Von Lukas Rosenthaler und Peter Fornaro

Was ist «Big Data» überhaupt? Der Begriff ist seit einiger Zeit omnipräsent, wenn es um Diskussionen zum Thema Internet, Speicherung oder um Chancen und Bedrohung durch die rasant fortschreitende Digitalisierung unserer modernen Gesellschaft im Informationszeitalter geht. Wir lesen oder hören fast täglich wie die grossen Internetkonzerne, beispielsweise Google oder Amazon, Daten über unser Verhalten im Netz sammeln. Zudem, wie dank Edward Snowden bekannt wurde, müssen wir stetig damit rechnen, dass die Geheimdienste überall mitlauschen und alle Informationen über unsere Aktivitäten im Internet speichern, derer sie habhaft werden können. Doch was ist Big Data wirklich?

In einem Satz erklärt, bezeichnet Big Data das Verfahren, kleine und kleinste Daten über eine grosse Menge von Objekten (z.B. Menschen) und deren Aktivitäten möglichst komplett zu erfassen und zu speichern, um sie zu einem späteren Zeitpunkt mit Hilfe vom Algorithmen zu durchsuchen und retrospektiv aufgrund gewisser Fragestellungen individuelle Verhaltensweisen zu bestimmen und Voraussagen über zukünftiges Verhalten zu machen. So hofft z.B. Amazon, auf Grund gesammelter Aktivitäten der Kunden, wie etwa getätigter Bestellungen oder der Wahl von Links auf der Webseite, Voraussagen über das Kaufverhalten bei neuen Produkten prognostizieren zu können und diese schon für den Versand vorzubereiten, bevor der Kunde die entsprechende Bestellung ausgelöst hat[1].

Damit solche (Alb-)Träume wahr werden können, müssen zwei technische Voraussetzungen erfüllt werden:

- Die Speichertechnologie muss in der Lage sein, riesige Datenmengen im mehrstelligen Petabyte-Bereich (1 Petabyte [PB] = 1000 Terabyte = 1 000 000 Gigabyte) effizient und schnell zugreifbar zu speichern. Dies ist heute mit modernen Festplatten und vernetzten Speichersystemen gegeben. Eine Manifestation davon sind die sog. Cloud-Speicher, bei denen die Daten auf verschiedene, grosse, vernetzte Datenfarmen verteilt werden. Der Zugriff erfolgt über das Internet, ohne dass bekannt sein muss, wo die Daten physisch gespeichert sind.
- Zweitens braucht es Analyseverfahren, welche auf riesige Datenmengen angewendet werden können. Dazu wurden in der Informatik neue, verteilt operierende Algorithmen entwickelt, welche diese enorm grossen Datenmengen effizient nach gewissen Mustern durchsuchen, um daraus Schlussfolgerung zu ziehen. Ein typischer Vertreter solcher Analyse- und Prognoseverfahren stellen die heutigen Methoden für die Wettervorhersage dar. Dabei wird eine sehr grosse Menge an Messwerten analysiert, um daraus das Wetter der kommenden Tage zu ermitteln.

Big Data, wie oben beschrieben, ist heute Realität. Die Datenspuren, welche wir beim Surfen im Internet, bei Online-Einkäufen und -Bestellungen etc. hinterlassen, werden gespeichert und analysiert. Durch die Kombination mit Handydaten, Punktesystemen der Grossverteiler, Kreditkarteninfos etc., aber auch durch das – mehr oder weniger – freiwillige und bewusste Freigeben von Information in den sozialen Netzwerken wie Facebook etc., werden wir – negativ ausgedrückt – alle in einem erschreckenden Grad zu «programmiert» agierenden Konsumenten mit einem weitgehend voraussagbaren Verhalten. Oder geschieht dies alles nur – wie uns dies die Prediger des digitalen Fortschritts und die Internetkonzerne glauben machen wollen – zu unserem Wohl, damit uns ein noch viel besserer individueller Service geboten werden kann? Wie oft haben wir uns schon gewundert, dass bei der Suche im Internet Werbung eingeblendet wird, welche durchaus einen gewissen Bezug zu unseren Interessen hat. Die Frage, ob dies letztendlich eine positive oder beängstigende Entwicklung ist, muss jede Nutzerin und jeder Nutzer des Internets für sich selbst beantworten. Rückgängig machen können wir diese Entwicklung wohl nicht mehr.

Sicher ist, unsere Daten bilden auch einen pekuniär wertvollen Rohstoff, mit dem ein florierender Handel betrieben wird. Nur so ist es überhaupt möglich, dass z.B. die Betreiber der Suchmaschinen ihre enorm aufwändigen Services «kostenlos» anbieten und dennoch ein stetes Wachstum und hohe Gewinne für die Aktionäre ausweisen können. Im Folgenden soll aufgezeigt werden, welchen Einfluss Big Data auf die Geisteswissenschaften und unser Kulturerbe haben könnte.

Big Data in den Geisteswissenschaften und für das Kulturerbe

Um die Frage, ob Big Data in den Geisteswissenschaften und für den Erhalt unseres Kulturerbes eine Rolle spielt, zu beantworten, muss zunächst einmal analysiert werden, was der heutige Stand der Digitalisierung in diesen beiden Bereichen ist:

- Die effiziente Digitalisierung von historischen Quellen und kulturellen Artefakten ist seit einigen Jahren technisch möglich.
- Die digitalen Kameratechniken übertreffen heute an Qualität und Effizienz die analogen Verfahren.
- 3D-Aufnahmeverfahren erlauben es seit kurzem, die Geometrie von Volumenobjekten mit vergleichsweise kleinem Aufwand genau zu erfassen.
- Für viele Textdokumente existieren brauchbare OCR-Verfahren (OCR = Optical Character Recognition, ein automatisiertes Zeichen- oder Texterkennungsverfahren), mit denen aus den Bildern der Seiten eigentliche Textdokumente hergestellt werden können (z.B. im TEI-Format[2]), die maschinell durchsuchbar und indexierbar sind.
- Digitale Speichermedien weisen heute grosse Kapazitäten auf und sie sind sehr billig geworden. Damit wird es möglich, grosse Datenmengen effizient und preiswert digital zu speichern.

[1] Siehe hierzu: Greg Bensinger. Amazon Wants to Ship Your Package Before You Buy It. In: Wall Street Journal, Jan 17 (2014). http://blogs.wsj.com/digits/2014/01/17/amazon-wants-to-ship-your-package-before-you-buy-it/ (Zugriff 28.10.2014).

[2] TEI (Text Encoding Initiative) ist ein auf der XML-Technologie basierende Auszeichnungssprache, welche zur Kodierung von z.B. literarischen Texten oder Manuskripten verwendet wird.

- Die Leistungsfähigkeit des Internet ist in den letzten Jahren so stark ausgebaut worden, dass auch sehr grosse Datenmengen in kürzester Zeit übertragen werden können.

Diese Entwicklungen haben dazu geführt, dass in den letzten Jahren grosse Mengen von Kulturgut digitalisiert worden sind (und immer noch digitalisiert werden), die letztlich auch für die geisteswissenschaftliche Forschung als Quellenmaterial interessant sind. Dies gilt nicht nur auf nationaler Ebene, sondern ist international zu beobachten. Wir möchten hier, stellvertretend für viele (nationale) Projekte, einige Beispiele nennen:

e-codices [3] – *Virtuelle Handschriftenbibliothek der Schweiz*

Das Ziel von e-codices ist es, sämtliche mittelalterlichen Handschriften der Schweiz und eine Auswahl der frühneuzeitlichen Handschriften als qualitativ hochstehende digitale Faksimiles online zur Verfügung zu stellen. Die Faksimiles werden durch reichhaltige Metadaten (z.B. wissenschaftlich redigierte Beschreibungen) ergänzt.

DoDis – *Diplomatische Dokumente der Schweiz* [4]

Die Internet-Datenbank erlaubt den freien Zugang zu einer grossen Zahl von digitalisierten Dokumenten (aus amtlichen Quellen), die für das Verständnis und die Rekonstruktion der aussenpolitischen Geschichte der Schweiz notwendig sind.

Schweizer Textkorpus [5]

Das Schweizer Textkorpus bietet ein Referenzkorpus mit ca. 20 Millionen Textwörtern zur Schweizer Standardsprache des 20. Jahrhunderts online an.

Montreux Jazz Digital Project [6]

In diesem noch in der Entstehung begriffenen Projekt des Metamedia Center der EPFL wird das persönliche Archiv von Claude Nobs, dem Begründer des Montreux Jazz Festival, komplett digitalisiert und so der Nachwelt zugänglich gemacht. Claude Nobs hat sämtliche Konzerte am Montreux Jazz Festival in teilweise experimentellen Videoformaten aufzeichnen lassen und in seinem persönlichen Archiv bewahrt. Diese Aufzeichnungen werden nun digitalisiert und sollen innerhalb des Campus der EPFL zugänglich gemacht werden.

Damit sind heute sehr grosse Datenmengen online verfügbar, welche einen direkten Bezug zur geisteswissenschaftlichen Forschung oder zu unserem kulturellen Erbe haben. Auch wenn diese Datenmengen vermutlich noch nicht ganz vergleichbar mit den durch die Internetkonzerne gesammelten Daten sind, so dürfen wir aus diesem Blickwinkel heraus für das Gebiet der Geisteswissenschaften durchaus von Big Data sprechen.

Wie steht es nun mit dem zweiten Punkt, der Durchforstung der Daten nach Mustern und Merkmalen, sowie der Gewinnung von neuen Aussagen aus dieser gezielten Datenanalyse, die zur eingangs gegebenen Definition von Big Data gehört? In diesem Bereich steht der Bereich der Kulturgut-Erhaltung vermutlich noch in den Anfängen. Grundsätzlich sind zwei Ebenen zu unterscheiden: Einerseits die Analyse der Daten an sich, wie z.B. die Analyse eines Textkorpus auf gewisse Merkmale, welche dann den einzelnen Werken, Kapiteln etc. zugeordnet werden können. Andererseits könnte auch analysiert werden, wie die Daten von den Forschenden und «Usern» genutzt werden. Während letzteres erst rudimentär eingesetzt wird (z.B. mit Google statistics), hat sich ersteres zu einer neuen Disziplin in den Geisteswissenschaften, den Digital Humanities (DH) entwickelt, wobei die DH noch viele weitere Aspekte beinhalten.

Digital Humanities: Big Data-Methoden für die Geisteswissenschaften?

Eine abschliessende Definition der Digital Humanities ist sehr schwierig. Grundsätzlich kann damit jener Bereich der Forschung und Methodenentwicklung bezeichnet werden, der im Grenzgebiet zwischen Geisteswissenschaften und Informatik liegt. Die DH bilden gewissermassen eine Brücke zwischen Geisteswissenschaften und Informatik, welche aber auch eigenständige Aspekte aufweist. Sie kann somit als eine Art moderne Hilfswissenschaft mit eigenständiger Forschung aufgefasst werden.

Ein wichtiger Bereich der DH ist es, Methoden zu erarbeiten, um grosse Datenmengen, welche einen Bezug zur Geisteswissenschaft (und damit oft auch zum kulturellen Erbe haben), auf gegebene Fragestellungen hin zu durchsuchen und zu analysieren. Im Bereich von Texten wird dabei oft der Begriff «distant reading» verwendet. Der Literaturwissenschaftler Franco Moretti (geb. 1950) vom Stanford Literary Lab prägte diesen Begriff, als er mit Hilfe von mathematischen Methoden (z.B. Korrelationen) Muster und Beziehungen in grossen Textkorpora zu finden suchte, welche beim genauen Lesen («deep reading») von einzelnen Werken verborgen blieben[7]. Während bei Texten solche Analysen plausibel sind, fehlen für andere Medien, wie z.B. für Bilder, die dazu notwendigen Verfahren weitgehend.

Big Data versus «Linked Open Data»

Eine andere Entwicklung, die sehr vielversprechend ist und von der DH-Community intensiv verfolgt wird, ist unter dem Begriff «Linked Open Data» (LOD) bekannt. Die meisten Datensammlungen in den Bereichen Kulturgut und Geisteswissenschaften sind zwar über Webportale zugänglich, doch leider ist es nur in seltenen Fällen möglich, die abgespeicherten Datenobjekte direkt über eine standardisierte Schnittstelle in eigene Applikationen und Analyseprogramme einzubinden. LOD ist ein Konzept, das zu diesem Zweck definiert wurde. Das Prinzip ist, die Datenobjekte in einer standardisierten, maschinell lesbaren Form frei auf dem Internet zur Verfügung zu stellen. Als Standards werden dabei offene, also nicht proprietäre, einfache Protokolle verwendet.[8] Damit wird die Verknüpfung von Quellen aus den verschiedensten digitalen Repositorien machbar, was völlig neue Möglichkeiten der Forschung und digitalen Präsentation eröffnet. So können beispielsweise Fragmente einer mittelalterlichen Handschrift virtuell zusammengeführt werden, auch wenn die Digitalisate in örtlich, organisatorisch und technisch verschiedenen Datenbanken liegen.

Nachhaltigkeit

«...digital information lasts forever — or five years, whichever comes first!»[9] Diese Aussage des Computerwissenschaftlers Jeff Rothenberg aus dem Jahre 1997 hat leider noch nichts an Gültigkeit verloren. Die nach wie vor extrem schnelle Obsoleszenz von Hard- und Software verunmöglicht eine langfristige, einfache Aufbewahrung von digitalen Daten. Da sich aber digitale Daten glücklicherweise verlustlos kopieren lassen, ist ein praktikabler Ansatz, die Daten in periodischen Abständen – oft nur wenige Jahre – auf neue Datenträger zu kopieren und sie, falls notwendig, so zu transformieren, dass sie mit der aktuellen Software kompatibel bleiben. Dieses Verfahren wurde durch das OAIS-Referenzmodel[10] eines digitalen Archivs standardisiert und gilt heute als «golden standard» der Langzeitarchivierung von digitalen Daten im Bereich der Kulturgüter und geisteswissenschaftlichen Quellen. Leider fehlen nicht selten – und nicht nur in der Schweiz – die entsprechenden Institutionen, die diese neue Aufgabe bewältigen sollen. Es ist daher sehr zu begrüssen, dass die kantonalen Staatsarchive und das Bundesarchiv daran arbeiten, entsprechende Infrastrukturen aufzubauen und teilweise auch schon in Betrieb haben. Viele Bibliotheken sind, beispielsweise unter dem Dach von e-lib,[11] ebenfalls daran, nachhaltige Infrastrukturen zur langfristigen Verfügbarkeit von digitalen Quellen aufzubauen.

Daher ist anzunehmen, dass in den nächsten Jahren genügend Kapazitäten und Institutionen vorhanden sein werden, um die wichtigsten digital vorliegenden Quellen und Objekte langfristig zu sichern. Dies um so mehr, als digitale Quellen in der Form von Digitalisaten meist eine sehr einfache Struktur (z.B. pro Bild eine TIFF-Datei mit einigen Metadaten) aufweisen. Solche Objekte lassen sich im OAIS-Modell einfach handhaben.

Schwieriger wird es für komplexe, genuin digitale Objekte wie z.B. Datenbanken, die Daten, Abfragelogik und Präsentation in einer untrennbaren Einheit verbinden. In diesem Bereich unterstützen das Staatssekretariat für Bildung, Forschung und Innovation SBFI und die Schweizerische Akademie der Geistes- und Sozialwissenschaften SAGW ein Pilotprojekt zur Schaffung eines Repositorys für «primäre Forschungsdaten» der Geisteswissenschaften, das auch die Behandlung solch komplexer Objekte umfasst.

Fazit

Wenn wir es genau nehmen, ist Big Data derzeit für die Geisteswissenschaften und den Erhalt des Kulturerbes nur von marginaler Bedeutung. Zwar sind in diesem Bereich durch die Digitalisierung von Quellen und kulturellen Artefakten sehr grosse Datenmengen entstanden, doch fehlen in den meisten Fällen die Methoden, um diese in ihrer Gesamtheit zu untersuchen. Damit wird das zweite Kriterium, die Analyse grosser Massen von Daten, kaum erfüllt. Dies kann sich jedoch mit der weiteren Entwicklung der Digital Humanities in Zukunft ändern. Ansätze wie «distant reading» zeigen, wie mächtig solche auf Big Data basierende Analysen sein könnten. Bis jedoch Methoden zur Verfügung stehen, die auf Bilder, Ton oder Film angewendet werden können, wird es noch einige Zeit und Forschungsanstrengungen brauchen.

Résumé

On nomme «mégadonnées» («big data») des collections d'informations résultant de la récolte et du stockage de données élémentaires portant sur une énorme masse d'individus (personnes, objets, biens culturels, etc.) et sur leurs «activités»; l'analyse de ces données vise à décrire le comportement habituel de ces individus, en répondant à certaines questions ciblées, et à prévoir leur comportement futur. Une telle analyse n'est possible qu'à deux conditions: 1. La technologie de stockage des données doit être en mesure de sauvegarder efficacement d'énormes quantités de données, auxquelles elle doit garantir un accès rapide (il s'agit de nombres de petaoctets à plusieurs chiffres; 1 petaoctet = 1000 téraoctets = 1 000 000 gigaoctets); c'est aujourd'hui possible, grâce aux disques durs modernes et aux systèmes de stockage en réseau. 2. On doit disposer de méthodes d'analyse applicables à des volumes de données gigantesques.

Ces dernières années, on a numérisé un grand nombre de biens culturels, et ce travail se poursuit, tant au niveau national que sur le plan international. Parmi les programmes en cours au niveau national, on mentionnera par exemple le projet «e-codices» et le «Montreux Jazz Digital Project». Aujourd'hui, on peut accéder en ligne à de grandes masses de données qui se rapportent directement à notre patrimoine culturel. Dans le domaine de la conservation des biens culturels, l'analyse ciblée de ces données et la récolte par ce biais de nouveaux résultats en sont encore à leurs balbutiements. Ce type d'analyse a donné naissance à une nouvelle discipline au sein des sciences humaines, les humanités numériques. Dans ce contexte, le développement des données ouvertes liées («linked open data, LOD») est particulièrement intéressant. Cette nouvelle technologie doit permettre aux chercheurs d'utiliser directement dans leurs propres applications et programmes d'analyse les données disponibles en ligne, grâce à une interface standardisée; par ce moyen, des données provenant des bibliothèques numériques les plus diverses pourront être combinées et comparées, ce qui ouvrira des perspectives radicalement nouvelles en matière de recherche et de présentations numériques.

[3] www.e-codices.unifr.ch

[4] www.dodis.ch

[5] www.schweizer-textkorpus.ch/index.php

[6] http://metamedia.epfl.ch/cms/site/metamedia/lang/en montreux_jazz_digital_project/about/mjdp

[7] Siehe beispielsweise: Kathryn Schulz. What is distant reading. In: New York Times, June 24 (2011). www.nytimes.com/2011/06/26/books/review/the-mechanic-muse-what-is-distant-reading.html?pagewanted=all (Zugriff 28.10.2014).

[8] Beispielsweise «Representational State Transfer REST» (www.ibm.com/developerworks/library/ws-restful/) oder SPARQL, die standardisierte Abfragesprache für das semantische Web.

[9] Jeff Rothenberg. Digital Information Lasts Forever – Or Five Years, Whichever Comes First. RAND Video V-079, 1997.

[10] Reference Model for an Open Archival Information System, (OAIS), Recommended Practice, Consultative Committee for Space Data Systems (CCSDS) 650.0-M-2 (Magenta Book) Issue 2, June 2012.

[11] www.e-lib.ch