arbido

2013/4 Linked Open Data Big Data - Tout en réseau

Schneider René,

Tags: documentation big data,

Dimensionen und Zusammenhänge grosser, verknüpfter, offener und wissenschaftlicher Daten

Nachdem die Flut der Diskussionen, Kommentare und Prophezeiungen das Social Web betreffend nun verebbt ist, bricht bereits die neue Flutwelle über die Dämme und überspült uns mit dem, was im Social Web und anderswo in Massen ohne Unterlass und Überlegung produziert wird: Daten!

Einleitung

Man gewinnt nicht nur den Eindruck, dass die verhältnismässig alte Bezeichnung der elektronischen Datenverarbeitung eine ganz neue Bedeutung erhalten hat, sondern dass die Daten gleichsam selbst zu Information geworden sind, auch wenn dies auf den ersten Blick als widersinnig erscheinen mag. Dieser paradoxe Eindruck verstärkt sich jedoch, wenn in einflussreichen Kreisen nicht mehr vom Informationszeitalter gesprochen wird, sondern vom Zeitalter der Daten. So wird bereits ein neues und viertes wissen schaftliches Paradigma [Jim Gray on e-Science: «A transformed Scientific Method», in: Hey, Tony, Tansley, Stewart, Tolle, Kristin: The Fourth Paradigm: Data Intensive Scientific Discovery. Microsoft Corporation. 2009, S. xviii http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_complete_lr.pdf] heraufbeschworen und nach den wissenschaftstheoretischen Vorläufern der empirischen Beschreibung von Naturphänomenen (1. Paradigma), der theoretischen Generalisierung (2. Paradigma) und der computerbasierten Simulation (3. Paradigma) nun das Zeitalter der datenintensiven wissenschaftlichen Entdeckung verkündet.

Aber was heisst das eigentlich? Sind die übrigen Paradigmen der Wissenschaftsgeschichte damit bedeu tungslos geworden? Und wie verhält es sich mit dem Kontext der Daten, durch den sie normalerweise zu Information werden? Oder sind Daten nicht mehr gleich Daten? Und was heisst dies für Bibliotheken, Archive und Dokumentationszentren? Sind sie oder ihre Daten betroffen? Ergeben sich neue Arbeitsaufträge für sie oder werden sie von einem weiteren Bedeutungsverlust betroffen sein?

Im Folgenden soll versucht werden, diese Fragen einzugrenzen und entsprechend dem derzeitigen Stande der Dinge zu beantworten. Dies soll anhand dreier unterschiedlicher Datentypen illustriert werden, die derzeit unterschiedlich stark diskutiert werden, sich trotz ihrer Unterschiede aber an einzelnen Punkten berühren bzw. in unterschiedlicher Beziehung zueinander stehen: Big Data, Linked Open Data und Research Data (d.h. grosse Datenmengen, verknüpfte und offene Datenmengen sowie Forschungsdaten). Diese dreifache Auffächerung von Datenverarbeitungsinitiativen beantwortet bereits unmittelbar eine der weiter oben formulierten Fragen und zeigt, dass Daten wirklich nicht mehr gleich Daten sind. Die Informationswissenschaft und alle damit verbundenen Institutionen stehen damit auch vor der Herausforderung, sich auf diese Daten und die damit verbundenen Aufgaben einzustellen.

Big Data

Den grössten Block (im doppelten Sinn) machen dabei die gemeinhin unter Big Data subsumierten Datensätze aus, von deren Existenz und Konsequenz mittlerweile auch die Normal bürger erfahren haben, was insofern gut ist, da sie selbst zu den Hauptproduzenten, bedauerlicherweise aber nicht zu den Nutzern gehören. Big Data wurde dabei letztlich durch zwei Phänomene möglich: erstens die weiterhin exponentiell steigende Leistungsfähigkeit der Prozessoren sowie die damit parallel verlaufende Steigerung der Speicherkapazitäten der Rechner, zweitens durch den Anstieg der täglich bzw. sekündlich produzierten und aufgezeichneten Datenmengen, die allerorts durch die Vernetzung der Endgeräte (vom Grossrechner bis zum Smartphone und den darauf befindlichen Apps) ins Web eingespeist und von dort weiterverarbeitet werden. Gespeichert wird dabei alles, was geht: Daten, mit oder ohne Metadaten, oder die Metadaten alleine, aus denen sich ja streng genommen der eigentliche Kontext der Daten bzw. ihr Verständnis ergibt. Entscheidend ist aber, dass aus diesen Daten ein neuer Kontext erzeugt wird und zwar mithilfe spezieller Algorithmen, genauer mit sogenannten Inferenzmechanismen, also Schlussfolgerungsregeln der Maschinen, die sie verwalten und bearbeiten. In den Schlussfolgerungsregeln des Big Data werden orthogonale, d.h. prinzipiell voneinander unabhängige Merkmale miteinander verknüpft, um neue oder bis dahin unbekannte Zusammenhänge aufzuzeigen[Dietmar Dath, Ranga Yogeshwar (2013): «Rechnen Sie damit, lebenslang ein Verdächtiger zu sein», FAZ vom 12.7.2013.]. Mit dieser an und für sich recht einfachen Vorgehensweise, die von Menschen täglich bewusst oder unbewusst ausgeführt wird, stellt sich der einfache, und von daher sehr erfolgsversprechende Business Case von Big Data dar, der nun aber von den Maschinen ausgeführt wird. Damit sollte auch klar sein, dass hinter Big Data zu einem sehr grossen Teil wirtschaftliche oder andere Interessen stehen, die nicht primär im Interesse der Produzenten sind, auch wenn die ethischen Implikationen an dieser Stelle nicht ausgeführt werden sollen.

Linked Open Data (LOD)

Eher unabhängig davon verläuft die Linked-Open-Data-Initiative[Bizer, Christian, Tom Heath, and Tim Berners-Lee (2009). «Linked data-the story so far.» International Journal on Semantic Web and Information Systems (IJSWIS) 5.3, 1–22.]. Unabhängig in dem Sinne, dass die hier verknüpften (Linked) und veröffentlichten (Open) Datensätze nicht notwendigerweise eine Teilmenge der grossen (Big) Daten sind, im Gegenteil: Der Umstand, dass LOD die gemeinnützige Publikation sowie eine Verknüpfung zur Conditio sine qua non macht, schliesst den orthogonalen Charakter der grossen Datensätze aus und eröffnet eine ganz andere Sichtweise. Im Gegensatz zu Big Data beschränkt sich LOD auf Daten, die im Web zur Verfügung gestellt werden, und nicht in irgendwelchen nicht-öffentlichen Daten banken oder Clouds gespeichert werden. Des Weiteren ist ihr Umfang, trotz der an und für sich unvorstellbaren Menge der Daten, die die LOD-Cloud mittlerweile enthält, vergleichsweise klein.

Im Rahmen von Linked Open Data soll ein Grundproblem gelöst werden, das darin besteht, dass der grösste Teil des Wissens, das im Web publiziert wird, in relationalen Datenbanken gespeichert, aber nur einem sehr flachen Format (genauer: HTML, der Hypertext Markup Language) repräsentiert wird. Dies führt dazu, dass die ursprünglich gesetzten Relationen verloren gehen und eine direkte Kommunikation (im Sinne eines reziproken Verständnisses) zwischen den Datenbanken nicht möglich ist. Dieses Problem wird nun dadurch gelöst, dass eine neue Daten ebene eingeführt wird, die Auskunft über die semantischen Relationen gibt. Der eigentliche Mehrwert von Linked Open Data besteht also darin, dass sie die Relationen aller Datenbanken in ein einheitliches Format überführen, was – genau wie im Fall von Big Data – dazu führen kann, dass die Maschine über sie räsonieren kann. Wiederum geht es also nicht um kognitive Prozesse von Menschen, sondern darum, diese in Rechnern zu repräsentieren und zu simulieren und den Maschinen zu erlauben, selbstständig Rückschlüsse aus den verknüpften Daten zu ziehen.

Forschungsdaten

Etwas im Abseits dieser grossen Bewegungen findet – auf den Bereich der Wissenschaften beschränkt – seit einigen Jahren eine weitere Diskussion statt, die die sogenannten Forschungsdaten zum Hauptgegenstand des Interesses gemacht hat.

Die gesamten Anstrengungen lassen sich sehr gut an zwei grundlegen den Modellen illustrieren: dem Modell des Datenlebenszyklus[Sarah Higgins (2008): «The DCC Curation Lifecycle Model». International Journal of Digital Curation 3(1), 134–148.], das von ersten Projektideen über die Ablage und dau erhafte Speicherung von Datensätzen bis hin zu deren Integration in Publi kationen und deren Zitation reicht, und dem Modell für das Datenkontinuum[Andrew Treloar (2011): Private Research, Shared Research, Publication, and the Boundary Transitions. http://andrew.treloar.net/research/diagrams/data_curation_continuum.pdf·], das eine kontinuierliche Weitergabe und den Austausch der Daten garantieren soll und zwar unabhängig davon, ob es sich um Rohdaten, strukturierte Daten, Metadaten oder die in späteren Publikationen eingebundenen Daten handelt. Der Mehrwert für den Forscher und die Öffentlichkeit ergibt sich daraus, dass beide quasi uneingeschränkt und über alle Soft und Hard warewechsel Zugang zu diesen Daten für eine Nachnutzung oder Überprüfung haben. Auch wenn man davon ausgehen kann, dass sich das Interesse dafür sowohl bei den meisten Forschern als auch der Gesellschaft, die diese Forschung finanziert, in Grenzen hält, wird eine Sicherung der Forschungsdaten mittlerweile zu den guten und festen Regeln für eine transparente Forschung gezählt. Der grösste Vorteil für den Forscher ergibt sich daraus, dass die Datensätze selbst, d.h. die Primärdaten, zitier und publizierbar werden. In einigen Disziplinen gibt es parallel dazu bereits «Peer Reviewed Data Paper», die allein die Datensätze enthalten.

Was das spezielle Verhältnis von Forschungsdaten zu Big und Linked Open Data betrifft, lässt sich sagen, dass sie – etwa im Fall des Large Hadron Colliders – sehr gross sein können. Sie können auch mit LOD verknüpft und angereichert werden, müssen es aber nicht.

Dabei sollte nicht ausser Acht gelassen werden, dass bei den Forschungsdaten eine ganz andere grundsätzliche Frage im Vordergrund steht: wie bewahren wir (Roh)Daten langfristig auf, sodass sie uns zu einem späteren Zeitpunkt zur Nachnutzung wieder problemlos zur Verfügung stehen? Es geht also in erster Linie um ganz grundsätzliche Fragen der Archivierung, der Kompatibilität und der Interoperabilität, im Grunde also um Fragen, die auch die Bereiche der grossen und der verknüpften Datenmengen irgendwann betreffen könnten. Man kann sich aber auch nicht des Eindrucks erwehren, dass gerade in der Debatte um die Forschungsdaten so prinzipielle Fragen angegangen werden, dass den Beteiligten noch gar nicht genau klar ist, wie diese gelöst werden sollen, auch wenn es zumindest in einzelnen Disziplinen bereits Lösungsansätze bzw. respektable Lösungsvorschläge gibt.

Bibliotheken und Daten: eine neue und notwendige Herausforderung

Bleibt die Frage nach der Rolle der Institutionen, die traditionell mit diesen Fragen beschäftigt waren. Zuvorderst handelt es sich dabei um die Bibliotheken, auch wenn die Archive und Dokumentationszentren in einzelnen Bereichen ein Wort mitsprechen könnten, sofern ihre Expertise noch gefragt ist.

Die grossen Datenmengen (Big Data) können die Bibliotheken, Archive und Dokumentationszentren schnell ausser Acht lassen, auch wenn sie traditionell damit beauftragt waren, grössere Informationsmengen aufzubewahren. An Big Data zeigt sich viel mehr, wie schnelllebig die Zeitläufte geworden sind. Wähnten sich die Bibliotheken (zumindest einige erlauchte unter ihnen) bis vor Kurzem noch an vorderster Front in der grössten Digitalisierungsinitiative und durften sie sich noch als strategisch wichtige Datenlieferanten der erfolgreichsten Suchmaschine ansehen, müssen sie nunmehr erkennen, dass sie vom Gehilfen des globalen Players zum staunenden Betrachter noch weitaus grösserer Daten verwalter geworden sind.

Anders verhält es sich bei der Linked-Open-Data-Initiative. Sollten die Bibliotheken und ihre Katalogisierungsabteilungen sowie die Verbundzentren dazu bereit sein, würde sich Linked Open Data als der notwendige, wenn nicht sogar zwingende kommende Schritt darstellen, um die Kataloge in ein Format zu bringen, von dem deren Anbieter selbst, aber auch viele an dere profitieren können. Schlussendlich könnten damit – neben vielen an deren positiven Seiteneffekten und neuen Nutzungsformen der Katalogdaten – vor allen Dingen die Probleme der Konvergenz unterschiedlicher Verbunddaten sowie die komplizierten Bemühungen für die Erstellung standardisierter Metakataloge ein Ende haben.

Ähnlich verhält es sich mit den Forschungsdaten, auch wenn hierbei weniger die Katalogisierenden als vielmehr die Kollegen aus dem Bereich des E-Publishing sowie die Betreiber der Open-Access-Plattformen eine Rolle spielen werden. Ihre Aufgabe wird es sein, neben den Langzeitarchivaren und im Verbund mit den Datenzentren die Datensätze, so gross oder (eher) klein sie auch sein mögen, mit einer persistenten Adressierbarkeit für eine Nachnutzung, Nanopublikation oder zitation zur Verfügung zu stellen.

Erinnerung, Betrachtung, Erwartung

Festzuhalten bleibt, dass es insgesamt sehr unterschiedliche Vorstellungen vom Umgang mit Daten gibt, die derzeit in drei unterschiedlich grossen Initiativen angegangen und diskutiert werden, wobei durchaus Berührungs punkte bestehen. Das gegenwärtige Bild der Dateninitiativen zeigt sich so dynamisch, dass derzeit nicht antizipierbar ist, ob die Dateninitiativen weiter zersplittern, konvergieren oder etwa zu einem Schichtenmodell zueinander finden können.

Die reine Betrachtung der Gegenwart liefert aber noch keine Antworten auf die zwei grundlegenden Fragen: Warum heben wir die Daten eigentlich auf ? Und: Warum interessieren wir uns so sehr für sie? Diese zwei Fragen lassen sich letztlich mit ganz grundsätzlichen menschlichen Bedürfnissen beantworten: erstens einem eher rück wärtsgewandten Bedürfnis, nämlich dem Sammeln und dem daraus resultierenden Bedürfnis nach Erinnerung, zweitens mit der Organisation bzw. der Repräsentation von Wissen, um es immer, d.h. gegenwärtig, zur Verfügung zur haben, und drittens dem eher vorwärtsgewandten Wunsch, Zukünftiges vorhersagen zu können. Aus dieser Perspektive betrachtet, lassen sich kurioserweise Parallelen zu den drei Dimensionen der Zeit erkennen, die Augustinus im elften Buch der Confessiones in seiner Reflektion über das Wesen der Zeit entwickelte. Interessanterweise sprach er nicht allein von Praeteritum, Praesens und Futurum, sondern von «praesens de praeteritis memoria, praesens de praesentibus contuitus, praesens de futuris expectatio»1, d.h., er assoziierte die Erfahrung der Zeit mit drei zentralen Begriffen: Memoria (Erinnerung), Contuitus (Betrachtung) und Expectatio (Erwartung).

So mag es für ein grundlegendes Verständnis hilfreich sein, die unterschiedlichen Konzepte zum Umgang mit Daten schlussendlich mithilfe dieser drei Dimensionen zu begreifen: der Memoria entsprechen die Bemühungen um die Archivierung und Nachnutzung von Daten; die LOD-Initiative zeigt das Bemühen, eine Datenschicht für die gegenwärtige Betrachtung, also den Contuitus, aufzubauen; die Bestrebungen von Big Data entsprechen schliesslich dem Wunsch, die Handlungs und Gedankenwege anderer vorherzusagen.

Nicht umsonst begann die Wissenschaft der Daten mit Euklids Geometrie, in denen zum ersten Mal von «Dedomena» (dem Gegebenen)Jakob Voss (2013): «Was sind eigentlich Daten?» Libreas. Library Ideas, 23, http://libreas.eu/ausgabe23/02voss·die Rede ist, die in der ersten Übersetzung ins Lateinische eben mit dem Wort Data wiedergegeben wurden. Euklid sah als Dedomena jenes Gegebene an, aus dem sich gesuchte, d.h. nicht bekannt geometrische Zusammenhänge erschliessen lassen. Nichts anderes ist – wenn wir den geometrischen Zusammenhang beiseitelassen – bei Big Data und Linked Open Data der Fall. Hier ist der Zusammenhang so einfach wie offensichtlich.

Wir sollten dabei auch nicht vergessen, dass seit Anbeginn der Wissenschaftsgeschichte praktisch alle namhaften Forscher darum bemüht waren, die Zukunft vorherzusagen, mit welchen Mitteln auch immer. Auch wenn wir es heute nicht mehr wahrhaben wollen, schrieb Newton, der gerne als Vater der modernen Wissenschaften angesehen wird, und darauf verwies «auf den Schultern von Giganten» zu stehen, mehrere Bücher zur Prophetie und der AlchimieJohn Freely (2009): Aladdin’s Lamp, Alfred E. Knopf, New York..

Diese beiden alles überragenden Interessen (Prophetie und Alchimie) führten letztlich zum Entstehen der ungleichen Schwestern Astronomie und Astrologie sowie der Ablösung der Alchimie durch die Chemie. Wissenschaftstheoretisch würde sich die ganze Diskussion um Daten also als nichts Weiteres herausstellen als die Suche nach dem Beweis, dass Prophetie nichts anderes als ein ganz rationales und berechenbares Unterfangen ist.

1 Augustinus, Confessiones, XI, 20, 26, http://www9.georgetown.edu/faculty/jod/latinconf/11.html.

René Schneider

Haute Ecole de Gestion de Gene?ve

Résumé

Français

Mais que peuvent donc signifier ces énormes quantités de données ouvertes et reliées entre elles? Les autres paradigmes de l'histoire de la science sont-ils donc devenus obsolètes? Ou bien les données ne sont-elles plus des données? Et qu'est-ce que cela signifie pour les bibliothèques, les archives et les centres de documentation? Sont-ils concernés ou est-ce leurs données qui le sont? En résulte-t-il de nouvelles tâches pour ces institutions ou leur importance s'en trouvera-t-il amoindrie?

Dans cet article, l'auteur tente de cerner ces questions et d'y répondre. Il le fait en discutant de trois types de données qui, malgré leurs différences, ont quelques points communs qui les relient d'une certaine manière, à savoir: big data, linked open data et research data (c.à.d. de grandes quantités de données ouvertes et reliées entre elles ainsi que des données de recherche). Ce triptyque d'initiatives de traitement des données répond d'ores et déjà à l'une des questions formulées ci-dessus et montre que les données ne sont plus tout à fait des données. Les sciences de l'information et toutes les institutions qui y sont liées se trouvent donc face à un défi: maîtriser ces données et les tâches qu'elles induisent.