www.jusletter-it.eu

Erich Schweighofer

Transparenzfiktion und Big Data

Big Data reflects the increased dynamics of today's world with enormous data needs. Transparency in this knowledge cloud remains a pre-fiction; despite "Google». Legal information provides a good example. On one hand, much is not yet available digitally; on the other hand, the available range of information is only usable if the semantics of legal language and legal meta data are handled. Efficient data use requires an appropriate and data protection friendly network of information with events, objects and anonymised person classes.

Category: Articles
Field of law: Big Data, Open Data & Open Government; Legal Informatics
Region: Austria

Citation: Erich Schweighofer, Transparenzfiktion und Big Data, in: Jusletter IT 21 May 2015

ISSN 1664-848X, http://jusletter-it.weblaw.ch, Weblaw AG, info@weblaw.ch, T +41 31 380 57 77 EDITIONS WEBLAW

Inhaltsübersicht

1 Einführung

[Rz 1] Die Wissens- und Netzwerkgesellschaft bringt eine immer grösSere Flut von Daten mit sich. Das sogenannte *Digitale Universum* wächst exponentiell.[1] Verantwortlich hierfür ist die wachsende Wissensrepräsentation, aber auch die Maschinen: Das Internet der Dinge als auch Sensordaten produzieren enorme Datenmengen.

[Rz 2] Mit dieser zunehmenden Digitalisierung wird einem erst bewusst, wie groSS die Datenmenge ist, um überhaupt zu überleben bzw. — noch schwieriger — die Welt gestalten zu können. Vieles ist bisher unbewusst geblieben oder wurde als selbstverständlich gesehen. Viele dürften sich gedacht haben, dass die Speicherkapazität des Gehirns maSSlos übertrieben wurde, aber dem ist nicht so. Wir brauchen sehr viele Daten, um zu leben, zu arbeiten oder zu gestalten. Big Data ist auch das Ergebnis der Bemühungen des Wissensmanagements, sämtliches Wissen in computerverfügbarer Form abzubilden und wieder zu verwerten.

[Rz 3] Je mehr Wissen täglich benötigt wird, desto wichtiger ist es, dass eine Transparenz dieses Wissens für alle Menschen gegeben ist. Die einfache Lösung der Gleichsetzung von Verfügbarkeit mit Transparenz des Wissens funktioniert nicht. Auch in der Welt der Suchmaschinen ist die Beherrschung des Wissens eine lebenslange Herausforderung. Am besten ist es, sich eine grundlegende Aussage der Erkenntnistheorie vor Augen zu halten. Jede Erkenntnis basiert auf Sinnesdaten, die durch den menschlichen Wahrnehmungsapparat gefiltert und fortwährend unbewusst interpretiert werden. Es gibt daher kein absolut sicheres Wissen und die Abbildung der Realität bleibt ein hypothetisches Modell.[2]

2 Der vielschichtige Begriff «Transparenz»

[Rz 4] Im Sinne einer demokratischen und offenen Gesellschaft ist es notwendig, dass möglichst alles Wissen transparent ist, d.h. für alle verfügbar, ohne wesentlichen zusätzlichen Aufwand nutzbar und natürlich auch für alle verständlich.

---

1 Futurezone, http://futurezone.at/digital-life/weltweite-datenmenge-verzehnfacht-sich-bis-2020/60.109.918, 11. April 2014 (alle Internetquellen zuletzt abgefragt am 9. Mai 2015); EMC Digital Universe Study, http://www.emc.com/leadership/digital-universe/2014iview/index.htm, April 2014.

2 Vgl. Schweighofer, Erich, Wissensrepräsentation und Rechtsinformatik. Springer, Wien (1999), S. 16 f., Schülein, Johann August / Reitze, Simon, Wissenschaftstheorie für Einsteiger, WUV UTB, 2. Aufl. (2005); Hassemer, Winfried / Kaufmann, Arthur / Neumann, Ulfrid (Hrsg.), Einführung in Rechtsphilosophie und Rechtstheorie der Gegenwart, 7., neubearb. und erweiterte Aufl., C.F. Müller UTB (2004); Eco, Umberto, Einführung in die Semiotik. W. Fink UTB, 7. Aufl. (1991); Wikipedia, http://de.wikipedia.org/wiki/Wissen.

[Rz 5] Mit dem Internet ist eine Verteilungsplattform gegeben, womit ohne wesentlichen Zusatzaufwand alles Wissen für alle zugänglich gemacht werden kann. Wissen zu schaffen kostet aber viel Geld; daher wird trotz der grundsätzlichen Informationsfreiheit wirtschaftlich verwertbares Wissen für einige Zeit nur über Bezahldienste verfügbar sein. Es ist die Aufgabe der Gesellschaft, d.h. der öffentlichen Institutionen und der Zivilgesellschaft, möglichst umfassend relevantes Wissen kostenlos bereitzustellen. Im Bereich des Rechtswissens gibt es mit dem WorldLII *(World Legal Information Institute*, Motto: *Free, independent and non-profit access to worldwide law)* eine hoch zu schätzende und sehr erfolgreiche Initiative.[3]

[Rz 6] Verfügbarkeit heiSSt aber nur Abrufbarkeit aus einem Repositorium, wobei naturgemäSS die Distanz unwichtig geworden ist. Die Lokalität des Repositoriums war und ist wichtig, auch wenn die Website nun wichtiger ist als der physische Ort.

[Rz 7] Wissen ist ein Bestand an gesicherten Erkenntnissen. Im täglichen Leben muss das Wissen zu nutzbringender Information verwertet werden. Diese Herausforderung, das Wissen immer für die jeweilige Situation abzurufen, anzupassen und umzusetzen, ist keinesfalls trivial. Dies ergibt sich schon durch die Existenz der Bibliotheks- bzw. der Informationswissenschaft.[4] Hier sind auch die wesentlichen Lösungsansätze zu finden; jetzt zeigt sich auch, dass Verfügbarkeit von Informationen noch lange nicht auch deren kontextbezogene und zeitgerechte Verwertung bedingt.

[Rz 8] Die Assoziation von Transparenz (nach lat. «*transparens*» durchscheinend) mit Durchsichtigkeit ist evident. Noch verständlicher ist das Gegenteil: Intransparenz ist Undurchsichtigkeit. Die wissenschaftliche Fachsprache hat zwei wesentliche Gruppen von Begriffsverwendungen entwickelt: Durchsichtigkeit, aber auch Nichtsichtbarkeit oder Nichtbemerkbarkeit.[5]

[Rz 9] In der Informatik ist eine Hardware oder Software transparent, wenn ihre Existenz für den Benutzer nicht erkennbar und relevant ist. In der Kommunikation ist ein Signal transparent, wenn es sich beim Empfänger nicht bemerkbar macht. Denkt man an die riesigen Mengen der Internet-Kommunikation, wird klar, dass das Bewusstsein über die Komplexität der Kommunikation über die fünf Schichten der Internet Protokoll-Familie, das Verschicken von Datenpaketen weltweit über irgendwelche Kanäle, die vielen Routing-Anfragen und Zusicherungen etc. etc. jede Freude an der Internet-Nutzung verderben würde. Der Nutzer will genau das, was er heute bekommt: Information, Wissen und manchmal auch eine Lieferung bzw. Dienstleistung, ohne dass er sich um diese vielen Details kümmern muss. Was für die Kommunikation funktioniert, soll auch für das Wissen selbst umgesetzt werden. Transparente Zurverfügungstellung des Wissens.

[Rz 10] Die zweite Bedeutung von Durchsichtigkeit wird bei Wissen vornehmlich in Verfügbarkeit und Nutzbarkeit verstanden. In der Physik bedeutet Transparenz Durchlässigkeit in Bezug auf elektromagnetische Wellen, insbesondere des Lichts. Die Computergraphik bezeichnet damit durchscheinend wirkende Elemente in einer Bilddatei. In der Akustik wird unter Transparenz die Unterscheidbarkeit zeitlich aufeinanderfolgender Töne bei musikalischen Schalldarbietungen verstanden. In der Politik steht der Begriff für von auSSen nachvollziehbare Vorgänge der

---

3 Website World Legal Information Institute, http://www.worldlii.org/. Die Charta des WorldLII ist die Montreal Declaration on Free Access to Law (2002); Änderungen: Sydney (2003), Paris (2004), Montreal (2007) und Ithaca (2012).

4 Kuhlen, Rainer / Semar, Wolfgang / Strauch Dietmar, Grundlagen der praktischen Information und Dokumentation, Handbuch zur Einführung in die Informationswissenschaft. De Gruyter, Berlin (2014).

5 Wikipedia, http://de.wikipedia.org/wiki/Transparenz.

Öffentlichkeit. Die Volkswirtschaftslehre spricht von Markttransparenz, wenn Informationen in und über einen Markt verfügbar sind. Daher sollte man auch von der Wissenstransparenz sprechen; d.h. Wissen ist bei zumutbarem Bildungsaufwand für alle verfügbar und nutzbar.

[Rz 11] Zusammenfassend bedeutet dies die Forderung nach Transparenz-Transparenz des Wissens. Es soll für alle verfügbar und nutzbar sein, aber der erforderliche IKT-Einsatz soll im Interesse der Benutzerfreundlichkeit möglichst wenig sichtbar sein.

3 Von Big Data und Daten zu Information und Wissen

[Rz 12] Big Data bezeichnet Datenmengen, die wegen GröSSe, Komplexität bzw. Dynamik nur mehr mit informatorischen Methoden ausgewertet werden können.[6] Big Data war immer schon da, nur nicht für die Menschen erfassbar bzw. verarbeitbar. Über die längste Zeit der Menschengeschichte mussten die Menschen die Datenmenge enorm reduzieren, um damit umgehen zu können. Formale Modelle bzw. metaphysische Ansätze waren nötig, um einigermaSSen mit der Komplexität der Welt zu Rande zu kommen. Je besser die Welt verstanden wird und je mehr Daten erfasst und verarbeitet werden können, desto mehr wird realisiert, dass die Verarbeitung von enormen Datenmengen notwendig ist, um die Welt wirklich verstehen zu können. Die erkenntnistheoretische Reduktion des menschlichen Wahrnehmungsapparats fällt mehr und mehr weg; dazu kommen aber die Restriktionen der digitalen Welt. Der Preis für das Mehr an Exaktheit ist die enorme Datenflut. Nachteilig ist ebenfalls, dass die Datenschutzkomponenten der Vergesslichkeit und Unschärfe fehlen.

[Rz 13] Der menschliche Wahrnehmungsapparat strukturiert die Datenmenge semantisch, d.h. nach Themen, Personen, Orten usw.; durch Aus- und Weiterbildung und Praxis wird diese begriffliche Strukturierung im menschlichen Gehirn optimiert.

[Rz 14] Schon fast ein Jahrzehnt nach den ersten Schritten den Internets, des World Wide Webs, hat dessen Gründer Tim Berners-Lee auf die Notwendigkeit einer semantischen Strukturierung hingewiesen; bisher aber leider noch nicht mit durchschlagendem Erfolg.[7]

[Rz 15] Big Data bringt es mit sich, dass nun sehr viele Daten — auch öffentlich — verfügbar sind (Stichwort Open Data). Dies wird einem immer mehr bewusst und natürlich auch geschätzt. Als wichtige Beispiele müssen nur Geoinformation, Rechtsinformation, Wikipedia, die vielen informativen Websites etc. angeführt werden.

[Rz 16] Daten, Information und Wissen sind Kern jeder Analyse des Wissenskreislaufs.[8] Begrifflich gibt es im allgemeinen Sprachgebrauch eine Vermengung zwischen Daten, Information und Wissen. Der Ausdruck «Wissen» stammt von Althochdeutsch «wizzan» ab.[9] Wissen wird als gesicherter Bestand von Fakten Theorien und Regeln verstanden, sodass von ihrer Gültigkeit ausgegangen werden kann und diese als wahr anzusehen sind. Es kann kein absolut sicheres Wissen

---

6 Mayer-Schönberger, Viktor / Cukier, Kenneth, Big Data. Murray, London (2013); Wikipedia, http://de.wikipedia.org/wiki/Big_Data; Hofstetter, Yvonne, Der Angriff der Intelligenz. Die Zeit, 10. September 2014, http://www.zeit.de/kultur/2014-09/yvonne-hofstetter-kuenstliche-intelligenz.

7 Berners-Lee, Tim / Hendler, James / Lassila, Ora, The Semantic Web. In: Scientific American, 17 May 2001, http://www.cs.umd.edu//LBSC690/SemanticWeb.html.

8 Nonaka, Ikujiro / Takeuchi, Hirotaka, Die Organisation des Wissens — Wie japanische Unternehmen eine brachliegende Ressource nutzbar machen. Aus dem Engl. von Friedrich Mader, Campus-Verlag, Frankfurt am Main (1997).

9 Vgl. FN 2.

geben, weil jede Erkenntnis auf Sinnesdaten passiert und diese durch den Wahrnehmungsapparat gefiltert und unbewusst interpretiert werden. Je mehr jedoch das Wissen interdisziplinär und umfassend ausgetauscht und verifiziert wird, umso mehr ist von einem gesicherten Wissen zu sprechen. Der Mensch selbst steht nun vor der Herausforderung, diese objektivierten Wissensmengen (vornehmlich Textkorpora) möglichst objektiv zu verstehen und für die jeweiligen subjektiven Zwecke einzusetzen.

[Rz 17] Die Datenwolke des Wissens besteht physikalisch gesehen aus Daten, d.h. eine syntaktische Repräsentation, binäre Zahlenketten, eine Sammlung von Nummern, Zeichen und Bildern in einer digitalen Welt. Für eine menschliche Nutzung bedarf es aber einer Semantik und Pragmatik.

[Rz 18] Semantik ist die Theorie der Bedeutung der Zeichen sowie deren Beziehungen.[10] Aufgrund der Bedeutung der sprachlichen Zeichen für das Recht ist diese eine wesentliche Teildisziplin der Linguistik. Pragmatik ist die Beschreibung der kontextabhängigen Bedeutungen von Ausdrücken in konkreten Situationen.[11] Der Datenwolke sollen semantische Bedeutungen beigegeben werden. Im Semantischen Web soll dies durch den Mark-up der jeweiligen Texte, Bilder, Audios und Videos erfolgen; besser wäre es natürlich, wenn dies situativ ohne Mark-up erfolgen könnte. Derzeit sind aber die IKT nur beschränkt in der Lage, Semantik und Pragmatik aus der Datenwolke eine Wissenssammlung zu machen, d.h. Daten mit einer semantischen Struktur im pragmatischen Kontext zu kreieren.

[Rz 19] Hier liegen das Problem und die Herausforderung für Big Data in der Datenwolke. Es ist zwar eine enorme Menge von Wissen verfügbar, aber ohne Unterstützung für die Menschen sind diese nur in einem bescheidenen MaSSe tatsächlich nutzbar. Die Herausforderungen der zweckmäSSigen Strukturierung, der Metadaten, der Suche, des Rankings der Antworten etc. zeigen, dass es sich hiermit um ein an sich fast unlösbares Problem handelt. Die erkenntnistheoretischen Probleme mit Wissen an sich multiplizieren sich.

[Rz 20] Eine Wissenswolke entspricht explizitem Wissen, d.h. es ist sprachlich, bildlich oder verbal ausgedrückt. In den Rechtswissenschaften ist sprachliches Wissen dominant. Daneben gibt es Wissen eines Menschen, sowohl explizit als auch implizit. Bei der Überwindung jedes Wissensproblems kommt es entscheidend darauf an, dass eigenes explizites Wissen, fremdes explizites Wissen und das eigene implizite Wissen in bestmöglicher Weise miteinander verschränkt werden.

[Rz 21] Information wird oft synonym mit Wissen gebraucht. In der Terminologie der Informationswissenschaft (früher Bibliothekswissenschaft) ist Wissen etwas Statisches, das in einer Wissenswolke oder als persönliches Wissen in menschlichen Gehirnen fixiert ist.[12] Information ist der dynamische Prozess der aktiven wie passiven Nutzung des Wissens. Die «Informationswissenschaft untersucht das Auswerten/Selektieren, ErschlieSSen, Bereitstellen/Wiederverwerten, Suchen, Vermitteln und Finden von relevantem (vorwiegend digital vorliegendem) Wissen, durch Informations- und Kommunikationsprozesse».[13] Genauer betrachtet, zielt der Begriff Informati-

[10] Wikipedia, http://de.wikipedia.org/wiki/Semantik; Rathert, Monika / Grewendorf, Günther (Eds.), Formal Linguistics and Law, Trends in Linguistics, Mouton de Gruyter, Berlin (2009).

[11] Wikipedia, http://de.wikipedia.org/wiki/Pragmatik_(Linguistik).

[12] Wikipedia, http://de.wikipedia.org/wiki/Informationswissenschaft.

[13] Stock, Wolfgang G., Information Retrieval. Informationen suchen und finden. München: Oldenbourg (2007); Manning, Christopher D. / Raghavan, Prabhakar / Schutze, Hinrich, Introduction to Information Retrieval. Cambridge University Press (2008); Croft, Bruce. W. / Metzler, Donald / Strohman, Trevor, Search Engines, Information Retrieval in Practice. Addison Wesley, Boston etc. (2010):

on auf den Aspekt der Kommunikation und Nutzung von Wissen. Technisch — nach der Informationstheorie — ist es die Zeichenkette, die von einem Empfänger A an einen Empfänger B gesandt wird.[14] Informationswissenschaftlich wird die Datenmenge beschrieben, die zur Lösung eines Informationsproblems erforderlich ist.

4 Die Bedingungen einer Transparenz des Wissens

[Rz 22] Um daher die Daten in der Wissenswolke nutzen zu können, bedarf es geeigneter Instrumente, sogenannter «Wissensaufbereiter» oder «Wissensmediatoren». Diese können automatisiert, semi-automatisiert oder manuell agieren. Bedeutsam sind Suchmaschinen, mit oder ohne Einsatz von Metadaten, natürlich die bisherigen Informationsdienstleister (Massenmedien, Experten, Bibliotheken etc.) bzw. Experten. In allen Fällen sind die erkenntnistheoretischen Probleme evident vorhanden, was natürlich immer die Gefahr von Manipulationen mit sich bringt. Transparenz der digitalen Datenwelt wird damit zu einer (fast) unlösbaren Aufgabe.

[Rz 23] Sehr intensiv haben sich die Informationswissenschaft sowie das Wissensmanagement mit dieser Frage beschäftigt. Der Prozess des Auffindens und der Verwendung der Information durch Informations- und Kommunikationsprozesse ist durch das Information Retrieval, d.h. den Einsatz von Suchmaschinen, wesentlich erleichtert worden. Ohne Überwindung des hermeneutischen Problems des Vorwissens an Problemverständnis und Terminologie kommt es aber nur zu ungenügenden Ergebnissen. Im Wissensmanagement wird die Wissensbasis des Unternehmens bzw. der jeweiligen Person untersucht und methodisch verbessert. Neben der Nutzung des Wissens, dem Aufbau der Wissensrepositorien und der Wissenskartografie wird auf die Optimierung des Wissenskreislaufs groSSer Wert gelegt.

[Rz 24] Transparenz wäre somit eine gesellschaftspolitische Aufgabe des Aufbaus, der Bereitstellung, und der Kartografie des Wissens; eingebunden ist einen dynamischen Wissenskreislauf.

[Rz 25] Dadurch wird evident, dass es bei weitem nicht ausreicht, dass Wissenssammlungen in Wolken für alle angeboten werden. Das Wissen ist noch lange nicht transparent, weil noch wesentliche Voraussetzungen zur Beherrschung nötig sind. Einerseits müssen Suchalgorithmen gegeben sein, um aus der Fülle der Dokumente jene auszuwählen, die für die Frage tatsächlich relevant sind. Des Weiteren bedarf es eines erheblichen hermeneutischen Vorverständnisses, damit diese Dokumente auch verstanden werden. Dies ist eine Aufgabe der Bildung im Wissens- und Netzwerkzeitalter, die noch ungenügend gelöst ist.

[Rz 26] In der Folge soll die Qualität der Suchmaschinen als auch die Aufgabe der Metadaten untersucht werden. Hier liegen die wesentlichen Ansätze darin, aus der Fülle des Wissens jene Teile auszuwählen, die für die Frage relevant sind.

[Rz 27] Ein Gegensatz wird sich nie auflösen lassen: Die Wissenswolke soll transparent durch Suchmethoden sein, aber auch transparent im Sinne von Nichtbemerkbarkeit sein. Die Qual der Beherrschung des Wissens soll eliminiert werden. Dieser «magischen Herausforderung» nehmen sich die Suchmaschinenbetreiber an und entwickeln immer bessere Lösungen dazu, ohne dieses Problem jedoch ganz eliminieren zu können.

[14] Weaver, Warren / Shannon, Claude Elwood, The Mathematical Theory of Communication. Univ. of Illinois Press (1963).

5 Die Wissenswolke als Multimediakorpus

[Rz 28] Die Wissenswolke stellt eine riesige Dokumentensammlung dar, die vornehmlich aus Sprachdokumenten, zunehmend aber auch Bilder, Audiodateien, Grafiken, Filme usw. enthält. Für die Beherrschung gibt es neben der bibliografischen Beschreibung die Suchmaschinen sowie das Strukturierungswerkzeug der Metadaten.

[Rz 29] Die bibliografische Beschreibung ist keine FleiSSaufgabe, sondern notwendige Grundlagenarbeit. Schon die Strukturierung nach Autoren, Jahren, Erscheinungsorten, Dokumenttypen etc. bringt eine wesentliche Erleichterung zur Beherrschung der Wissenswolke.

[Rz 30] Am Beispiel des Rechts soll nun die Entwicklung wie der Status kurz beschrieben werden. Derzeit ist etwa die Hälfte des Rechtswissens der letzten 10 Jahre digital verfügbar. Es fehlen Kommentare, Lehrbücher, Rechtsgutachten von Anwaltskanzleien, Rechtsabteilungen etc. sowie Verwaltungsentscheidungen. Es gibt mehrere hundert Typen von Dokumenten in diesem Korpus. Die Unterscheidung von Autoren ist höchst relevant für die Relevanzbewertung des Dokuments; Haft spricht hier von Autoritäten des Rechts, deren ÄuSSerungen wesentliche Informationen für die juristische Arbeit sind.[15]

6 Suchmaschinen

[Rz 31] Zur Recherche im digitalen Dokumentkorpus einer Wissenssammlung ist die wesentlichste Unterstützung eine sogenannte Suchmaschine. Dies ist ein Programm zur Recherche von Dokumenten in digitalen Archiven. Eine Suchmaschine ist nur eine andere Bezeichnung für ein Information Retrieval-System. Kern jeder Suchmaschine ist ein Algorithmus zur Berechnung der Ähnlichkeit zwischen der Suchanfrage und dem Dokumentenkorpus. Am bekanntesten und am meisten verwendet ist die Boolesche Suche mit Distanzoperatoren. Während Experten damit durchaus zufrieden sind, «leiden» normale Menschen unter der formalen Suchsprache und ihrer unzureichenden Beherrschung der Terminologie.

[Rz 32] Bei der Frage des Wissenserwerbs kommt man heutzutage nicht mehr an Google vorbei. Wenn auch Google bei weitem nicht das gesamte Internet indiziert (es fehlt das Deep Web, das verstecke Internet[16] sowie das Darknet[17]), so ist doch eine Wissensmenge verfügbar, die weit über die einer kommerziellen Enzyklopädie hinausgeht. Dazu trägt natürlich auch und insbesondere die freie Enzyklopädie Wikipedia bei. Ein weiteres Problem besteht im Verzicht auf die Objektivierung des jeweiligen Wissens bzw. der Website. Daher muss überprüft werden, ob das Wissen auch tatsächlich authentisch und als wahr angesehen werden kann. Trotz wesentlicher Verbesserungen liegt immer noch eine Schwachstelle in diesem System der kooperativen Wissensschaffung.

[Rz 33] Die gleichnamige Suchmaschine des US-amerikanischen Unternehmens Google ist seit

---

15 Haft, Fritjof, Juristische Schreibschule, Anleitung zum strukturierten Schreiben. Normfall (2009).

16 Wikipedia, http://de.wikipedia.org/wiki/Deep_Web. Das sichtbare Web besteht aus den über Suchmaschinen zugänglichen Webseiten und wird auch Visible Web oder Surface Web genannt. Beim Deep Web handelt es sich vornehmlich um themenspezifische Datenbanken und Websites mit Indexierungsschutz (Zugangsbeschränkung oder Indexierungsverbot).

17 Wikipedia, http://de.wikipedia.org/wiki/Darknet. Hierbei wird zwar das Internet Protokoll genutzt, aber die Verbindungen selbst werden durch ein Peer-to-Peer-Overlay-Netzwerk manuell zwischen den Teilnehmern hergestellt.

dem 27. September 1998 online (Vorläufer: BackRub).[18] Wesentliches Merkmal und Vorteil von Google ist, neben der Schnelligkeit der Suche, die Qualität der Trefferliste. Diese wird nach einem Relevanzalgorithmus sortiert, welcher wesentlich besser als jene bei früheren Suchmaschinen wie AltaVista funktioniert. Hierbei wird das patentierte Verfahren namens PageRank verwendet, welcher die Popularität der Website und des Dokuments repräsentiert. Je höher die Website gewichtet ist und je mehr andere wichtige Websites auf diese Webseite bzw. das Dokument zeigen, desto höher ist der PageRank-Wert. Dazu verwendet Google noch über 200 weitere Faktoren für die Sortierung, wie beispielsweise den Ort des Auftretens der Suchbegriffe (z.B. Titel, Text, Überschriften). Es erfolgt eine laufende Anpassung der Algorithmen; zuletzt mit den Updates Google Panda (2011) und Google Penguin (2012). Im Dezember 2012 wurde der Knowledge Graph eingeführt: Dieser zeigt bei bestimmten Suchbegriffen (z.B. Tiere, Orte, Bauwerke und Menschen) auf der rechten Seite eine Detailansicht mit Daten; desgleichen werden ähnliche Suchbegriffe bzw. Objekte angezeigt. Google kann nun auch einfache Fragen beantworten.

[Rz 34] 2013 wurde die bisher gewichtigste Modifikation des Suchalgorithmus mit dem Algorithmus Hummingbird vorgenommen. Hummingbird bewertet neben der Linkpopularität auch die Suchanfrage selbst. Hier wird die Beziehung der Wörter der Suchanfrage, deren semantische Beduetung, verschiedene Schreibweisen und Aussprache, der Ort des Benutzers und die ausgewählte Sprache einbezogen, desgleichen wird eine Personalisierung der Suchergebnisse vorgenommen. Da Google im sogenannten Webprotokoll die durchgeführten Suchanfragen und die darauf besuchten Seiten protokolliert, können einerseits Wortdisambiguierungen vorgenommen werden (Golf als Sportart bzw. der PKW Golf).

[Rz 35] Dieses zunehmend semantische Verständnis der Suchanfrage birgt für den Nutzer wesentliche Vorteile. Der wesentlichste Nachteil liegt darin, dass die vorgenommene personenbezogene Speicherung der Suchgeschichte datenschutzrechtlich ohne ausdrückliche Zustimmung nicht möglich ist.[19]

7 Metadaten

[Rz 36] Metadaten sind Beschreibungsinformationen über Inhalte und Merkmale anderer Daten.[20] Die Prinzipien der formalen Beschreibung und Verweisung sind schon jahrhundertelange bibliothekarische Praxis. Die Digitalisierung hat das Instrument durch den Einsatz von Datenbanken, strukturierter Abfragesprache sowie Hyperlinks wesentlich verstärkt und die Beschränkungen des kartengebundenen Index überwunden. Die wesentlichen Metadaten sind:

- Bibliographische Daten (Dublin Core)
- Dokumentkategorie (Klassifikation)
- Beschreibung (Thesaurus)
- Zeitschichten (Geltung, Änderung)
- Verweise (Position im Netzwerk des [Rechts-]Systems)

[18] Wikipedia, http://de.wikipedia.org/wiki/Google.

[19] Schmitt, Stefan, Automatisch vorsortiert. In: Die Zeit 26/2011, 23. Juni 2011: «Wenn wir mit Google suchen oder Neuigkeiten bei Facebook lesen, passt das Netz sich unmerklich unseren Vorlieben an. Was bedeutet diese Verengung der Welt?».

[20] Wikipedia, http://de.wikipedia.org/wiki/Metadaten.

- Zusammenfassungen & Textextraktion (Verarbeitbarkeit)

[Rz 37] Ohne Verdichtung der Informationen durch den Einsatz von Metadaten ist kein optimales Handling von Big Data möglich; unsere Erkenntniswerkzeuge sind für die vielen Daten zu schwach. Man kann nicht alles Lesen, Anschauen oder auch nur Anhören. Diese Erfahrung zeigt sich in der überschieSSenden und extremen Nutzung von Big Data im Rahmen der Kommunikationsüberwachung der «Five Eyes».[21]

[Rz 38] Wichtig ist es jedoch auch, nicht nur die Suchmaschinen und die Metadaten zu betrachten, sondern auch, sich deren Qualität für die Verlinkung mit Fragestellungen anzusehen. Fragestellungen oder Probleme sind Lebenssituationen, Sachverhalte, Handlungen oder auch Unterlassungen.

[Rz 39] Es ist in der Wissenschaft anerkannt, dass die hochkomplexe Google-Suche die derzeit beste Lösung für die Beherrschung von Wissenswolken darstellt. Die Sprachsuche eignet sich gut, wenn die Frage gut formuliert werden kann. Dies ist sehr oft gegeben, wenn bereits eine grobe Kenntnis des Ergebnisses vorhanden ist (z.B. das Wissen von einem japanischen Restaurant in einem bestimmten Stadtteil, aber ohne Kenntnis des Namens und der Adresse etc.). Schwierig wird es, wenn die sprachliche Umschreibung des Suchproblems unscharf und ungenügend bleibt. Hier kann die Suche durch den Suchmaschinenbetreiber verbessert werden, d.h. es werden Begriffe insbesondere Synonyme dazu gefügt bzw. aufgrund des Kontextes und der persönlichen Interessen eine Disambiguierung der Begriffsausprägungen vorgenommen.

[Rz 40] Bei der Rechtssuche spielen Metadaten eine entscheidende Rolle. Ansonsten können Suchergebnisse nicht strukturiert werden; desgleichen sind diese für die Relevanzbewertung unverzichtbar. Metadaten können aber durch Aggregieren und die Bewertung des persönlichen Nutzerverhaltens teilweise ersetzt werden. Der Vorteil liegt darin, dass der Nutzer die komplexen Metadaten ignorieren kann. Es ist sehr interessant zu beobachten, dass bei häufig verwendeten Rechtsanfragen die Google Suche gute Ergebnisse liefert. Aufgrund der Verknüpfung und Verdichtung von früheren Anfragen mit Informationen ist eine hohe Qualität der Recherche gegeben. Beispielsweise liefert Google bei der Suche nach der Datenschutzrichtlinie sofort das relevante Dokument. Desgleichen schafft es Google, bei dieser Suche und aufgrund des Profils als deutschsprechender und in Wien ansässiger Mensch eine zielgerichtete Werbung für ein Datenschutzbuch zu machen. Die Vorschläge zur Ergänzung der Suchanfrage liefern wertvolle Vorarbeit für die Disambiguierung des Begriffes.

[Rz 41] Auf die datenschutzrechtliche Problematik dieser Suchmethodik wird unten noch im Detail eingegangen werden. Vorab: An sich bedarf es der Zustimmung des jeweiligen Nutzens, weil personenbezogene Daten gespeichert werden, woran doch gezweifelt werden darf. Es wäre natürlich wünschenswert, dass Google durch Akkreditierung der Daten bzw. Vertauschung der Daten die Erfordernisse an die Wahrung der Anonymität erzielt. Dies kann im Rahmen dieses Beitrags aber nicht untersucht werden.

[Rz 42] Für die eigentliche juristische Recherche schaut die Sache schon wesentlich komplexer aus. Es geht ja nicht um die Frage nach einem Dokument oder einer schon oft gelösten und daher nicht mehr strittigen Rechtsfrage. Vielmehr sollen insbes. die strittigen Rechtsfragen auf Basis des bestehenden Rechts gelöst bzw. auf ihre Übereinstimmung mit grundlegenden Prinzipien der Rechtsordnung und der Grundrechte überprüft werden. Damit wird die gesamte rechtliche Wis-

[21] Wikipedia, http://en.wikipedia.org/wiki/Five_Eyes.

senswolke in ihrer komplexen Struktur der Rechtsordnung zu berücksichtigen sein. Schon das geistige Bild der Millionen von Dokumenten lässt an einen riesigen Wald denken: Viele Bäume, aber wie finde ich mich dort zurecht?

[Rz 43] Zur ErschlieSSung werden Rechtsinformationssysteme oder besser Rechtsretrievalsysteme verwendet. Diese funktionieren sehr ähnlich wie Internet Suchmaschinen, waren auch schon vorher da; zeichnen sich aber durch wesentlich komplexere Suche unter Einbeziehung der Metadaten aus. Desgleichen ist die Suchfrage eine wesentlich differenzierte. Es geht nicht darum, ein bestimmtes — sehr oft redundant vorhandenes — Wissen zu finden, sondern zu einer komplexen Rechtsfrage relevante Dokumente zu finden, und zwar mit einer Nachweisquote von 100%. Das Suchspiel mit vielen Begriffen und Metadaten mit regelmäSSigen Kontrollschleifen mit Experten bzw. Handbüchern und Kommentaren ist praktische Realität.

[Rz 44] Für die Standardfälle der Rechtsordnung wurden schon in den 1990er Jahren sogenannte Bürgerinformationssysteme aufgebaut. Diese orientieren sich an Lebenssituationen, beschreiben die Rechtslage und bieten eine oft unverzichtbare Hilfe im Behördenweg bzw. in der Bereitstellung der notwendigen Formulare. Aufgrund der doch geringen Anzahl der relevanten Lebenssituationen ist diese Hilfe gut anwendbar und letztlich unverzichtbar für den Bürger geworden.[22]

[Rz 45] Für die juristische Suche jedoch muss gesagt werden, dass es ohne hermeneutisches Vorwissen nicht geht. Es wird eine ausreichende Kenntnis der Rechtsordnung, ihrer Verfahren, der Rechtssprache bzw. des verwendeten Vokabulars etc. vorausgesetzt. Eine Suche ist natürlich auch ohne dieses Vorwissen möglich, das Ergebnis ist hilfreich in der weiteren Suche, es kann aber nur schwerlich bewertet werden, ob auch alle relevanten Dokumente gefunden wurden oder ob die gefundenen Dokumente auch relevant sind. Die Rechtsinformation erfordert daher eine exakte Beherrschung der sprachlichen Vielfalt, wobei bisher nur eine recht bescheidene Unterstützung von Seiten der jeweiligen Anbieter geboten wird. Hier ist anzumerken, dass Google schon viel weiter ist und eine sehr hilfreiche Suchergänzung anbietet.

[Rz 46] Am Beispiel des Rechts zeigt sich, dass bei weitem nicht alles Wissen transparent ist und schon gar nicht die Rechtsinformation. Obwohl letztlich jeder davon betroffen ist, bedarf es zum Auffinden der relevanten Dokumente eines erheblichen Aufwandes an speziellen Kenntnissen und Fähigkeiten. Auch Big Data hat daran nichts geändert; wohl aber ist zu erwarten, dass durch derartige Methoden eine wesentlich bessere Analyse des Dokumentenkorpus vorhanden ist. Der Aufwand der Suche wird verringert, wenn mehr Metadaten und mehr Kenntnisse über die verwendete Rechtssprache vorhanden sind. Effizientes Ranking, wie bei Google Standard, ist bei juristischen Rechtsinformationssystemen noch ungenügend gelöst. Aus diesem Grund ist eine dynamische Datenanalyse der Rechtskorpora so bedeutsam. Ist diese Untersuchung — mit einem neuen Wort Rechtsdatalystik genannt — erfolgreich, können die Ergebnisse für eine zielgerichtete Suche verwendet werden. Im Ergebnis könnte dies zu einer intelligenten Suche führen.[23]

---

[22] Krenmayer, Andreas / Traunmüller, Roland, Bürgerinformationssysteme, Neue Vorstellungen. In: Schweighofer, Erich, Kummer, Franz, Hötzendorfer, Walter (Hrsg./eds.), Kooperation, Tagungsband des 18. Internationalen Rechtsinformatik Symposions IRIS 2015, 26.—28. Februar 2015, books@ocg.at, Wien 2015, S. 227—234 (2015); bzw. Krenmayer, Andreas / Traunmüller, Roland, Bürgerinformationssysteme, Neue Vorstellungen, in: Jusletter IT 26. Februar 2015.

[23] Schweighofer, Erich, Rechtsdatalystik — Versuch einer Teiltheorie der Rechtsinformatik. In: Schweighofer, Erich, Kummer, Franz, Hötzendorfer, Walter (Hrsg./eds.), Kooperation, Tagungsband des 18. Internationalen Rechtsinformatik Symposions IRIS 2015, 26.—28. Februar 2015. books@ocg.at, Wien 2015, 61—72 (2015); bzw. Schweighofer, Erich

8 Datenschutzgerechte Vernetzungs-Identifikation

[Rz 47] In der Verwendung von Nutzerdaten wird ein groSSes Potential für die Rechtsrecherche der Zukunft gesehen. Während Westlaw[24] sich der datenschutzrechtlichen Problematik offensichtlich bewusst ist, fordert die Praxis von Google bzw. Facebook den Datenschutz heraus. Die Suchanfragen werden mit Lebenssituationen und Menschen vernetzt. Es scheint nunmehr so, dass durch die Notwendigkeiten von Big Data es keinen Ausweg mehr gibt, dieser Vernetzung von Daten mit Menschen zu entgehen. Die Situation kann aber nicht hingenommen werden, weil immer mehr Menschen Big Data nicht als Chance sondern als Bedrohung ansehen.[25]

[Rz 48] Die Frage ist daher, wie Daten für den jeweiligen Zweck effektiv und verhältnismäSSig miteinander vernetzt werden können. Der Nutzer soll die personalisierten Suchergebnisse bekommen; aber nicht hinnehmen müssen, dass die Suche auf ihn persönlich zurückgeführt werden kann. Ein Bild der Nutzer-Präferenzen zeigen auch der Browser, der Domain Name Server. Als Ziel gilt daher die Anonymisierung der Suchanfragen. Dies wird in der Praxis aber nicht mehr erreichbar sein. Daher wird vorgeschlagen, dass eine Vielzahl von Pseudo-Anonymitäten verwendet wird, die zwar durch föderierte Identitätssysteme miteinander vernetzt sind; aber aufgrund der hohen Kosten können diese im Normalfall nicht zusammengeführt werden. Für meine Person bedarf es nicht des Namens sondern es reicht fast immer die Beschreibung als Wiener Rechtswissenschaftler über 50 Jahre mit der Spezialisierung in der Rechtsinformatik[26]. Dies reicht für eine bessere Suche, aber es wird auch dem Datenschutz Rechnung getragen.

9 Schlussfolgerungen

[Rz 49] Für die Transparenz von Wissen sind drei Voraussetzungen notwendig: eine leistungsfähige Informations- und Kommunikationsinfrastruktur, der Aufbau von Wissensrepositorien sowie die Verfügbarkeit von leistungsstarken «Wissensmediatoren» wie Suchmaschinen. In den letzten 20 Jahren hat sich die Situation wesentlich verbessert. Sehr viel Wissen ist über das Internet verfügbar; Big Data ist da und damit auch eine riesige Wissenswolke. Was fehlt, sind semantische Suchmaschinen; boolesche Suchlogik ohne Semantik und Pragmatik liefert unzureichende Ergebnisse. Derzeit ist das Internet noch eher Textspeicher mit Suchmaschine, aber kein Wissensspeicher. Daher ist Big Data nicht transparent im Sinne der Sozialwissenschaften und auch nicht transparent im Sinne der Informatik.

[Rz 50] Am Beispiel des Rechtssystems lässt sich gut illustrieren, dass noch sehr viel Arbeit zu erledigen ist. Die besten Lösungsansätze liefert Google mit einem hochkomplexen Suchalgorithmus, der schon semantische Elemente aufweist. Die besten Erfolge scheint Google aber mit der assoziativen Vernetzung von Suchanfragen mit Dokumenten zu erzielen. Hier ist aber die Frage des Datenschutzes noch nicht ausreichend geklärt. Die Herausforderung der Vernetzungs-Identifikation muss gelöst werden. Dann wird auch eine intelligente Big Data Nutzung möglich und auch vom Nutzer akzeptiert werden.

---

[24] Westlaw verwendet die Nutzerdaten nur bei Überschreiten einer hohen Aggregationsschwelle von Nutzern, womit eine Anonymisierung realisiert wird.

[25] Knyrim, Rainer, Datenschutzrecht. 2. Aufl., Manz, Wien (2013).

[26] Dies wird im weiten Sinne verstanden: Rechtsinformation, AI & Recht, Automatisierung des Rechts und IT-Recht. Davon gibt es etwa 3 bis 4; mit jeweils unterschiedlichem Hauptfach.

Ao. Universitätsprofessor Erich Schweighofer, Universität Wien, Leiter der Arbeitsgruppe Rechtsinformatik. Rechtswissenschaftliche Fakultät, Institut für Europarecht, Internationales Recht und Rechtsvergleichung, Abteilung für Völkerrecht. Erich.Schweighofer@univie.ac.at; http://rechtsinformatik.univie.ac.at. Herausgeber von Jusletter IT — Die Zeitschrift für IT und Recht.

Den Podcast zum Vortrag von Erich Schweighofer «Die Transparenzfiktion in der Big Data Welt», gehalten am 5. November 2014 bei der Tagung Informatik und Recht zum Thema Big Data Governance, finden Sie in dieser Ausgabe von Jusletter IT:

- Erich Schweighofer, Die Transparenzfiktion in der Big Data Welt (Podcast), in: Jusletter IT 21. Mai 2015