Adieu Einwilligung?

Neue Herausforderungen für die informationelle Selbstbestimmung im Angesicht von Big Data-Technologien

Max-R. Ulbricht und Karsten Weber

1. Einleitung

Die Bundesregierung weist in ihrem Strategiepapier *Digitale Agenda 2014–2017* darauf hin, dass es notwendig sei, Big Data-Technologien »weiter zu erschließen«, um die deutsche Wirtschaft durch die Gewinnung neuer Erkenntnisse aus der Analyse großer Datenbestände produktiver zu gestalten. Im selben Atemzug wird vor den Risiken und Gefahren gewarnt, welche diese Technologien begünstigen können.[1] Um künftig die mit diesen Technologien verbundenen Potenziale nutzen zu können und die Risiken zu minimieren, wurde Anfang des Jahres 2015 das IT-Sicherheits-Forschungsrahmenprogramm *Selbstbestimmt und sicher in der digitalen Welt 2015–2020* vorgestellt, welches (unter anderem) zum Ziel hat, Konzepte zu entwickeln, welche dafür Sorge tragen sollen, informationelle Selbstbestimmung auch bei Big Data-Anwendungen gewährleisten zu können.[2]

In diesem Diskussionsbeitrag soll der Frage nachgegangen werden, wie dies erreicht werden könnte. Hierzu wird die Entwicklung wichtiger Prinzipien zur Wahrung der informationellen Selbstbestimmung als Reaktion auf den technologischen Fortschritt skizziert, anschließend grundlegende Big Data-(bzw. Data Mining-)Architekturen und die damit einhergehenden Herausforderungen für mögliche datenschutzfreundliche Erweiterungen betrachtet und nachfolgend bereits

* Max-R. Ulbricht | Technische Universität Berlin | mu@ise.tu-berlin.de
Prof. Dr. habil. Karsten Weber | Ostbayerische Technische Hochschule Regensburg | Karsten.Weber@oth-regensburg.de
Teile dieses Textes sind (in englischer Übersetzung) genutzt worden, um die skizzierte Plattform im Rahmen eines Workshops auf der *IEEE International Conference on Cloud Engineering 2016* vorzustellen (Ulbricht und Pallas, »CoMaFeDS«).

1 Bundesregierung, *Digitale Agenda 2014 – 2017*, S. 4-5.

2 Bundesministerium für Bildung und Forschung, *Selbstbestimmt und sicher in der digitalen Welt*, S. 20.

M. Friedewald et al. (Hrsg.), *Informationelle Selbstbestimmung im digitalen Wandel*, DuD-Fachbeiträge, DOI 10.1007/978-3-658-17662-4_16

existierende Lösungsansätze auf ihre Eignung für die angestrebten technischen Weiterentwicklungen analysiert. Aus den daraus gewonnen Erkenntnissen wird ein Konzept für technische Erweiterungen bzw. Ergänzungen von Big Data-Architekturen entworfen, welches es ermöglicht, die Vorteile von Big Data zu nutzen, ohne die informationelle Selbstbestimmung aufgeben zu müssen.

2. Informationelle Selbstbestimmung: Wandel eines Rechts auf Privatsphäre im Zuge technischer Entwicklungen

Eine ausführliche Würdigung der Genese der Ideen zur Schutzwürdigkeit der Privatsphäre und zum Datenschutz als einem Werkzeug, um diesen Schutz zu erreichen, kann hier nicht geleistet werden. Ein historischer Abriss kann jedoch bereits verdeutlichen, dass einige der derzeit geführten Debatten weit in der Vergangenheit verankert sind und schon zu früheren Zeiten sehr oft Technik Ausgangspunkt entsprechender Überlegungen waren. Eine entscheidende Zielsetzung ist dabei häufig, das Wohl bzw. das Glück der Menschen zu steigern.

Um dieses Ziel zu erreichen zu können, muss man entweder die Menschen zur Einsicht in die wahre Natur des Glücks führen oder ihr Handeln so steuern, dass sie – womöglich ohne es zu merken – das Richtige tun; so kann man einige Ideen zur Glückssteigerung zusammenfassen. Angestoßen durch Thalers und Sunsteins 2009 erschienenes Buch *Nudge* wird die zweite Variante heute unter der Bezeichnung »libertarian paternalism«[3] diskutiert: Ziel ist, die Freiheit der Menschen zu erhalten und trotzdem zu erreichen, dass sie das Richtige tun. In Platons *Politeia* hingegen ist die beschriebene Gemeinschaft durch ubiquitäre Unfreiheit charakterisiert; Platon wollte jeglichen Aspekt des menschlichen Lebens kontrollieren. Was wir heute als Bestandteil unserer Privatsphäre erachten, namentlich die Rückzugsmöglichkeit vor Eingriffen und neugierigen Blicken anderer Menschen, Unternehmen, Institutionen oder staatlichen Autoritäten, lehnte Platon ab, da sie Quelle von Unfrieden und Unglück sei. In der einschlägigen Literatur zur Genese des modernen Begriffs der Privatsphäre wird nach dem Rekurs auf Platon meist ein großer zeitlicher Sprung vollzogen und dann auf den Artikel *The Right to Privacy* von Samuel D. Warren und Louis D. Brandeis[4] aus dem Jahr 1890 verwiesen. Privatsphäre bedeutet dort das Recht, in Ruhe gelassen und nicht gestört zu werden: »the right to be let alone«. Im Gegensatz zu der unfreien Gemeinschaft in der *Politeia* – so kann man Warren und Brandeis verstehen – gewährt eine liberale, auf rechtsstaatliche Prinzipien aufbauende Gesellschaft ihren Mitgliedern einen unverletzbaren Rückzugsraum, der ausschließlich der Willkür der jeweiligen Person unterworfen ist. Damit wird

[3] Thaler und Sunstein, *Nudge*.

[4] Warren und Brandeis, »The Right to Privacy«.

die Argumentation John Stuart Mills[5] für die individuelle und gesellschaftliche Bedeutung der Privatsphäre als Rückzugsraum aufgenommen: es ist die Privatsphäre, die auch exzentrische Lebensentwürfe ermögliche und damit überhaupt erst Autonomie gewähre.

Warren und Brandeis nahmen das Aufkommen einer neuen Informations- und Kommunikationstechnologie zum Anlass, über Privatsphäre nachzudenken, denn Ende des 19. Jahrhunderts verbreitete sich die Fotografie nicht zuletzt als Werkzeug der Presseberichterstattung. Für Deutschland gilt Ähnliches, denn im deutschen Kaiserreich gab es gegen Ende des 19. Jahrhunderts einen dramatischen Fall, der zusammen mit einer schon länger andauernden rechtswissenschaftlichen Debatte zur juristischen Kodifizierung des Rechts am eigenen Bild beitrug: Reporter drangen in das Sterbezimmer des ehemaligen Reichskanzlers Otto von Bismarck ein, fotografierten den Leichnam und versuchten die Fotos zu verkaufen.[6]

Tatsächlich jedoch taucht das Thema und die entsprechende Kurzformel in der US-amerikanischen juristischen Literatur schon auf, als Thomas M. Cooley 1879 die Verletzung der Privatsphäre in seinem Werk *Treatise on the law of torts* behandelt;[7] in §101 bespricht Cooley das Problem der Verletzung der Privatsphäre durch Personenbildnisse zum Zweck der Werbung. Er verweist darauf, dass verschiedene Gerichte zu unterschiedlichen Urteilen gekommen seien: Der New York *Court of Appeals* habe ein Recht auf Privatsphäre vor allem mit dem Hinweis abgelehnt, dass es völlig ausufern würde, weil keine klare Grenze dessen zu ziehen sei, was privat sei; der *Supreme Court of Georgia* wiederum aber habe ein Recht auf Privatsphäre mit einer Analogie zum Eigentumsschutz bestätigt (knapp 130 Jahre später nimmt Lessig[8] diesen Topos wieder auf). Von Beginn der modernen Debatte an sind also die Begründung der Privatsphäre und damit deren Legitimation wie auch die Legitimation des Eingriffs in die Privatsphäre umstritten. Der letzte Satz, den Cooley aus dem Urteil des Supreme Court of Georgia zitiert, fasst viele der bis heute relevanten Diskussionsstränge zusammen: »[1] The right would be conceded if she had sat for her photograph; [2] but if her face or her portrait has a value, the value is hers exclusively; [3] until use be granted away to the public.«[9] Der erste Teil des Zitats hebt auf die Notwendigkeit einer Einwilligung der jeweiligen Person ab, über die Informationen gesammelt werden sollen – das Fotografieren ist hier als Paradigma des Informationseingriffs zu verstehen. Diese zentrale Forderung (ebenso wie die Zweckbindung) findet sich heute bspw. in der Datenschutz-Grundverordnung

5 Mill, *On Liberty / Über die Freiheit*.
6 Gerhardt und Steffen, *Kleiner Knigge des Presserechts*, S. 206.
7 Cooley, *A Treatise on the Law of Torts*.
8 Lessig, »Privacy as Property«.
9 Cooley, *A Treatise on the Law of Torts*, S. 194f.

(DS-GVO) auf EU-Ebene,[10] den OECD Privacy Guidelines,[11] der EU Grundrechtecharta[12] oder auch der ISO/IEC 29100[13] wieder. Im zweiten Teilsatz wird erneut die Autonomie der Person betont, der zufolge sie allein über die Verwendung personenbezogener Informationen bestimmt; die heutige Forderung nach Zweckbindung der erhobenen Informationen könnte hieraus abgeleitet werden, wenn man die Art der Verwendung als Teil der Nutzungsvereinbarung ansieht, der eine Person zustimmt. Bis hier klingen die getroffenen Aussagen bzgl. aktueller Debatten sehr vertraut. Der dritte Teil des Zitats hingegen birgt gerade in Zeiten von Social Media und vor allem Big Data erhebliches Konfliktpotenzial, da man die darin enthaltende Aussage so deuten kann, dass eine einmal gegebene Einwilligung zur Nutzung von Informationen zur Folge habe, dass damit eine endgültige Preisgabe der Informationen einherginge – das wäre das Gegenteil der Idee der Zweckbindung. Ein »Rückholen« im Rahmen eines »Rechts auf Vergessen« ist, folgt man dem Urteil des *Supreme Court of Georgia*, ebenfalls nicht begründbar. Die Parallele, die durch ein »Recht allein gelassen zu werden« und ein »Recht vergessen zu werden« nahegelegt wird, findet damit also keine Zustimmung; in der DS-GVO wird eine abgeschwächte Variante als »Recht auf Löschung« gefordert.

Mit diesen Texten und den darin enthaltenen Begründungsmustern geht eine (möglicherweise unvermeidbare und notwendige) Begriffsverschiebung einher,[14] die bis heute die Diskussion prägt, denn es geht nicht mehr zentral um das »private life«, von dem Mill spricht, sondern um Informationen über das Private. Mit dieser Verschiebung geht einher, dass meist von Datenschutz gesprochen wird, wenn der Schutz der Privatsphäre verhandelt wird. Ob dieser Bedeutungswandel sinnvoll ist, kann hier nicht weiter untersucht werden; vermutlich trägt er aber dazu bei, dass viele Autoren konstatieren, dass begriffliche Unsicherheiten existieren. Marx[15] sieht dies als Ursache, dass oft nicht ausreichend klar zwischen privater und öffentlicher Sphäre unterschieden werden würde; Michelfelder[16] wiederum vermutet, dass sich jene, die sich mit Privatsphäre auseinandersetzen, sich schnell in einem Begriffswirrwarr wiederfänden, bei dem kaum Einigkeit über den Diskussionsgegenstand hergestellt werden könne. Die Gefahr, die der *New York Courts of Appeals* im 19. Jahrhundert gegen ein Recht auf Privatsphäre anführte, ist also zumindest auf der begrifflichen Ebene gegeben.

---

10 »Verordnung (EU) 2016/679 (Datenschutz-Grundverordnung)«.
11 *OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data*.
12 »Charta der Grundrechte der Europäischen Union«.
13 *ISO/IEC 29100:2011 – Information technology – Security techniques – Privacy framework*.
14 Moor, »Towards a Theory of Privacy in the Information Age«.
15 Marx, »Murky conceptual waters: The public and the private«, S. 161.
16 Michelfelder, »The moral value of informational privacy in cyberspace«, S. 129.

Es ist aber zu vermuten, dass alle, die für ein Recht auf Privatsphäre plädieren, dem Zusammenhang von Privatsphäre auf der einen und Freiheit bzw. Autonomie auf der anderen Seite, den Alan F. Westin[17] etabliert hat, zustimmen werden. Damit ist als Minimalbedingung für Datenschutz und Privatsphäre die Zustimmung bei Eingriffen zu sehen; dies ist Grundlage der Idee des Menschen als autonomes Wesen. Folgt man einer »limitation theory of privacy« oder einer »control/restriction access theory of privacy«,[18] so lassen sich daraus die Prinzipien der Zweckbindung und auch der Datensparsamkeit ableiten. Entscheidend ist, dass die wichtigsten heute diskutierten (philosophischen) Theorien der Privatsphäre davon ausgehen, dass Personen bei *jedem* Informationseingriff neu entscheiden, ob sie dem zustimmen wollen. Es ist jedoch offensichtlich, dass dies der Lebenswelt nicht entsprechen kann. Zustimmung, Zweckbindung und andere Prinzipien des Schutzes der Privatsphäre und des Datenschutzes müssen also an gesellschaftliche und nicht zuletzt technische Veränderungen angepasst werden. Entscheidend bei diesem Wandel wird sein, ob die grundsätzliche Idee dabei erhalten bleiben kann oder am Schluss eine Haltung steht, wie sie vor mehr als 100 Jahren der *New York Court of Appeals* äußerte und heute durch die Post-Privacy-Bewegung vertreten wird – dass ein Recht auf Privatsphäre aus juristischer Sicht nicht gewährt werden könne oder aus sozialer Sicht nachteilig sei. Im Folgenden soll jedoch aufgezeigt werden, dass die Aufgabe der Privatsphäre nicht notwendig ist, um neue Entwicklungen gesellschaftlich verträglich nutzen zu können.

3. Risiken und Nebenwirkungen von Big Data

Was in jüngerer Zeit unter dem Stichwort »Big Data« von Politik, Wirtschaft und Medien als Neuigkeit gefeiert wird, kann aus technologischer Perspektive durchaus als »alter Wein in neuen Schläuchen« bezeichnet werden. Unter den Begriffen »Knowledge Discovery in Databases« (KDD)[19] bzw. »Data Mining«[20] wurden einige der Konzepte und Technologien bereits vor Jahrzehnten entwickelt. Die große öffentliche Aufmerksamkeit bzgl. Big Data ist, neben der Verfeinerung der wirkenden Mechanismen, dem Umstand geschuldet, dass die momentan verfügbare Hardware signifikant leistungsfähiger ist, als es zur Zeit der Entwicklung der grundlegenden Ideen und Konzepte der Fall war. Die Performancesteigerung moderner Rechnersysteme sowie ständig verfügbare Hochgeschwindigkeitsnetzwerke, welche diese Rechnersysteme miteinander verbinden, sorgen dafür, dass es

17 Westin, *Privacy and freedom*.
18 Tavani, »KDD, data mining, and the challenge for normative privacy«.
19 Fayyad, Piatetsky-Shapiro und Smyth, »From Data Mining to Knowledge Discovery in Databases«.
20 Chen, Han und Yu, »Data mining«.

heute möglich ist, praktisch in Echtzeit große Datenmengen aus verschiedensten Quellen zu (re-)kombinieren, zu analysieren und daraus neue Erkenntnisse zu generieren. In diesem Grundprinzip von Big Data, der Kombination verschiedenster Datensätze aus unterschiedlichen Quellen, liegen nicht nur die größten Gefahren und Risiken begründet, sondern auch enorme Herausforderungen an angestrebte Regulierungsvorhaben, sowohl rechtlicher als auch technischer Natur.

Im Folgenden sollen die Herausforderungen bezüglich der datenschutzfreundlichen Integration heterogener, autonomer und verteilter Datenquellen zur Anwendung von Big Data-Analysen mithilfe eines einfachen Szenarios anschaulich skizziert werden, anhand dessen kurz grundlegende Data Mining Ansätze vorgestellt sowie die technischen und rechtlichen Problemfelder aufzeigt werden.

3.1. Szenario

Man stelle sich ein Institut für Verkehrsforschung vor, welches vor dem Hintergrund städtebaulicher Entwicklung daran interessiert ist, die Sicherheit von Verkehrsflüssen dahingehend zu optimieren, den Stresslevel beteiligter Verkehrsteilnehmer durch städtebauliche Maßnahmen positiv zu beeinflussen und damit die Wahrscheinlichkeit auftretender Unfälle zu vermindern.

Um dieses Vorhaben zu realisieren, wäre eine Messung diverser Stressindikatoren wie Pulsschläge, Hautwiderstand und -feuchtigkeit unter verschiedenen typischen Situationen bei der Teilnahme am Straßenverkehr wünschenswert. So ließe sich beispielsweise ermitteln, inwiefern sich unterschiedliche straßenbauliche Gegebenheiten wie Kreuzungen mit oder ohne Ampel im Gegensatz zu Kreisverkehren auf die oben genannten Stressindikatoren von Auto-, Motorrad- oder Fahrradfahrern, aber auch von Fußgängern auswirken. Der Einfluss des Wetters als Stressfaktor sollte, wenn möglich, ausgeschlossen werden können.

Um derlei Korrelationen zu finden, wäre es in einer kontrollierten wissenschaftlichen Studie nötig, Teilnehmer zu rekrutieren, diese mit diversen medizinischen Messgeräten sowie Fahrzeugen auszustatten und anschließend die Indikatoren für Stress während der Teilnahme an typischen Straßenverkehrssituationen zu messen und aufzuzeichnen. Anschließend würden die erhobenen Daten bezüglich der obigen Fragestellungen analysiert, um die erwarteten Korrelationen zu bestätigen oder zu widerlegen. Dieses wissenschaftliche Standardverfahren erzeugt aber aufseiten des Instituts nicht unerheblichen Aufwand und damit Kosten. Um einen signifikanten Teil der Kosten der Datenerhebung zu minimieren, könnte man auf die Idee kommen, bereits vorhandene Daten, welche eigentlich zu anderen Zwecken erhoben wurden, zu nutzen.

Informationen über Verkehrsflüsse ließen sich beispielsweise aus den Datenbanken von Anbietern von Navigationslösungen beziehen. So bieten u. a. Google

(über seinen Dienst »Maps«[21]) oder TomTom (mit seinem »MyDrive«-Service[22]) detaillierte Informationen zu Verkehrsflüssen, welche sich aus den Bewegungsdaten der genutzten (vernetzten) Navigationslösungen des jeweiligen Anbieters errechnen. Daten zum Stresslevel der beteiligten Verkehrsteilnehmer sind prinzipiell in den Datenbanken von Herstellern sogenannter Wearables zu finden. Diese am Körper getragenen Geräte zeichnen teilweise Körperdaten wie Puls, Hautfeuchtigkeit und -widerstand auf, um es ihrem Träger zu ermöglichen, seinen Körperzustand auf Basis der bereitgestellten Informationen zu optimieren.[23] Über das Mobiltelefon des Trägers, mit welchem sich die genannten Geräte üblicherweise verbinden, um die gemessenen Daten mit der Datenbank des jeweiligen Anbieters zu synchronisieren, ließe sich der notwendige Ortsbezug herstellen. Auch die relevanten Wetterdaten zu den untersuchten Orten liegen grundsätzlich vor. Diese könnten aus den Datenbanken von Anbietern vernetzter, privat betriebener Wetterstationen wie netatmo[24] bezogen werden.

All diesen Daten gemein ist, dass sie öffentlich vorliegen[25] und prinzipiell auch über geeignete Programmierschnittstellen abfragbar wären. Fraglich bleibt hingegen, ob die skizzierte Kombination dieser verschiedenen, unabhängigen Datenquellen in einem rechtlich zulässigen Verfahren realisierbar ist.

3.2. Technische Herausforderungen

Wenn als wesentliche Charakteristik neuartiger Big Data-Verfahren die Verknüpfung bereits vorhandener Datensätze zur Gewinnung neuer Erkenntnisse aus eben dieser neuen Kombination angenommen wird, kommt deren Integration bzw. Zusammenführung ein besonderer Stellenwert zu.

Für das verteilte Data Mining (z. B. Big Data-Analysen) lassen sich prinzipiell zwei grundlegende Architekturen identifizieren. Einerseits kann die Analyse der Daten lokal auf den einzelnen Datenquellen erfolgen und anschließend die dort gefundenen Zusammenhänge zu einem globalen Modell kombiniert werden. Dieses Verfahren hat allerding den Nachteil, dass sich die gerade spannenden Zusammenhänge über verschiedene Datenquellen hinweg nur schwer bis gar nicht erkennen lassen. Die Fragestellungen des oben beschriebenen Szenarios ließen sich mit diesem Ansatz beispielweise nicht beantworten. Der zweite Ansatz kombiniert daher

21 https://www.google.de/maps/

22 https://mydrive.tomtom.com

23 Swan, »Sensor Mania!«, »The Quantified Self«.

24 https://www.netatmo.com/weathermap

25 Während die erwähnten Verkehrs- und Wetterdaten öffentlich zugänglich sind, ist dies bei den skizzierten Körperdaten davon abhängig, welche Plattform bzw. welche Anbieter zur Aggregation genutzt und welche Einstellungen vorgenommen werden.

vor der eigentlichen Analyse die gewünschten Datensätze, sodass der eigentliche Mining-Vorgang dann auf dieser nun breiteren Datenbasis erfolgt.[26]

Big Data-Analysen, wie jene im oben skizzierten Szenario, versuchen komplexe Korrelationen innerhalb großer Datenmengen, welche in heterogenen und autonom betriebenen Datenquellen gehalten werden, die wiederum unter verteilter, dezentraler Kontrolle stehen, zu finden.[27] Für die dazu notwendige Integration der verschiedenen Datenquellen lassen sich zwei grundlegende Strategien verfolgen, welche verschiedene Aspekte der oben genannten Eigenschaften Heterogenität, Autonomie und Verteiltheit[28] bedienen.

Während die sogenannte »materialisierte Integration«, oft auch »Data Warehousing« genannt, vorhandene Daten aus den zu integrierenden Quellen in eine neue Datenbank, das Warehouse, transferiert und für weitere Verarbeitungsprozesse über eine einheitliche Schnittstelle zugreifbar macht, belässt eine »virtuelle Integration« die Daten bei den ursprünglichen Datenquellen. Dabei sorgt eine Zwischenschicht (realisiert als Mediator[29] oder Föderation[30]), welche das Wissen enthält, welche Daten wo lagern und wie darauf zugegriffen werden kann, dafür, dass sich für Nutzer oder Applikationen die kombinierten Datenquellen wie eine einzelne »große« Datenbank abfragen lassen.[31]

Beide Integrationsstrategien lösen zwar sowohl die Heterogenität und Verteiltheit der integrierten Datenquellen für Analysevorhaben auf, verlangen aber von den Haltern der ursprünglichen Datenquellen signifikant verschiedene Maße an Kooperation. Beim »Data Warehousing« geben die originären Datenhalter die Kontrolle über die von ihnen zur Verfügung gestellten Daten komplett ab, da sie nach deren Transfer künftig keinerlei Einfluss auf die weitere Verarbeitung nehmen können. Demgegenüber bleibt bei der »virtuellen Integration« die Autonomie der integrierten Datenquellen erhalten, da hier keine Übertragung der Daten stattfindet, sondern lediglich (temporärer) Datenzugang gewährt wird.

Es lässt sich schlussfolgern, dass für institutionenübergreifende Analysevorhaben, wie sie im obigen Szenario skizziert wurden, eine virtuelle Integrationsstrategie für alle Beteiligten vorteilhaft wäre. Die Daten bleiben zwar vollständig unter Kontrolle der datenhaltenden Institutionen, aber es wird gleichzeitig möglich, per Big Data-Analyseverfahren Korrelationen zwischen diesen autonomen Datenquellen zu finden und damit je nach Fragestellung einen gesellschaftlichen und/oder wirtschaftlichen Mehrwert zu generieren.

---

26 Park und Kargupta, »Distributed Data Mining«.
27 Wu u. a., »Data Mining with Big Data«.
28 Hasselbring, »Information System Integration«.
29 Wiederhold, »Mediators in the Architecture of Future Information Systems«.
30 Sheth und Larson, »Federated Database Systems«.
31 Doan, Halevy und Ives, *Principles of Data Integration*.

3.3. Gefahr der De-Anonymisierung

Eine zentrale Säule des Schutzes der informationellen Selbstbestimmung ist die Anonymisierung. Wenn es gelingt, die Repräsentation einer Person, egal ob im realen oder virtuellen Raum, von ihren Identitätsmerkmalen zu »entknüpfen« bzw. »entkoppeln«, sodass von ihr innerhalb einer Gruppe von sonst »Gleichartigen« keinerlei Alleinstellungsmerkmale existieren, diese Person also innerhalb der Gruppe nicht mehr identifiziert werden kann,[32] so kann von einer Anonymisierung gesprochen werden.

Sowohl rechtliche Regularien als auch Leitlinien und Rahmenvereinbarungen zum Datenschutz verlangen und/oder empfehlen, wo immer möglich Verfahren zur Anonymisierung einzusetzen, also aus Datensätzen identifizierenden Merkmale von natürlichen Personen zu entfernen und somit den Personenbezug zu beseitigen. Ohne weitere Vorgaben muss ein solches Verfahren aber nicht zwingend zielführend sein.

Gerade in Zusammenhang mit Big Data-Technologien und den oben skizzierten Integrationen verschiedener Datenquellen ist die Anonymisierung immer weniger tragfähig,[33] da sich durch Zusammenführung und Auswertung mehrerer (scheinbar) anonymer Datensätze einzelne Daten de-anonymisieren lassen, was, wie zahlreiche Beispiele[34] verdeutlichen, dazu führt, dass auch eine Re-Identifikation der sie betreffenden Individuen möglich wird.[35]

3.4. Grenzen der Einwilligung

Leicht nachvollziehbar ist, dass die zentralen Datenschutzprinzipien der Einwilligung und Zweckbindung bei den angestrebten Nutzungsweisen von Big Data-Technologien kritisch zu hinterfragen sind.[36] Dies lässt sich am oben skizzierten Szenario recht plastisch nachvollziehen. Wenn ein privater Betreiber einer Wetterstation die von dieser Station gemessenen Daten für die Verknüpfung mit den Daten anderer Wetterstationen freigibt, um daraus von einem Portalbetreiber eine Wetterkarte erstellen zu lassen, muss er für genau diesen Zweck seine Einwilligung abgeben. Für die (Weiter-)Verarbeitung seiner Daten durch das Forschungsinstitut liegt seine Einwilligung natürlich mitnichten vor, da dieser neue spezifische Verarbeitungszweck bei der Datenerhebung ja noch gar nicht absehbar und damit auch

[32] Pfitzmann und Köhntopp, »Anonymity, Unobservability, and Pseudonymity«.

[33] Barocas und Nissenbaum, »Big Data's End Run Around Procedural Privacy Protections«.

[34] Narayanan und Shmatikov, »Robust De-anonymization of Large Sparse Datasets«; Montjoye u. a., »Unique in the Crowd«.

[35] Sweeney, *Simple Demographics Often Identify People Uniquely*.

[36] Tene und Polonetsky, »Big data for all«, S. xxvii; Solove, »Privacy Self-Management and the Consent Dilemma«, S. 1880.

nicht »einwilligungsfähig« war. Analog verhält es sich bei den anderen skizzierten Datenquellen.

Wenn also eine fundamentale Eigenschaft von Big Data in der Kombination und Analyse existierender Datensätze besteht, müssten nach heutigem juristischen Verständnis alle Betroffenen, von denen sich Angaben in diesen Datensätzen finden, um ihre Einwilligung zur (Weiter-)Verarbeitung zu einem neu zu spezifizierenden Zweck ersucht werden. Alternativ müssten bei der »Ersterhebung« alle künftig denkbaren Verarbeitungszwecke durch die gegebene Einwilligung abgedeckt werden. Beide Ansätze sind mit heutigen Mitteln kaum realisierbar. Daher bedarf es in Bezug auf Einwilligung und Zweckbindung in Big Data-Kontexten sowohl technischer Lösungen, welche Nutzer dazu ermächtigen ihre Rechte wahrzunehmen, als auch einer Diskussion darüber, inwiefern die juristischen Anforderungen an Einwilligungen zu spezifischen Verarbeitungszwecken heute noch zielführend und vor allem praktikabel umzusetzen sind und welche Alternativen denkbar wären.

4. Lösungsansätze

Für die angestrebte technische Durchsetzung der Beachtung von Einwilligungen sowie von Zweck- bzw. Kontextbindung[37] ist es erforderlich technische Systeme derart zu gestalten, dass eine Kontrolle und Steuerung sowohl von Datenflüssen als auch von Zugriffen auf einzelne Daten ermöglicht wird. Zur technischen Ausgestaltung solcher Kontroll- und Steuerungsmechanismen existieren einige Konzepte aus der Datenbankforschung und verwandten Bereichen.

Diese Lösungskonzepte sind darauf ausgelegt, Einwilligung bzw. Zugriff bezüglich einzelner Daten zu regeln. Im Folgenden soll die Untersuchung einer Eignung dieser Ansätze zur Zugriffssteuerung und -kontrolle in Big Data-Kontexten geleistet werden, um zu ergründen, welche Komponenten für eine Erweiterung von Data Mining-Architekturen nutzbar wären. Neben den hier aufgeführten Konzepten wurden u. a. auch verschiedene Ansätze zum Thema *Distributed Usage Control*[38] betrachtet. Da sich diese aufgrund der Notwendigkeit umfangreicher technischer und organisatorischer Veränderungen bestehender Systeme und Prozesse als grundsätzlich ungeeignet für den angestrebten Nutzungsfall der Integration autonom betriebener heterogener Datenquellen herausstellten, werden sie im Rahmen dieses Textes nicht weiter aufgeführt.

---

[37] Bundesministerium für Bildung und Forschung, *Selbstbestimmt und sicher in der digitalen Welt*, S. 21.

[38] Pretschner, Hilty und Basin, »Verteilte Nutzungskontrolle«, »Distributed Usage Control«; Lovat und Pretschner, »Data-centric Multi-layer Usage Control Enforcement: A Social Network Example«.

4.1. Hippokratische Datenbanken

Ein Konzept, welches die Durchsetzung von Einwilligung und Zweckbindung durch technische Maßnahmen umsetzen soll, wurde 2002 als *Hippocratic Databases*[39] vorgestellt. Es orientiert sich am hippokratischen Eid, der seit Jahrhunderten von Medizinern abgelegt wird und auch einen Passus zu Schweigepflichten enthält, welche dazu verpflichten, über medizinische Aspekte eines Patienten Stillschweigen zu bewahren und damit ein Vertrauensverhältnis zwischen Arzt und Patient zu etablieren.

Hippokratische Datenbanken sollen zehn Prinzipien folgen, welche sich an der Datenschutzgesetzgebung und diesbezüglichen Richtlinien orientieren, die neben der Begrenzung der Datensammlung, -nutzung, -veröffentlichung und -haltung als wichtigste Prinzipien die Einwilligung des Individuums, welches durch die Daten repräsentiert wird, zu spezifischen Verarbeitungszwecken festschreiben und dies durch technische Komponenten innerhalb der Datenbank umsetzen. Damit soll sichergestellt werden, dass spezifische Einzelangaben nur zu spezifizierten Verarbeitungszwecken zugänglich bzw. abfragbar sind.

Um die skizzierten Zugangsbeschränkungen durchsetzen zu können, wird bei Hippokratischen Datenbanken die Architektur klassischer Datenbanken dergestalt erweitert, dass zusätzliche Schutzmechanismen zwischen die datenhaltenden Tabellen und diejenigen Personen oder Applikationen, welche Inhalte der Datenbank von außen anfragen, eingesetzt werden, um nichtautorisierte Zugriffe zu unterbinden. Die Basis dieser Mechanismen stellen Anfrage-Modifikationen, welche anhand zusätzlicher Datenbanktabellen entscheiden, wann ein Zugriff auf bestimmte Datenfelder erlaubt ist, aber auch wann er unterbunden werden muss.

Die angesprochenen zusätzlichen Tabellen können innerhalb des Systems einerseits die Datenschutz-Präferenzen der Individuen kodifizieren, deren Daten im System gespeichert sind und die hierin festlegen können, welche Einzelangaben zu bestimmten Verarbeitungszwecken durch spezifische Datenverwender zugänglich sein sollen. Andererseits werden aber auch die Datenschutz-Richtlinien der Institution, welche die Datenbank betreibt, in den Tabellen dergestalt abgebildet, dass mögliche Verwender (bspw. verschiedene Fachabteilungen) sowie die zugehörigen Verarbeitungszwecke festgeschrieben werden. Damit enthalten die resultierenden Tabellen Einträge, in denen sowohl alle zulässigen Verarbeitungszwecke als auch potentielle Datenverwender aufgeführt sind. Durch die Auswertung dieser Tabellen vor einem Zugriff auf die gespeicherten Einzelangaben können jegliche Zugriffe, welche keine valide Kombination aus Verarbeitungszweck und Datenverwender darstellen, unterbunden werden.

[39] Agrawal u. a., »Hippocratic Databases«.

Hippokratische Datenbanken können so die informationelle Selbstbestimmung stärken, da sie die Durchsetzung von Einwilligung und Zweckbindung technisch erzwingen, indem nur bestimmten Verwendern der Zugriff auf Daten zu vorher festgelegten Verarbeitungszwecken gewährt wird.

Vorteile	Nachteile
+ Technische Umsetzung von Zweckbindung und Einwilligung + bei korrekter Implementierung keinerlei Zugriffe von Unbefugten möglich	– In der Praxis recht selten genutzt (lediglich akademische Prototypen-Implementierungen[40]) – Bei föderierten Datenquellen müsste jede einzelne hippokratisch organisiert sein – Nachträgliche Umstrukturierungen erzeugen Aufwand

4.2. Sticky Policies

Die Grundidee der *Sticky Policies* besteht darin, anfallende Daten schon bei der Erhebung mit einer Richtlinie zu versehen, welche die Nutzung der Daten regelt. Dabei ist diese Richtlinie als sogenanntes Meta-Datum zu einem spezifischen Datensatz zu verstehen, welches Informationen darüber enthalten kann, zu welchem Zweck und unter welchen Umständen auf die Daten innerhalb des Datensatzes zugegriffen werden darf. Als Zweck sind verschiedene Kategorien wie Forschung, Abwicklung von Transaktionen oder Ähnliches denkbar. Des Weiteren kann vorgesehen werden, Einschränkungen bezüglich des Umfeldes des Datenzugriffs festzulegen. So ließe sich bestimmen, dass ein Zugriff nur innerhalb von Netzwerken mit spezifizierten Sicherheitsvorkehrungen oder auch nur innerhalb eines festgelegten Teilnetzes einer genannten Institution erfolgen darf. Auch andere Zugriffseinschränkungen wie Ablaufdaten oder vertrauenswürdige Institutionen, an welche eine Datenweitergabe erfolgen kann, könnten in den Policies festgeschrieben werden.

Um den in den Richtlinien festgelegten, bestimmungsgemäßen Zugriff auf die Daten auch durchsetzen zu können, bedarf es eines Mechanismus, welcher den Zugriff außerhalb der Spezifikationen unterbindet. Hierfür wird starke Verschlüsselung genutzt. Nach der Erhebung des Datensatzes wird parallel zur Erstellung

[40] Laura-Silva und Aref, *Realizing Privacy-Preserving Features in Hippocratic Databases*; Azemović, *Data Privacy in SQL Server based on Hippocratic Database Principles*.

der Richtlinie der gesamte Datensatz verschlüsselt und der zugehörige Schlüssel an eine sogenannte »Trusted Authority« (TA) übersandt.

In die zugehörige Richtlinie muss die Information aufgenommen werden, wo der Schlüssel zu finden ist.[41] Ab diesem Zeitpunkt ist der Datensatz ohne den zugehörigen Schlüssel sicher vor unberechtigten Zugriff und ohne diesen nicht mehr sinnvoll nutzbar. Er kann nun bedenkenlos transferiert werden. Eine Institution, welche die Daten nutzen möchte, muss mittels der Informationen aus der angehangenen Richtlinie bei einer der angegebenen TAs den zugehörigen Schlüssel zur Freigabe der Informationen anfordern. Sie muss sich verpflichten, mit den Daten ausschließlich gemäß der Spezifikation der entsprechenden Richtlinie zu verfahren. An dieser Stelle ist auch eine weitere Überprüfung der gegebenen Rahmenbedingungen des Zugriffs (wie die oben skizzierten infrastrukturellen Einschränkungen) angedacht, welche sich beispielsweise durch technische Systeme (»remote software verification«) oder Audit-Verfahren realisieren lassen. Sind alle vorgegebenen Bedingungen erfüllt, gibt die TA den zum Datensatz passenden Schlüssel frei und die anfragende Institution kann die gewünschten Daten weiterverarbeiten.[42]

Vorteile	**Nachteile**
+ direkt an Datensätzen zu findende Privatspären-Präferenzen	– Verfahren abhängig von dritter Instanz (TA)
+ kein Zugriff ohne Zustimmung der TA möglich	– Neben Richtlinienerstellung weitere Vorarbeiten (Verschlüsselung, Transfer der Schlüssel zu den TAs, etc.) der datenerhebenden Institution nötig
+ alle Zugriffen werden (durch die Anfragen bei den TAs) protokolliert	

4.3. Dynamic Consent

Um auf geänderte Umstände adäquat reagieren zu können, sollte es für Individuen, welche diese Umstände betreffen und deren Einwilligung für die Verarbeitung der

[41] Die »Trusted Authority« (TA) muss hierbei keine zentrale Institution sein, welche alle zu einer Datenbank gehörenden Schlüssel verwaltet. Aus einer Perspektive der Informations- und Systemsicherheit heraus, und bei Betrachtung des Missbrauchspotenzials einer solchen zentralen »Schlüsselverwaltung«, wäre es durchaus wünschenswert, wenn eine Zentralisierung vermieden würde. Es könnten so prinzipiell ebenso viele TAs wie verschlüsselte Datensätze existieren. Wichtig ist lediglich, dass die zu einem verschlüsselten Datensatz gehörende Richtlinie genau spezifiziert, wo und wie der entsprechende Schlüssel zur Nutzung zugänglich ist.

[42] Mont, Pearson und Bramhall, »Towards accountable management of identity and privacy«; Pearson und Mont, »Sticky policies«.

sie betreffenden Daten bereits vorliegt, die Möglichkeit geben, die Einwilligung den Umständen entsprechend anzupassen. Die kürzlich verabschiedete, künftig EU-weit geltende Datenschutzgrundverordnung[43] verlangt beispielsweise in Artikel 7(3), dass eine einmal gegebene Einwilligung jederzeit widerrufen werden können muss, wobei der Widerruf ebenso einfach gestaltet sein sollte wie die Abgabe der Einwilligung.

Ein Ansatz zur nachträglichen Anpassung von Einwilligungen zu sich verändernden Datenverarbeitungszwecken, welcher die oben genannten Anforderungen erfüllt, ist das aus der Forschung zu biomedizinischen Forschungsdatenbanken stammende Konzept der *dynamischen Zustimmung*.[44] Hiermit wird das Problem adressiert, dass vorhandene medizinische Daten zwar für verschiedene Forschungsprojekte nutzbar sein sollten, gleichzeitig aber eine »allgemeine Zustimmung« möglichst vermieden werden soll, da diese nur bedingt der gesetzlichen Anforderung der Einwilligung zu einem spezifischen Verarbeitungszweck entspricht.

Um einem Individuum, das seine Daten für die medizinische Forschung freigeben möchte oder schon freigegeben hat, eine Möglichkeit zu geben, diese Einwilligung dynamisch anzupassen, wird ihm eine personalisierte Schnittstelle zur Kommunikation mit Forschenden zur Verfügung gestellt. Diese bietet einerseits den Forschenden die Möglichkeit nach einer Erweiterung/Änderung der vormals für ein anderes Forschungsprojekt gegebenen Einwilligung zu ersuchen, andererseits ermächtigt die Schnittstelle aber auch das Individuum dazu, seine gegebene Einwilligung jederzeit zu widerrufen oder entsprechend seiner momentanen Wünsche und Vorstellungen anzupassen. Der Kommunikationskanal ist dabei frei wählbar und reicht von Papierbriefen bis zu zahlreichen elektronischen Kommunikationsmedien wie SMS, E-Mail oder auch Soziale Netzwerke.

Vorteile	**Nachteile**
+ Ermöglicht dynamische Anpassung von Nutzerpräferenzen (auch bezüglich neuer unvorhergesehener Verarbeitungszwecke)	– hoher Implementierungsaufwand – Kommunikation erzeugt Zeitaufwand, daher nur bedingt echtzeittauglich
+ Verminderung der Notwendigkeit von »broad consent«	

[43] »Verordnung (EU) 2016/679 (Datenschutz-Grundverordnung)«.
[44] Kaye u. a., »Dynamic consent«.

5. Konzept: Einwilligungsmanagement für föderierte Datenquellen

Wie die vorangegangene Untersuchung bestehender technischer Konzepte zur Durchsetzung rechtskonformer Datenverarbeitung basierend auf Einwilligungen und Zweckbindung gezeigt hat, existieren Ansätze, welche für einzelne Datensätze oder auch Datenbanken die Einhaltung der Zweckbindung mit vorhandener Einwilligung auf technologischem Weg erzwingen können, wobei bezüglich der anvisierten Nutzung für die Integration autonom betriebener, verteilter Datenquellen einige nachteilige Eigenschaften offenbar wurden. Im Folgenden wird auf Grundlage dieser Erkenntnisse das Konzept einer Plattform für das Management von Einwilligungen zu verschiedenen Datenverarbeitungszwecken skizziert, welche die Vorteile der untersuchten Ansätze nutzt, um eine datenschutzkonforme Integration föderierter Datenquellen zu ermöglichen und die gefundenen Nachteile durch die zugrundeliegende Architektur umgeht.

5.1. Vorbedingungen

Geht man von den im Szenario angedeuteten Vorbedingungen der dynamischen Integration unabhängiger heterogener Datenquellen für Big Data-Analysen aus, zeichnet sich ab, dass Privatsphären-Präferenzen, insbesondere die Einwilligung zur Datenverarbeitung zu verschiedenen Zwecken, ähnlich wie im Konzept der *Sticky Policies*, an oder bei den entsprechen Datensätzen zu finden sein sollten. Mithilfe dieser Präferenzen hat ein Individuum, dessen Daten erhoben werden, die Möglichkeit, schon im Voraus genau zu spezifizieren, zu welchen verschiedenen Zwecken und für welche potenziellen verarbeitenden Institutionen die entsprechenden Daten freizugeben sind.

Um die Entscheidung darüber zu vereinfachen, sollten potenzielle Verarbeitungszwecke sowie mögliche verarbeitende Institutionen kategorisiert und zur Vorauswahl angeboten werden. Dies würde es erlauben, Einwilligungen der Art »Meine Daten dürfen von unabhängigen Forschungsinstituten für den Zweck demografischer Erhebungen genutzt werden, von staatlichen Institutionen zum Zweck der Steuerschätzung jedoch nicht« abzugeben. Um solche Präferenzen abbilden zu können, müssen diese in ein wohl-definiertes Format transferiert werden, welches eine präzise Spezifikation von Kategorien für Zwecke und »Verarbeiter« erlaubt und dabei ermöglicht, feingranular beliebig viele Subkategorien ebendieser anzulegen.

Solcherlei kodifizierte Präferenzen sollten zu jedem zu integrierenden Datensatz vorliegen.[45]

Des Weiteren ist die angestrebte Plattform davon abhängig, wissen zu müssen, wo spezifische Datensätze zu finden sind. Interessierte Institutionen, welche ihre Datenbanken für die Analyse durch Dritte freigeben wollen, sollten eine Beschreibung oder Spezifikation ihrer Datenbank erzeugen, welche Details zu den enthaltenen Datensätzen und die spezifische interne Struktur der Datenbank enthält. Zu jeder potentiellen Datenquelle sollte also ein maschinenlesbares Dokument vorliegen, welches spezifiziert, wo welche Arten von Daten zu finden sind und wie auf diese zugegriffen werden kann.

5.2. Architektur

Da das angestrebte System zum Einwilligungsmanagement so flexibel und universell einsetzbar wie möglich sein sollte, wird es als Plattform im weiteren Sinne bzw. als sogenannte Middleware entworfen, welche es ermöglicht, sowohl als gehosteter Webservice in der Cloud betrieben zu werden, aber auch als eigenständige Software-Komponente bestehende Datenanalysewerkzeuge zu erweitern. Abbildung 1 zeigt die generelle Architektur eines Data Mining-Systems, welches eine solche Plattform zum Management von Einwilligung nutzt. Die Plattform ist als Verbindung zwischen Data Mining-Anwendungen und den von diesen zur Analyse genutzten Datenquellen konzeptioniert und bietet standardisierte Schnittstellen in beide Richtungen.

Entscheidet sich eine Institution ihre Datenbank freizugeben, um diese von einer externen Organisation analysieren und nutzen zu lassen, wird die Datenbank über die standardisierten Schnittstellen mit der skizzierten Plattform verbunden. Während dieses »Kopplungsvorgangs« werden die Beschreibungen der Datensätze zur internen Struktur der Datenbank sowie die korrespondierenden Privatsphären-Präferenzen, welche potentielle Datennutzer und mögliche Verarbeitungszwecke spezifizieren,[46] zur Plattform transferiert. Die Plattform integriert die übertragenen Informationen über Datenbank und -sätze in ein föderiertes Schema oder einen Ontologie-basierten Wissensgraphen, welcher das Wissen über vorhandene Datensätze sowie deren technische Zugangsmöglichkeiten enthält und leicht verarbeitbar ist, um externen Anfragern dieses interne Wissen zur Verfügung zu stellen.

---

[45] Zu evaluieren wäre, ob hierzu einer der vorhandenen Standards für »Privacy Preference Languages« (wie P3P, XPref, APPEL o. ä., evtl. mit einer Erweiterung des Sprachumfangs) sinnvoll nutzbar ist oder eine Neuentwicklung erfolgen muss.

[46] Die genannten Spezifikationen sind natürlich bei der Datenerhebung vom Individuum, welches durch die entsprechenden Daten repräsentiert wird, abzufragen und festzulegen.

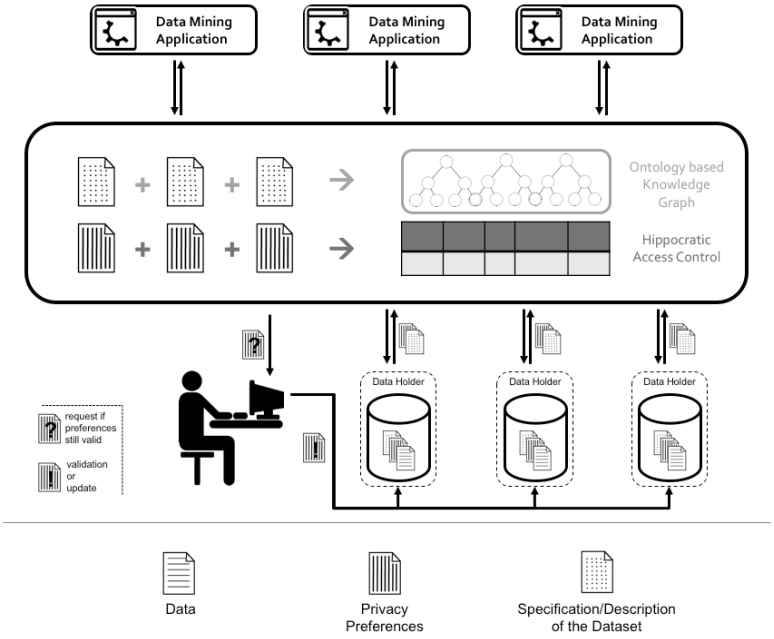

Abbildung 1.: Generelle Architektur einer Plattform zum Einwilligungsmanagement (aus Ulbricht und Pallas, »CoMaFeDS«, S. 110; © IEEE)

Die Privatsphären-Präferenzen werden genutzt, um ein »Hippokratisches Integrationsmodell« zu erzeugen. Dies kann mittels dedizierter Datenbanktabellen oder anderen Speicherstrukturen geschehen, welche geeignet sind, Informationen darüber vorzuhalten, welche potentiellen Datennutzer zu spezifizierten, wohl-definierten Verarbeitungszwecken auf ein bestimmtes Datum zugreifen dürfen. Ähnlich wie Hippokratische Datenbanken führt dieser Entwurf zu einem System, welches alle Datenzugriffe ohne eine valide Kombination von Datennutzer und Verarbeitungszweck unterbindet.

Eine Organisation oder Institution, welche Datensätze für eine Big Data-Analyse sucht, hat nun die Möglichkeit die Plattform über die Schnittstellen zu kontaktieren und den obig beschriebenen Wissensgraphen zu befragen, ob für ihren Analysezweck brauchbare Daten von einer der verbundenen Datenquellen durch die Plattform zur Verfügung gestellt werden. Sind die gewünschten Daten verfügbar, muss der Plattform sowohl die Identität der Organisation als auch der anvisierte Verarbeitungszweck übermittelt werden. Entspricht diese Kombination den Voraussetzungen

der korrespondierenden Regeln des »Hippokratischen Integrationsmodells«, wird der Zugriff gestattet und die Analyse kann durchgeführt werden.

Die Trennung von Wissensgraph und Hippokratischem Integrationsmodell würde es darüber hinaus erlauben, den durch die Daten repräsentierten Individuen eine Form der *dynamischen Zustimmung* als Erweiterung (bspw. als Webservice) anbieten zu können. Findet eine Institution für ihre geplante Analyse potenziell nutzbare Daten durch Abfrage des Wissensgraphen und eine Einwilligung zum angestrebten Verarbeitungszweck liegt nicht vor, sollte die Plattform in der Lage sein, nach einer geänderten oder neu zu erteilenden Einwilligung zu fragen. Das Individuum hat dann die Möglichkeit, seine ursprünglich abgegebenen Präferenzen direkt bei der ursprünglichen Datenquelle zu ändern. Die Plattform sucht bei dieser in periodischen Intervallen nach einer Änderung und wenn diese vorhanden ist, erfolgt eine Anpassung des Hippokratischen Integrationsmodells.

5.3. Diskussion

Als potenzieller Nachteil der hier kurz skizzierten Plattform ist sicherlich anzuführen, dass ein boshafter potenzieller Datennutzer falsche Angaben bezüglich seiner Identität oder des angestrebten Verarbeitungszwecks präsentieren kann. Um dies zu verhindern, könnte ein vor der Nutzung der Plattform zu durchlaufender Akkreditierungsprozess etabliert werden, in welchem die gemachten Angaben überprüft werden. Verläuft dieser Prozess zufriedenstellend, kann ein elektronisches Zertifikat für die Nutzung der Plattform erstellt werden. Da solch ein Akkreditierungsprozess auf allen Seiten Aufwand erzeugt, ist er in dieser ersten Phase der Konzeptionierung nicht berücksichtigt worden.

6. Fazit

Wie gezeigt wurde, sind Einwilligung und Zweckbindung wichtige Säulen der informationellen Selbstbestimmung, welche für die datenschutzkonforme Verarbeitung personenbezogener Daten unbedingt zu beachten sind. Neuartige Technologien, deren Entwicklung ohne Berücksichtigung dieser Prinzipien verlaufen ist, tendieren dazu, die informationelle Selbstbestimmung in erheblichem Maße zu schwächen.

Big Data-Technologien und deren Anwendung besitzen nicht nur das Potenzial, zusätzlichen ökonomischen und gesellschaftlichen Mehrwert aus bereits vorhandenen Daten zu generieren; sie bergen gleichzeitig das Risiko, grundlegende Prinzipien der informationellen Selbstbestimmung bzw. des Datenschutzes auszuhebeln. Die Verknüpfung und Analyse bereits existierender Datenquellen sorgt dafür, dass Anonymisierung bedeutend schwerer zu gewährleisten ist und lässt es, sobald personenbezogene Daten involviert sind, welche auf Basis von Einwilligung und

Zweckbindung erhoben wurden, zumindest fragwürdig erscheinen, ob diese Analyse, die ja einen neuen Verarbeitungszweck darstellt, zu welchem nicht eingewilligt werden konnte, datenschutzrechtlich überhaupt zulässig ist.

Der vorliegende Text hat gezeigt, dass sich technische Mechanismen finden lassen, welche die Einhaltung der Prinzipien von Einwilligung und Zweckbindung auch bei der Verknüpfung bereits vorhandener Datenquellen erzwingen könnten. Dazu bedarf es aber auch einer juristischen Neubewertung dieser Prinzipien, die es einem Individuum erlaubt, im Voraus Einwilligungen zu generalisierten Kategorien von möglichen Verarbeitungszwecken und potenziellen verarbeitenden Institutionen abzugeben.

Literatur

Agrawal, Rakesh u. a. »Hippocratic Databases«. In: *Proceedings of the 28th International Conference on Very Large Data Bases*. VLDB '02. Hong Kong, China: VLDB Endowment, 2002, S. 143–154.

Azemović, Jasmin. *Data Privacy in SQL Server based on Hippocratic Database Principles*. Microsoft MVP Award Program Blog. 7. 30 2012. URL: http://blogs.msdn.com/b/mvpawardprogram/archive/2012/07/30/data-privacy-in-sql-server-based-on-hippocratic-database-principles.aspx (besucht am 04. 12. 2015).

Barocas, Solon und Helen Nissenbaum. »Big Data's End Run Around Procedural Privacy Protections«. In: *Communications of the ACM* 57.11 (2014), S. 31–33. DOI: 10.1145/2668897.

Bundesministerium für Bildung und Forschung. *Selbstbestimmt und sicher in der digitalen Welt 2015-2020 – Forschungsrahmenprogramm der Bundesregierung zur IT-Sicherheit*. Bonn und Berlin, Jan. 2015. URL: http://www.bmbf.de/pub/Forschungsrahmenprogramm%5C_IT%5C_Sicherheit.pdf (besucht am 14. 09. 2016).

Bundesregierung. *Digitale Agenda 2014 – 2017*. Berlin, Aug. 2014. URL: https://www.digitale-agenda.de/Webs/DA/DE/Home/home_node.html (besucht am 14. 09. 2016).

»Charta der Grundrechte der Europäischen Union«. In: *Amtsblatt der Europäischen Union* C 364 (18. Dez. 2000), S. 1–22. URL: http://www.europarl.europa.eu/charter/pdf/text_de.pdf.

Chen, Ming-Syan, Jiawei Han und P.S. Yu. »Data Mining: An Overview from a Database Perspective«. In: *IEEE Transactions on Knowledge and Data Engineering* 8.6 (1996), S. 866–883. DOI: 10.1109/69.553155.

Cooley, Thomas. *A Treatise on the Law of Torts or the Wrongs Which Arise Independent of Contract*. Chicago: Callaghan, 1879.

Doan, AnHai, Alon Halevy und Zachary G. Ives. *Principles of Data Integration*. Waltham, MA: Morgan Kaufmann, 2012.

Fayyad, Usama, Gregory Piatetsky-Shapiro und Padhraic Smyth. »From Data Mining to Knowledge Discovery in Databases«. In: *AI Magazine* 17.3 (1996), S. 37–54.

Gerhardt, Rudolf und Erich Steffen. *Kleiner Knigge des Presserechts: Wie weit Journalisten zu weit gehen dürfen – mit Extrateil »Recht im Bild«*. Berlin: Berliner Wissenschafts-Verlag, 2009.

Hasselbring, Wilhelm. »Information System Integration«. In: *Communications of the ACM* 43.6 (Juni 2000), S. 32–38. DOI: 10.1145/336460.336472.

ISO/IEC 29100:2011 – Information technology – Security techniques – Privacy framework. Geneva: International Organization for Standardization, 2011.

Kaye, Jane u. a. »Dynamic consent: a patient interface for twenty-first century research networks«. In: *European Journal of Human Genetics* 23.2 (Feb. 2015), S. 141–146. DOI: 10.1038/ejhg.2014.71.

Laura-Silva, Yasin und Walid Aref. *Realizing Privacy-Preserving Features in Hippocratic Databases*. Computer Science Technical Reports 06-022. Purdue University, Department of Computer Science, Dez. 2006. URL: http://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=2664&context=cstech (besucht am 16. 09. 2016).

Lessig, Lawrence. »Privacy as Property«. In: *Social Research* 69.1 (2002), S. 247–269.

Lovat, Enrico und Alexander Pretschner. »Data-centric Multi-layer Usage Control Enforcement: A Social Network Example«. In: *Proceedings of the 16th ACM Symposium on Access Control Models and Technologies*. SACMAT '11. New York, NY, USA: ACM, 2011, S. 151–152. DOI: 10.1145/1998441.1998467.

Marx, Gary T. »Murky conceptual waters: The public and the private«. In: *Ethics and Information Technology* 3.3 (Sep. 2001), S. 157–169.

Michelfelder, Diane P. »The moral value of informational privacy in cyberspace«. In: *Ethics and Information Technology* 3.2 (Juni 2001), S. 129–135.

Mill, John Stuar. *On Liberty / Über die Freiheit (1859)*. Hrsg. von Bernd Gräfrath. Übers. von Bruno Lemke. Stuttgart: Reclam, 2009.

Mont, Marco Casassa, Siani Pearson und Pete Bramhall. »Towards Accountable Management of Identity and Privacy: Sticky Policies and Enforceable Tracing Services«. In: *14th International Workshop on Database and Expert Systems Applications, 2003. Proceedings*. Sep. 2003, S. 377–382. DOI: 10.1109/DEXA.2003.1232051.

Montjoye, Yves-Alexandre de u. a. »Unique in the Crowd: The privacy bounds of human mobility«. In: *Scientific Reports* 3 (2013). DOI: 10.1038/srep01376.

Moor, James H. »Towards a Theory of Privacy in the Information Age«. In: *ACM SIGCAS Computers and Society* 27.3 (Sep. 1997), S. 27–32. DOI: 10.1145/270858.270866.

Narayanan, Arvind und Vitaly Shmatikov. »Robust De-anonymization of Large Sparse Datasets«. In: *Proceedings of the IEEE Symposium on Security and Privacy, Oakland, California, May 18-21, 2008*. Los Alamitos: IEEE Computer Society, 2008, S. 111–125. DOI: 10.1109/SP.2008.33.

OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data. Paris: OECD, 1980.

Park, Byung-Hoon und Hillol Kargupta. »Distributed Data Mining: Algorithms, Systems, and Applications«. In: *The Handbook of Data Mining*. Hrsg. von Nong Ye. Mahwah, N.J.: Lawrence Erlbaum, 2003, S. 341–358.

Pearson, Siani und Marco Casassa Mont. »Sticky Policies: An Approach for Managing Privacy across Multiple Parties«. In: *IEEE Computer* 44.9 (2011), S. 60–68.

Pfitzmann, Andreas und Marit Köhntopp. »Anonymity, Unobservability, and Pseudonymity – A Proposal for Terminology«. In: *Designing Privacy Enhancing Technologies*. Hrsg. von Hannes Federrath. Lecture Notes in Computer Science 2009. Berlin und Heidelberg: Springer, 2001, S. 1–9.

Pretschner, Alexander, Manuel Hilty und David Basin. »Distributed Usage Control«. In: *Communications of the ACM* 49.9 (Sep. 2006), S. 39–44. DOI: 10.1145/1151030.1151053.

– »Verteilte Nutzungskontrolle«. In: *Digma: Zeitschrift für Datenrecht und Informationssicherheit* 7.4 (2007), S. 146–149.

Sheth, Amit P. und James A. Larson. »Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases«. In: *ACM Computing Surveys* 22.3 (Sep. 1990), S. 183–236. DOI: 10.1145/96602.96604.

Solove, Daniel J. »Privacy Self-Management and the Consent Dilemma«. In: *Harvard Law Review* 126 (2012), S. 1880–1903.

Swan, Melanie. »Sensor Mania! The Internet of Things, Wearable Computing, Objective Metrics, and the Quantified Self 2.0«. In: *Journal of Sensor and Actuator Networks* 1.3 (Nov. 2012), S. 217–253. DOI: 10.3390/jsan1030217.

– »The Quantified Self: Fundamental Disruption in Big Data Science and Biological Discovery«. In: *Big Data* 1.2 (Juni 2013), S. 85–99. DOI: 10.1089/big.2012.0002.

Sweeney, Latanya. *Simple Demographics Often Identify People Uniquely*. Data Privacy Working Paper 3. Pittsburgh: Carnegie Mellon University, 2000, S. 1–34. URL: http://dataprivacylab.org/projects/identifiability/paper1.pdf (besucht am 14. 09. 2016).

Tavani, Herman T. »KDD, data mining, and the challenge for normative privacy«. In: *Ethics and Information Technology* 1.4 (Dez. 1999), S. 265–273. DOI: 10.1023/A:1010051717305.

Tene, Omer und Jules Polonetsky. »Big Data for All: Privacy and User Control in the Age of Analytics«. In: *Northwestern Journal of Technology & Intellectual Property* 11.5 (2013), S. 239–273.

Thaler, Richard H. und Cass R. Sunstein. *Nudge: Improving Decisions about Health, Wealth, and Happiness*. New Haven: Yale University Press, 2008.

Ulbricht, Max-R. und Frank Pallas. »CoMaFeDS - Consent Management for Federated Data Sources«. In: *Proceedings of the 2016 IEEE International Conference on Cloud Engineering Workshops*. Los Alamitos: IEEE Computer Society, 2016, S. 106–111. DOI: 10.1109/IC2EW.2016.30.

»Verordnung (EU) 2016/679 des Europäischen Parlaments und des Rates vom 27. April 2016 zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten, zum freien Datenverkehr und zur Aufhebung der Richtlinie 95/46/EG (Datenschutz-Grundverordnung)«. In: *Amtsblatt der Europäischen Union* L 119 (4. Mai 2016), S. 1–88. URL: http://eur-lex.europa.eu/legal-content/DE/TXT/?uri=OJ:L:2016:119:TOC.

Warren, Samuel D. und Louis D. Brandeis. »The Right to Privacy«. In: *Harvard Law Review* 4.5 (1890), S. 193–220.

Westin, Alan F. *Privacy and freedom*. New York: Atheneum, 1967.

Wiederhold, G. »Mediators in the Architecture of Future Information Systems«. In: *IEEE Computer* 25.3 (März 1992), S. 38–49. DOI: 10.1109/2.121508.

Wu, Xindong u. a. »Data Mining with Big Data«. In: *IEEE Transactions on Knowledge and Data Engineering* 26.1 (Jan. 2014), S. 97–107. DOI: 10.1109/TKDE.2013.109.