Privatheit und selbstbestimmtes Leben
in der digitalen Welt

Data Sharing: Datenkapitalismus by Default?

Posterproceedings – Forum Privatheit 2023

Impressum

Data Sharing: Datenkapitalismus by Default?

Posterproceedings – Forum Privatheit 2023

Herausgeber:innen

Michael Friedewald, Murat Karaboga

Institut(e)

Fraunhofer-Institut für System- und Innovationsforschung ISI, Karlsruhe

Schriftenreihe

Forum Privatheit und selbstbestimmtes Leben in der digitalen Welt

ISSN-Print	2199-8906
ISSN-Internet	2199-8914
DOI	10.24406/publica-1887

Veröffentlicht

Oktober 2023, 1. Auflage
Fraunhofer-Institut für System- und Innovationsforschung ISI, Karlsruhe

Zitierempfehlung

Friedewald und Karaboga (Hrsg.) (2023): Data Sharing: Datenkapitalismus by Default? Posterproceedings – Forum Privatheit 2023. Forum Privatheit und selbstbestimmtes Leben in der digitalen Welt, Karlsruhe: Fraunhofer ISI.

Hinweise

Inhaltsverzeichnis

1 Technische Aspekte eines Privatsphäre-beachtenden Datenaustauschnetzwerks

Sascha Schiegg, Armin Gerl und Harald Kosch[1]

Künstliche Intelligenzen verdanken ihren Wissensschatz zumeist Lernverfahren auf der Basis gesammelter Daten. Diese Daten werden dabei von öffentlichen Quellen zusammengetragen, oft aber auch unter Datenhaltern geteilt. Personenbezogene Daten müssen auf Grund gesetzlicher Vorgaben vor Ihrer Weitergabe anonymisiert werden, falls keine explizite Einwilligung eingeholt wurde. Diese Anonymisierung vermindert dabei die Qualität eines verwendeten Datensatzes. Versuche zeigen, dass eine präzisere Anonymisierung von Vorteil für diese Qualität wäre (Schiegg & Gerl, 2022). Die folgende Arbeit stellt technische Möglichkeiten für ein Datenaustauschnetzwerk vor, das die Privatsphäre-Anforderungen der Nutzerinnen und Nutzer achtet, Unternehmen gleichzeitig das Teilen von Informationen mit bisher nicht bekannten Akteuren ermöglicht und die Datenqualität verbessert.

1.1 Motivation

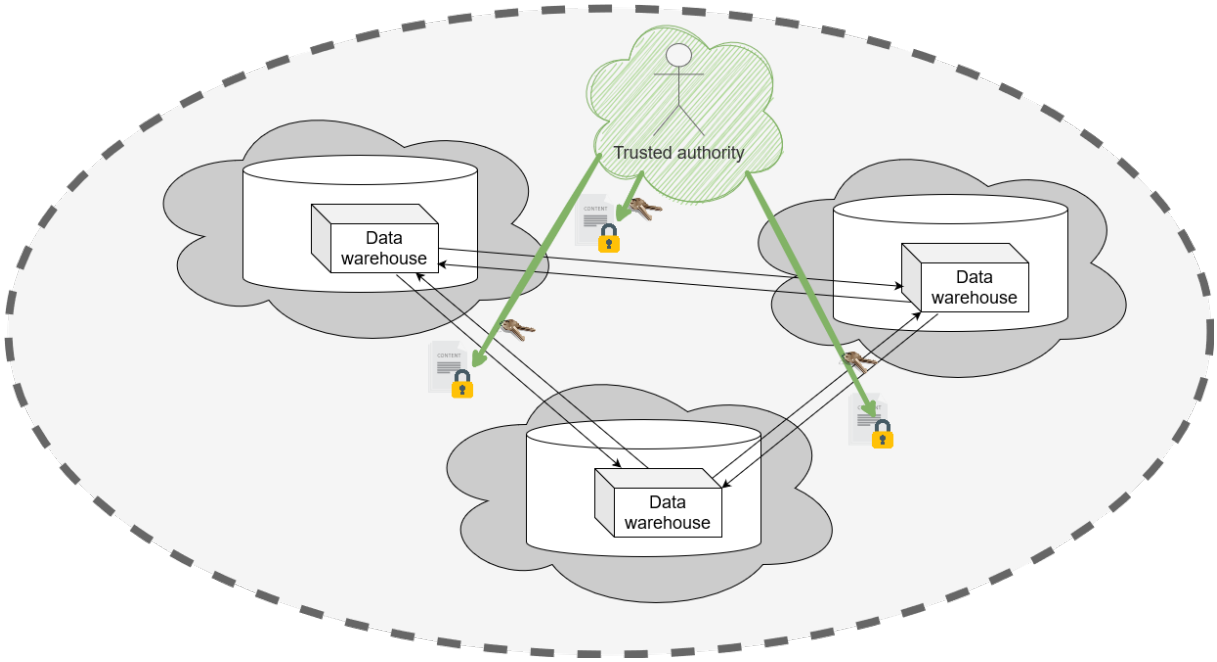
Künstliche Intelligenzen (KI) und deren abgeleitete Nutzungsmöglichkeiten sind das derzeit bestimmende Thema in einem interdisziplinären Umfeld aus Informatik, Ökonomie, Didaktik und vielen weiteren Domänen. Die Präzision von KI-Antworten basiert zumeist auf Daten, die einem Lernverfahren zur Verfügung gestellt wird. Ein wichtiger Faktor dabei ist der Qualitäts- und Quantitätsgrad der Trainings-Daten. Die Nachfrage nach Daten steigt daher eminent. Unternehmen sind somit bedacht, Daten zu akquirieren und mit Geschäftspartnern zu teilen, um den strategischen Vorteil von KI zu nutzen. Verfügen die Daten über personen-identifizierende Merkmale unterliegen sie den Bestimmungen der einschlägigen Datenschutzgesetze, wie der Datenschutzgrundverordnung (DSGVO). Letztere gestattet die Weitergabe nur, wenn die entsprechenden Personen ihr explizites Einverständnis erteilt haben oder als nicht mehr wiedererkennbar gelten. Letzteres wird u.a. durch Anonymisierung erreicht. Verfahren zur Anonymisierung wie Generalisierung limitieren den Qualitätsgrad der Daten bis zu einem gewissen Grad, der nicht vorgegeben wird. Da Unternehmen diesen somit selbst festlegen müssen, kann dieser zu hoch aber auch zu gering gewählt werden, was das Risiko einer Re-identifizierung entweder übermäßig erhöht oder unnötig geringhält. Um den Wünschen der betroffenen Personen gerecht zu werden, könnte diesen die Möglichkeit gegeben werden, selbst über den Grad Ihrer Anonymisierung zu bestimmen. Der Datensatz müsste dann hinreichend des größten Nutzerwunsches anonymisiert werden. Im vorgeschlagenen Datenaustauschnetzwerk werden eben jene individuellen Wünsche durch Konzepte wie *Sticky Policies* und *Personal Privacy* realisiert. Wir untersuchen, wie effizient technische Maßnahmen zur Vertrauensmitigierung umgesetzt und der Qualitätsgrad dabei aufrechterhalten werden kann.

1.2 Verwandte Arbeiten

Datenschutzerklärungen als nicht vereinheitlichte Form der Repräsentation einer Vereinbarung mit Nutzerinnen und Nutzern können nur schwer digital verarbeitet werden. Forschungsarbeiten, die diese Vereinbarungen in Maschinen-lesbaren Formaten repräsentieren, sind u.a. *PPL*, *XACML* und die *Layered Privacy Language* (LPL). Mit *LPL* führte Gerl die Möglichkeit für betroffene Personen ein, Vorgaben hinsichtlich der erlaubten Verwendungszwecke in Kombination mit Minimalanforderungen bzgl. des Anonymisierungsgrades vorzugeben (*Personal Privacy*) (Gerl, 2019).

[1] Universität Passau, Lehrstuhl für Verteilte Informationssysteme, E-Mail: {vorname.nachname}@uni-passau.de

Abbildung 1.1: Schematischer Aufbau des vorgestellten Datenaustauschnetzwerks



Betroffene Personen können ihre Wünsche zur Datenverarbeitung damit klarstellen, diese Rechte jedoch nicht überprüfen, sobald sie ihren Beeinflussungsraum verlassen, beispielsweise indem ein Unternehmen diese Informationen mit einem dritten Partner teilt. Dann kann technisch nicht mehr sichergestellt werden, ob sich dieser weitere Akteur ebenfalls an die Vereinbarung hält. Das von Karjoth et al. (Karjoth et al., 2002) vorgeschlagene und von Mont et al. (Mont et al., 2003) und Pearson und Casassa-Mont (Pearson & Casassa-Mont, 2011) implementierte *Sticky Policies* Konzept sieht hierfür die Verschlüsselung der Daten und das Anbringen der Privatsphäredefinition an jeder individuellen Aufzeichnung vor. Beim Teilen der Informationen wird eine Kopie der Definition sowie die Informationen zur Entschlüsselung an einen von allen Beteiligten vertrauten Akteur (*Trusted Authority* (TA)) übermittelt. Versucht der Empfänger an die Daten zu gelangen, muss er der TA die Konformität seiner Datenverarbeitung gemäß der Definition nachweisen. Hiernach werden ihm die Informationen zur Entschlüsselung mitgeteilt.

Datenbanken werden gemäß ihres Anforderungsprofils in transaktions- und analyseorientierte Anwendungsbereiche unterschieden. Analyseoptimierte Datenbanken werden auch als *Data warehouses* bezeichnet. Im Datenaustauschnetzwerk gehen wir davon aus, dass *Data warehouses* unter sich Daten austauschen, um den Datenbestand einer jeden teilnehmenden Datenbank zu erhöhen, und Anfragen von Analysten zu bearbeiten.

Praktische Anwendungsszenarien im Bereich der Verarbeitung von Datenbankanfragen mit *Sticky Policies* wurden u.a. von Trabelsi und Senor (Trabelsi & Sendor, 2012) präsentiert. Ulbricht und Pallas integrieren den Ansatz in verteilte Informationsstrukturen (Ulbricht & Pallas, 2016). Keine der untersuchten Arbeiten geht dabei auf die Möglichkeit ein, *Personal Privacy* zu berücksichtigen. Die Anwendung der Anonymisierungsmethoden kann an verschiedenen Stellen im Lebenszyklus des *Data Warehouse* erfolgen, wie von Fabian und Göthling (Fabian & Göthling, 2015) untersucht wurde.

1.3 Forschungsansatz

Um die Einhaltung der mittels *LPL* definierten Anonymisierungslevel und den Informationsfluss zwischen den beteiligten Partnern zu garantieren, wird auf *Sticky Policies* gesetzt. Da Datensätze verschlüsselt keine Informationen preisgeben, können sie beliebig zwischen Teilnehmern ausgetauscht werden. Erst, wenn eine Partei die Daten verarbeiten möchte, wird über die *TA* die entsprechende

Einhaltung der Regelungen klargestellt. Die damit einhergehende Anonymisierung folgt den Regeln der *Personal Privacy*. Verhält sich ein Teilnehmer untreu, kann ihm eine weitere Entschlüsselung durch die *TA* verweigert werden. Ein schematischer Aufbau des somit realisierten Datenaustauschnetzwerks findet sich in Abbildung 1.1.

Um die Regeln der *Personal Privacy* einzuhalten, müssen Akteure des Netzwerks das minimal zu erfüllenden Anonymisierungslevel des verwendeten Datensatzes ermitteln und entsprechend anonymisieren. Mit der Größe des Datensatzes steigt auch die Wahrscheinlichkeit, einen Eintrag mit besonders hohem Level zu beinhalten. Im angedachten Netzwerk soll die Ermittlung des benötigten Datensatzes Anfrage-basiert erfolgen, da Schiegg und Gerl auf die Effizienz dieses Ansatzes hindeuten (Schiegg & Gerl, 2022). Wir werden dies in zukünftigen Arbeiten mittels experimenteller Evaluation untersuchen.

Entschlüsselungen sind sehr intensive Operationen, die verhältnismäßig viel Rechenleistung und -zeit in Anspruch nehmen. Verarbeiter von Big Data Umgebungen stellt dies vor Herausforderungen, da beides Kosten- und Umweltfaktoren beeinflusst. Einen Ansatz zu finden, der die anfallenden Entschlüsselungen effizienter gestaltet, gilt daher als weiteres Ziel unserer Forschung. Da jeder Teilnehmer des Datenaustauschnetzwerks von den Aufwänden direkt selbst profitiert, ist zumindest bereits sichergestellt, dass diese fair verteilt sind.

1.4 Zusammenfassung und Ausblick

Mit dem vorgestellten Ansatz lässt dich das Teilen von personenbezogenen und damit besonders schützenswerten Daten unter Berücksichtigung der Interessen aller beteiligten Gruppen realisieren. Im Zuge künftiger Arbeiten wird die Kontrolle der Nutzerinnen und Nutzer über die für Sie zufriedenstellenden Verwendungszwecke und Anonymisierungslevel gesteigert werden. Datenhalter können die so erhaltenen Informationen gezielter anonymisieren und mit bisher unbekannten Akteuren teilen. Dies soll dazu führen, dass mehr Daten für Lernverfahren künstlicher Intelligenzen zur Verfügung stehen und diese gleichzeitig rechtssicher verarbeitet werden.

1.5 Literatur

Fabian, B., & Göthling, T. (2015). Privacy-preserving data warehousing. *International Journal of Business Intelligence and Data Mining*, 10(4), 297. https://doi.org/10.1504/IJBIDM.2015.072210

Gerl, A. (2019). Modelling of a Privacy Language and Efficient Policy-based De-identification [Universität Passau]. https://tel.archives-ouvertes.fr/tel-02900624/document

Karjoth, G., Schunter, M., & Waidner, M. (2002). Privacy-enabled services for enterprises. In *Proceedings 13th International Workshop on Database and Expert Systems Applications*, 483–487. https://doi.org/10.1109/DEXA.2002.1045944

Mont, M. C., Pearson, S., & Bramhall, P. (2003). 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* (S. 377–382). IEEE Comput. Soc. https://doi.org/10.1109/DEXA.2003.1232051

Pearson, S., & Casassa-Mont, M. (2011). Sticky Policies: An Approach for Managing Privacy across Multiple Parties. *Computer*, 44(9), 60–68. https://doi.org/10.1109/MC.2011.225

Schiegg, S., & Gerl, A. (2022). Trade-off between Privacy, Quality and Risk: Anonymization Strategy Evaluation for Data Warehouses. In *2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC)*, 1555–1560. https://doi.org/10.1109/COMPSAC54236.2022.00247

Trabelsi, S., & Sendor, J. (2012). Sticky policies for data control in the cloud. In *2012 Tenth Annual International Conference on Privacy, Security and Trust* (S. 75–80). IEEE. https://doi.org/10.1109/PST.2012.6297922

Ulbricht, M.-R., & Pallas, F. (2016). CoMaFeDS: Consent Management for Federated Data Sources. In *2016 IEEE International Conference on Cloud Engineering Workshop (IC2EW)* (S. 106–111). IEEE. https://doi.org/10.1109/IC2EW.2016.30

2 Project "PrivacyUmbrella": Ensuring data privacy by providing comprehensive anonymization processes

Lena Wiese[1], Despina Tawadros[1], Pronaya Prosun Das[1], Robert Zifrid[2], Isa Wasswa Musisi[2], Tim Bormann[2], Fihmi Mousa[2], Holger Storf[3], Axel Zieschank[3], Jens Göbel[3], Torsten Panholzer[4]

2.1 Background and Motivation

Thanks to the constant development of communication technologies, such as portable mobile devices and wearables, the concept of "smart health" - and with it the continuous improvement of personal health - has become an essential element of modern life for many people. Through the constant collection of data, smart health can promote a healthier lifestyle and be used to identify possible issues. For example, this technology might lead to a patient starting treatment early or at least in good time. However, as health data become more detailed and accessible to multiple parties, they also become vulnerable to attacks on individual patients' privacy. However, the challenge is not only to protect data privacy, but also to ensure that the shared data is informative enough to be useful for data analysis in terms of personalized medicine.

2.2 Project Objectives

PrivacyUmbrella aims to develop novel concepts for data analysis that complies with data protection requirements. The core objective is the development of a comprehensive anonymization system that combines diverse anonymization techniques, establishes standardized privacy metrics, and integrates these methods into an open-source demonstrator. Additionally, the project focuses on personalized medicine by utilizing anonymized wearable data for tailored and proactive health management. Therefore, the new anonymization procedures preserve the usability of medical data for analysis and empower patients to willingly share their data for medical research through mobile devices. Furthermore, we address the lack of standardization by establishing links with medical data integration efforts, encompassing the core data set[5] and other data formats such as Fast Healthcare Interoperability Resources (FHIR) endorsed by Health Level Seven International (HL7) with an open CC0 license. A simplified data integration, involving anonymization, aims to eliminate the need for pseudonym management and patient consent, thereby simplifying data sharing for research purposes. Project goals hence consist of:

- Establishing novel complex anonymization rules for standardized medical data formats (e.g., core data set/FHIR).
- Combining and optimizing a variety of anonymization techniques (k-anonymity, differential privacy, record linkage, homomorphic encryption, data separation) into an overarching system to prevent attacks and reduce overhead associated with anonymization.
- Defining the trade-off between resilience against data mining attacks and data analysis utility.

[1] Fraunhofer-Institut für Toxikologie und Experimentelle Medizin ITEM, Berlin. E-Mail: {vorname.nachname}@item.fraunhofer.de

[2] MCS Datalabs, Berlin. E-Mail:

[3] Institut für Medizinische Informatik, Goethe-Universität Frankfurt

[4] IMBEI-Medizinische Informatik, Universitätsmedizin Mainz

[5] https://www.medizininformatik-initiative.de/de/der-kerndatensatz-der-medizininformatik-initiative

- Employing Homomorphic encryption and other anonymization methods on genetic data to achieve cryptographically secure analysis.
- Developing an optimization method that allows data anonymization on mobile devices before being collected in a repository.

2.3 Methods

The challenge in the project is to combine several previously isolated techniques into a single integrated open-source demonstrator and to show that the anonymization properties (privacy metrics) of the individual methods can be maintained while deriving a holistic combined definition of privacy or anonymization. In particular, the project team is concerned with identifying and optimizing anonymization combinations, embedding these methods into existing open-source systems, and investigating their vulnerability through machine learning. The different components integrated in the system are described in the following paragraphs.

Decision system based on an optimization model for anonymization techniques

One aspect of privacy preservation is anonymizing shared data to protect individuals' privacy. Initially, removing direct identifiers such as personal ID, name and address is standard practice. However, the challenge arises when dealing with other attributes, known as quasi-identifiers, which, when combined, could lead to unique individual identification, such as gender and age. In this context, the goal is to generalize the quasi-identifiers, ensuring that individuals' information merges seamlessly with a larger group, making it harder to single out specific individuals. Nevertheless, it is essential to consider the trade-off between privacy and information loss. While extensive generalization enhances privacy by making individual identification a challenging task, it comes at the cost of reduced data utility. Excessive generalization can result in substantial loss of valuable information. For instance, in critical applications like disease diagnosis using machine learning models, overly coarse-grained data may fail to capture essential nuances and variations required for accurate predictions. This information loss due to high generalization granularity can hinder the effectiveness of data-driven applications, impacting decision-making processes and potentially compromising the quality of outcomes. Therefore, striking a balance between privacy and data utility when determining the degree of generalization for quasi-identifiers becomes crucial. However, the search problem for selecting a suitable generalization degree for the quasi-identifiers is inherently NP-hard. This complexity arises from constructing hierarchical generalization trees, resulting in a large search space comprising optimal combinations of generalized quasi-identifiers across different trees with various levels of generalizations.

To efficiently navigate this extensive solution space, we employ simulated annealing (Rutenbar, 1989) — a heuristic inspired by material annealing processes. This stochastic exploration is instrumental in preventing the algorithm from becoming trapped in local minima, allowing it to explore broader areas of the solution space and converge towards the global optimum. In our context, the cost function in the simulated annealing process has a dual purpose: it evaluates the average quasi-identifier generalization across different granularities, addressing privacy concerns, and quantifies the proportion of suppressed rows due to k-anonymity, indicating information loss and impact on data utility. The optimization problem is formulated as:

$$\text{Minimize } C(q) = \alpha \cdot G_{avg}(q) + (1-\alpha)L_{(q)} \text{ subject to } 0 \leq \alpha \leq 1$$

Where α is a tunable parameter to control the trade-off between privacy and utility and q the vector of selected generalization levels for the quasi-identifiers. The cost function $C(q)$ combines the privacy metric $G_{avg}(q)$ and the data utility metric $L_{(q)}$; the former is defined as $G_{avg}(q) = \frac{1}{Q} \cdot \sum_{i=1}^{N} G_i$, where G_i represents the generalization degree of the i^{th} quasi-identifier. It calculates the average

generalization degree across all quasi-identifiers, which reflects the extent to which individual records are made indistinguishable from one another. The latter is defined as $(q) = \frac{N-Suppressed}{N}$, where N total number of rows and $Suppressed$ is the count of suppressed rows. It measures the proportion of suppressed rows due to privacy-preserving techniques, such as k-anonymity, relative to the total number of records.

Data integration and record linkage

When personal data from various locations are to be merged, there is often a requirement to recognize a person who exists at several locations (record linkage). However, after an anonymization step, identifying the person and merging their data is no longer possible. We therefore check whether Bloom filters (or variants thereof; see Bloom (1970), Broder (2004), Heidt (2021)) can be created and stored before anonymization. A Bloom filter can return false positives, but not false negatives. Multiple identifiers such as last name, first name, date of birth can be included in one filter. Generation and storage of these filters can happen centrally, for example, in a cross-site trust center. But it can also take place decentrally at the locations because Bloom filters produce the same result if they are applied in the same way. After the anonymization step, we would still be able to recognize a person appearing more than once.

Nested medical data formats

The Fast Healthcare Interoperability Resources (FHIR)[6] provides a structured set of principles and norms for exchanging electronic healthcare data. This standard outlines data structures, referred to as "resources", and offers an API for transmitting electronic health records (EHR). It was developed by Health Level Seven International (HL7), a key player in healthcare standardization. In our setup, the HAPI-FHIR server[7], recognized for its integral compatibility with JSON and XML formats, is used in conjunction with the Open Source Registry System for Rare Diseases (OSSE) (Storf, 2017; Muscholl 2014) to anonymize nested medical data related to rare diseases.

Federated learning on distributed data sources

Federated learning combines decentralized data processing and differential privacy to train machine learning models collaboratively while protecting data privacy. Data stays on local devices, and model updates are periodically aggregated on a central server, ensuring individual data remains confidential. This approach is ideal for scenarios requiring data decentralization, confidentiality, and privacy. We deploy the Flower federated learning framework[8] for on-device execution, reducing bandwidth, power, and costs. Tensorflow Federated Learning[9] and the dp-accounting library[10] are used for large-scale pre-deployment experiments, calibrating the differential privacy process. On-device learning enhances personalized medicine by fine-tuning models for individual patients. This technology applies research concepts practically, initially focusing on vital parameter monitoring and non-invasive blood glucose estimation. Future possibilities include meal prediction from photos and fine-tuning Large Language Models (LLM) for medical letter analysis.

[6] https://www.hl7.org/fhir/

[7] https://hapifhir.io/

[8] https://flower.dev/

[9] https://www.tensorflow.org/federated

[10] https://pypi.org/project/dp-accounting/

2.4 Conclusion

Safeguarding data privacy is crucial for broad adoption of personalized medicine. Anonymization processes can protect privacy in such a way that individuals can no longer be identified. To account for the privacy concerns relating to personal medical data, the objective of our project is to develop highly specialized anonymization solutions based on modern data analysis. Striking a balance between data protection and analysis feasibility is complex, as unprotected processing compromises confidentiality, while a comprehensive protection can overly generalize data. Our approach ensures that privacy is upheld while maintaining data utility. The integration of wearable data addresses challenges related to privacy, analysis, and performance.

2.5 References

Bloom BH. Space/Time Trade-offs in Hash Coding with Allowable Errors. *Communications of the ACM*. 1970, 13(7), 422–426.

Broder A, Mitzenmacher M. Network applications of Bloom filters: A survey. *Internet Mathematics*. 2004, 1(4), 485-509.

Heidt CM, Hund H, Fegeler C. A Federated Record Linkage Algorithm for Secure Medical Data Sharing. *Stud Health Technol Inform*. 2021, 278,142-149.

Storf H, Schaaf J, Kadioglu D, Gobel J, Wagner TOF, Uckert F. [Registries for rare diseases: OSSE - An open-source framework for technical implementation]. *Bundesgesundheitsblatt* 2017/03/16 ed. 2017 May;60(5):523–31. doi: 10.1007/s00103-017-2536-7.

Muscholl M, Lablans M, Wagner TOF, Ückert F. OSSE – open source registry software solution. *Orphanet J Rare Dis.* 2014 9(1):O9. doi: 10.1186/1750-1172-9-S1-O9.

Rutenbar, Rob A. "Simulated annealing algorithms: An overview." *IEEE Circuits and Devices magazine* 5.1 (1989): 19-26.

3 Responsible Research and Innovation und die Utopie des gemeinnützigen Datenteilens

Daniela Fuchs und Margit Hofer[1]

3.1 Einleitung

Forschung und Entwicklung verantwortungsvoll zu betreiben ist eines der Prinzipien der Europäischen Forschungsförderung. Konzepte um Forschung und Entwicklung sozialverträglich zu gestalten existieren schon lange (Shanley, 2021). Konkrete Ansätze, wie z.B. Responsible Research and Innovation (RRI) machen es sich zur Aufgabe, Forschung und Entwicklung nach bestimmten Aspekten, nämlich antizipativ, responsiv und inklusiv zu gestalten (Owen et al., 2021, Owen et al., 2013, von Schomberg, 2013). Hierbei stellt die Beteiligung einer Vielzahl von Akteuren im Innovationsprozess ein zentrales Element dar (Bauer et al., 2021). Entsprechend entwickelte die Europäische Kommission für ihr Forschungsrahmenprogramm Horizon2020 konkrete Dimensionen um RRI besser in der Forschung zu etablieren (Ethik, Open Science, Wissenschaftsbildung, Inklusion, Gender-Gerechtigkeit). Im Forschungsrahmenprogramm Horizon Europe Rahmenprogramm sollen diese Dimensionen in Innovationsprozesse integriert und gestärkt werden.

Eine solche Annäherung passiert beispielsweise im Horizon Europe Projekt DATAMITE[2]. Das Projekt basiert auf der Annahme, dass europäische Unternehmen von technischen Lösungen zur Monetarisierung, Interoperabilität, zum Handel und Austausch von Daten profitieren können. Um dies zu unterstützen entwickelt DATAMITE eine modulare, mehrere Domänen abdeckende Open Source Rahmenumgebung, um entsprechende Lösungen in Form von Softwaremodulen, Schulungen und Geschäftsmaterialien anzubieten. Diese Rahmenumgebung ermöglicht Nutzer:innen, das Qualitätsmanagement ihrer Daten basierend auf den FAIR-Prinzipien zu verbessern, und sich zu technischen und geschäftlichen Aspekten weiterzubilden, welche die Zuverlässigkeit der Daten erhöhen. Zusätzlich generiert die Rahmenumgebung neue Einnahmequellen und Interaktionsmöglichkeiten mit anderen Akteursgruppen. Das Projekt DATAMITE plant eine Architektur, mit der digitale Innovationszentren „Sandkastentests" durchführen können, um KMUs besser auf ihren Eintritt in die Datenwirtschaft vorzubereiten.[3]

Damit verfolgt DATAMITE vorrangig technische und ökonomische Interessen. Gleichzeitig versucht es, auch Praktiken verantwortungsvoller Forschung und Entwicklung wie RRI einzubinden. Dazu gehört, nicht-monetäre Effekte des Datenteilens für Firmen wie KMUs, aber auch darüber hinaus für die Allgemeinheit, zu explorieren.

In diesem Zusammenhang stellen wir uns die Frage, welche solcher nicht-monetären Aspekte aus der Sicht von Stakeholdern (Pilotstudien) im Businesskontext als relevant gelten, welche Ideen von Gemeinnützigkeit in diesem Kontext propagiert werden, und wie sie in etablierten oder adaptierten Praktiken Berücksichtigung finden können. Konkret bietet DATAMITE also die Möglichkeit, die Reflexion kritischer Aspekte in Firmen zu erhöhen und in der Folge technische Entwicklungen, aber auch organisationale Praktiken entsprechend anzupassen.

---

[1] Zentrum für Soziale Innovation, Wien. E-Mail: {fuchs,hofer}@zsi.at

[2] Das Projekt Datamite erhält Unterstützung von der Europäischen Kommission im Rahmen von Horizon Europe unter der Fördervereinbarungsnummer 101092989.

[3] Vgl. https://cordis.europa.eu/project/id/101092989/de (Zugriff: 02.08.2023)

3.2 Methode

Im Rahmen des Projekts DATAMITE wird ein mehrstufiger Foresight-Prozess zur Visionsgenerierung entwickelt. Das übergeordnete Ziel ist, Visionen zu gesamtgesellschaftlichen Vorteilen aber auch Risiken von DataSharing gemeinsam mit diesen Praxis-Partnern zu generieren und damit einen Reflexionsprozess über DataSharing anzuregen.

Basierend auf einer Literaturanalyse zu gesellschaftlichen Impact-Dimensionen und möglichen Auswirkungen wurde im Juni 2023 ein partizipativer Stakeholder-Workshop durchgeführt. In einem „Brainwalk" wurden die erarbeiteten Dimensionen validiert, ergänzt und konkrete Auswirkungen zugeordnet. Zudem wurden Gemeinsamkeiten bzw. Unterschiede zwischen Sektoren identifiziert.

3.3 Ergebnisse

Die finale Auswahl der Dimensionen adressierte gesellschaftliche, umweltbezogene, sicherheitstechnische, wissenschaftliche, konsumenten-bezogene und makroökonomische Auswirkungen. Partner ergänzten außerdem eine Dimension von „Risiken" (nicht in Abbildung 3.1).

Gesellschaftliche Vorteile

Adressierte gesellschaftliche Auswirkungen von DataSharing bezogen sich einerseits auf die Open Source Community selbst. Andererseits wurde DataSharing als Möglichkeit gesehen, zu einer nachhaltigeren und faireren Gesellschaft beizutragen, entweder durch die Förderung von Innovation und die Entwicklung besserer bzw. neuer Services oder durch neue Ansätze der sozialen Gerechtigkeit (z.B. Verteilung des Gewinns entlang der Nahrungsmittelproduktionskette durch erhöhte Transparenz). In Bezug auf Energie wurde die Stärkung des Bewusstseins von Konsument:innen bezüglich ihres eigenen Energieverbrauchs sowie einer fairen Energieverteilung auf Gemeindeebene hervorgehoben. Damit solle auch politische Entscheidungsfindung erleichtert werden. Generell gingen Teilnehmer:innen davon aus, dass mehr Daten eine bessere politische oder administrative Entscheidungsfindung unterstützten, eine Behauptung, die aus mehrfachen Gründen kritisch zu betrachten ist (für ein Beispiel siehe z.B. Allhutter et al., 2020). So argumentierten die DATAMITE-Pilots, dass DataSharing für eine effizientere Unterstützung sozial Benachteiligter sorgen oder bei der Bekämpfung von Kriminalität unterstützen könne; sozial fragwürdige Praktiken, Missbrauchs- oder Fehlerszenarien wurden allerdings nicht berücksichtigt.

Ökologische Vorteile

Ökologische Auswirkungen werden vorrangig in Bezug auf die Kreislaufwirtschaft und effiziente Energienutzung bzw. grüne Energieerzeugung verstanden. Konkret wird daran gedacht, die Produktlogistik zu optimieren, um Prozesse unter Berücksichtigung ökologischer Gesichtspunkte zu verbessern und gleichzeitig effizient nachvollziehen zu können. Auf diese Weise strebt man an, eine ressourceneffiziente und nachhaltige Nutzung mithilfe von datenbasierten Ansätzen zu ermöglichen. Beispiele hierfür sind Digital Product Passports oder eine effiziente Umsetzung der Corporate Sustainability Reporting Directive (CSRD), die im November 2022 in Kraft trat. Weitere ökologische Vorteile beziehen sich auf eine energieeffiziente Automatisierung und eine intelligente Nutzung natürlicher Ressourcen. Auch Nachhaltigkeitspotenziale in Bezug auf DataSharing selbst wurde adressiert: Bessere Datenqualität benötigt weniger Ressourcen und daher weniger Energie in der Prozessierung (z.B. in der Entwicklung von KI-Modellen).

Des Weiteren wurde DataSharing auch als eine unterstützende Maßnahme für die Forschung zum Klimawandel erkannt. Die Verfügbarkeit von leicht zugänglichen Datensätzen könnte dazu beitragen, das Monitoring des Klimas zu vereinfachen und Auswirkungen des Klimawandels besser zu managen.

Sicherheitsvorteile

Hinsichtlich der Sicherheit gibt es verschiedene Auslegungen, die zu differenzieren sind. Zum einen liegt der Fokus auf Versorgungssicherheit, die vorrangig im Bereich der kritischen Infrastruktur eine Rolle spielt. Hier trägt DataSharing zu einem besseren Verständnis über die Auswirkungen von Energienutzung und -herstellung im Energiesystem bei.

Zum anderen wird Sicherheit im Kontext Cybersecurity verstanden. Einzelne Unternehmen erhofften sich von besserem Datenmanagement einen versierteren Umgang mit den eigenen Daten. Weiters wurde auf die notwendige Stärkung der Businesssicherheit gegen Industriespionage und Sabotage verwiesen. Beispielsweise können neue Datentools die Zahl der Sicherheitsmechanismen erhöhen, die von Firmen, Nutzern und Datenprovidern verwendet werden bzw. Kriminalfälle durch eine Rückverfolgbarkeit von Daten besser untersucht werden, beispielsweise durch sektorenübergreifende Datanalysen.

Wissenschaftliche Auswirkungen

Hier erschloss DataSharing vorrangig neuartige und erweiterte Möglichkeiten, Forschung zu betreiben, beispielsweise im Bereich interdisziplinärer Forschung. Besonders bedeutend war hierbei die erleichterte Verfügbarkeit von hochwertigen Daten (Open Data), einschließlich des Zugangs zu strukturierten, konsolidierten und validierten Daten aus vielfältigen Quellen.

Gleichzeitig hoben die Teilnehmer:innen neue Forschungsmöglichkeiten und Ansätze hervor, die auf intensivierten DataSharing beruhten (z.B. High-Performance Computing). Teilnehmer:innen betonten die Wichtigkeit von effizientem Datentransfer zwischen Forschungsfeldern und Sektoren, sowie die Automatisierung von Datenmanagement und -transformation. Hier spielen auch feste Regeln für ein faires Datenteilen (FAIR Prinzipien) eine Rolle, wobei man sich eine effizientere sektorenübergreifende Kollaboration zwischen Forschung, öffentlichem und privaten Sektor in Bezug auf Daten erhoffte. Dies wiederum könnte in weiterer Folge zur Entwicklung und Anwendung von Künstlicher Intelligenz sowie zur Schaffung neuer Dienstleistungen führen.

Vorteile für Konsument:innen

Hinsichtlich der Konsument:innen sollte DataSharing und Synergien zwischen Datensätzen helfen, neue Produkte und Services zu kreieren, zu adaptieren bzw. zu verbessern. So erleichtere DataSharing im Energiesektor die Entwicklung und Verbreitung neuer Verbrauchsmodelle (z.B. Prosumer), indem es ein besseres Verständnis der (eigenen jeweiligen) Möglichkeiten schaffe.

Ein weiterer bedeutender Vorteil, der herausgestellt wurde, besteht in der Steigerung der Transparenz von Produkten. So könne DataSharing entlang der Lieferkette die Nachvollziehbarkeit von Produkten und Produktionsprozessen verbessern, z.B. (aber nicht ausschließlich) in Nahrungsmittelproduktionsketten.

Makroökonomische Auswirkungen

Makroökonomische Vorteile erstrecken sich über mehrere Ebenen. Insbesondere im Energiesektor wurde die Wichtigkeit der mit DataSharing verbundenen Monitoring-Funktion hervorgehoben. So könne der nationale Energiehaushalt (Energieimport vs. -export), sowie der Energieverteilung zwischen Akteuren besser kontrolliert und gegebenenfalls korrigiert werden. Damit könne eine Reduktion des privaten Energieverbrauchs Raum für andere Akteure schaffen (z.B. Industrie), wobei vor allem (technische und administrative) Effizienz in der Energieproduktion und -distribution industrielle Aktivität und ökonomisches Wachstum sichern sollten.

In Bezug auf Governance wird erwartet, dass DataSharing und Open Data zu nachhaltigeren ökonomischen Entscheidungen auf politischer Ebene führen und die Kooperation zwischen privatem

und öffentlichem Sektor verbessern werde. Dies könnte beispielsweise durch die Förderung von Daten-Mehrfachnutzung auf verschiedenen Ebenen (Nutzer, kleine und mittlere Unternehmen, Verwaltung) geschehen. Dadurch erhofften Teilnehmer:innen auch, die Vorreiterrolle der EU in Bezug auf Datenökonomie zu stärken („making the European Approach to data concrete").

Auf Unternehmensebene erschließe ein weitreichenderes DataSharing neue Märkte wie z.B. für Nahrungsmittelhersteller und -verarbeiter und stärke den Wettbewerb und die Wettbewerbsfähigkeit zwischen EU-Unternehmen. Zudem könne DataSharing entlang der Lieferkette von Produkten dazu beitragen, gesellschaftliche Probleme zu adressieren (z.B. CO_2-Ausstoß in der Produktion, Menschenrechte, Regulierungen und Standards).

Risiken

Neben den aufgezeigten nicht monetären Dimensionen des Datenaustauschs wurden von den Teilnehmer:innen auch kritisch Risiken in Betracht gezogen. Diese Risiken konzentrierten sich vorrangig auf mögliche Gefahren für einzelne Unternehmen, die sich insbesondere aus dem Missbrauch von DataSharing und dem unbefugten Zugang zu Daten ergaben, z.B. Veröffentlichung von geschützten Daten, Cybersecurity Attacken, kriminelle Angriffe, unangemessener Umgang mit Daten etc.). Ebenso wurden legale und Datenschutzaspekte angesprochen. Auf Unternehmensebene standen auch Reputationsschäden durch die Verwendung falscher Daten und ein Verlust des Wettbewerbsvorteils im Mittelpunkt. Risiken mit weitreichenderen gesellschaftlichen Folgen wurden vor allem in der Diskriminierung von Konsument:innen durch gezielte Nutzung bestimmter Daten gesehen und in weiterer Folge eine mögliche Gefährdung von Menschenrechten.

Abbildung 3.1: Nicht-monetäre Impact Dimensionen und Auswirkungen von DataSharing aus Sicht der DATAMITE-Pilots

Quelle: DATAMITE, Fuchs & Hofer 2023

3.4 Analyse und Ansatzpunkte für weitere Diskussion

Unsere Analyse zeigt, dass ein grundsätzliches Bewusstsein über die gesellschaftlichen Auswirkungen von DataSharing auf unterschiedlichen Ebenen existiert, allerdings muss dieses weiter konkretisiert werden, um konkrete Empfehlungen für Praktiken des DataSharings formulieren zu können.

Dies verdeutlicht erneut die Herausforderung, über gesetzliche Vorgaben für den Umgang mit personenbezogenen Daten hinaus allgemeinere Schlussfolgerungen und Leitlinien für verantwortungsvolles DataSharing zu entwickeln. Im Gegensatz dazu sind die Auswirkungen auf bestimmte Branchen, allen voran in Bezug auf die wissenschaftliche Verwendung von Daten, bereits besser erkannt.

Konzepte wie Responsible Research and Innovation propagieren, Forschungs- und Entwicklungsprozess zu begleiten, wobei die eigentliche Erarbeitung reflexiver Praktiken Sektor spezifisch variiert. Ein frühzeitig angesetzter Foresight-Prozess zu nicht-monetären Effekten kann als eine Einbindung reflexiver Elemente interpretiert werden, die wiederum das antizipative Element verantwortungsvoller Forschung und Entwicklung stärken. Dennoch stellt sich dieser Beitrag die Frage, wie solche Aspekte in technisch-ökonomisch orientierten Projekten in entsprechenden Förderschienen besser gestärkt werden können, ohne „responsible washing" zu fördern. Dies ist besonders relevant als Konzepten wie RRI selbst für ihre inhärente Innovationsorientierung kritisiert werden (vgl. von Schomberg und Blok, 2021).

Um reflexive Praktiken zu stärken, lassen die Ergebnisse der DATAMITE Erhebung klar die Notwendigkeit von weitere qualitativer Datenerhebung erkennen, die sich vermehrt auf die unterschiedlichen Einflüsse in den jeweiligen spezifischen Branchen fokussiert.

3.5 Literatur

Allhutter, D., Cech, F.; Fischer, F.; Grill, G.; Mager, A. (2020): Algorithmic Profiling of Job Seekers in Austria: How Austerity Politics Are Made Effective. Frontiers in Big Data 3:5. https://doi.org/10.3389/fdata.2020.00005.

Bauer, A.; Bogner, A.; Fuchs, D. (2021): Rethinking societal engagement under the heading of Responsible Research and Innovation: (novel) requirements and challenges, *Journal of* Responsible Innovation, 8:3, 342-363. https://doi.org/10.1080/23299460.2021.1909812.

DATAMITE (2023): Datamite: Monetization, Interoperability, Tradinng & Exchange. EU funded project (2023). https://datamite-horizon.eu/

European Horizon Programme (2023): What is Horizon Europe? https://research-and-innovation.ec.europa.eu/funding/funding-opportunities/funding-programmes-and-open-calls/horizon-europe_en

Owen, R.; Stilgoe, J.; Macnaghten, P.; Gorman, M.; Fisher, E.; Guston, D. (2013): A Framework for Responsible Innovation. In: Owen, R.; Bessant, J.; Heintz, M. (eds.) *Responsible Innovation*. https://doi.org/10.1002/9781118551424.ch2.

Owen, R.; von Schomberg, R.; Macnaghten, P. (2021): An unfinished journey? Reflections on a decade of responsible research and innovation, *Journal of Responsible Innovation*, 8:2, 217-233. https://doi.org/10.1080/23299460.2021.1948789.

Shanley, D. (2021): Imagining the future through revisiting the past: the value of history in thinking about R(R)I's possible future(s), *Journal of Responsible Innovation*, 8:2, 234-253. https://doi.org/10.1080/23299460.2021.1882748.

von Schomberg, L.; Blok, V. (2021): Technology in the Age of Innovation: Responsible Innovation as a New Subdomain within the Philosophy of Technology. *Philosophy and Technology* 34, 309–323. https://doi.org/10.1007/s13347-019-00386-3 .

von Schomberg, R. (2013): A Vision of Responsible Research and Innovation. In *Responsible Innovation: Managing the Responsible Emergence of Science and Innovation in Society*, edited by R. Owen, J. Bessant, and M. Heintz, 51–74. London: Wiley.

4 Datenzugangsschutz als neues Vermögensrecht im Rahmen des EU Data Act

Johannes Kevekordes[1]

4.1 Ausgangslage

Die Verfügbarkeit von Daten bedeutet wirtschaftliche Macht. Nicht ohne Grund erkennt das deutsche Kartellrecht in §18 Abs. 3 Nr. 3 GWB den Zugang zu Daten als Kriterium für die Beurteilung einer marktbeherrschenden Position an und erlaubt §19a GWB bereits zuvor eine Kontrolle datenmächtiger Unternehmen, wenn sie eine „überragende marktübergreifende Bedeutung für den Wettbewerb" erlangt haben.

In Daten werden insbesondere Erfahrungen und Abläufe repräsentiert, aus denen wertvolle Rückschlüsse für die Entwicklung, Verbesserung und ideale Anpassung unterschiedlichster Produkte und Dienstleistungen gewonnen werden können. Die Entwicklung von Machine Learning-Models wie ChatGPT etwa gelingt nur durch die Konzentration immenser Mengen an Daten, durch die ein Machine Learning-Model Korrelationen lernen und eine „künstliche Intelligenz" ausbilden kann. Zudem ermöglichen riesige Datensammlungen Unternehmen, in jeden bestehenden Markt einzudringen und dort in hoher Geschwindigkeit marktmächtige bis -beherrschende Positionen einzunehmen.[2]

Gleichzeitig sind Daten derzeit in keiner Weise zwischen verschiedenen Marktakteuren gleich verteilt. Statt Daten auf einem prosperierenden Datenmarkt anzubieten, werden Daten von einzelnen großen Marktteilnehmern gleichsam in Datensilos faktisch exklusiv gehalten, zuvorderst von großen meist amerikanischen Technologieunternehmen, insbesondere von Google (heute Alphabet), Amazon, Apple, Facebook (heute Meta) und Microsoft. Dies ist insbesondere vor dem Hintergrund problematisch, dass Daten ein sog. non-rivales Wirtschaftsgut darstellen. Die gleichzeitige Nutzung eines solchen Guts führt zu keiner Einschränkung für die einzelnen Nutzer, sodass non-rivale Wirtschaftsgüter in der Theorie maximal wohlfahrtsteigernd eingesetzt werden, wenn sie von so vielen Akteuren wie möglich gleichzeitig genutzt werden.

4.2 Vorhaben der Europäischen Union

Die Europäische Union hat die mit der faktischen Exklusivität von Daten verbundene Problematik erkannt und will diese u.a. mit dem sog. Data Act lösen[3], der nach Abschluss der Triloge nur noch vom Rat und dem Europäischen Parlament bestätigt werden muss. Ziel des europäischen Gesetzgebers ist es insbesondere, durch die Regelung von Zugangsansprüchen im Data Act den Produktsekundärmarkt, also u. a. Reparatur- und Wartungsdienstleistungen für Drittanbieter, offenzuhalten.[4] Demgegenüber versuchen Produkthersteller, bspw. große Automobilhersteller, den

[1] Dr. Johannes Kevekordes ist Rechtsanwalt im IT & Datenschutzteam bei Taylor Wessing in Berlin. E-Mail: j.kevekordes@mail-box.org

[2] *Schweitzer/Haucap/Kerber u. a.*, Modernisierung der Missbrauchsaufsicht für marktmächtige Unternehmen, 2018, S. 59ff.

[3] Siehe daneben die bereits von der EU erlassenen Verordnungen des Data Governance Act, VO 2022/868, ABl. L 152, 3.6.2022, S. 1–44, des Digital Market Act, VO 2022/1925, ABl. L 265, 12.10.2022, S. 1–66 und des Digital Services Act, VO 2022/2065, ABl. L 277, 27.10.2022, S. 1–102.

[4] Vgl. ErwGr. 15 der letzten Fassung des Data Act bei *Europäisches Parlament*, Provisional Agreement resulting from Interinstitutional Negotiations, 14.07.2023, https://www.europarl.europa.eu/RegData/commissions/itre/inag/2023/07-14/ITRE_AG(2023)751822_EN.pdf (abgerufen am 13.09.2023).

Sekundärmarkt gegenüber nachgelagerten Wettbewerbern abzuschotten, indem sie den Zugang zu Daten des jeweiligen Produkts, die für die Wartung essentiell sind, blockieren.

Nach den Plänen der EU soll der sog. Data Holder deshalb gegenüber dem sog. User gem. Art. 4 Abs. 1 Data Act verpflichtet sein, diesem diejenigen Daten unverzüglich, kostenlos und gegebenenfalls kontinuierlich und in Echtzeit zugänglich machen, die der User selbst durch seine Nutzung eines Produkts oder einer Dienstleistung erzeugt und auf die er keinen direkten Zugriff hat. Erfasst werden sollen sowohl Privatnutzer als auch gewerbliche Nutzer.

Data Holder ist gem. Art. 2 Nr. 6 der letzten Fassung des Data Act eine juristische oder natürliche Person, die nach dem Data Act, nach anderem Unionsrecht oder nach nationalen Rechtsvorschriften, die Unionsrecht umsetzen, berechtigt oder verpflichtet ist, bestimmte Daten zu nutzen oder bereitzustellen, einschließlich, sofern vertraglich vereinbart, Produktdaten oder Daten über damit verbundene Dienstleistungen, die der Data Holder während der Erbringung einer damit verbundenen Dienstleistung erzeugt hat. Der in der früheren Fassung vorhandene Zusatz, dass im Falle nicht personenbezogener Daten Data Holder weiterhin die natürliche oder juristische Person sei, die durch die Kontrolle über die technische Konzeption des Produktes und damit verbundener Dienste in der Lage ist, bestimmte Daten bereitzustellen[5], ist weggefallen.

Neben dem direkten Zugangsanspruch kann der User vom Data Holder gem. Art. 5 Abs. 1 Data Act verlangen, dass dieser die jeweiligen Daten einem dritten Marktteilnehmer direkt zur Verfügung stellt. Dafür soll der Data Holder gem. Art. 9 Data Act eine monetäre Entschädigung von dritten Marktteilnehmer erhalten.

4.3 Analyse

Die Definition des Data Holder knüpft nicht mehr direkt an die faktische Kontrollmöglichkeit über Daten an. Dennoch muss diese zwingend mitgedacht werden. Damit eine Person in der Lage ist, Daten zu erfassen und zu nutzen, muss diese Person eine irgendwie geartete Kontrolle über diese Daten ausüben, wie schon der Begriff „Data Holder" selbst ausdrückt. Zudem käme es ansonsten zu einem teilweisen Zirkelschluss: Die Person des Data Holder kann nicht aus Zugangs- und Nutzungsrechten aus dem Data Act bestimmt werden, als diese ja gerade dem erst zu bestimmenden Data Holder zugewiesen werden. Insbesondere der letzte Zusatz der Definition des Data Holder macht deutlich, dass seine Person vielmehr im Wege vertraglicher oder gesetzlicher Nutzungs- und Zugangsrechte bestimmt werden soll. Aus einer solchen Berechtigung folgt aber zwangsläufig eine faktische Kontrollmöglichkeit. Die Person des Data Holder ist damit zweifelsohne eine Rechtsfigur, die es ohne Bezug zu einer faktischen Zugangsmöglichkeit nicht geben kann.

Diese Zwitterstellung des Data Holder als ein faktisch-rechtliches Konstrukt erinnert stark an den berechtigten Besitzer iSd bürgerlichen Rechts. Seine Person wird zunächst gem. §854 Abs. 1 BGB anhand der „tatsächlichen Gewalt" über die Sache bestimmt, der sog. Sachherrschaft. Hinzu tritt sein Recht zum Besitz, das ihm über §1007 Abs. 3 S. 3 BGB sämtliche Ansprüche aus dem Eigentümer-Besitzer-Verhältnis gegenüber schlechter berechtigten Besitzern zugesteht.

Auch wenn laut Erwägungsgrund 5 des Data Act ausdrücklich keine neuen Rechte an den Data Holder verliehen werden sollen, lassen sich bereits aus der Definition des Data Holder vermögensrechtliche Befugnisse ableiten, die zu einem subjektiven Recht führen. Data Holder ist, wer ein Recht zur Datennutzung innehat. Der Zugang und damit die Nutzung dieser Daten soll jenseits von Art. 4 Data Act nicht jedermann erlaubt sein. Zudem soll der Data Holder von dritten Marktteilnehmern eine monetäre Entschädigung für die Bereitstellung des Datenzugangs erhalten.

[5] Vgl. *Europäische Kommission*, Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates über harmonisierte Vorschriften für einen fairen Datenzugang und eine faire Datennutzung (Datengesetz) COM(2022) 68 final, 23.02.2022, S. 47.

Im Sinne der Definition subjektiver Rechte durch Jürgen Schmidt[6] als Aktions- und Vermögensberechtigungen werden der Person des Data Holder durch den Data Act somit insgeheim Vermögensberechtigungen zugewiesen: Außerhalb bestimmter Zugangsansprüche hat der Data Holder ein Recht, „seine" Daten zu nutzen sowie diese als Wirtschaftsgut verwerten dürfen.

Wie dem berechtigten Besitz kommt der Person des Data Holder ein Vermögenszuweisungsgehalt zu, als sich in ihr insbesondere die relative vertragliche Berechtigung dinglich-absolut, nämlich in Form der tatsächlichen Kontrollmöglichkeit, verkörpert.[7]

Das in Erwägungsgrund 5 des Data Act formulierte Vorhaben, kein neues Recht schaffen zu wollen, erinnert an den Geschäftsgeheimnisschutz. Auch bei diesem wird stets betont, dass er zu keinem neuen Ausschließlichkeitsrecht führe[8], obwohl gleichzeitig sowohl umfassende Abwehransprüche gem. Art. 7 GeschGehRL als auch bereicherungsrechtliche Ansprüche aus Eingriffskondiktion gem. Art. 13 GeschGehRL für den Geschäftsgeheimnisinhaber geschaffen wurden. Die Parallelen zwischen Data Holder und Geschäftsgeheimnisinhaber sind damit offensichtlich, denn für beide lassen sich eindeutige Vermögensberechtigungen ableiten.

Während der Data Act für den Data Holder dafür ein Recht zur Datennutzung verlangt, knüpft die Geschäftsgeheimnisrichtlinie in Art. 2 Nr. 3 an die rechtmäßige Kontrolle über die jeweilige geheime Information an. Aus dem jeweiligen Wortlaut ergibt sich sowohl für Data Holder als auch Geschäftsgeheimnisinhaber das Erfordernis einer Nutzungsberechtigung sowie einer Kontrolle über den jeweiligen Schutzgegenstand.

Für beide ist der berechtigte Besitzer aus dem Sachenrecht Vorbild: Eine faktische Position wird mit endgültigen rechtlichen Befugnissen aufgeladen und das Bestehen dieser Rechte zugleich an das Bestehen dieser faktischen Position geknüpft.[9] Auch die Figur des Data Holder knüpft an eine solche faktische Position an. Es würde dem Wunsch des EU-Gesetzgebers, keine neuen Rechte zu normieren, nämlich widersprechen, wenn der Data Holder sogar jenseits faktischer Kontrollmöglichkeiten nur aufgrund seines vertraglichen Nutzungsrechts eine vermögensrechtlich anerkannte absolute Position innehätte.

4.4 Vorschlag

Die aufgezeigte rechtlich anerkannte Position des Data Holder wirft die Frage auf, ob das vom Data Act implizit vorausgesetzte Datenvermögensrecht nicht explizit normiert werden sollte. Zu berücksichtigen ist dabei auch die Funktion parlamentarischer Gesetzgebung als transparenter Vermittler zwischen Bürger und Staat, der eine größere Legitimität eines Datenvermögensrechts und zugleich mehr Rechtssicherheit für Marktteilnehmer schaffen kann.[10] Zu diesem Zweck wird vom Verfasser ein Datenzugangsschutzregime vorgeschlagen, das ergänzend zum Geschäftsgeheimnisschutz die Datenebene von Informationen schützt.[11]

---

[6] Siehe *J. Schmidt*, Aktionsberechtigung und Vermögensberechtigung, 1969, S. 54.

[7] Siehe zum Vermögenszuweisungsgehalt von Besitz ausführlich *Kevekordes*, Daten als Gegenstand absoluter Zuordnung, 2022, S. 197ff.

[8] Siehe ErwGr. 16 Richtlinie (EU) 2016/943, ABl. L 157, 15.6.2016, S. 1–18 (GeschGehRL).

[9] Vgl. *Ohly*, GRUR 2014, 1(3), der von der Absicherung eines faktischen Zustands ähnlich wie beim berechtigten Besitz spricht; ebenso *Zech*, Information als Schutzgegenstand, 2012, S. 241.

[10] *Kevekordes*, Daten als Gegenstand absoluter Zuordnung, S. 69ff.; *Fezer*, Repräsentatives Dateneigentum, 2018, S. 81f.

[11] Siehe ausführlich bei *Kevekordes*, Daten als Gegenstand absoluter Zuordnung, S. 324ff.

Parallele Geschäftsgeheimnisschutz

Der Geschäftsgeheimnisschutz erfasst nur geheime Informationen. Dabei knüpft er an die sog. semantische Ebene von Informationen an, also ihren einzigartigen Bedeutungsinhalt, der dem Verständnis des Menschen unterliegt. Der Geschäftsgeheimnisschutz erstreckt sich also nur auf solche Information, dessen Bedeutung als solche geheim ist. Bei den typischen Big Data-Sachverhalten, insbesondere beim Training von Machine Learning-Models, ist die Geheimheit der Bedeutung aber nicht entscheidend. Entscheidend sind vielmehr die Korrelationen, die zwischen Millionen von bekannten Informationen gefunden werden können. Der Schutz der Daten selbst als Geschäftsgeheimnis ist daher umstritten.[12] Daten sind keine einzigartigen Bedeutungsgehalte, sondern eine Repräsentation dieser Gehalte durch binäre, maschinenlesbare Zeichen, durch die auf effiziente Art und Weise automatisierte Analysen und Operationen durchgeführt werden können. Daten betreffen somit die sog. syntaktische Informationsebene der Repräsentation von Bedeutungsgehalten.[13]

Datenbankherstellerrecht

Das einzige Recht, das in die Richtung eines Schutzes dieser Informationsebene weist, stellt das Recht des Datenbankherstellers aus Art. 7 der Datenbankrichtlinie der EU von 1996 dar.[14] Sowohl der Gesetzgeber als auch der Europäische Gerichtshof haben jedoch betont, dass das Recht des Datenbankherstellers nicht den Schutz der Daten selbst, sondern nur den Schutz der Investition in die Erstellung einer Datenbank umfasse[15] – ein Rechtssatz, der bei der heutigen Leichtigkeit der Erstellung und Anordnung von Datenbanken überholt erscheint, aber bis heute Bestand hat.[16]

Datenzugangsschutz

Grundlage des Datenzugangsschutzes ist der Datenbesitz, also eine irgendwie geartete Herrschaft über Daten bzw. den Zugang zu diesen. Bezugspunkt des Datenbesitzes sind die jeweiligen einzelnen Daten als Repräsentation von Bedeutungsgehalten. Selbst eine faktische Zuweisung des Datenbesitzes kann nur erfolgen, wenn eine bestimmte Kontrolle über Daten möglich ist. Daher bezieht sich der Datenbesitz und damit auch Datenzugangsschutz nur auf solche Daten, deren Zugang besonders gesichert ist, z. B. durch Verschlüsselung oder Passwörter. Erfasst werden insbesondere auch Datenpools, bei denen mehrere Unternehmen Daten untereinander teilen. Da der Datenzugangsschutz die Rechtsposition des Data Holder ausgestalten soll, kann er nur für den berechtigten Datenbesitzer bestehen. Nur diesem stehen Abwehransprüche gegen die unberechtigte (non-rivale) Nutzung von Daten zu. Die Berechtigung ergibt sich aus vertraglicher oder gesetzlicher Regelung, insbesondere ist derjenige berechtigt, der einen Zugangsanspruch auf die Daten hat.[17]

Im Einzelfall kann es sehr schwer sein, die unbefugte Nutzung von Daten durch einen Dritten nachzuweisen. Gerade der Nachweis, dass von einem Dritten genutzte Daten aus der Quelle des Data Holder stammen, ist sehr aufwändig. Selbst digitale Wasserzeichen können ohne weiteres

---

[12] *Krüger/Wiencke/Koch*, GRUR 2020, 578 (581).

[13] Zur Unterteilung des Informationsbegriffs in verschiedene Ebenen siehe mwN *Zech*, Information als Schutzgegenstand, S. 25ff.; *Kevekordes*, Daten als Gegenstand absoluter Zuordnung, S. 35ff.

[14] Richtlinie 96/9/EG, ABl. L 77, 27.3.1996, S. 20–28 (Datenbank-Richtlinie).

[15] Vgl. EuGH, GRUR 2005, 244 (247); EuGH, GRUR 2005, 252 (253); EuGH, GRUR 2005, 254 (256); siehe auch ErwGr. 9,10, 12 Datenbank-Richtlinie.

[16] Siehe insofern *Europäische Kommission*, Evaluation of Directive 96/9/EC on the legal protection of databases SWD(2018) 146 final, S. 35ff., 46f., die eine Reform des sui generis-Rechts für unverhältnismäßig aufwändig hält.

[17] Data Holder wird aber nur derjenige, der gleichzeitig auch tatsächlicher Datenbesitzer ist.

beseitigt werden. Insofern bietet es sich an, den Maßstab des Nachweises einer unbefugten Nutzung abzusenken. Vorbild könnte hier das Urheberrecht sein, das für den Nachweis der unbefugten Privatkopie in §53 Abs. 1 UrhG nur eine „offensichtliche Rechtswidrigkeit" der Informationsquelle verlangt. Der Begriff der Offensichtlichkeit sollte insofern zur höheren Rechtssicherheit objektiv ausgelegt werden.

Der Datenzugangsschutz böte in Anlehnung an den berechtigten Besitz auch die Möglichkeit von Zugangsbruchteilsgemeinschaften gem. §§743 ff. BGB, die die inneren Befugnisse von Datenpools regeln sowie vermögensmäßige Zuweisungen beinhalten könnten.

Ein expliziter Datenzugangsschutz könnte zudem die freiwillige marktmäßige Teilung von Daten verstärken. Die Schaffung von Immaterialgüterrechten wird klassischerweise neben der These, Anreize zur Erzeugung von Informationen zu geben, dadurch gerechtfertigt, zur verstärkten Teilung und effizienten Allokation von Information beizutragen.[18] Durch den größeren Rechtsschutz und größere Rechtssicherheit seien Marktteilnehmer eher bereit, Informationen mit anderen zu teilen. Die Schaffung eines besitzähnlichen Datenzugangsschutzes könnte somit die freiwillige Teilung von Daten auf Datenmärkten verstärken. Anspruch des Datenzugangsschutzes ebenso wie der Europäischen Union muss es gleichzeitig sein, durch vermögensrechtliche Regelungen nicht die bestehende faktische Exklusivität von Daten zu zementieren und die Gemeinfreiheit von Informationen unverhältnismäßig zu beschneiden. [19]

4.5 Fazit

Der Data Act der EU macht deutlich, dass diese vor der Regulierung der digitalen Wirtschaft nicht zurückschreckt und einer Regelung von allgemeinen Datenzugangsanspüchen offen gegenübersteht. Die geplanten Zugangsansprüche begründen gleichzeitig unweigerlich eine vermögensrechtliche Ebene von Daten. Gerade angesichts der Absicht der EU, den Data Act nur als eine Art Grundgerüst für weitere sektorspezifische Datengesetze konzipieren zu wollen, sollte diese vermögensrechtliche Ebene schon im Interesse einer transparenten politischen Auseinandersetzung durch einen Datenzugangsschutz explizit gemacht werden.

[18] Grundlegend *Kitch*, 20 (1977) J. Law Econ. 265 (276 ff.).

[19] Siehe für eine ausführliche rechtspolitische Diskussion über die Schaffung eines Datenzugangsschutzes *Kevekordes*, Daten als Gegenstand absoluter Zuordnung, S. 362 ff.

5 Data Sharing im Kontext digitaler Selbstvermessung[1]

Simone Salemi[2], Bianca Steffes[3] und Nils Wiedemann[4]

5.1 Einleitung

Die digitale Selbstvermessung durch Smart Watches und vergleichbare Wearables erfreut sich einer großen Beliebtheit.[5] Bei diesen Geräten steht meist die systematische Erfassung und Speicherung von Daten, die das eigene Leben und den (Gesundheits-)zustand betreffen, im Vordergrund. Solche Geräte werden auch bereits im Arbeitskontext eingesetzt.[6] Im Rahmen des „Workloggings" werden mithilfe von Wearables oder anderen IoT-Geräten einzelne Arbeitsvorgänge zur Effizienzsteigerung sowie zur Erhöhung von Arbeits- und Gesundheitsschutz vermessen.[7] Regelmäßig werden hierbei personenbezogene Daten der Arbeitnehmer, auch sensible personenbezogene Daten wie etwa Gesundheitsdaten, mithilfe von Wearables oder anderen IoT-Geräten verarbeitet. Ein Konflikt mit der DSGVO ist daher denkbar, weshalb sich die Frage stellt, wie der Schutz dieser Daten entsprechend der Vorgaben der DSGVO sichergestellt werden kann.[8] Darüber hinaus steht zu erwarten, dass der sich derzeit im Entwurfsstadium befindliche Data Act[9] weitreichende Auswirkungen für den Einsatz von Wearables im Arbeitskontext mit sich bringen könnte. Der Beitrag erörtert die datenschutzrechtlichen Probleme und zeigt technische Lösungen für eine praktische und datenschutzgerechte Umsetzung auf. Zudem werden die möglichen Auswirkungen des geplanten Data Acts untersucht.

5.2 Datenschutzrechtliche Probleme

Die Verarbeitung personenbezogener Daten im Arbeitskontext unterliegt aufgrund des Machtungleichgewichts und des Abhängigkeitsverhältnisses zwischen Arbeitgeber und Arbeitnehmer besonderen Anforderungen. Für den Beschäftigtendatenschutz existiert in der DSGVO eine Öffnungsklausel, die es den Mitgliedstaaten erlaubt, spezifische Regelungen hierzu zu erlassen (Art. 88 Abs. 1 DSGVO).[10] Der deutsche Gesetzgeber hat diese und weitere Öffnungsklauseln mit § 26 BDSG ausgefüllt.[11] § 26 BDSG bietet indes verschiedene Rechtsgrundlagen: So erlaubt § 26 Abs. 1 BDSG

[1] Diese Arbeit ist im Kontext des durch das BMBF geförderten Projekts WearPrivate (16KIS1512) entstanden.

[2] Saarbrücker Zentrum für Recht und Digitalisierung, E-Mail: simone.salemi@zrd-saar.de

[3] Universität des Saarlandes, E-Mail: bianca.steffes@uni-saarland.de

[4] Universität des Saarlandes, E-Mail: nils_torben.wiedemann@uni-saarland.de

[5] Im Jahr 2022 wurden in Deutschland knapp 7,2 Millionen Wearables abgesetzt, https://de.statista.com/statistik/daten/studie/551366/umfrage/absatz-von-wearables-in-deutschland/ (alle Links wurden zuletzt am 20.09.2023 abgerufen).

[6] Etwa in wissenschaftlichen Arbeiten (bspw. Mengru Xue, Rong-Hao Liang, Jun Hu, Bin Yu, and Loe Feijs. 2022. Understanding How Group Workers Reflect on Organizational Stress with a Shared, Anonymous Heart Rate Variability Data Visualization. In Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems (CHI EA '22). Association for Computing Machinery, New York, NY, USA, Article 27, 1–7. https://doi.org/10.1145/3491101.3503576) oder in Praxislösungen (bspw. https://wearhealth.com/).

[7] *Schröter, Welf*, Virtuelle Identitäten im „Worklogging" – Impulse zur sozialen Gestaltung der Arbeitswelt in der „Industrie 4.0" in: Selke, Stefan, Lifelogging: S. 193-214 (204).

[8] Siehe: *Leibenger Dominik/Möllers, Frederik/Petrlic, Anna/Petrlic, Ronald/Sorge, Christoph*, Privacy Challenges in the Quantified Self Movement – An EU Perspective, In Proceedings on Privacy Enhancing Technologies, no. 4, pp. 315–334, 2016.

[9] Vorschlag für eine Verordnung des Europäischn Parlaments und des Rates über harmonisierte Vorschriften für einen fairen Datenzugang und eine faire Datennutzung (Datengesetz), vom 23.2.2022, COM (2022) 68 final.

[10] *Seifert* in: Simitis/Hornung/Spiecker gen. Döhmann, Datenschutzrecht, 1. Auflage 2019, Art. 88 DSGVO Rn. 2.

[11] *Zöll* in: Taeger/Gabel DSGVO – BDSG – TTDSG, § 26 BDSG, Rn. 6, 8.

die Verarbeitung von Beschäftigtendaten für Zwecke des Beschäftigungsverhältnisses, nach § 26 Abs. 2 BDSG gibt es die Möglichkeit, eine Einwilligung der Beschäftigten einzuholen und § 26 Abs. 3 BDSG erlaubt unter bestimmten Umständen auch die Verarbeitung besonders sensibler Daten zum Zwecke des Beschäftigungsverhältnisses. Geht es um die Verarbeitung von Daten im Rahmen des Workloggings ist danach zu unterscheiden, ob direkt auf die Daten vom Wearable des Arbeitnehmers zugegriffen wird oder ob diese Daten nachträglich ausgewertet werden. Beim Zugriff auf die auf dem Wearable gespeicherten Daten ist die ePrivacy-Richtlinie der EU zu beachten und es ist im Einklang mit Erwägungsgrund 32 zum Data Act grundsätzlich eine Einwilligung des Arbeitnehmers erforderlich. Die weitere Auswertung der Daten richtet sich allerdings nach § 26 BDSG. Aufgrund eines aktuellen Urteils des EuGH[12] zur fast wortgleichen Vorschrift im hessischen Datenschutzgesetz (§ 23 Abs. 1 S. 1 HDSIG) wird derzeit allerdings die Europarechtskonformität des § 26 Abs. 1 S. 1 BDSG in Frage gestellt, weshalb es fraglich ist, ob § 26 Abs. 1 S. 1 BDSG zukünftig noch als Rechtsgrundlage dienen kann. Grundsätzlich besteht für Mitgliedstaaten zwar die Möglichkeit, zusätzliche, strengere oder einschränkende, nationale Vorschriften vorzusehen, und ihnen steht ein Ermessen hinsichtlich der Art und Weise der Durchführung dieser Bestimmungen zu.[13] Allerdings lassen beide nationalen Vorschriften laut dem EuGH geeignete und besondere Maßnahmen zur Wahrung der menschlichen Würde, der berechtigten Interessen und der Grundrechte der betroffenen Person missen, weshalb eine Vereinbarkeit mit Art. 88 Abs. 1 und 2 DSGVO zu bezweifeln ist.[14] Das Urteil legt es nahe, dass es künftig einer Überarbeitung des § 26 Abs. 1 S. 1 BDSG bedarf. Nicht eingeschränkt wird in diesem Zusammenhang die Möglichkeit, mit einer Einwilligung der Beschäftigten zu arbeiten oder die Verarbeitung auf § 26 Abs. 3 BDSG zu stützen. Die letztgenannte Vorschrift dürfte regelmäßig auch einschlägig sein, wenn im Rahmen des Workloggings besonders sensible Daten verarbeitet werden. Neben den rechtlichen Schwierigkeiten stellen sich allerdings auch technische Probleme, sobald es zum Einsatz von Wearables im Arbeitskontext kommt.

5.3 Technische Betrachtung

Zumeist wird in der Selbstvermessung ein Zusammenspiel aus einem Wearable, einem oder mehreren Smartphones und einem Server genutzt.[15] Das Wearable erhebt dabei üblicherweise die Daten und sendet sie an ein Smartphone, welches sie für den Nutzer zugänglich macht und die Daten zentral an einen Server schickt. Durch den Server wird einerseits die parallele Nutzung mehrerer Smartphones mit denselben Daten, andererseits auch eine Langzeitspeicherung der Daten ermöglicht, denn die erhobene Datenmenge kann bei wenig vorhandenem lokalen Speicherplatz ab einem gewissen Punkt zu groß sein, um auf dem Smartphone gespeichert zu werden. Diese Aufteilung dient jedoch nicht vornehmlich dem Ziel der Privatheit: Die Nutzung der Dienste ist in der Regel nur mit einem Account beim Dienstanbieter möglich[16], der weitere personenbezogene Daten benötigt. Zudem sind die meisten Dienste, welche das Erheben von Wearable-Daten ermöglichen, auf die Bereitstellung von verschiedenen Fitness- und Gesundheitsanalysen ausgelegt, die erst auf den Servern des Dienstes ausgeführt werden.[17] Dementsprechend müssen Nutzer in der Regel ihre

[12] EuGH, Urt. v. 30.3.2023 – C-34/21, NZA 2023, 487.

[13] EuGH, Urt. v. 30.3.2023 – C-34/21, NZA 2023, 487, (489 Rn. 51).

[14] EuGH, Urt. v. 30.3.2023 – C-34/21, NZA 2023, 487 (490 Rn. 64 f.; 491 Rn. 74).

[15] *Saifuzzaman, M., Ananna, T.N., Chowdhury, M.J.M. et al.* A systematic literature review on wearable health data publishing under differential privacy. *Int. J. Inf. Secur.* 21, 847–872 (2022). https://doi.org/10.1007/s10207-021-00576-1.

[16] Apple setzt die Verknüpfung mit einer Apple-ID voraus (https://support.apple.com/de-de/HT204351) und Fitbit die Nutzung eines Fitbit- oder Google-Kontos (https://www.fitbit.com/global/de/legal/terms-of-service).

[17] Beispiele dafür sind etwa Polar Flow (https://flow.polar.com/) oder Garmin Connect (https://www.garmin.com/de-DE/p/125677).

Wearable-Daten dem Diensteanbieter gegenüber offenlegen. Eine datenschutzfreundlichere Umsetzung ist dabei jedoch möglich: Einige Dienste erlauben bereits die Nutzung eines Wearables komplett ohne die Kopplung an Smartphone und Server, auch wenn dies zurzeit nur mit begrenzter Funktionalität möglich ist.[18] Andererseits ist es auch denkbar, bloß auf die Synchronisierung mit dem Server zu verzichten oder die Daten durch eine Verschlüsselung für den Server nicht lesbar zu machen.[19] Ebenfalls ist die Nutzung von Methoden der Anonymisierung, die dazu führen würden, dass Server nur anonymisierte Daten der Nutzer erhalten, die jedoch nur einen beschränkten Funktionsumfang ermöglichen könnten, oder auch ein kompletter Verzicht auf den Server denkbar, wenn etwaige Berechnungen lokal ausgeführt werden würden. Auch das Teilen der Daten ist stark vom Dienstanbieter abhängig: Oftmals wird das Teilen der Daten an Dritte ebenfalls über den Server ausgeführt[20], was vor allem bei großen Datenmengen für den Nutzer von Vorteil ist. Viele Anbieter ermöglichen auch bereits ein einfaches Teilen von Daten mit Dritten. Dabei handelt es sich jedoch stets um das Teilen mit anderen Personen[21] oder Diensten, die eine Ergänzung zum eigenen Dienst darstellen[22]. Ein Teilen von Daten für die Weiternutzung in einer anderen App mit gleichem Funktionsumfang scheint zumeist nicht vorgesehen zu sein. Auch das manuelle Exportieren der Daten mit anschließender Weiterleitung an einen neuen Dienst gestaltet sich als schwierig: Obwohl regelmäßig die Möglichkeit eines Datenexports angeboten wird[23], bieten nur sehr wenige Dienste auch eine Importfunktion an[24]. Dies ist wohl dem Problem geschuldet, dass selbst bei der Nutzung eines strukturierten, gängigen und maschinenlesbaren Formats (so bspw. das Recht auf Datenübertragbarkeit nach Art. 20 Abs. 1 DSGVO) keine einheitlichen Datenstrukturen genutzt werden müssen. Um nicht das Datenformat jedes Anwenders implementieren und stetig aktualisieren zu müssen, ist eine Vereinheitlichung dieser Datenformate für ein möglichst unkompliziertes Teilen von Daten zwischen Diensten notwendig.[25] Um einen hohen Schutz der Daten der Nutzer zu erhalten wäre es zudem sinnvoll, die Synchronisation der Daten entweder mit verschlüsselten Daten durch die Server oder vollkommen durch die Client Anwendungen durchzuführen: Möchte der Nutzer etwa von einer App auf seinem Smartphone auf eine andere App umsteigen, könnten die Daten lokal ausgetauscht werden und es würden keine hohen Übertragungskosten anfallen.

5.4 Möglicher Einfluss des Data Acts

In diesem Zusammenhang werden sich weitere Herausforderungen durch die Vorschriften des zukünftigen Data Acts[26] (DA) ergeben. Das Ziel des DA ist es, Datensilos aufzubrechen und den Zugang sowie das Teilen von Daten zu erleichtern.[27] Danach soll das bereits angesprochene Recht auf

[18] Ein Beispiel dafür bietet der Hersteller Polar, der einen "autonomen Modus" für seine Wearables anbietet. Weitere Informationen finden sich hier: https://www.polar.com/de/legal/terms-of-use#toc5.

[19] Apple Health etwa sichert die Daten in der iCloud. Diese Synchronisierung kann jedoch auch deaktiviert werden und die Daten durch eine Ende-zu-Ende-Verschlüsselung ab iOS 12 geschützt werden: https://support.apple.com/de-de/HT204351.

[20] *Saifuzzaman et al.* 2022.

[21] Apple erlaubt etwa in den USA das Teilen von Daten mit Ärzten: https://support.apple.com/en-us/HT212629.

[22] Beispielsweise zu sehen bei Fitbit (https://help.fitbit.com/articles/en_US/Help_article/1742.htm) oder Garmin (https://support.garmin.com/en-US/?faq=dOYUNh48g67aPgYTfBJPh7).

[23] Hier etwa für Apple (https://support.apple.com/de-de/guide/iphone/iph5ede58c3d/ios) oder Fitbit (https://help.fitbit.com/articles/de/Help_article/1133.htm).

[24] *Kuebler-Wachendorff, S., Luzsa, R., Kranz, J. et al.* The Right to Data Portability: conception, status quo, and future directions. Informatik Spektrum 44, 264–272 (2021), https://doi.org/10.1007/s00287-021-01372-w.

[25] Die Data Transfer Initiative etwa arbeitet an Verfahren zur Vereinheitlichung der verschiedenen Datenmodelle: https://dtinit.org/.

[26] Europäische Kommission, COM (2022) 68 final.

[27] Europäische Kommission, COM (2022) 68 final. S. 1 ff.

Datenübertragbarkeit nach Art. 20 DSGVO durch die Vorschriften der Art. 3 ff. DA ergänzt und den Nutzern ein Zugang zu den bei der Nutzung der Wearables generierten Daten in einem deutlich weiteren Umfang ermöglicht werden.[28] Neben der Verpflichtung nach Art. 3 DA zur Konzipierung und Herstellung von Produkten und verbundenen Diensten, die standardmäßig einen vereinfachten Zugang ermöglichen, regeln die Art. 4 ff. DA den Zugang des Nutzers zu den nutzergenerierten Daten gegenüber dem Dateninhaber und die Weitergabe dieser Daten an Dritte. Die praktische Ausgestaltung wird sich aber voraussichtlich als problematisch erweisen, da der Entwurf die tatsächliche Umsetzung der Interoperabilität nur begrenzt adressiert.[29] So stellen Art. 3 ff. DA im Gegensatz zu Art. 26 und Art. 28 DA keine expliziten Anforderungen an Struktur oder Format der Daten. Diese fehlende Vereinheitlichung könnte – wie bereits bei Art. 20 DSGVO[30] – die praktische Bedeutung des Teilens von Daten nach Art. 3 ff. DA erheblich beeinträchtigen.

Darüber hinaus stellt sich die Frage, in welchem Umfang und gegen wen das Recht des Nutzers auf Zugang und Weitergabe der Daten besteht.[31] Die Begriffsbestimmung des Dateninhabers nach Art. 2 Nr. 6 DA-E ist unscharf und kann zu erheblichen Abgrenzungsschwierigkeiten führen.[32] So könnten beispielsweise sowohl der Arbeitgeber als auch der Hersteller des Produktes Dateninhaber sein, sofern diese entsprechend den Voraussetzungen des Art. 2 Nr. 6 DA die Daten bereitstellen können. Zudem enthält der DA als "horizontaler Vorschlag" nur grundlegende Vorschriften für alle Sektoren, weshalb der Zugang und Nutzung der Daten auf "sektoraler Ebene" unterschiedlich geregelt werden können.[33] Für "elektronische Gesundheitsdaten" enthält beispielsweise der Vorschlag der Kommission für eine Verordnung über den europäischen Raum für Gesundheitsdaten (EHDS) sektorspezifische Verpflichtungen zur Bereitstellung dieser Daten an eine nationale Zugangsstelle gemäß Art. 33 Abs. 1 lit. f, Art. 2 Abs. 2 lit. o EHDS, die neben dem Hersteller der Geräte zur digitalen Selbstvermessung unter Umständen auch den Arbeitgeber als Dateninhaber im Sinne von Art. 2 Abs. 2 lit. y EHDS treffen könnten.

5.5 Fazit

Es ist zu erwarten, dass Methoden zur Selbstvermessung im Arbeitskontext künftig häufiger eingesetzt werden, da mit der steigenden Verbreitung der entsprechenden Geräte auch die Vorteile für den Arbeitskontext immer deutlicher sichtbar werden. Hiermit ist wie beschrieben eine Vielzahl an Herausforderungen verbunden. So ist derzeit weder geklärt, welchen Einfluss künftige Rechtsnormen, wie der Data Act haben werden, noch inwiefern derzeit bestehende Rechtsgrundlagen künftig anwendbar sein werden. Darüber hinaus zeigt sich auch, dass die technische Umsetzung derzeit nicht die Privatheit der Nutzer fokussiert wird, obwohl es durchaus Möglichkeiten gäbe, auch diesem Aspekt gerecht zu werden. Daher bleibt es Aufgabe der Forschung, die bestehenden Probleme zu adressieren und sowohl nutzer- als auch datenschutzfreundliche Lösungen zu entwickeln.

[28] Europäische Kommission, COM (2022) 68 final. S. 2.

[29] Stellungnahme des Max-Planck-Instituts für Innovation und Wettbewerb vom 25.05.2022; https://www.ip.mpg.de/de/forschung/meldungen-aus-der-forschung/stellungnahme-zum-datengesetz-data-act-der-eu.html, Rn. 295 ff.

[30] *Von Lewinski* in: Wolff/Brink, BeckOK Datenschutzrecht, 41. Edition, Art. 20 Rn. 1.1.

[31] Siehe beispielsweise: *Paal/Götz*, Aktuelle Fragen zur Datenübertragbarkeit aus Art. 20 DS-GVO, ZD 2023, 67; *Steinrötter*, Verhältnis von Data Act zur DS-GVO, GRUR 2023, 216; *Metzger/Schweitzer*, Shaping Markets: A Critical Evaluation of the Draft Data Act, ZEuP 2023, 42.

[32] *Bomhard/Merkle*, Der Entwurf eines Data Acts, RDi 2022, 168; Specht-Riemenschneider, Data Act – Auf dem (Holz-)Weg zu mehr Dateninnovation?, ZRP 2022, 137

[33] Europäische Kommission, COM (2022) 68 final, S. 2

6 Reciprocity of Data Sharing Infrastructures: A Conceptual Norms Framework

Frederik M. Metzger and Greta Runge[1]

6.1 Introduction

The current discussions on the development of national and European Data Infrastructures and Data Spaces are based on the premise of finding a balance between the common good and individual interests in the handling and use of data. In a society where an increasing amount of data is generated, collected, and made accessible to various actors, the way in which this data is used must be driven by the interests of the individual and be in line with European values, fundamental rights, and regulations (European Commission, 2020). Prominent policy initiatives, such as GAIA-X, the Mobility Data Space, or the European Health Data Space, currently show that the sharing of data between different actors is becoming increasingly important. In this context, data sharing can be characterized as the collaborative use of data for a common goal and goes beyond the exchange of data itself (Otto et al., 2022). This raises a number of questions including the interests and reasons for participation of the actors involved in the development of data-based infrastructures, the way data sharing is structured, and the conditions under which it takes place (Beverungen et al., 2022; Curry et al., 2022; Wagner, 2019).

Retrieval of shared data will be crucially dependent on both data availability and the balance between contribution and retrieval. This holds true in both quantitative and qualitative terms. If the pie is not large enough to be distributed and shared, data sharing infrastructures will be condemned to collapse. Similarly, if data contributed to sharing infrastructures is poor in quality, motivation for using it as a valuable source may be low. Reciprocity is seen here as an individual's behavior of giving back either in the same way or in different terms to the sender what he or she has received from the latter. Typically, reciprocity is seen as a moral norm guiding individual social behavior (Gouldner, 1960). Researchers have already pointed out that information-sharing behavior in online consumption communities rests upon the norm of reciprocity (Pai & Tsai, 2016), while other academics have focused on the question what steers reciprocity in online communities of the sharing economy (Proserpio et al., 2018).

However, these studies shed light on the topic of reciprocity in online communities and focus on the relationship between individuals and for-profit organizations. The role of contributing organizations remains untouched. Furthermore, the norm of reciprocity is seen as a whole, not permitting to distinguish between different motivators. Reciprocity is considered either as a black box or as a mere "social" mechanism, thus not allowing to identify where the motivation for reciprocating behavior comes from. For the sake of making data sharing infrastructures a burgeoning place of economic activity, it may become important to disentangle norms in use to incentivize by the most efficient levers. In addition, the question remains open whether new norms are needed or the known mechanisms of existing norms can be relied upon. Hence, in the present study we ask: how can reciprocity be described in the data sharing context? What are differences to reciprocity at the individual level?

Our study aims at providing both theoretical and practical guidance in the field of data sharing. The results can have important consequences on the design and incentives set to data-sharing infrastructures. In order to answer the research questions, we integrate the phenomenon of reciprocity

[1] Fraunhofer-Institut für System- und Innovationsforschung ISI, Karlsruhe. E-Mail: {vorname.nachname}@isi.fraunhofer.de

at an individual's level to the group decision level. Furthermore, we combine the three levels of moral, social, and legal norms into one framework to formulate a new "reciprocity norm of data sharing." Additionally, our study identifies important contingencies that mark the difference between the phenomenon of reciprocity described by anthropologists and sociologist and as needed in the present context of data sharing.

With our study, we contribute to the field of data sharing infrastructures by proposing a norms-based framework of data sharing reciprocity. The major contribution lies in transferring the concept of reciprocity from the individual to the group level and from the exchange of goods and services to the data context. The study proposes looking at three levels of norms—moral, social, and legal. By doing so, we are able to detect the underlying motivators of actors in the data-sharing context. Furthermore, we detect four important contingencies of the norm of reciprocity in the data-sharing sphere. These contingencies contrast the data-sharing context to the individuals' context, previously used in the study of both norms and reciprocity.

To answer the research question, we proceed as follows. First, we shed light on the phenomenon of reciprocity and discern what exactly we are looking at in the present study. Next, the three levels of moral, social, and legal norms are explained and exemplified. To conclude the theory section, we formulate four important contingencies of a "reciprocity norm of data sharing." Finally, our short article ends by proposing a conceptual norms framework of reciprocity in data sharing.

6.2 Theory

Reciprocity

Reciprocity is characterized by a unilateral act of giving, without knowing when, to what extent, and whether there will be a reply by the receiver in the future (Molm et al., 2000). In his seminal work, Gouldner (1960) conceptualized reciprocity as a moral norm having effects in social life. He recognized that there is an inherent force to reciprocate, which he refers to the moral obligation of reciprocity. While prior anthropological research had focused on the description of reciprocity in a broad sense, as the exchange of tangible and intangible goods and gifts (Malinowski, 1922; Mauss, 1923/1924), Gouldner (1960) called this exchange "complementarity." It consists of Ego's right and Alter's duty, or Ego's duty and Alter's right. However, what is not captured by these descriptions of complementarity is that rights and duties can also exist on each side of the exchanging parties. The latter case is what Gouldner (1960) called reciprocity, and what is reciprocity in the narrow sense.

Reciprocity can be translated into the form: if one person helped the other, the latter will help the former and will not harm him or her. According to Gouldner (1960, p. 170) this obligation is regardless of the status and the role one person fulfills or plays. Thus, it is not contingent on social roles, but rather inherent to persons and their belief system. It, therefore, can be observed across all cultures, albeit in different strength (Gouldner, 1960, p. 171). Reciprocity, thus, is regarded as positive reciprocity here. In contrast, negative reciprocity would follow the rule: if one person does not help the other, the latter will not help or even punish the former.

Furthermore, direct and indirect reciprocity have been the matter of scholarly work over the past years. Whereas direct reciprocity refers to the exchange between two parties, indirect reciprocity can be observed in a situation where three or more actors are involved. Indirect reciprocity is the act of not giving back to the initiating party, but to a third party who eventually gives back to the initial party. We will concentrate on direct reciprocity in this paper.

Since data sharing bears the term "sharing" in its denomination, the question arises why it is necessary to talk of reciprocity in the data-sharing context. If acknowledging that there is a difference between the two, this immediately leads to the question to which degree sharing and reciprocity differ. Belk (2010), in his theoretical contribution within consumer behavior, contrasts commodity

exchange and gift giving to sharing. Commodity exchange suffices the rules of acquisition in the marketplace and is valued as being highly reciprocal. Gift giving has implicit non-reciprocity, but is reciprocal in practice. This may be explained by the moral obligation pointed out by Gouldner (1960) that obliges an individual to reciprocate a gift. What unites the two is the transfer of ownership. In contrast, Belk (2010) sees sharing as a practice that is non-reciprocal and having shared ownership based on close social links to others with whom sharing is taking place. These are probably the three most representative characteristics out of a series of distinguishing criteria between the three phenomena.

When we look more precisely at data sharing, none of the three criteria we selected from Belk (2010) do apply to this phenomenon: neither close social links nor shared ownership nor non-reciprocity are given. Instead, data sharing bases on highly utilitarian motivations. Drawing on these characteristics, we choose reciprocity as one of the criteria of which we believe that it makes data sharing infrastructures work. We think that the success of data sharing infrastructures will depend on both data availability and the balance between contribution and retrieval.

Moral, Social, and Legal Norms

The use of the term "norms" must be disambiguated in order to better understand what is meant by its different meanings. Reciprocation and its forms will also depend on the social structures surrounding the act of reciprocity. Social structures build the evaluative framework on which to assess whether an act of giving will be valued as such. While the norm of reciprocity will rely upon an inner urge to reciprocate, social norms will determine how and under what conditions the act of reciprocation will take place. Furthermore, the legal environment will exert a formative role on the use of reciprocation.

Moral norms regulate behavior not by social convention, but rather motivated by an inner conviction of right and wrong that we identify to be moral responsibility. Thus, they are much less dependent on what others do or are perceived to think one own should do, compared to social norms. An example of moral norms is to prevent injuring another person, since this is seen as morally wrong, no matter what others do and say (Mackie et al., 2014, p. 26). Attaching to moral norms happens based on interior convictions, regardless of exterior coercion, and is the matter of inner dialogue or speech without considering exterior sources. Bicchieri (2014) argues that there is no sharp distinction between moral and social norms. However, she admits that a distinguishing feature of moral norms in contrast to social norms is that we unconditionally follow them. A moral norm is unconditional, as it has universal content, like the proscription of harming others without a reason. Moral norms are acknowledged to be internalized, what becomes clear when facing violation against them, since allegiance to them is perceived as unconditional. Bicchieri (2014) debates the possibility that social norms become so much internalized that they are perceived as moral norms. At the same time, it remains unsolved whether we come into being with a "moral organ," i.e., whether morality is implanted when we come into being. In contrast to social norms moral norms can be qualified as not being a product of groups and societies. We argue that moral norms are connected to both positive and negative internal rewards, as the conscience can express both types. Moral norms, at their base, are unwritten and thus informal.

Social norms are the informal rules that govern social behavior in groups and societies. They can best be described as "the grammar of social interactions," as Bicchieri (2006) calls them. Social norms specify what is good use and bad use defined by groups and societies. And like grammar in language, they are not the product of human design (Bicchieri et al., 2018), but rather come into being through other ways. An example of social norms is the rule of driving the vehicle on the right-hand side of the street because non-compliance would harm others and they would disapprove of the person for putting others in risk (Mackie et al., 2014, p. 26). In contrast to moral norms, they are not unconditionally followed and internalized (Bicchieri, 2014). In contrast to legal norms, social

norms have no discernable origin. Protracted social interactions made them emerge as good social strategies that are stable at a given moment in time. However, they can change rapidly, as the example of smoking in restaurants has shown. Social norms imply that we get an expectation on good and desirable behavior. Alter's behavior thus is measured against an implicit scale of desirable behavior, since we have an ingrained tendency to parallel the "what is" to "what ought to be" and conclude that "what is" behavior must be right or good (Bicchieri, 2014). To sum up, social norms are typically informal but externally coerced.

When talking of legal norms, the institutional level is concerned. Legal norms can be seen as a form of behavior imposed by the state and mostly occur in an explicit form, as they are to be enforced by coercion, mostly by governmental agencies. An example of legal norms is the prohibition of robbing since penalties are put upon this behavior and the belief in this enforcement will prevent people from doing so (Mackie et al., 2014, p. 26). In contrast to social norms, legal norms usually proscribe behavior, while social norms also prescribe behavior. The explicit form of a legal norm regulates in which condition it is implemented, the subjects concerned by it, their mutual rights and duties, as well as the sanctions for failing to obey the duty (Bicchieri, 2014). Human behavior, by nature, can be motivated by more than one reason. This is why this contrasting distinction can also bear a mix of motives, such as personal attitude, population regularity, social proof, or legal norms (Mackie et al., 2014, p. 26). Overall, legal norms are externally coerced and have formality.

Contingencies

With regard to data sharing infrastructures, context factors compared to the initial concepts are different and must be accounted for. These contingencies are depicted in table 6.1 and will be briefly described in the following. Contingencies are influencing factors of the outcomes of a possible "reciprocity norm of data sharing."

First, in contrast to the studies of reciprocity and of norms cited above, the entity of analysis will not be individuals, but rather organizations and groups of decision makers. This entails important consequences with regard to the decision-making processes. Depending on the configuration of the groups and their dynamics, there will be multi-party processes to consider. This brings to the fore both the social and legal norms levels, and outshines the moral level of the individual.

Second, the need to distinguish between quantitative and qualitative contributions or reciprocation becomes important. There is a difference whether participants transfer and share the amount of data or whether they do so depending on the type of data. The former referring to the quantity of data implies the broadening of information, whereas the latter refers to the quality of data transmitted. Both can be disambiguated with regard to the phenomena of datasets versus data rows. While the former means there are different types, i.e., quality of, data, the latter implies the higher amount of equivalent data entries. The norms of reciprocity introduced above refer to both types—quantity and quality.

Third, the intangible character of data in contrast to tangibility of goods must be taken into consideration. With reference to the three levels of norms, it makes a difference whether the goods shared and reciprocated are tangible or not. Tangibility is connected to the senses of seeing, touching, taste, or smell (Miller & Foust, 2003). While all these characteristics are not given for data, the latter may become more ephemeral compared to tangible goods. This may have important consequences on the use of the norm of reciprocity at the three levels.

Fourth, data have an infinitely divisible character. They can be copied and distributed at the marginal cost close to zero, and there is no loss of information while copying. This character of virtuality is inherent in data and is supposed to have a major impact in reciprocating behavior. Since goods and services, which have been the matter of exchange between parties, are typically undividable,

their relative value will be coupled to their scarcity. However, the effort of copying data and making them duplicate has nearly no marginal cost, which will impact the reciprocating behavior of actors.

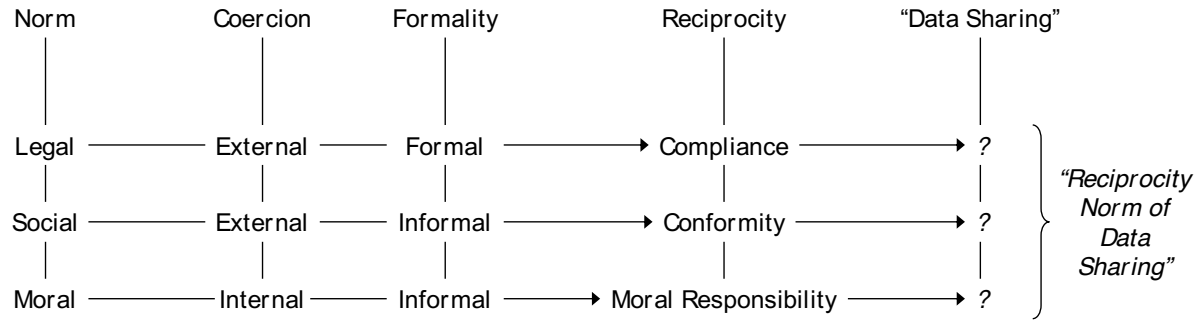
Table 6.1: Contingencies of a "Reciprocity Norm of Data Sharing"

Entity	Organizations and groups, instead of individuals
Quality	Quantitative versus qualitative contributions
Intangibility	Intangible, instead of tangible character of data
Divisibility	Infinitely divisible character of data

6.3 Conclusion: Proposing a Conceptual Norms Framework

What remains unclear at this point in time, is the constellation of a "reciprocity norm of data sharing." This means, the question remains open at which of the three levels the norm will be located. When taking the above-mentioned norms into account, a three-partite conceptual norms framework, as shown in figure 6.1 arises. Norms as the binding mechanisms underlying human behavior can be separated into the categories of moral, social, and legal norms (Bicchieri, 2014; Mackie et al., 2014). Moral norms are delineated as behavior motivated by interior coercion and no formality. Social norms are seen as external, social coercion and nor formality. Finally, legal norms are seen as external, governmental coercion combined by formality. As the framework will show, one or a combination of these norms can motivate reciprocating behavior. At the moral level, reciprocity is seen as a moral responsibility; at the social norms level as conformity; and at the legal level as compliance.

Figure 6.1: Proposition of a Norms Reciprocity Framework



We conclude by showing that the three levels of norms can be motivators for reciprocity on data sharing infrastructures. As prior cases show, there is an interaction between all three levels (Mockus, 2002). We demonstrate the explanatory power of the framework by describing that there is also a hierarchy of norms, whereby the moral norm is hypothesized to be the most powerful in general. However, in the data sharing example, there will be an abstraction and alienation from personal and moral foundations. As we identified, four important contingencies could apply in this context. The decision entity as well as the quality, the intangibility, and divisibility of data are rooted in the impersonal information-processing activity of organizations and the character of data differing from goods and services. There is no unique norm yet explicitly developed for reciprocity in data sharing or the handling of data-related activities. Knowing which norm level will be the strongest allows for the development of efficient data sharing norms. We thus call for the development of a "reciprocity norm of data sharing," namely data-compliant norms to be researched and formulated and eventually internalized to the subordinate norms levels.

6.4 References

Belk, R. (2010). Sharing. *Journal of Consumer Research*, 36(5), 715–734. https://doi.org/10.1086/612649

Beverungen, D., Hess, T., Köster, A., & Lehrer, C. (2022). From private digital platforms to public data spaces: Implications for the digital transformation. *Electronic Markets*, 32(2), 493–501. https://doi.org/10.1007/s12525-022-00553-z

Bicchieri, C. (2006). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Bicchieri, C. (2014). Norms, Conventions, and the Power of Expectations. In N. Cartwright & E. Montuschi (Eds.), *Philosophy of social science: A new introduction* (pp. 208–229). Oxford University Press.

Bicchieri, C., Muldoon, R., & Sontuoso, A. (2018). Social Norms. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2018). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2018/entries/social-norms/

Curry, E., Scerri, S., & Tuikka, T. (Eds.). (2022). *Data Spaces: Design, Deployment and Future Directions*. Springer. https://doi.org/10.1007/978-3-030-98636-0

European Commission. (2020). A European Strategy for Data. COM(2020) 66 final. Brussels. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0066

Gouldner, A. W. (1960). The Norm of Reciprocity: A Preliminary Statement. *American Sociological Review*, 25(2), 161. https://doi.org/10.2307/2092623

Mackie, G., Moneti, F., Denny, E., & Shakya, H. (2014). What are Social Norms? How are they Measured? [Working Paper]. https://www.youthpower.org/sites/default/files/YouthPower/files/resources/Mackie_2014_What%20are%20Social%20Norms.pdf

Malinowski, B. (1922). *Argonauts of the Western Pacific*. Routledge & Kegan Paul.

Mauss, M. (1923/1924). Essai sur le don: Forme et raison de l'échange dans les sociétés archaïques. *L'Année sociologique* (1896/1897-1924/1925), 1, 30–186.

Miller, D. W., & Foust, J. E. (2003). Classifying Services by Tangibility/Intangibility of Attributes and Benefits. *Services Marketing Quarterly*, 24(4), 35–55. https://doi.org/10.1300/J396v24n04_03

Mockus, A. (2002). Co-existence as Harmonization of Law, Morality and Culture. *Prospects*, 32(1), 19–37. https://doi.org/10.1023/A:1019740325436

Molm, L. D., Takahashi, N., & Peterson, G. (2000). Risk and Trust in Social Exchange: An Experimental Test of a Classical Proposition. *American Journal of Sociology*, 105(5), 1396–1427. https://doi.org/10.1086/210434

Otto, B., Ten Hompel, M., & Wrobel, S. (Eds.). (2022). *Designing Data Spaces: The Ecosystem Approach to Competitive Advantage*. Springer. https://doi.org/10.1007/978-3-030-93975-5

Pai, P., & Tsai, H.-T. (2016). Reciprocity norms and information-sharing behavior in online consumption communities: An empirical investigation of antecedents and moderators. *Information & Management*, 53(1), 38–52. https://doi.org/10.1016/j.im.2015.08.002

Proserpio, D., Xu, W., & Zervas, G. (2018). You get what you give: Theory and evidence of reciprocity in the sharing economy. *Quantitative Marketing and Economics*, 16(4), 371–407. https://doi.org/10.1007/s11129-018-9201-9

Wagner, H. (2019). Daten—Räume—Datenräume. Schriften zur Kultur- und Mediensemiotik | Online, 313–326. https://doi.org/10.15475/SKMS.2018.2.13

7 Open Personal Data: Anonymisierung im Spannungsfeld zwischen Informationsgehalt und Robustheit

Sebastian Wilhelm, Jakob Folz und Florian Wahl[1]

7.1 Motivation

Daten bilden die Grundlage von Entscheidungen in vielfältigen Bereichen der Politik, Wissenschaft und Gesellschaft. Daten führen zu Innovationen und unterstützen zahlreiche Unternehmensprozesse, indem sie Möglichkeiten offenlegen, neue Dienste zu entwickeln, Einsparpotenziale zu identifizieren und Prozesse zu optimieren (European Commission, 2022; Tran & Scholtes, 2015). Der Zugang und die Verfügbarkeit von frei zugänglichen und lizenzfreien Daten (Open Data) erhält daher weltweit eine wachsende Bedeutung, wird zu einem immer bedeutenderen Wirtschaftsfaktor und ist Bestandteil einer modernen Gesellschaft und öffentlichen Infrastruktur (Bundesministerium des Innern und für Heimat, 2021)

Die Veröffentlichung von Open Data birgt jedoch Herausforderungen in Bezug auf Datenschutz, Sicherheit und Persönlichkeitsrechte (Demary et al., 2019; Tran & Scholtes, 2015). Um den rechtlichen Bestimmungen zu genügen, dürfen öffentlich zugänglich gemachte Daten, sofern keine anderweitige explizite Rechtsgrundlage vorliegt, im Allgemeinen keinen Personenbezug offenbaren. Anonymisierung oder Entfernung von identifizierbaren Personeninformationen ist der konventionelle Weg, um diese rechtlichen Bestimmungen einzuhalten und Datenschutzbedenken zu berücksichtigen (Murray Jr et al., 2021). Diverse Handlungsleitfäden beschreiben daher, dass Daten vor einer Veröffentlichung als Open Data anonymisiert werden müssen (GovData, 2020; Klessmann et al., 2012; Kompetenzzentrum Open Data, 2020; Open Data Charter, 2015). Konkrete Anweisungen, wie diese Anonymisierung umzusetzen ist, werden jedoch nicht gegeben. Vielmehr wird die Zuständigkeit an die jeweiligen Datenschutzbeauftragten delegiert (Kompetenzzentrum Open Data, 2020).

7.2 Spannungsfeld zwischen Informationsgehalt und Robustheit der Anonymisierung

Die Anonymisierung von Datensätzen selbst steht jedoch im Spannungsfeld zwischen Informationsgehalt und Robustheit. Einerseits ist es wichtig, dass die Anonymisierung robust genug ist, um eine Rückführung der Daten auf individuelle Personen und somit eine De-Anonymisierung zu verhindern. Andererseits sollte der Informationsgehalt der Daten nicht unnötig reduziert werden, um den für Analysen und Anwendungen notwendigen Informationsgehalt der Daten zu erhalten.

Ein Problem besteht darin, dass bei der Anonymisierung von Daten häufig Zielparameter definiert werden müssen, wie bei der *k-Anonymisierung*. Die Definition dieser Parameter bestimmt anschließend den Grad der Anonymisierung – konkret den erforderlichen Aufwand oder die Menge an Informationen, die von extern zugespielt werden müssen, um die Daten zu de-anonymisieren. Dieser Grad dient gemäß Art. 26 DSGVO als wichtiges Kriterium, um zu entscheiden, ob ein Datensatz als anonym anzusehen ist und daher als Open Data veröffentlicht werden darf. Es ist jedoch nicht eindeutig, wie diese Parameter gewählt werden sollen, dass eine robuste und rechtssichere Anonymisierung gewährleistet ist. Gleichzeitig soll der Informationsgehalt der Daten nicht unnötig weit reduziert werden. Folglich befinden sich Datenhaltende bei der Anonymisierung

[1] Technische Hochschule Deggendorf, Campus Grafenau E-Mail: {vorname.nachname}@th-deg.de

von Daten, die als Open Data veröffentlicht werden sollen, im Spannungsfeld zwischen „*zu wenig Anonymisieren*“, also einem zu geringen Anonymisierungsgrad (Manske, 2016; Mirani, 2014; Sweeney et al., 2013), mit der Folge, dass die Daten potenziell zu leicht de-anonymisiert werden könnten und „*zu viel Anonymisieren*“, also einem hohen Grad an Aggregation/Reduktion (Manske, 2016), mit der Folge, dass der Datensatz an Nutzwert verliert. Es bedarf einer sorgfältigen Abwägung und eines fundierten Verständnisses der Datenschutzanforderungen und der spezifischen Anwendungskontexte, um diesen Balanceakt zu meistern.

Das Spannungsfeld ist in **Fehler! Verweisquelle konnte nicht gefunden werden.** und **Fehler! Verweisquelle konnte nicht gefunden werden.** exemplarisch dargestellt. Beide Tabellen zeigen denselben fiktiven Datensatz der einmal mit Fokus auf die Robustheit (**Fehler! Verweisquelle konnte nicht gefunden werden.**) und einmal mit dem Fokus des Erhalts des Informationsgehalts (**Fehler! Verweisquelle konnte nicht gefunden werden.**) anonymisiert wurden. Es wird unmittelbar deutlich des der mit Fokus auf Robustheit anonymisierte Datensatz einen sehr hohen Verlust an Informationsgehalt aufweist. Analysen und Statistiken welche basierend auf diesem Datensatz erstellt werden (z. B. zum Durchschnittsgehalt nach Altersgruppen und Beruf), haben eine viel geringere Detailtiefe als dieselben Analysen und Statistiken, die auf dem Datensatz auf **Fehler! Verweisquelle konnte nicht gefunden werden.** erstellt werden. Jedoch ist die Gefahr einer Re-Identifikation von Einzelpersonen und somit einer De-Anonymisierung auf dem Datensatz in **Fehler! Verweisquelle konnte nicht gefunden werden.** deutlich höher.

Tabelle 1: Fiktiver Beispieldatensatz mit Fokus auf Robustheit anonymisiert

Name	Alter	Geschl.	Arbeitsbereich	PLZ	Std./Woche	Einkommen
***	20-29	m	Dienstl. und Handwerk	5****	31-40	30.000 - 39.999
***	30-39	m	Dienstl. und Handwerk	4****	>40	> 50.000
***	50-59	w	Büro und Verwaltung	2****	31-40	> 50.000
***	30-39	w	Büro und Verwaltung	5****	31-40	30.000 - 39.999
***	30-39	w	Kunst und Kultur	4****	21-30	< 19.999

Tabelle 2: Fiktiver Beispieldatensatz mit Fokus auf Informationsgehalt anonymisiert

Name	Alter	Geschl.	Arbeitsbereich	PLZ	Std./Woche	Einkommen
W. E****	24	m	Postbote	56284	39	34.100
F. W******	30	m	sebstst. Handwerker	42113	50	64.300
J. S*****	56	w	Steuerberaterin	24626	40	79.540
S. B***	38	w	Bürokauffrau	54450	39	38.500
K. H***	33	w	Musikerin	43221	20	19.800

Die rechtliche Situation trägt zusätzlich zur Komplexität bei. Die Gesetzgebung legt mit Erwägungsgrund Nr. 26, Satz 4 der DSGVO fest, dass bei der Bewertung, ob ein Datensatz anonym ist, auch zukünftige technische Entwicklungen berücksichtigt werden müssen. Dies stellt für Einzelpersonen, die Daten veröffentlichen möchten, eine umfassende Herausforderung dar. Technologische Fortschritte und die Entwicklung neuer Methoden zur De-Anonymisierung könnten die Wirksamkeit der Anonymisierung beeinflussen und erfordern daher eine kontinuierliche Bewertung und Anpassung der Anonymisierungsverfahren.

Insgesamt ist die Anonymisierung von personenbezogenen Daten, die als Open Data veröffentlicht werden sollten, ein komplexes und anspruchsvolles Thema. Das Spannungsfeld zwischen

Informationsgehalt und Robustheit erfordert eine gründliche und individuelle Abwägung der Anonymisierungsmethoden und -parameter sowohl im rechtlichen, ethischen, als auch technischen Diskurs. Die rechtlichen Anforderungen und die Dynamik der technologischen Entwicklungen stellen zusätzliche Herausforderungen dar. Dennoch ist es von entscheidender Bedeutung, dass die Anonymisierung von Daten sorgfältig umgesetzt wird, um den Schutz der Privatsphäre zu gewährleisten und gleichzeitig das Potenzial von Open Data für Innovationen und gesellschaftlichen Fortschritt zu nutzen.

7.3 Ansatz im Projekt „*EAsyAnon*“

Ein möglicher Ansatz, um dem genannten Spannungsfeld zu begegnen und Personen, welche einen Datensatz als Open Data veröffentlichen möchten, bei der Findung eines geeigneten Anonymisierungskonzepts zu unterstützen und dabei die Robustheit der Anonymisierung und den Datensatz in ein zweckmäßiges und ausgewogenes Verhältnis zu bringen wird im Projekt *EAsyAnon*[2] erforscht. Es soll ein zweistufiges System entwickelt werden, bestehend aus einem Empfehlungssystem und einem Auditsystem.

Im *Empfehlungssystem* sollen Datenhaltende Personen zunächst darin unterstützt werden, ein geeignetes Anonymisierungskonzept für einen spezifischen Datensatz zu erhalten, welches einerseits eine robuste Anonymisierung sicherstellt und andererseits den Informationsverlust minimiert. Dazu werden basierend auf einer Meta-Beschreibung des Datensatzes, also einer Beschreibung der Form und des Inhalts des Datensatzes ohne konkrete Inhalte preiszugeben, unter Zuhilfenahme von Methoden der Künstlichen Intelligenz (KI) und der Statistik geeignete Anonymisierungsmethoden samt geeigneter Zielparameter empfohlen. Die Algorithmik betrachtet dabei den Kontext des Datensatzes, die einzelnen im Datensatz enthaltenen Attribute sowie eine vom Datenhaltenden erstelltes Attribut-Scoring, also einer Einschätzung der Relevanz einzelner Attribute im Hinblick auf den Informationsgehalt. Basierend auf dieser Empfehlung kann die Datenhaltende Person anschließend die Anonymisierung des Datensatzes selbst vornehmen.

Im zweiten Schritt des vorgeschlagenen Systems erfolgt die Überprüfung des anonymisierten Datensatzes. Nachdem ein Datensatz basierend auf dem vorgeschlagenen Konzept anonymisiert wurde, sollte es den Datenhaltenden Personen ermöglicht werden, eine Überprüfung des Datensatzes durch externe zu erhalten. Dazu soll zunächst automatisiert überprüft werden, ob die vorgeschlagenen Zielparameter, beispielsweise *k* bei der *k-Anonymisierung*, eingehalten wurden. Ist diese automatisierte Überprüfung erfolgreich, so soll der Datensatz in einem zweiten Überprüfungsschritt vor einer allgemeinen Veröffentlichung verschiedenen Expert:innen aus unterschiedlichen Fachdisziplinen vorgelegt werden, die den Datensatz hinsichtlich der Robustheit der Anonymisierung und dem Informationsgehalt bewerten. Erst nachdem die Expert:innen den Datensatz bewertet haben, erfolgt, sofern die Robustheit der Anonymisierung ausreicht, und ausreichen Informationsgehalt quittiert wurde, eine allgemeine Veröffentlichung des anonymen Datensatzes als Open Data.

Die Bewertungen der Expert:innen können in den KI-Algorithmus zur Empfehlung von Anonymisierungskonzepten für künftige Datensätze mit einfließen, sodass dadurch auch künftige Empfehlungen das Spannungsfeld zwischen Robustheit und Informationsgehalt auch basierend auf den Meinungen der Expert:innen berücksichtigen.

Das vorgeschlagene System kann zwar das Spannungsfeld nicht auflösen aber wenigstens Datenhaltende Personen, die bereit sind einen Datensatz als Open Data zu veröffentlichen eine fundierte Unterstützung bieten, um mit der Abwägung zwischen Robustheit und Informationsgehalt umzugehen.

[2] siehe www.easyanon.de (abgerufen am 13.10.2023)

Danksagung

Das Projekt „EAsyAnon – Empfehlung- und Auditsystem zur Anonymisierung“ wird gefördert durch das Bundesministerium für Bildung und Forschung (BMBF) und finanziert durch die Europäische Union (NextGenerationEU).

Literatur

Bundesministerium des Innern und für Heimat. (2021). *Open Data*. Bundesministerium des Innern und für Heimat. https://www.bmi.bund.de/DE/themen/moderne-verwaltung/open-government/open-data/open-data-node.html

Demary, V., Azkan, C., Fritsch, M., Goecke, H., Korte, T., Krotova, A., Lichtblau, K., & Schmitz, E. (2019). Readiness Data Economy–Bereitschaft der deutschen Unternehmen für die Teilhabe an der Datenwirtschaft. *Institut der deutschen Wirtschaft Köln e. V., Köln*.

European Commission. (2022). *Wertschöpfung durch Open Data*. European Commission. https://data.europa.eu/elearning/de/module2/#/id/co-01

GovData. (2020). *Fragen und Antworten: GovData*. https://www.govdata.de/faq

Klessmann, J., Denker, P., Schieferdecker, I., & Schultz, S. E. (2012). *Open Government Data Deutschland*. Budesministerium des Inneren. https://www.verwaltung-innovativ.de/SharedDocs/Publikationen/eGovernment/open_government_data_deutschland_langfassung.pdf?__blob=publicationFile&v=5

Kompetenzzentrum Open Data. (2020). *Open Data Handbuch*. https://www.bva.bund.de/SharedDocs/Downloads/DE/Behoerden/Beratung/Methoden/open_data_handbuch.pdf?__blob=publicationFile&v=8

Manske, J. (2016). *Offene Daten und der Schutz der Privatsphäre*. https://www.stiftung-nv.de/sites/default/files/impulse_julia_manske_offene_daten_privatsphare.pdf

Mirani, L. (2014). *London's bike-share program unwittingly revealed its cyclists' movements for the world to see*. http://qz.com/199209/londons-bike-share-program-unwittingly-revealed-its-cyclists-movements-for-the-world-to-see/

Murray Jr, J., Mashhadi, A., Lagesse, B., & Stiber, M. (2021). Privacy Preserving Techniques Applied to CPNI Data: Analysis and Recommendations. *arXiv preprint arXiv:2101.09834*.

Open Data Charter. (2015). *International Open Data Charter*. https://opendatacharter.net/wp-content/uploads/2015/10/opendatacharter-charter_F.pdf

Sweeney, L., Abu, A., & Winn, J. (2013). Identifying Participants in the Personal Genome Project by Name. *arXiv preprint arXiv:1304.7605*. https://doi.org/10.2139/ssrn.2257732

Tran, E., & Scholtes, G. (2015). Open~Data Literature Review. *Barkeley School of Law, University of California*.

Plattform < Privatheit >

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

PROJEKTPARTNER

Fraunhofer

Natur
Technik
Kultur
Gesellschaft

UNIKASSEL
VERSITÄT