie zu verwenden. Eine solche Bibliothek hat keinen bestimmten Ort, sondern viele. Sie ist nicht in einer Sprache abgefasst, sondern in hunderten. Und ihre Leserinnen und Leser sind über die Welt verteilt. Gemeinsame Standards der Metadaten, Text- und Bildauszeichnung ermöglichen es, dass die Daten zu einer Bibliothek zusammengeführt werden können. Von Big Data hier zu sprechen, ist keine Übertreibung, wenn schon eine Spezialsammlung wie die Perseus Digital Library³ Hunderttausende von Texten und Zehntausende von archäologischen Objekten für die Klassische Philologie bereit hält und jeden Monat mehr als 300.000 Leser diese Bibliothek nutzen, unter ihnen mehr als 35.000, die die Texte in Latein oder Griechisch lesen. Aber nicht nur von den großen Daten ist zu reden. So verweist Darnton auf das Emily Dickinson Archive seiner Universität Harvard,⁴ das aus unterschiedlichen lokalen Archiven und verschiedenen Ausgaben eine Bibliothek fügt, die wiederum Teil einer weltweiten Bibliothek ist. Da in den USA fast jedes Kind mit den Gedichten Dickinsons in Kontakt kommt, kann es hier die unterschiedlichen Ausgaben vergleichen, Scans der Handschriften, TEI-ausgezeichnete Gedichttexte und Metadaten zu jedem einzelnen Gedicht studieren. Zugespitzt gesagt, bedingen Weltbibliothek und Digitalität einander.

VOM KLASSISCHEN SCHREIBTISCH ZUM DIGITALEN KORPUS

Das alles ist ermutigend, weitert und vertieft das, was Bibliotheken immer schon ausgemacht haben. Und doch könnten und sollten wir einen Schritt weiter gehen. Und dieser Schritt betrifft nun die Bibliotheken in ihrem Selbstverständnis als Infrastrukturen des Wissens. Kurz gesagt, wäre es mindestens aus Sicht der



Gerhard Lauer

Robert Darntons
Plädoyer für eine digitale
Weltbibliothek

neue Wege des wissenschaftlichen Publizierens

Wissenschaft wünschenswert, wenn Bibliotheken die Bildung von Korpora unterstützen würden. Denn alle modernen Wissenschaften beruhen auf Daten, das mögen Beobachtungsdaten in den Naturwissenschaften sein, können in den Geisteswissenschaften Texte oder Objekte sein, in einem Fall etwa möglichst viele und verlässliche Belege für rotfigurige Vasen, wie sie die klassische Archäologie ausgegraben hat und in verschiedenen Sammlungen über die Welt verstreut zu finden sind, in einem anderen Fall handelt es sich um möglichst viele und verlässlich erstellte Belege für Märchenmotive in den literarischen Überlieferungen der Welt. Diese Daten – oft sind es dann auch Big Data – sind zumeist nicht an einem Ort und doch müssen sie für die jeweilige Forschungsfrage zu einem Untersuchungskorpus zusammengestellt werden. Dazu müssen die Daten aus verschiedenen Bibliotheken und Sammlungen der Welt über standardisierte Datenmodelle und Schnittstellen an einem Ort zusammengeführt werden können. Wenn diese Daten dann nach Abschluss der Forschung für weitere Arbeiten oder auch für die Validierung der Ergebnisse weiterhin zur Verfügung stünden, wäre dies genau das, was Forschungsinfrastrukturen zu leisten haben. Insofern sind Bibliotheken seit den Tagen Francis Bacons Teil der Modernisierung des Wissens, auch wenn sie gerne als Gegenfantasie inszeniert werden.

Editionen wie das Emily Dickinson Archive sind sprechend erst in den Zusammenhängen, in denen sie stehen. Zumeist bildet man bei Editionen auch im digitalen Zeitalter ab, was man schon früher in Bibliotheken getan hat. Zu diesem Gedicht Dickinsons wird man vielleicht jene Ausgabe der Gedichte ihres Zeitgenossen Walt Whitmann oder die Lyrik des englischen Metaphysikers George Herbert aus dem 17. Jahrhundert heranziehen und so einen Schreibtisch digital nachbilden, der so auch in einer analogen Bibliothek hätte stehen können. Das ist natürlich sinnvoll und wird unverändert getan. Aber die spezifischen Möglichkeiten einer Weltbibliothek, von der Robert Darn-ton in seinem Artikel redet, sind damit nur zum Teil genutzt. Statt nur die bislang möglichen Fragen an das Werk Dickinsons zu richten, kann die lesende Neugierde oder wissenschaftliche Fragestellung auf noch wenig begangene Wege gebracht werden. Wie unterscheidet sich Dickinsons lyrische Sprache von der ihrer Zeit, welchen Texten ähnelt sie? Wann wird Emily Dickinson zu einer Autorin, deren Gedichte in Schulbüchern gedruckt werden? Wo auf der Welt werden ihre Gedichte in anderen Sprachen gelesen und zitiert? Gehört ihr Werk zum kulturellen Gedächtnis der Welt, das in Wikipedia-Artikeln, in Twitter-Nachrichten oder Fanfiction aufgerufen wird? Solche und ähnliche Fra-

gen zu beantworten setzt voraus, dass die Texte Dickinsons Teil eines größeren, je nach Fragestellung zu bauenden Korpus werden. Noch genauer eines Korpus, das am klassischen Schreibtisch nicht zusammengeführt werden kann, entweder weil es zu groß ist, um überblickt werden zu können oder weil es zu divers ist, um es in einem Raum aufstellen zu können. Dickinsons Gedichte mit der lyrischen Produktion im Amerika des langen 19. Jahrhunderts vergleichen zu können, ihre stilistischen Eigenheiten oder die Besonderheiten ihrer Themen mit Verfahren wie Stylometrie oder Topic Modeling zu untersuchen oder ihren Einfluss auf spätere Lyriker bis hin zu Paul Celan zu ermitteln, setzt voraus, dass die Gedichte aus dem Emily Dickinson Archive mit anderen Sammlungen zu einem Korpus zusammengeführt werden. Diese Korpora können sehr groß sein, größer als Sammlungen, die man noch vollständig lesen kann, und sie können aus verschiedenen Archiven oder Editionen auf der Welt stammen. Texte mag man sich dabei noch irgendwie in einen Raum passend vorstellen. Aber wie sieht es mit Vertonungen ihrer Gedichte aus, Adaptionen in der Kunst wie etwa Judith Chicagos *Dinner Party*-Installation?⁵

KOMBINATION UND REKOMBINATION

Eine Weltbibliothek im digitalen Zeitalter ist nicht mehr so strikt an die räumlichen Bedingungen gebunden, wie noch ihre Vorgängerin. Um daher ihr besonderes Potential zu heben, muss sie es ermöglichen, dass ihre Bestände mit denen anderer Bibliotheken, Archive und Sammlungen zu Korpora verknüpft werden können, digital und konzeptuell. Wie groß das Korpus ist, wie heterogen es unterschiedliche Gattungen und Medien integriert, all das hängt an der jeweiligen Forschungsfrage. Wenn daher mit einigen guten Gründen so viel von der Weltbibliothek die Rede ist, dann kann das nicht nur eine irgendwie größere Bibliothek als die uns vertrauten meinen. Big Data ist bestenfalls die Hälfte der Innovation. Von der Weltbibliothek kann nur dann sinnvoll die Rede sein, wenn diese ihre Bestände für die verschiedenen Anfragen so zur Verfügung stellt, dass eine Rekombination zu jeweils neuen Korpora unterstützt wird und die Ergebnisse einschließlich der Primärdaten vorgehalten werden. Manche wie der Gründer der Creative Commons Licence Lawrence Lessig sprechen von »Remix« als Kennzeichen digitaler Kultur⁶ und heben dabei darauf ab, dass hergebrachte Verknüpfungen kultureller Gegenstände aufgelöst und neue Verbindungen möglich werden. Das gilt auch für Bibliotheken, die von ihren Schatzkammern wohl dann erst Gebrauch machen, wenn sie den Remix neuer und auch unerwarteter Korpora ermöglichen.

Bei dem Begriff des Korpus denkt man zumeist an große Textsammlungen für sprachwissenschaftliche Untersuchungen, etwa an das Korpus der Zeitungssprache des Instituts für Sprache in Mannheim und vergleichbare Korpora.⁷ Sie kennzeichnet es, dass ihre Sprachdaten strukturiert sind, grammatische Kategorien wie Nomen oder Verb durch Part of Speech-Tagger ausgezeichnet und auffindbar gemacht worden, Tiefenstrukturen in Baumbanken abgebildet sind und so in sehr großen Datenmengen Abfragen gestartet werden können, etwa über die Entwicklung der Verb-Endstellung im Deutschen. Was für die Linguistik gilt, gilt auch für verwandte Korpora anderer Disziplinen. Die Archäologie, die längst schon eine digitale Archäologie ist, verzeichnet Kataloge mit Tausenden von Objekten, seien es etwa Vasen oder Statuen.⁸ Die Musikwissenschaft, die Noten mit den Standards der Music Encoding Initiative annotiert,⁹ baut vergleichbare Korpora in ihrem Feld auf, die es erlauben, die Entwicklung rhythmischer oder melodischer Muster über große Zeiträume zu untersuchen. Die Kunstgeschichte nutzt über digitale Edition wie etwa des Werkes Vincent van Goghs¹⁰ oder Leonardo da Vincis¹¹ im wachsenden Maße digitale Verfahren, um Farbähnlichkeiten, Röntgenaufnahmen oder Kontrastverhältnisse zur Grundlage von Sammlungen zu machen.

Mag auch die Geschwindigkeit, mit der computerbasierte Verfahren in den Geisteswissenschaften selbstverständlicher Teil dieser Disziplinen werden, unterschiedlich sein, die Entwicklung selbst ist kaum zu unterschätzen. Und das gilt nicht nur für die Wissenschaften im engeren Sinn. Die Wiedereröffnung des Amsterdamer Rijksmuseums geht mit einem digitalen Rijksstudio einher, das es auch Laien ermöglicht, die Amsterdamer Sammlung mit ihren mehr als 150.000 Werken mit computergestützten Methoden zu erkunden, persönliche Sammlungen anzulegen, und das auch für nicht-akademische Zwecke.¹² Sammlungen, wie sie das Rijksstudio ermöglicht, sind Korpora, was immer der Zweck ist, der zu ihrer Zusammenstellung geführt hat. Von einer Weltbibliothek können wir aber erst reden, wenn solche Korpora nicht nur innerhalb einer bestehenden Institution wie dem Rijksmuseum in Amsterdam, sondern über die Grenzen solcher Einrichtungen hinweg erstellt werden können. Darin erst liegt die Zukunft der Weltbibliothek. Deshalb sind Korpora der Prüfstein, ob Bibliotheken infrastrukturell mehr aus ihren Beständen zu machen wissen.

NEUE KORPORA FÜR NEUES WISSEN

Bibliotheken hatten und haben viele Funktionen und haben deshalb viele Formen in ihrer Geschichte angenommen, seien es klösterliche Studiensäle oder repräsentative Wunderkammern, Bildungsstätten für die Vielen und Laborstätten für die Wenigen und oft Mischungen aus alle dem. Der Komparatist Jeffrey Schnapp argumentiert zusammen mit seinem Kollegen Matthew Battles in ihrem Buch *The Library beyond the Books*, dass die Bibliothek der Zukunft Bücher mit digitalen Büchern, analoge mit digitalen Formaten und Papier mit Pixeln verbinden werde.¹³ Der Remix der Formate und Medien, der Gattungen und der Inhalte, hat aber nur Sinn, wenn damit je neue Korpora aufgebaut werden können, die erst die lesende Neugierde hervorbringt. Reservoirfunktionen und Remix schließen sich nicht mehr aus. Weltbibliothek ist dann nicht nur eine versehentlich im digitalen Zeitalter zu groß geratene Bibliothek, sondern eine Erkundungsstätte für neues Wissen.

Begriff des Korpus

Reservoirfunktion und Remix

Korpora in den Kunstwissenschaften

¹ Robert Darnton, A World Digital Library Is Coming True, in: The New York Review of Books 61, 9 (22. Mai 2014), www.nybooks.com/articles/archives/2014/may/22/world-digital-library-coming-true/ (zuletzt aufgerufen 15.05.2014).

² Vgl. John Houghton et al., Economic Implications of Alternative Scholarly Publishing Models. Exploring the Costs and Benefits, JISC E-ASPM Project, Report to the Joint Information Systems Committee (JISC) (UK), CSES and Loughborough University, January 2009, www.jisc.ac.uk/publications/reports/2009/economicpublishingmodels/finalreport.aspx (zuletzt aufgerufen 15.05.2014).

³ Perseus Digital Library, www.perseus.tufts.edu (zuletzt aufgerufen 15.05.2014).

⁴ Emily Dickinson Archive, Harvard University, www.edickinson.org (zuletzt aufgerufen 15.05.2014).

⁵ Judith Chicago, The Dinner Party, 1974–79, www.brooklynmuseum.org/eascfa/dinner_party/place_settings/emily_dickinson.php (zuletzt aufgerufen 15.05.2014).

⁶ Lawrence Lessig, Remix. Making Art and Commerce Thrive in the Hybrid Economy, New York 2008, <http://remix.lessig.org/> (zuletzt aufgerufen 15.05.2014).

⁷ Das Deutsche Referenzkorpus – DeReKo, www1.ids-mannheim.de/kl/projekte/korpora/ (zuletzt aufgerufen 15.05.2014).

⁸ Vgl. z. B. den Blog archäologiedigitale, <http://archdigi.hypothesess.org/> (zuletzt aufgerufen 15.05.2014).

⁹ Music Encoding Initiative (MEI), <http://music-encoding.org/> (zuletzt aufgerufen 15.05.2014).

¹⁰ Vincent van Gogh, *The Letters*, www.vangoghletters.org/vg/ (zuletzt aufgerufen 15.05.2014).

¹¹ Universal Leonardo, www.universalleonardo.org/ (zuletzt aufgerufen 15.05.2014).

¹² Rijks Studio, <https://www.rijksmuseum.nl/en/rijksstudio> (zuletzt aufgerufen 15.05.2014).

¹³ Jeffrey Schnapp / Matthew Battles, *The Library Beyond the Book*, Cambridge/MA 2014.

DER VERFASSER

Prof. Dr. Gerhard Lauer, Professor für Deutsche Philologie, Universität Göttingen, Seminar für Deutsche Philologie, Käte-Hamburger-Weg 3, 37073 Göttingen, Tel.: 0551 – 39 7526, E-Mail: gerhard.lauer@phil.uni-goettingen.de