Bachelorarbeit

Provenance und Privacy in ProSA

Eingereicht von: Nic Scharlau

Eingereicht am: 18. Februar 2020

Gutachter: Prof. Dr. rer. nat. habil. Andreas Heuer
Dr.-Ing. Holger Meyer

Betreuerin: Tanja Auge, M. Sc.

Zusammenfassung

In Anbetracht von **Big Data** entstanden in den letzten Jahren diverse neue Herausforderungen im Datenmanagement, die es zu bewältigen gilt. Eine davon ist das Zusammenspiel von **Data Provenance** – dem Zurückverfolgen von Anfrageergebnissen – mit dem **Datenschutz** und der **Privatheit** eines jeden Menschen. Ziel dieser Arbeit ist es, zu untersuchen, welche **Privacy**-Probleme bei der Anwendung von Data Provenance entstehen, wie der Datenschutz umgekehrt der Provenance im Wege steht und erste Ideen zu diskutieren, wie beide Aspekte miteinander kombiniert werden können. Wir beschränken uns dabei auf **where**-, **why**- und **how**-Provenance sowie hauptsächlich auf **extensionale Provenance-Antworten**. Dazu muss jedoch zunächst geklärt werden, was allgemein unter „Provenance“ und „Privacy“ verstanden wird, welche Daten im Forschungsalltag anfallen und wie mit ihnen umgegangen wird und welche Interessen überhaupt bestehen, **Forschungsdaten** (nicht) zu veröffentlichen. Aus diesem Grund haben wir **Interviews** mit 20 Personen – sowohl aus dem wissenschaftlichen als auch aus dem nicht-wissenschaftlichen Bereich – geführt und die Ergebnisse zusammengetragen und diskutiert. So werden wir beispielsweise feststellen, dass insbesondere der Privacy-Begriff gar nicht so eindeutig ist, wie es zunächst scheint.

Abstract

In regard to **big data**, many new challenges considering data management that need to be addressed have arose in recent years. One of them is the combination of **data provenance** – the ability to trace back certain query results – with the **privacy** of every person. In this bachelor thesis, we researched which privacy-driven problems can occur whilst using data provenance, in which cases privacy might impede our provenance abilities, and how both provenance and privacy can be combined without violating the principles of either. We mainly focus on **where**, **why** and **how** provenance as well as **extensional provenance answers**. First and foremost however, we need to clarify what people associate with the terms “provenance” and “privacy”, what types of data are being generated in everyday research and how they are managed, and what reasons scientists might have to (not) publish their research data. Therefore, we have interviewed 20 people – scientists and non-scientists – and evaluated their answers for discussion. For example, we concluded that the term “privacy” is not as unique as one would think at first glance.

Inhaltsverzeichnis

Änderungen gegenüber der Print-Version vom 18. Februar 2020

- Seite 21: Das Prinzip der k-Anonymität wurde nicht, wie zuvor behauptet, 1998 von *Latanya Sweeney* und *Pierangela Samarati* vorgestellt, sondern erst 2001 von *Samarati* im Paper [Sam01]. Zwar taucht dieser Begriff bereits in einem Paper aus dem Jahr 1998 auf, dieses wurde allerdings nie veröffentlicht.
- Seite 47: Das permutierte Provenance-Polynom zu Tabelle 4.14 wurde korrigiert. In der ursprünglichen Fassung wurden nur die Werte vertauscht. Korrekt ist ein Vertauschen der gesamten Teilpolynome, wie es auch im Text beschrieben wird.
- Seite 73: Der Absatz „Transkribierte Interviews“ wurde um eine Erläuterung des Aufbaus der jeweiligen Transkripte erweitert.

Letzte Änderung: 28. März 2020

1. Einleitung

Die vorliegende Bachelorarbeit „Provenance und Privacy in ProSA“ befasst sich mit dem Zusammenspiel von Data Provenance und Privacy (Datenschutz) im an der *Universität Rostock* entwickelten Projekt *ProSA*[1] („Provenance Management durch Schema-Abbildungen und Annotationen“). Dieses nutzt neben Data Provenance ein universales Werkzeug namens *CHASE*[2], um mittels inverser Schema-Abbildungen eine minimale Teildatenbank einer gegebenen Forschungsdatenbank zu berechnen. Das Hauptziel des Projektes liegt somit in der Reduzierung von Forschungsdaten sowohl im Sinne der Speicherplatzausnutzung als auch im Hinblick auf Privacy-Aspekte. Als praktischer Anwendungsfall dient hierbei stets das *Leibniz-Institut für Ostseeforschung in Warnemünde (IOW)*, welches bereits seit geraumer Zeit mit der *Universität Rostock* zusammenarbeitet.

Doch was ist überhaupt mit Data Provenance gemeint?
Data Provenance bezeichnet die Rückverfolgung von (aggregierten) Datensätzen bis zu den originalen Datensätzen, also dem Ursprung der Daten, um nach einer Datenbankanfrage den originalen Datensatz aus dem Ergebnis zu rekonstruieren. Hierzu werden neben der Ergebnisrelation auch eine Reihe weiterer zusätzlicher Daten wie eine (minimale) **Zeugenmenge** bzw. **-basis** und sogenannte **Provenance-Polynome** gespeichert. Diese dienen der Beantwortung der im Forschungsdatenmanagement relevanten Fragen, woher die Daten stammen (**where**-Provenance), warum ein bestimmtes Ergebnis zustande kam (**why**-Provenance) und wie dieses konkret berechnet wurde (**how**-Provenance).

Je nachdem, wie umfangreich diese Zusatzdaten ausfallen, ist es möglich, einige Teile der ursprünglichen Datenbank wiederherzustellen. Dabei gilt, dass die Ausgangsdatenbank umso präziser rekonstruiert werden kann, je mehr Provenance-Daten gespeichert werden. Aggregiert man beispielsweise numerische Werte, indem man deren Durchschnitt bildet, kann man je nach Art der Provenance-Anfrage (**where**, **why** oder **how**) tatsächlich nur den Durchschnittswert, den Durchschnittswert sowie die Anzahl der Tupel oder aber jeden einzelnen Wert der originalen Spalte rekonstruieren. Es muss also allgemein zwischen dem Grad der gewünschten Genauigkeit und der Menge an Speicherplatz, die dafür verwendet werden soll, abgewogen werden.

Allerdings ist es gar nicht immer sinnvoll, die Datenbank genauestens rückberechnen zu können. Soll beispielsweise die Durchschnittsnote aller Studierenden berechnet werden, die eine bestimmte Klausur geschrieben haben, ist es unter Umständen nicht nur irrelevant, welche Person welche Note erhalten hat, sondern hinsichtlich des Datenschutzes sogar bedenklich. Hierauf werden wir im weiteren Verlauf der Arbeit immer wieder verweisen.

An dieser Stelle stellt sich natürlich auch die Frage, was Datenschutz eigentlich ausmacht:
Datenschutz, oder **Privacy**, beschäftigt sich mit dem Schutz personenbezogener Daten eines Individuums. Diese sind laut *Artikel 4 der Europäischen Datenschutz-Grundverordnung (DSGVO)* definiert als „Informationen, die sich auf eine identifizierte oder identifizierbare natürliche Person [...] beziehen“, wobei eine Person „direkt oder indirekt, insbesondere mittels Zuordnung zu einer Kennung [...] oder zu einem

[1] `https://dbis.informatik.uni-rostock.de/forschung/aktuelle-projekte/prosa/`
[2] Der *CHASE* ist ein ursprünglich von *Maier*, *Mendelzon* und *Sagiv* entwickelter Fixpunkt-Algorithmus, welcher in der Datenbanktheorie unter anderem für Anfrageoptimierung, Untersuchung von Äquivalenzen und der Darstellung von funktionalen Abhängigkeiten in Datenbanken verwendet wird.

oder mehreren besonderen Merkmalen, die Ausdruck der physischen, physiologischen, genetischen, psychischen, wirtschaftlichen, kulturellen oder sozialen Identität dieser natürlichen Person sind, identifiziert werden kann“. Rückschlüsse auf einzelne Personen sollen in einem veröffentlichten Datensatz somit unter keinen Umständen möglich sein. Zwischen Data Provenance und Privacy besteht also ein natürlicher Interessenskonflikt: Ersteres hat den Anspruch, Daten bis zur Quelle rückverfolgen zu können, während letzteres zum Ziel hat, ebendies zu verhindern.

Ein Ziel dieser Bachelorarbeit ist es, gegenwärtige Datenschutzaspekte wie Anonymitätsmaße und Datenschutzmethoden hinsichtlich ihrer Verträglichkeit mit Provenance-Anfragen zu untersuchen und ein Konzept zu entwickeln, mit dem der Datenschutz auch bei einer (Teil-)Rekonstruktion der Datenbank gewährleistet werden kann.

Ein weiteres Ziel der Arbeit ist es, den Umgang mit Forschungsdaten in der Praxis zu untersuchen. Dazu werden Forscherinnen und Forscher, aber auch Personen aus nichtwissenschaftlichen Bereichen interviewt, welche in ihrem konkreten Alltag mit Forschungsdaten arbeiten. Auf diese Weise soll ermittelt werden, welche Anforderungen beim Forschungsdatenmanagement bestehen, welche Daten überhaupt langfristig benötigt werden und welche überhaupt rekonstruierbar sein sollen. Tatsächlich ist es nicht immer im Sinne von Forscherinnen und Forschern, ihre Daten vollumfänglich zu veröffentlichen. Liegt beispielsweise ein exklusives Nutzungsrecht erhobener Daten vor, sei es auch nur für einen gewissen Zeitraum, besteht ein Interesse daran, nicht mehr Daten preiszugeben als es für Publikationen notwendig ist. Der Privacy-Begriff bezieht sich hierbei also nicht auf die Privatheit einzelner Personen, sondern auf den Schutz der Forschungsdaten an sich. Letzteres ist insbesondere im Zusammenhang mit wissenschaftlichen Fördervereinen wie der *Deutschen Forschungsgemeinschaft* interessant, die im Sinne wissenschaftlicher Reproduzierbarkeit bestimmte Anforderungen an Projekte und Veröffentlichungen haben.

Abweichung von der ursprünglichen Aufgabenstellung Der ursprüngliche Fokus dieser Bachelorarbeit lag auf dem *ProSA*-Projekt und dem Untersuchen von Datenschutztechniken im Zusammenspiel mit Provenance-Anfragen. Dieser Teil ist auch nach wie vor Bestandteil der Arbeit. Allerdings stellte sich bereits zu Beginn heraus, dass zunächst geklärt werden muss, welchen Bezug Forscherinnen und Forscher in der Praxis überhaupt zu „Privacy“ haben, da Personendaten im Forschungsalltag keinesfalls Standard sind. Forschungsdaten anderer Art können allerdings dennoch schützenswert sein. Dies führt zu der Frage, was der Privacy-Begriff eigentlich abdeckt und was nicht und inwiefern Provenance, konkreter Data Provenance, dort unterstützend angewandt werden kann oder vielleicht sogar hinderlich ist. Ergänzend zur eigentlichen Aufgabenstellung wurde deshalb eine empirische Befragung durchgeführt. Dazu waren zunächst zehn Expertinnen und Experten vorgesehen, später wurde der Kreis jedoch auf 20 ausgeweitet, um qualitativere und vielfältigere Daten zu erhalten. Die ursprüngliche Aufgabe – das Untersuchen von Provenance und Privacy speziell im Projekt *ProSA* – verlor dadurch an Priorität.

1.1. Der Beispieldatensatz

Als Beispieldatensatz dient in dieser Arbeit die Datenbank einer Universität mit Studentinnen und Studenten, diversen Vorlesungen und Prüfungsergebnissen. Die dafür benötigten Tabellen seien wie folgt definiert:

- **Tabelle 1.1**: STUDENT = {MatNr, Name, Vorname, Geburtstag, Straße, PLZ, Ort, Stadtteil, Studiengang} mit
  - $\mathbb{D}(\text{MatNr}) := \{10001, 10002, \ldots, 10017\}$
  - $\mathbb{D}(\text{Name}) := \{\text{Bach, Deckert, Fieber}, \ldots, \text{Wolter, Zimmermann}\}$
  - $\mathbb{D}(\text{Vorname}) := \{\text{Damian, Daniel, Erich}, \ldots, \text{Sarah, Ulrike}\}$
  - $\mathbb{D}(\text{Geburtstag}) := \{\text{26.11.1991, 27.11.1991, 13.12.1992}, \ldots, \text{23.07.1999, 01.02.2000}\}$
  - $\mathbb{D}(\text{Straße}) := \{\text{Albert-Einstein-Straße 26, Bertolt-Brecht-Straße 14}, \ldots, \text{Ährenkamp 3}\}$
  - $\mathbb{D}(\text{PLZ}) := \{18057, 18059, 18106, 18119, 18146, 18147\}$
  - $\mathbb{D}(\text{Ort}) := \{\text{Rostock}\}$
  - $\mathbb{D}(\text{Stadtteil}) := \{\text{Biestow, Dierkow, Evershagen}, \ldots, \text{Südstadt, Warnemünde}\}$
  - $\mathbb{D}(\text{Studiengang}) := \{\text{Elektrotechnik, Informatik, Wirtschaftsinformatik}\}$
- **Tabelle 1.2**: MODUL = {ModNr, Name} mit
  - $\mathbb{D}(\text{ModNr}) := \{1, 2, \ldots, 16\}$
  - $\mathbb{D}(\text{Name}) := \{\text{Algorithmen und Datenstrukturen, Betriebssysteme}, \ldots, \text{Vortragsseminar}\}$
- **Tabelle 1.3**: PRUEFUNG = {MatNr, ModNr, Semester, Note} mit
  - $\mathbb{D}(\text{MatNr}) := \{10001, 10002, \ldots, 10017\}$
  - $\mathbb{D}(\text{ModNr}) := \{1, 2, \ldots, 16\}$
  - $\mathbb{D}(\text{Semester}) := \{\text{WS 19/20, SS 20, WS 20/21}\}$
  - $\mathbb{D}(\text{Note}) := \{1.0, 1.3, 1.7, 2.0, 2.3, 2.7, 3.0, 3.3, 3.7, 4.0, 5.0\}$

Natürlich entspricht diese Datenbank keiner real existierenden und ist stark abstrahiert, sowohl konzeptuell als auch inhaltlich.

1.2. Aufbau der Arbeit

Beginnen wir die Bachelorarbeit damit, uns einige theoretische Grundlagen anzueignen (siehe Kapitel 2). Dazu klären wir in Abschnitt 2.1 zunächst, was es mit dem Begriff der Provenance auf sich hat. Der Schwerpunkt liegt hierbei auf dem Thema Data Provenance, während der verwandte Themenbereich der Workflow Provenance nur kurz angeschnitten wird. Anschließend beschäftigen wir uns in Abschnitt 2.2 mit Datenschutz (Privacy) und dort insbesondere mit Anonymisierungstechniken. Weiterhin setzen wir uns in Abschnitt 2.3 mit empirischen Befragungen auseinander, insbesondere mit qualitativen Befragungen, da diese im zweiten Teil der Bachelorarbeit eine große Rolle spielen werden. Quantitative Befragungen hingegen werden nur kurz behandelt. Zu guter Letzt klären wir dann in Abschnitt 2.4, was Forschungsdaten sind und worin deren Management besteht.

In Kapitel 3 geht es um den aktuellen Stand der Forschung in den Bereichen Provenance und Privacy sowie in deren Kombination. Dabei werden wir feststellen, dass es bislang nur wenige Veröffentlichungen zu Privacy bei Data Provenance gibt.

Im vierten Kapitel untersuchen wir deshalb selbst, inwiefern sich Datenschutz (Privacy) und Data Provenance miteinander vereinbaren lassen. Dabei diskutieren wir zunächst das Prinzip der Differential Privacy, bevor wir uns mit intensionalen Provenance-Antworten und einigen weiteren Ansätzen auseinandersetzen. Im fünften Kapitel führen wir eine qualitative Befragung von Forscherinnen und Forschern durch und versuchen auf diese Weise zu ermitteln, wie in der Praxis mit Forschungsdaten umgegangen wird, welche Daten überhaupt anfallen, was allgemein unter den Begriffen „Provenance“ und „Privacy“ verstanden wird und welches Verständnis die Forscherinnen und Forscher von Datenschutz und Open Science haben.

Zu guter Letzt beinhaltet das letzte, sechste Kapitel ein Fazit zu den Erkenntnissen dieser Bachelorarbeit und liefert einen Ausblick auf offene Fragen, die es in der Zukunft zu beantworten gilt.

$ID_{Student}$	MatNr	Name	Vorname	Geburtstag	Straße	PLZ	Ort	Stadtteil	Studiengang
S_1	10001	Fieber	Fabian	08.03.1998	Albert-Einstein-Straße 26	18059	Rostock	Südstadt	Informatik
S_2	10002	Sonnenschein	Sarah	21.10.1993	Robert-Koch-Straße 3	18059	Rostock	Südstadt	Informatik
S_3	10003	Müller	Max	22.10.1994	Karl-Marx-Straße 62	18057	Rostock	Hansaviertel	Wirtschaftsinformatik
S_4	10004	Deckert	Luisa	22.10.1994	Dethardingstraße 12	18057	Rostock	Hansaviertel	Wirtschaftsinformatik
S_5	10005	Gebauer	Ulrike	01.02.2000	Thomas-Morus-Straße 1	18106	Rostock	Evershagen	Wirtschaftsinformatik
S_6	10006	Zimmermann	Jonas	23.07.1999	Maxim-Gorki-Straße 14	18106	Rostock	Evershagen	Elektrotechnik
S_7	10007	Bach	Franziska	08.03.1998	Fährstraße 63	18147	Rostock	Gehlsdorf	Elektrotechnik
S_8	10008	Kemper	Moritz	27.11.1991	Erich-Weinert-Straße 25	18059	Rostock	Südstadt	Informatik
S_9	10009	Wolter	Franziska	22.10.1994	Richard-Wagner-Straße 52	18119	Rostock	Warnemünde	Wirtschaftsinformatik
S_{10}	10010	Jansen	Jana	27.11.1991	Ährenkamp 3	18059	Rostock	Biestow	Informatik
S_{11}	10011	Wegner	Daniel	19.01.1995	Nobelstraße 31	18059	Rostock	Südstadt	Informatik
S_{12}	10012	Wegner	Laura	13.12.1992	Nobelstraße 31	18059	Rostock	Südstadt	Wirtschaftsinformatik
S_{13}	10013	Scholz	Erich	03.10.1993	Bertolt-Brecht-Straße 14	18106	Rostock	Evershagen	Elektrotechnik
S_{14}	10014	Müller	Mira	05.10.1994	Uhlenweg 2	18146	Rostock	Dierkow	Informatik
S_{15}	10015	Miller	Mia	12.09.1994	Uhlenweg 19	18146	Rostock	Dierkow	Informatik
S_{16}	10016	Freiberg	Damian	01.02.2000	Ehm-Welk-Straße 7	18106	Rostock	Evershagen	Informatik
S_{17}	10017	Schmidt	Hans	26.11.1991	Parkstraße 64	18119	Rostock	Warnemünde	Informatik

Tabelle 1.1.: Die STUDENT-Relation

$\mathbf{ID_{Modul}}$	**ModNr**	**Name**
M_1	1	Vortragsseminar
M_2	2	Logik und Berechenbarkeit
M_3	3	Imperative Programmierung
M_4	4	Rechnernetze und Datensicherheit
M_5	5	Theoretische Informatik
M_6	6	Algorithmen und Datenstrukturen
M_7	7	Digitale Systeme
M_8	8	Datenbanken
M_9	9	Softwaretechnik
M_{10}	10	Modellierung und Simulation
M_{11}	11	Computergrafik
M_{12}	12	Smart Computing
M_{13}	13	Betriebssysteme
M_{14}	14	Verteilte Systeme
M_{15}	15	Mathematik für Informatik
M_{16}	16	Mathematik für Elektrotechnik

Tabelle 1.2.: Die MODUL-Relation

$\mathbf{ID_{Pruefung}}$	**MatNr**	**ModNr**	**Semester**	**Note**
P_1	10001	15	WS 19/20	1.3
P_2	10002	15	WS 19/20	2.0
P_3	10003	15	WS 19/20	2.7
P_4	10004	15	WS 19/20	1.3
P_5	10005	15	WS 19/20	1.7
P_6	10008	15	WS 19/20	4.0
P_7	10009	15	WS 19/20	3.3
P_8	10010	15	WS 19/20	5.0
P_9	10011	15	WS 19/20	5.0
P_{10}	10012	15	WS 19/20	3.0
P_{11}	10014	15	WS 19/20	1.3
P_{12}	10015	15	WS 19/20	2.3
P_{13}	10016	15	WS 19/20	1.0
P_{14}	10017	15	WS 19/20	3.3
P_{15}	10002	8	SS 20	1.3
P_{16}	10003	8	SS 20	1.3
P_{17}	10005	6	SS 20	1.0
P_{18}	10007	8	SS 20	3.0
P_{19}	10008	7	SS 20	3.7
P_{20}	10011	11	SS 20	1.7
P_{21}	10014	8	SS 20	1.0
P_{22}	10016	3	SS 20	1.0
P_{23}	10003	8	WS 20/21	5.0
P_{24}	10002	5	SS 20	3.3
P_{25}	10002	2	SS 20	2.0
P_{26}	10002	4	SS 20	1.0
P_{27}	10003	7	SS 20	1.7
P_{28}	10005	13	SS 20	2.0
P_{29}	10005	1	SS 20	3.0

Tabelle 1.3.: Die PRUEFUNG-Relation

2. Grundlagen

In diesem Kapitel beschäftigen wir uns mit den fundamentalen Grundlagen dieser Bachelorarbeit. Dazu werden wir in Abschnitt 2.1 zunächst klären, was Provenance allgemein bedeutet, warum es sinnvoll ist, sich damit zu beschäftigen und welche Arten und Methoden es dabei gibt. Ein Schwerpunkt wird dabei die Data Provenance darstellen, während das Thema Workflow Provenance nur kurz angeschnitten wird. Weiterhin beschäftigen wir uns in Abschnitt 2.2 mit Privacy und deren Varianten. Dazu gehören vor allem diverse Anonymitätsmaße wie die k-Anonymität und deren Weiterentwicklung, die l-Diversität. Als drittes beschäftigen wir uns in Abschnitt 2.3 dann mit Empirie, konkret mit empirischen Umfragen und Interviews. Hierbei wird erläutert, wieso Interviews für unser Vorhaben sinnvoller sind als Fragebögen und welche allgemeinen Anforderungen, Kriterien und Richtlinien an ein gutes empirisches Interview bestehen. Zu guter Letzt beschäftigen wir uns in dieser Bachelorarbeit mit dem Thema Forschungsdaten und ihrem Management und klären dabei, was Forschungsdaten überhaupt sind, warum es sinnvoll ist, diese zu verwalten und langfristig zu archivieren sowie mit den wichtigen Herausforderungen, welche hierbei gemeistert werden müssen (siehe Abschnitt 2.4).

2.1. Provenance

Provenance, zu Deutsch „Herkunft“ oder „Ursprung“, beschäftigt sich im wissenschaftlichen Kontext mit der Herkunft bzw. dem Zustandekommen eines Ergebnisses einer wissenschaftlichen Auswertung. Damit kann beispielsweise das Resultat einer Aggregation über einen Datensatz gemeint sein, oder aber das Ergebnis einer Selektionsanfrage. Zu unterscheiden ist dabei vor allem zwischen Workflow und Data Provenance: Ersteres hat den Anspruch, den Arbeitsablauf zurückzuverfolgen, der für ein bestimmtes Ergebnis gesorgt hat, während letzteres ermitteln möchte, welche Daten zu einem Ergebnis beigetragen haben [HDB17].

2.1.1. Data Provenance

Das Thema der **Data Provenance** gewann vor allem in den letzten zwei Jahrzehnten an Bedeutung. Beispielsweise veröffentlichten *Cheney et al.* 2009 den Artikel *„Provenance in Databases: Why, How and Where“* [CCT09], welcher die drei Anfragearten auf über 90 Seiten ausführlichst erläutert. **Big Data**, also das Verwalten riesiger Datenmengen, die mit konventionellen Methoden nicht mehr zu bewältigen sind, rückt angesichts vernetzter Geräte – dem **Internet of Things** – und Interesse am **Cloud Computing** immer weiter in den Vordergrund. Daraus entsteht ein Interesse, Daten, die langfristig archiviert werden müssen, so gering wie möglich zu halten, um Ressourcen wie etwa Speicherplatz zu sparen. Gleichzeitig soll die Herkunft bzw. der Ursprung von Ergebnisdaten weiterhin nachvollziehbar sein. Genau dies ist die Aufgabe von Data Provenance – das Minimieren von (Forschungs-)Daten, ohne dabei die Möglichkeit einer (teilweisen) Rekonstruktion der Originaldaten zu verlieren.

In der Datenbanktheorie kann Data Provenance wie folgt verstanden werden: Führen wir eine Anfrage Q (engl. *query*) auf eine Datenbank D aus, erhalten wir eine Ergebnisrelation $R = Q(D)$. Allerdings wissen

wir ohne zusätzliche Informationen nichts über den Ursprung dieses Ergebnisses. Ziel der Data Provenance besteht daher darin, die folgenden vier Fragen zu beantworten:

1. Woher kommen die Daten? (**where**)
2. Wieso kam dieses Ergebnis zustande? (**why**)
3. Wie genau wurde das Ergebnis berechnet? (**how**)
4. Warum ist ein konkreter Datensatz nicht im Ergebnis enthalten? (**why not**)

Entsprechend dieser Fragen unterscheidet man bei der Data Provenance vier Anfragearten: Die **where**-, **why**-, **how**- und **why-not**-Provenance. Zusätzlich dazu existieren vier verschiedene Antworttypen auf diese Fragen (vgl. [Aug17]):

Antworttyp	**Ergebnis**	**Antwort auf**
extensional	Beteiligte Tupel der Originaldaten	**where · why**
intensional	Beschreibung der Daten	**where · why · how**
anfragebasiert	Selektionsprädikate	**why · how**
modifikationsbasiert	Minimaler Änderungsvorschlag der Auswertung	**why not**

Tabelle 2.1.: Antworttypen auf Provenance-Anfragen und deren Anwendungsbereich

Zunächst werden alle Tupel jeder Relation mit einer ID versehen – unabhängig von ihren Schlüsseln und sonstigen Attributen. Diese ID dient der Identifikation eines jeden Tupels. Wir wählen an dieser Stelle den jeweils ersten Buchstaben einer Relation und eine fortlaufende natürliche Zahl im Index. Der dritte Eintrag in der Relation `STUDENT` bekommt so beispielsweise die ID S_3, wie es auch in den Beispieltabellen zu sehen ist.

where-Provenance Die erste Frage – Woher kommen die Daten? – kann mit der sogenannten **where**-Provenance beantwortet werden [AH19]. Dies kann sowohl tupel- als auch relationenorientiert erfolgen. Wir betrachten dazu folgende Anfrage Q, welche uns beantwortet, welche Prüfungen (Modulbezeichnung) im Sommersemester 2020 von welcher Person geschrieben wurden, sowie deren Ergebnisrelation R:

```
SELECT DISTINCT STUDENT.Name, STUDENT.Vorname, MODUL.Name as Modulbezeichnung
FROM STUDENT
    JOIN PRUEFUNG ON (STUDENT.MatNr = PRUEFUNG.MatNr)
    JOIN MODUL ON (PRUEFUNG.ModNr = MODUL.ModNr)
WHERE Semester = 'SS 20' OR Semester = 'WS 20/21'
ORDER BY Modulbezeichnung
```

Anfrage 2.1: Wer schrieb im SS 2020 oder WS 2020/21 welche Prüfung?

Die Frage nach dem **where** kann nun auf zwei Arten beantwortet werden: Relationenorientiert lautet die Antwort {`STUDENT`, `MODUL`}, da die Daten sowohl aus der Relation `STUDENT` (die Attribute `Name` und `Vorname`) als auch aus der Relation `MODUL` (das Attribut `Modulbezeichnung`) stammen, wie man der Anfrage in diesem Fall auch direkt entnehmen kann. Bei der tupelorientierten Antwort ist es nicht ganz so einfach. Wir sehen, dass das erste Ergebnistupel aus den Tupeln S_5 und M_6 entstammt, das zweite aus S_{11} und M_{11} und so weiter. Die Antwort lautet für das erste Tupel also $\{S_5, M_6\}$, für das zweite dementsprechend $\{S_{11}, M_{11}\}$ und so weiter. Hierbei handelt es sich um sogenannte **Zeugenlisten**; die einzelnen für das Ergebnis relevanten Tupel werden **Zeugen** genannt. S_5 und M_6 „bezeugen“ also exemplarisch die Daten des Tupels R_1.

	Name	Vorname	Modulbezeichnung	
R_1	Gebauer	Ulrike	Algorithmen und Datenstrukturen	$S_5 \cdot M_6 \cdot P_{17}$
R_2	Wegner	Daniel	Computergrafik	$S_{11} \cdot M_{11} \cdot P_{20}$
R_3	Sonnenschein	Sarah	Datenbanken	$S_2 \cdot M_8 \cdot P_{15}$
R_4	Müller	Max	Datenbanken	$S_3 \cdot M_8 \cdot (P_{16} + P_{23})$
R_5	Bach	Franziska	Datenbanken	$S_7 \cdot M_8 \cdot P_{18}$
R_6	Müller	Mira	Datenbanken	$S_{14} \cdot M_8 \cdot P_{21}$
R_7	Kemper	Moritz	Digitale Systeme	$S_8 \cdot M_7 \cdot P_{19}$
R_8	Freiberg	Damian	Imperative Programmierung	$S_{16} \cdot M_3 \cdot P_{22}$

Tabelle 2.2.: Ergebnis der Anfrage „Wer schrieb im SS 2020 oder WS 2020/21 Prüfungen in welchem Modul?“ inklusive Informationen für **where**- und **why**-Provenance

why-Provenance Die **why**-Provenance geht einen Schritt weiter und möchte nicht nur wissen, woher die Daten stammen, sondern wieso dieses Ergebnis zustande kam. In unserem Beispiel stellt sich also zum Beispiel die Frage, wieso Max Müller die Modulbezeichnung „Datenbanken“ zugeordnet wurde.

Wie aus der Anfrage ersichtlich wird, erfolgt ein Verbund der Relationen `STUDENT` und `PRUEFUNG` über das gemeinsame Attribut `MatNr` (Matrikelnummer) sowie ein Verbund mit `MODUL` über das Attribut `ModNr` (Modulnummer). Max Müller wurde also dem Modulnamen „Datenbanken“ zugeordnet, weil der Student mit der `MatNr` 10003 (Max Müller) zwei Prüfungen in dem Modul mit der `ModNr` 8 (Datenbanken) geschrieben hat. Diese Informationen wurden mittels der `PRUEFUNG`-Relation miteinander verknüpft, konkret mithilfe der Tupel P_{16} und P_{23}. Zeile R_4 der Ergebnisrelation kam also aufgrund der Tupel S_3, M_8 und P_{16} sowie aufgrund von S_3, M_8 und P_{23} zustande. $\{S_3, M_8, P_{16}\}$ und $\{S_3, M_8, P_{23}\}$ bilden hierbei die beiden **Zeugenmengen** (siehe [CCT09]); die Menge aller Zeugenmengen, hier also $\{\{S_3, M_8, P_{16}\}, \{S_3, M_8, P_{23}\}\}$, bildet die **Zeugenbasis** und dient der Beantwortung der Frage nach dem **why**. Analog dazu existiert zum Beispiel zu Zeile R_1 die Zeugenmenge $\{S_5, M_6, P_{17}\}$ und die Zeugenbasis $\{\{S_5, M_6, P_{17}\}\}$. Idealerweise sollte diese minimal sein, um den Speicherplatzbedarf der Provenance-Daten zu reduzieren. Eine Zeugenbasis einer Anfrage Q ist dabei **minimal**, wenn keine ihrer Teilmengen bereits als Zeugenbasis für Q genügt.

Gibt man die Zeugenbasis einer Relation an, so ergibt sich die Menge aller Zeugenbasen der einzelnen Tupel und somit eine Menge von Mengen von Mengen. In unserem Beispiel sieht diese wie folgt aus:

$$\begin{aligned} &\{\{\{S_5, M_6, P_{17}\}\}_{R_1}, \{\{S_{11}, M_{11}, P_{20}\}\}_{R_2}, \{\{S_2, M_8, P_{15}\}\}_{R_3}, \\ &\{\{S_3, M_8, P_{16}\}, \{S_3, M_8, P_{23}\}\}_{R_4}, \\ &\{\{S_7, M_8, P_{18}\}\}_{R_5}, \{\{S_{14}, M_8, P_{21}\}\}_{R_6}, \{\{S_8, M_7, P_{19}\}\}_{R_7}, \{\{S_{16}, M_3, P_{22}\}\}\}_{R_8}. \end{aligned}$$

Die Indizes verweisen dabei jeweils auf das Tupel, welches die Grundlage der jeweiligen Zeugenbasis ist. $\{\{S_5, M_6, P_{17}\}\}_{R_1}$ stellt also beispielsweise die Zeugenbasis des Tupels R_1 dar.

how-Provenance Möchte man nun nicht nur wissen, wieso ein Ergebnis zustande kam, sondern auch, wie dieses berechnet wurde, kann die **how**-Provenance zu Rate gezogen werden. Wir betrachten dazu folgendes, neues Beispiel, bei welchem wir die Durchschnittsnote aller Wirtschaftsinformatik-Studentinnen und -Studenten wissen wollen, welche im Wintersemester 2019/2020 die Klausur „Mathematik für Informatik“ geschrieben haben:

```
SELECT AVG(Note) AS Durchschnitt
FROM STUDENT NATURAL JOIN PRUEFUNG
WHERE ModNr = 15
AND Semester = 'WS 19/20'
AND Studiengang = 'Wirtschaftsinformatik'
```

Anfrage 2.2: Die Durchschnittsnote aller Studentinnen und Studenten der Wirtschaftsinformatik, welche im WS 19/20 die Mathematik-Prüfung absolvierten

Name	Vorname	Note	
Müller	Max	2.7	$S_3 \cdot P_3$
Deckert	Luisa	1.3	$S_4 \cdot P_4$
Gebauer	Ulrike	1.7	$S_5 \cdot P_5$
Wolter	Franziska	3.3	$S_9 \cdot P_7$
Wegner	Laura	3.0	$S_{12} \cdot P_{10}$

Durchschnitt
2.4

Die Zeugenliste ist in diesem Fall $\{S_3, S_4, S_5, S_9, S_{12}, P_3, P_4, P_5, P_7, P_{10}\}$. Dies sind all jene Tupel, welche notwendig sind, um obiges Ergebnis zu berechnen (S_i liefert dabei die Personen und P_j die Noten). Die Zeugenbasis der gesamten Relation lautet

$$\{\{\{S_3, P_3\}\}, \{\{S_4, P_4\}\}, \{\{S_5, P_5\}\}, \{\{S_9, P_7\}\}, \{\{S_{12}, P_{10}\}\}\}$$

und besteht aus allen Zeugenbasen der Ergebnistupel, wobei jede Zeugenbasis wiederum aus Tupeln der `STUDENT`- und einem Tupel der `PRUEFUNG`-Relation besteht. Dabei weisen beide Tupel jeweils die gleiche `MatNr` auf. Die exakte Berechnungsvorschrift des Ergebnisses wird nun in einem **Provenance-Polynom** festgehalten.

Definition 2.1. Provenance-Polynome sind Polynome, die auf dem kommutativen Halbring $\mathbb{N}[X] = (\mathbb{N}[X], +, \cdot, 0, 1)$ definiert sind, wobei X eine Menge von Tupel-Identifikatoren einer Datenbankinstanz I ist [Aug17] [GT17]. □

Der Begriff des Halbrings wird an dieser Stelle als bekannt vorausgesetzt und nicht weiter erläutert, es sei aber auf [GT17] verwiesen. Der „+“-Operator wird genutzt, um Duplikate zu eliminieren – in der relationalen Datenbanktheorie für Projektionen und Vereinigungen – und der „·“-Operator, um Tupel miteinander zu verbinden, also für Verbundoperationen (Joins), Selektionen und Schnittmengenbildungen (Intersections). Mithilfe des Tensorprodukts $\odot$ und der direkten Summe $\oplus$ lassen sich diese Polynome dann mit einem konkreten Wert ($\odot$) oder aber miteinander ($\oplus$) verbinden.

Unser Provenance-Polynom besteht konkret aus dem Quotienten von Summe (`SUM(Note)`) und Anzahl (`COUNT(Note)`) aller Noten, da für den Durchschnitt von Merkmalsausprägungen $x_1, x_2, \ldots, x_n$ allgemein $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ gilt. Möchte man die Summe aller Noten berechnen, also $2.7 + 1.3 + 1.7 + 3.3 + 3.0 = 12$, so lautet das dazugehörige Provenance-Polynom

$$([S_3 \cdot P_3] \odot 2.7) \oplus ([S_4 \cdot P_4] \odot 1.3) \oplus ([S_5 \cdot P_5] \odot 1.7) \oplus ([S_9 \cdot P_7] \odot 3.3) \oplus ([S_{12} \cdot P_{10}] \odot 3.0),$$

wobei die Klammern ausschließlich der Übersichtlichkeit dienen. Das Polynom für die Berechnung der Anzahl, also $|\{2.7, 1.3, 1.7, 3.3, 3.0\}| = 5$, lautet entsprechend

$$(S_3 \cdot P_3) \oplus (S_4 \cdot P_4) \oplus (S_5 \cdot P_5) \oplus (S_9 \cdot P_7) \oplus (S_{12} \cdot P_{10}).$$

Somit ergibt sich für das vollständige Provenance-Polynom mittels $\frac{\texttt{SUM(Note)}}{\texttt{COUNT(Note)}}$:

$$\frac{(S_3 \cdot P_3 \odot 2.7) \oplus (S_4 \cdot P_4 \odot 1.3) \oplus (S_5 \cdot P_5 \odot 1.7) \oplus (S_9 \cdot P_7 \odot 3.3) \oplus (S_{12} \cdot P_{10} \odot 3.0)}{(S_3 \cdot P_3) \oplus (S_4 \cdot P_4) \oplus (S_5 \cdot P_5) \oplus (S_9 \cdot P_7) \oplus (S_{12} \cdot P_{10})}.$$

Aus dem Provenance-Polynom kann nun direkt die Zeugenbasis abgelesen werden, da alle Zeugenmengen direkt ersichtlich sind; beispielsweise enthält das Polynom $S_3 \cdot P_3 \odot 2.7$ die Zeugenmenge $\{S_3, P_3\}$. Die **why**-Provenance ergibt sich also aus der **how**-Provenance; es gilt: **why** $\preceq$ **how**. Ferner kann aus der **why**-Provenance auf die (tupelorientierte) **where**-Provenance geschlossen werden (und von der tupelorientierten auf die relationenorientierte), indem die Zeugenbasis, welche formal eine Menge von Mengen ist, zu einer einfachen Menge reduziert wird – der Zeugenliste. Es gilt also ebenfalls **where** $\preceq$ **why** und somit insgesamt **where** $\preceq$ **why** $\preceq$ **how**.

Konkret wissen wir nun also, dass wir aus der `STUDENT`-Relation die Tupel mit den IDs S_3, S_4, S_5, S_9 und S_{12} und aus der `PRUEFUNG`-Relation die Tupel mit den IDs P_3, P_4, P_5, P_7 und P_{10} aufbewahren müssen, um das Ergebnis rekonstruieren zu können. Was für unser Beispiel trivial erscheint, kann für Berechnungen von Ergebnissen eines Datensatzes mit mehreren Tausend Zeilen eine essentielle Erkenntnis sein, um den Speicherplatzbedarf drastisch zu minimieren.

2.1.2. Workflow Provenance

Workflow Provenance und Data Provenance haben einiges gemeinsam: Beispielsweise beschäftigen sich beide mit der Frage nach dem Zustandekommen gewisser Forschungsergebnisse. Während Data Provenance dabei vor allem die Herkunft der Daten berücksichtigt, versucht die Workflow Provenance, ein Ergebnis mittels des Arbeitsablaufes zu erklären und dabei exemplarisch die folgenden Fragen zu beantworten [DF08]:

- Wer hat die Daten generiert und wann?
- Wann wurden die Daten von wem modifiziert?
- Auf welche Weise und mit welchem Prozess wurden die Daten generiert?
- Wurden zwei Ergebnisdaten aus denselben Rohdaten produziert?

Dabei werden Modelle und Notationen verwendet, die in der Informatik allgemein bekannt sind, wie beispielsweise Datenflussdiagramme, welche Teil der *Unified Modeling Language (UML)* sind. Eine weitere Methode ist die Modellierung in Form von gerichteten, azyklischen Graphen. Hierbei stellen die Knoten die jeweiligen Module bzw. Prozesse (Software, Arbeitsschritte, ...) und die Kanten den Datenfluss zwischen den jeweiligen Prozessen dar [DKR+11]. Die Module können dabei mit Namen, Beschreibungen und Stichworten versehen werden, um sie genauer zu erklären. Ein Beispiel dafür stellt die Abbildung 2.1 dar, in der der Prozess einer Datenvisualisierung mittels Workflow Provenance festgehalten wurde.

Trotz einiger Gemeinsamkeiten mit der Data Provenance ist die Workflow Provenance für diese Bachelorarbeit nicht von Relevanz und wird daher nicht näher behandelt. Die interessierte Leserin/der interessierte Leser sei aber auf [HDB17] verwiesen.

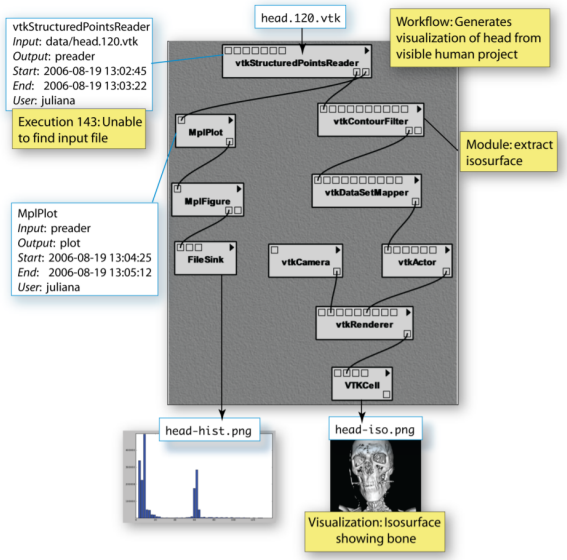


Abbildung 2.1.: Ein Beispiel für Workflow Provenance [DF08]

2.2. Privacy

Im folgenden Abschnitt bezeichne $\cdot_b$ *(engl.* bag*) eine Multimenge. Multimengen sind dabei Mengen, in denen Duplikate zugelassen sind, das heißt, dass jedes Element auch mehrfach vorkommen kann. Insbesondere in Kombination mit dem* π*-Operator der relationalen Algebra bedeutet dies, dass* $\pi.(\cdot)_b$ *einer Multimenge und somit dem Mengen-Standard in* SQL *entspricht, während* $\pi.(\cdot)$ *wie gewohnt eine einfache Menge bezeichnet.*

Privacy, zu Deutsch Datenschutz, bezeichnet den Schutz personenbezogener Daten vor unzulässiger Erhebung, Speicherung und Veröffentlichung. Insbesondere soll die Privatsphäre einzelner Personen geschützt werden. Datenschutz ist allerdings keinesfalls so trivial, wie es zunächst scheint und wie folgendes Beispiel von *Latanya Sweeney* anschaulich aufzeigt [Swe02b]:

Die *Group Insurance Commission (GIC)* veröffentlichte[1] im US-Bundesstaat Massachusetts einen Datensatz von 135 000 Patientinnen und Patienten, welcher unter anderem die Ethnie, die Medikation, die Postleitzahl, das Geburtsdatum, das Geschlecht und die diagnostizierte Krankheit dieser Menschen enthielt. Da die Sozialversicherungsnummern und Namen, welche ebenfalls erhoben worden, nicht in diesem Datensatz enthalten waren, hielt die *GIC* ihn für anonym. Gleichzeitig ist es in Massachusetts allerdings möglich, vor einer Wahl eine Kopie des Wählerverzeichnisses zu erwerben, welches unter anderem den Namen der Person, deren Adresse und Parteizugehörigkeit, aber eben auch die Postleitzahl, das Geburtsdatum und das Geschlecht beinhaltet. Aufgrund dieser Schnittmenge war es möglich, dem Gouverneur seine Diagnose zuzuordnen – im gesamten Datensatz gab es nur drei Männer mit seinem Geburtstag und davon nur einen mit seiner Postleitzahl.

Das bloße Entfernen eindeutig personenbezogener Attribute wie identifizierende Nummern oder den Namen der Person reicht also noch nicht aus, um die Anonymität eines Datensatzes zu gewährleisten. Unter dem Heranziehen externen Wissens, beispielsweise anderer Datensätze, ist es eventuell möglich, einzelne Personen dennoch zu identifizieren. Im weiteren Verlauf dieses Abschnitts werden wir uns deshalb ansehen, wie man Datenschutz bei Datensätzen mit Personenbezug dennoch verwirklichen kann.

[1] Der Datensatz wurde Forscherinnen und Forschern frei zur Verfügung gestellt und Unternehmen zum Kauf angeboten, war aber nicht öffentlich zugänglich.

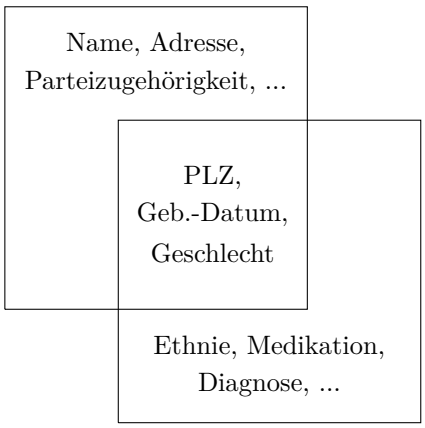


Abbildung 2.2.: Die Schnittmenge der beiden Datensätze lässt einen Rückschluss auf (mindestens) eine konkrete Person zu (vgl. [Swe02b])

Zunächst wird dafür geklärt, wann eine Zeile in einer Tabelle (ein Tupel) überhaupt als identifizierbar gilt. Anschließend werden diverse Maße eingeführt, welche ein Messen des Datenschutzes ermöglichen.

2.2.1. Identifikatoren und Quasi-Identifikatoren

Aus der Theorie relationaler Datenbanken ist bekannt, dass jedes Tupel einer Relation durch einen Schlüssel (**Identifikator**) eindeutig identifizierbar sein muss. *Heuer*, *Saake* und *Sattler* definieren dabei zunächst den Begriff der identifizierenden Attributmenge [HSS18]:

Definition 2.2. Eine **identifizierende Attributmenge** für ein Relationenschema R ist eine Menge $K := \{B_1, \ldots, B_k\} \subseteq R$, so dass für jede Relation $r(R)$ gilt:

$$\forall t_1, t_2 \in r : t_1 \neq t_2 \Rightarrow \exists B \in K : t_1(B) \neq t_2(B)$$

□

Es existiert also eine Kombination aus Attributen, die für jedes Tupel eindeutig ist, so dass man von der Attributmenge auf das Tupel schließen kann. Ist K bezüglich $\subseteq$ zusätzlich minimal, bezeichnet man K als **Schlüssel** einer Relation. Schwächt man dieses Prinzip nun so weit ab, dass eine Kombination von Attributen genügt, um mithilfe von externem Wissen ein Tupel eindeutig identifizieren zu können, auch wenn es in der Relation zunächst nicht eindeutig identifizierbar ist, spricht man von einem **Quasi-Identifikator**.

Definition 2.3. Ein **Quasi-Identifikator** ist eine Attributkombination, die in Verbindung mit extern verfügbaren Informationen die eindeutige Identifikation einer Person ermöglicht [PS17]. □

Der Begriff des Quasi-Identifikators ist allerdings bewusst nicht eindeutig definiert, weshalb es auch verschiedene Möglichkeiten gibt, diesen zu ermitteln. Als Beispiel sei das Verhältnis zwischen der Anzahl verschiedener Tupel und der Anzahl aller Tupel eines Datensatzes genannt (**Distinct Ratio**) [Gru19].

MatNr	Name	Vorname	Geburtstag	PLZ	Stadtteil	Note
10010	Jansen	Jana	27.11.1991	18059	Biestow	5.0
10014	Müller	Mira	05.10.1994	18146	Dierkow	1.3
10015	Miller	Mia	12.09.1994	18146	Dierkow	2.3
10005	Gebauer	Ulrike	01.02.2000	18106	Evershagen	1.7
10016	Freiberg	Damian	01.02.2000	18106	Evershagen	1.0
10003	Müller	Max	22.10.1994	18057	Hansaviertel	2.7
10004	Deckert	Luisa	22.10.1994	18057	Hansaviertel	1.3
10001	Fieber	Fabian	08.03.1998	18059	Südstadt	1.3
10002	Sonnenschein	Sarah	21.10.1993	18059	Südstadt	2.0
10008	Kemper	Moritz	27.11.1991	18059	Südstadt	4.0
10011	Wegner	Daniel	19.01.1995	18059	Südstadt	5.0
10012	Wegner	Laura	13.12.1992	18059	Südstadt	3.0
10009	Wolter	Franziska	22.10.1994	18119	Warnemünde	3.3
10017	Schmidt	Hans	26.11.1991	18119	Warnemünde	3.3

Tabelle 2.3.: Das Ergebnis der Anfrage. Gelb hinterlegte Attributkombinationen sind hierbei uneindeutig und von links nach rechts zu interpretieren.

Im Folgenden sei R eine Relation, A eine Menge von Attributen dieser Relation und π der Projektionsoperator in der relationalen Algebra, welche an dieser Stelle als bekannt vorausgesetzt wird[2]. Der Quotient q wird dann wie folgt berechnet:

$$q = \frac{|\pi_A(R)|}{|R_b|}.$$

Übersteigt $q \in (0, 1]$ einen gewissen, zuvor festgelegten Schwellwert, so dienen die Attribute, auf die projiziert wird (die Menge A), als Quasi-Identifikator. Je höher dieser Schwellwert ist, desto mehr Tupel können durch A eindeutig identifiziert werden. Ist q unwesentlich größer als 0, so ist nahezu keinerlei Identifikation möglich; ist $q = 1$, so ist A eine identifizierende Attributmenge für R.

Beispiel 2.1. Wir betrachten den in Tabelle 2.3 dargestellten Datensatz. Die gelben Markierungen in der Tabelle 2.3 markieren hierbei Attributkombinationen, die **nicht** eindeutig sind und sind von links nach rechts zu interpretieren. So kann `Jana Jansen` beispielsweise nicht allein durch ihren `Geburtstag` (Der Wert „`27.11.1991`“ kommt zwei Mal vor) und auch nicht durch die Kombination aus `Geburtstag` und `PLZ` identifiziert werden (auch `(„27.11.1991", „18059")` kommt noch zwei Mal vor), wohl aber durch die Kombination aus `Geburtstag`, `PLZ` und `Stadtteil` (denn `(„27.11.1991", „18059", „Biestow")` existiert in dieser Kombination genau ein einziges Mal). Wir legen weiterhin fest, dass eine Attributmenge genau dann als Quasi-Identifikator dient, wenn $q \geq 0.85$ ist. In Tabelle 2.3 ist jede Person eindeutig über die Matrikelnummer sowie über ihren Vor- und Nachnamen identifizierbar. Doch auch, wenn man diese Attribute entfernt, lassen sich bereits nur durch das Attribut „Geburtstag“ die Hälfte der Personen eindeutig identifizieren ($q = {}^{10}/_{14} \approx 71.4\%$). Nimmt man zusätzlich das Attribut „PLZ“ hinzu, ergibt sich für $q = {}^{11}/_{14} \approx 78.6\%$ und 8 der 14 Tupel sind eindeutig identifizierbar. Inklusive „Stadtteil“ sind es dann sogar 10 von 14, wobei $q = {}^{12}/_{14} \approx 85.7\%$. Die Kombination aus Geburtstag, Postleitzahl und Stadtteil stellt also einen Quasi-Identifikator mit $q \approx 0.86$ dar.

2.2.2. Anonymitätsmaße

Wie im vorherigen Abschnitt aufgezeigt, sind weitere Maßnahmen erforderlich, um einen Datensatz tatsächlich zu anonymisieren. Dieser Abschnitt befasst sich deshalb mit zwei Anonymitätsmaßen, die sich im Privacy-Bereich etabliert haben: die k-Anonymität als grundlegendes Maß und die l-Diversität als Erweiterung dessen.

[2] Als Literatur sei auf das Buch *„Datenbanken: Konzepte und Sprachen“* von *Heuer et al.* verwiesen [HSS18].

MatNr	Name	Vorname	Geburtstag	PLZ	Stadtteil	Note
*	*	*	1990 – 1994	18146	Dierkow	1.3
*	*	*	1990 – 1994	18146	Dierkow	2.3
*	*	*	2000 – 2004	18106	Evershagen	1.7
*	*	*	2000 – 2004	18106	Evershagen	1.0
*	*	*	1990 – 1994	18057	Hansaviertel	2.7
*	*	*	1990 – 1994	18057	Hansaviertel	1.3
*	*	*	1990 – 1994	18059	Südstadt	2.0
*	*	*	1990 – 1994	18059	Südstadt	4.0
*	*	*	1990 – 1994	18059	Südstadt	3.0
*	*	*	1995 – 1999	18059	Südstadt	1.3
*	*	*	1995 – 1999	18059	Südstadt	5.0
*	*	*	1990 – 1994	18119	Warnemünde	3.3
*	*	*	1990 – 1994	18119	Warnemünde	3.3

Tabelle 2.4.: Unser k-anonymer Datensatz mit $k = 2$

k-Anonymität Das Prinzip der **k-Anonymität** wurde 2001 erstmals von *Pierangela Samarati* vorgestellt [Sam01]. Es besagt, dass mindestens k Tupel eines Datensatzes hinsichtlich des Quasi-Identifikators jeweils identisch sein müssen. Ist diese Bedingung erfüllt, kann man die Tupel einer Relation R in Äquivalenzklassen, sogenannte q^*-Blöcke, unterteilen. Da es zur k-Anonymität bislang keine formale Definition gibt, welche die relationale Algebra nutzt, definieren wir sie an dieser Stelle wie folgt:

Definition 2.4. Sei q_i^* $(i = 1, 2, \dots)$ die i-te Äquivalenzklasse einer Relation r und A eine Attributmenge, die als Quasi-Identifikator dient. k-Anonymität ist genau dann erfüllt, wenn gilt:

$$\forall q_i^* \in R : |\pi_A(q_i^*)| = 1 \wedge |(q_i^*)_b| \geq k.$$

□

Vereinfacht gesagt müssen also zu jedem Tupel mindestens $k-1$ weitere Tupel existieren, die den gleichen Quasi-Identifikator aufweisen (vgl. [PS17]). Weiterhin erfüllt eine Relation k-Anonymität genau dann, wenn jeder q^*-Block der Relation mindestens k-Anonymität erfüllt.

k-Anonymität kann erreicht werden, indem zusätzliche Tupel (Rauschen) hinzugefügt, mehrere Attributwerte zu einem verdichtet (generalisiert) oder einzelne abweichende Tupel entfernt (unterdrückt) werden. Allerdings beeinflussen alle drei Methoden den Informationsgehalt des Datensatzes, was oft ein negativer Nebeneffekt ist – der Datenschutz sollte also so gewährleistet werden, dass möglichst viele Informationen erhalten bleiben. Es wird jedoch angenommen, dass dies allgemein ein NP-schweres Problem darstellt, wie *Petrlic* und *Sorge* in ihrem Buch erwähnen und *Meyerson* und *Williams* 2004 in ihrem Paper *„On the Complexity of Optimal K-Anonymity“* für die Generalisierung und Unterdrückung von Tupeln mit konstantem k bewiesen [PS17] [MW04].

Beispiel 2.2. Unser Beispiel in Tabelle 2.3 weist nach dem Entfernen aller identifizierenden Attribute momentan nur eine 1-Anonymität auf, da beispielsweise nur ein einziges Tupel mit `(27.11.1991, 18059, Biestow)` als Quasi-Identifikator existiert. Um zumindest 2-Anonymität herzustellen, bietet es sich zunächst an, das einzelne Tupel mit `Stadtteil = Biestow` zu unterdrücken. Anschließend können die Geburtstage generalisiert werden: anstelle des exakten Tages wäre eine Unterteilung in 5-Jahres-Intervalle sinnvoll. Der neue, 2-anonyme Datensatz ist in Tabelle 2.4 zu sehen.

l-Diversität Wenn wir uns das Beispiel allerdings genauer ansehen, fällt auf, dass alle Tupel des untersten q^*-Blocks auch 3.3 als Note besitzen und somit denselben sensiblen Wert aufweisen (gelb hinterlegt). Eine eindeutige Zuordnung eines Tupels zu einer Studentin oder einem Studenten ist somit dank k-Anonymität zwar nach wie vor nicht möglich, allerdings auch gar nicht notwendig: Egal, welche der beiden Zeilen die richtige ist, die Note ist in jedem Fall 3.3; der Person kann also zweifelsfrei ihr sensibles Attribut zugeordnet werden. Dieses Beispiel stellt einen sogenannten *Homogenitätsangriff* auf die k-Anonymität dar [PS17]. Um ebendiesen zu vermeiden, gibt es mit der **l-Diversität** ein weiteres Anonymitätsmaß, welches die k-Anonymität erweitert: Neben der Eigenschaft, dass jede Äquivalenzklasse mindestens k identische Werte im Quasi-Identifikator besitzen muss, muss das sensible Attribut innerhalb dieser Äquivalenzklasse nun auch mindestens l verschiedene Werte aufweisen. Formal bedeutet dies:

Definition 2.5. Sei q_i^* $(i = 1, 2, \dots)$ die i-te Äquivalenzklasse einer Relation r, A eine Attributmenge, die als Quasi-Identifikator dient und S ein sensibles Attribut (formal eine Attributmenge mit $|S| = 1$). l-Diversität ist genau dann erfüllt, wenn gilt:

$$\forall q_i^* \in r : |\pi_A(q_i^*)| = 1 \land |(q_i^*)_b| \geq k \land |\pi_S(R)| \geq l.$$

□

Beispiel 2.3. An dieser Stelle haben wir im Wesentlichen nur zwei Optionen: Wir können den `(1990-1994, 18119, Warnemünde)`-q^*-Block komplett entfernen. Allerdings bedeutet dies natürlich eine Reduktion des Datensatzes und somit einen erheblichen Informationsverlust. In der Realität sind Datensätze jedoch wesentlich umfangreicher, weshalb das Entfernen eines q^*-Blockes dort weniger ins Gewicht fällt. Alternativ können wir den Datensatz aber auch noch stärker generalisieren, beispielsweise durch eine Reduktion des Attributs `PLZ` auf die ersten drei Stellen bei gleichzeitiger Erweiterung der `Geburtstag`-Intervalle. Die q^*-Blocke mit `Stadtteil = Warnemünde` und `Stadtteil = Dierkow` könnten dann etwa zusammenfallen. Aus exemplarischen Gründen entscheiden wir uns jedoch für das Unterdrücken des kritischen Blocks. Der neue, 2-anonyme und 2-diverse Datensatz ist in Tabelle 2.5 zu sehen:

Geburtstag	PLZ	Stadtteil	Note
1990 – 1994	18146	Dierkow	1.3
1990 – 1994	18146	Dierkow	2.3
2000 – 2004	18106	Evershagen	1.7
2000 – 2004	18106	Evershagen	1.0
1990 – 1994	18057	Hansaviertel	2.7
1990 – 1994	18057	Hansaviertel	1.3
1990 – 1994	18059	Südstadt	2.0
1990 – 1994	18059	Südstadt	4.0
1990 – 1994	18059	Südstadt	3.0
1995 – 1999	18059	Südstadt	1.3
1995 – 1999	18059	Südstadt	5.0

Tabelle 2.5.: Unser Datensatz, nun 2-anonym und 2-divers

So wie die l-Diversität eine Erweiterung der k-Anonymität darstellt, gibt es auch eine Verschärfung der l-Diversität selbst: die t-Closeness. Diese besteht genau dann, wenn die Verteilung der Attributwerte innerhalb eines q^*-Blocks höchstens um den Faktor t von der Verteilung der Attributwerte in der gesamten Relation abweicht. Da sie allerdings für die Arbeit nicht von Relevanz ist, wird an dieser Stelle nicht weiter darauf eingegangen.

2.3. Empirische Befragungen

Das folgende Kapitel basiert auf dem Buch *„Interview und schriftliche Befragung: Grundlagen und Methoden empirischer Sozialforschung“* von *Horst Otto Mayer* [May13].

In diesem Kapitel beschäftigen wir uns mit den Grundlagen empirischer Befragungen. Diese können im Wesentlichen auf zwei Weisen geschehen: quantitativ und qualitativ. Diese Bachelorarbeit stellt beide Herangehensweisen kurz vor, setzt sich mit ihnen auseinander und vergleicht sie miteinander. Empirische Untersuchungen sind insbesondere in Geistes- und Sozial-, aber auch in Wirtschaftswissenschaften eine häufig verwendete Forschungsmethode, da es dort selten bis gar nicht möglich ist, mathematisch-formale Beweise anzuführen. Aber auch in der Informatik, speziell im Bereich der Mensch-Computer-Interaktionen, werden empirische Evaluationen eingesetzt. Ziel empirischer Befragungen ist ein Informationsgewinn aus real existierenden Gegebenheiten, beispielsweise dem menschlichen Verhalten, um daraus neue Erkenntnisse ableiten zu können. So kann beispielsweise untersucht werden, ob und wie stark zwei Tätigkeiten miteinander korrelieren.

Eine Herausforderung der Empirie besteht in dem Schluss von einzelnen Personen auf die Menge aller relevanten Personen – in der Statistik spricht man von einer **Stichprobe einer Grundgesamtheit**. Während es bei quantitativen Befragungen (Kapitel 2.3.1) noch möglich ist, die Nähe der Stichprobe zur Grundgesamtheit mit mathematischen Methoden zu untersuchen, muss diese bei qualitativen Befragungen (Kapitel 2.3.2) argumentativ begründet werden.

2.3.1. Quantitative Befragungen

Quantitative Befragungen haben eine möglichst große Menge an Probanden zum Ziel. Es gilt allgemein: Je mehr Probanden an einer quantitativen Befragung teilnehmen, desto hochwertiger sind die erhobenen Daten. Statistisch gesehen ist dies trivial – je umfangreicher die Stichprobe ist, desto näher ist sie an der Grundgesamtheit. Auch das *Gesetz der großen Zahlen* aus der Stochastik, welches als bekannt vorausgesetzt wird, kann hier als Argument herangezogen werden. Als Methode dient hier meist ein standardisierter Fragebogen.

Die folgende Unterteilung in nominale, ordinale und kardinale Merkmale basiert auf [BW15].

Die einzelnen Variablen eines Fragebogens, so zum Beispiel das Alter, Geschlecht, Einkommen etc., werden **(statistische) Merkmale** genannt. Diese werden näher bestimmt, um sie anschließend leicht verarbeiten und mathematisch auswerten zu können. Jedes dieser Merkmale gehört dabei einer bestimmten Art an – insgesamt existieren nominale, ordinale und kardinale Merkmale. **Nominale Merkmale** sind Merkmale, auf die nur die Äquivalenzrelation angewendet werden kann. Für zwei Merkmale m_1, m_2 gilt also:

$$m_1 \theta m_2, \quad \theta \in \{=, \neq\}.$$

Es ist also weder zugelassen, Hierarchien aufzustellen noch Produkte oder andere mathematische Verknüpfungen zu bilden. Ein Beispiel für nominale Merkmale sind die biologischen Geschlechter „männlich“ und „weiblich“: zwischen ihnen kann unterschieden werden, Aussagen wie „männlich ist größer/höherwertig als weiblich“ sind allerdings unzulässig.

Ordinale Merkmale hingegen lassen auch hierarchische Vergleiche zu. Aussagen wie „m_1 ist größer als m_2“ sind also zulässig. Somit gilt für ordinale Merkmale:

$$m_1 \theta m_2, \quad \theta \in \{=, \neq, <, >, \leq, \geq\}.$$

Ordinale Merkmale stellen somit eine Erweiterung der nominalen Merkmale dar. Als Beispiel dienen Bildungsabschlüsse: die mittlere Reife ist ein niedrigerer Abschluss als das Abitur, welches wiederum einem Hochschulabschluss untergeordnet ist. Aussagen wie „Ein Hochschulabschluss ist doppelt so gut wie ein Abitur“ sind aber nicht möglich.

Kardinale Merkmale sind nun Merkmale, die die ordinalen Merkmale um die Möglichkeiten zur Addition, Subtraktion, Multiplikation und Division erweitern. Beispielsweise ist das Alter einer Person kardinal: $14 \neq 28$, $14 < 28$ und 28 ist doppelt so alt wie 14 ($14 \cdot 2 = 28$). Somit gilt:

$$m_1 \theta m_2, \quad \theta \in \{=, \neq, <, >, \leq, \geq, +, -, \cdot, \div\}.$$

Bemerkung: numerische Merkmale sind nicht zwingend kardinal – beispielsweise ist die Schulnote 4 nicht „doppelt so schlecht“ wie die Note 2.

Im Mittelpunkt der quantitativen Befragungen steht nun der **standardisierte Fragebogen**, welcher vor allem – aber nicht nur – nach solchen nominalen, ordinalen oder kardinalen Merkmalen fragt. Der Aufwand, der zur Auswertung eines einzelnen Fragebogens betrieben werden muss, ist vergleichsweise gering: jede Antwort (aus den obigen Kategorien) besitzt einen wohldefinierten, meist endlichen Wertebereich, beispielsweise eine natürliche oder reelle Zahl oder einige vorgegebene Möglichkeiten. Die Auswertung der gesamten Befragung erfolgt dann mittels statistischer Tests. Dabei wird eine eindeutige **Nullhypothese** formuliert, welche überprüft und anschließend angenommen oder verworfen wird. Wir möchten an dieser Stelle einige solcher Tests vorstellen:

- Die **Varianzanalyse** dient dazu, die Varianzen innerhalb mehrerer Stichproben mit der zwischen den Stichproben zu untersuchen, wobei sich die Stichproben in einem bestimmten Merkmal voneinander unterscheiden. Kleine Varianzen innerhalb der Proben bei gleichzeitig großen Varianzen zwischen den Proben sprechen nämlich dafür, dass dieses Merkmal einen Einfluss auf die gemessenen Werte hat.
- Der **Chi-Quadrat-Anpassungstest** dient der Untersuchung, ob ein gewisses Merkmal einer bestimmten Verteilung folgt, beispielsweise der Normalverteilung oder der Exponentialverteilung.
- Der **Chi-Quadrat-Homogenitätstest** gibt Auskunft darüber, ob die Verteilungen verschiedener Stichproben identisch sind.

Es ist mittels standardisierter Fragebögen ebenfalls möglich, Korrelationen zwischen je zwei Merkmalen zu untersuchen, das heißt, zu prüfen, ob es zwischen zwei Merkmalen einen mathematisch messbaren Zusammenhang gibt. Je nach Skalierung (nominal, ordinal, kardinal) können hierbei andere Koeffizienten verwendet werden; eine ausführlichere Erläuterung würde an dieser Stelle aber zu weit führen. Es sei allerdings auf [BW15] verwiesen, wo sich zudem weitere Informationen zu den oben genannten und weiteren statistischen Tests finden lassen.

2.3.2. Qualitative Befragungen

Im Unterschied zu quantitativen Befragungen stützen sich qualitative Befragungen auf eine oft wesentlich kleinere Menge von Probanden, welche dafür allerdings umfangreicher befragt werden. Im Gegensatz zum standardisierten Fragebogen kommen hier **Leitfadeninterviews** zum Einsatz. Diesen liegt, wie der Name bereits verrät, ein Leitfaden aus (verschiedenen) Fragen zu Grunde, welche im Laufe des Gesprächs gestellt werden. Es ist allerdings jederzeit möglich, von diesen Fragen abzuweichen und beispielsweise Folgefragen zu stellen, die in einem Fragebogen gar nicht vorgesehen wären. Weiterhin kann der Grad der Vertiefung einer Antwort jederzeit spontan angepasst werden, indem die fragende Person die befragte

Person unterbricht und mit einer anderen Frage fortfährt – oder sie eben ausreden lässt. Durch ein solches Interview können Antworten mit einem bedeutend umfangreicheren Informationsgehalt gewonnen werden, welche allerdings auch schwieriger zu quantifizieren sind, sodass eine systematische Verarbeitung und Auswertung somit erschwert, wenn nicht sogar unmöglich ist.

Aufbau des Leitfadens Der Leitfaden besteht aus exemplarischen Fragen, die den Experten und Experten zu stellen sind. Dabei ist es zunächst wichtig, eine Klassifikation der befragten Person zu ermöglichen, indem nach gewissen klassifizierenden Merkmalen gefragt wird. Dazu können beispielsweise der Beruf, das Alter, das Einkommen, der Familienstand oder die Zugehörigkeit zu einer bestimmten Gruppe (Partei, Religion, ...) zählen. Ausgehend davon kann der Leitfaden dann in verschiedene Themenkomplexe unterteilt werden, welche jeweils unterschiedliche Fragen und diverse Folgefragen für bestimmte Personenklassen beinhalten können. Die interviewende Person wird dadurch auch während der Interviews entlastet.

Wichtig ist auch, trotz der gegebenen Offenheit das Ziel der Datenerhebung im Auge zu behalten und bei starken Abweichungen vom Gesprächsrahmen entsprechend in das Gespräch einzugreifen, um zu vermeiden, dass die Interviews unnötig lang oder zu oberflächlich werden. Dies ist auch insbesondere deshalb relevant, da die qualitative Auswertung eines einzelnen Interviews sehr zeit- und/oder kostenintensiv sein kann.

Die Probanden werden in ihrer Rolle als Experten eines bestimmten Gebietes befragt. Der Begriff des Experten wird dabei wie folgt definiert:

Definition 2.6. Als **Experte** wird angesprochen, wer in irgendeiner Weise Verantwortung trägt für den Entwurf, die Implementierung oder die Kontrolle einer Problemlösung oder wer über einen privilegierten Zugang zu Informationen über Personengruppen oder Entscheidungsprozesse verfügt (*Meuser und Nagel*, via [May13]). □

Die Fragen des Leitfadens sind im Gegensatz zu denen eines Fragebogens offen, das heißt, dass es keinen vordefinierten „Wertebereich“ für die Antworten gibt (beispielsweise eine numerische Skala oder eine binäre Unterteilung in „gut“ und „schlecht“). Der Vorteil hierbei ist, dass die Antworten freier und ausführlicher und die erhobenen Daten somit qualitativer sein können.

Sobald eine erste Version des Leitfadens erarbeitet wurde, erfolgt in der Regel ein sogenannter **Pretest**. Dabei werden probeweise einige wenige Personen interviewt, um unklare, zu komplexe oder anderweitig ungeeignete Fragen zu erkennen und daraufhin zu ersetzen, verbessern oder entfernen. Sobald sich der Leitfaden als tauglich erwiesen hat, können die eigentlichen Interviews beginnen.

Mayer etabliert in seinem Buch noch einen weiteren Begriff: den des Gatekeepers [May13]. Als **Gatekeeper** wird dabei eine Person oder Institution bezeichnet, welche „von der Stellung her in der Lage ist, dem Forscher Zugang zum Feld zu verschaffen“, also den Kontakt zwischen Interviewer und entsprechenden Experten herstellt.

Die eigentlichen Interviews werden dann entsprechend durchgeführt. Dabei bietet es sich an, diese aufzuzeichnen, damit sich die interviewende Person vollständig auf das Gespräch und den Inhalt der Antworten konzentrieren kann und ggf. weitere Fragen stellen oder zu einem anderen Themenbereich wechseln kann. Zudem sorgt ein Gespräch, welches nicht direkt niedergeschrieben wird, für eine angenehmere und persönlichere Atmosphäre während des Interviews.

Auswertung der Interviews Sofern die Interviews aufgezeichnet wurden, müssen sie zunächst transkribiert werden. Dabei ist zu beachten, dass der gesprochene Inhalt nicht verfälscht und originalgetreu wiedergegeben wird. Sofern erforderlich, sollten Besonderheiten in der Tonlage oder Sprache hervorgehoben werden. Gleichzeitig ist es nicht notwendig, Füllwörter, abgebrochene Sätze, Wiederholungen etc. zu transkribieren – es bietet sich sogar an, diese Dinge der Einfachheit halber zu ignorieren. Ein weiterer Aspekt ist der des Datenschutzes. Sichert man den Expertinnen und Experten zu, dass sie nicht anhand ihrer Antworten identifizierbar sein werden, so muss das Interview hinterher anonymisiert werden. Dazu werden Namen – sowohl von Personen als auch von Unternehmen, Institutionen etc. – pseudonymisiert oder vollständig entfernt. Auch Aussagen, die sehr spezifisch sind, sind mit Vorsicht zu behandeln und müssen eventuell verkürzt oder gar entfernt werden. Liegen die Interviews dann transkribiert und anonymisiert vor, kann mit der Auswertung der Antworten begonnen werden. *Mayer* stellt in seinem Buch zwei wesentliche Vorgehensweisen vor: ein sechsstufiges pragmatisches Verfahren, entwickelt von *Mühlfeld et al.*, und ein fünfstufiges Verfahren von *Meuser und Nagel*.

Das pragmatische Auswertungsverfahren von *Mühlfeld et al.* hat das Ziel, offenkundige und direkte Aussagen zu ermitteln und nur diese zur Interpretation heranzuziehen. Es ist also nicht notwendig, jeden einzelnen Satz zu interpretieren. Im ersten Schritt werden dazu sämtliche Aussagen markiert, die die Fragen des Leitfadens unmittelbar beantworten. Im zweiten Schritt werden die ermittelten Aussagen dann gewissen, üblicherweise vorher festgelegten Kategorien zugeteilt. Der dritte Schritt besteht im Herstellen einer inneren Logik, also den Zusammenhängen zwischen den jeweiligen Textstellen, auch aus anderen Interviews. Der Fokus liegt dabei sowohl auf Aussagen, die sich gegenseitig widersprechen, als auch auf Aussagen, die inhaltlich miteinander einhergehen. Diese innere Logik wird anschließend im vierten Schritt verschriftlicht, bevor die Interviews dann im fünften Schritt ausgewertet und diese Auswertung im sechsten Schritt in Form eines Berichtes präsentiert wird.

Das Auswertungsverfahren von *Meuser* und *Nagel* ist hingegen laut *Mayer* etwas aufwändiger. Hier werden die Transkripte im ersten Schritt paraphrasiert, um unwichtige Textpassagen zu entfernen und die Texte auf die relevanten Aussagen zu konzentrieren. Relevant sind hierbei nicht nur direkte Antworten auf die Fragen des Leitfadens, sondern auch neue Aussagen, die keiner Frage direkt zugeordnet werden können. Im zweiten Schritt erfolgt eine thematische Zuordnung der Textpassagen zu einzelnen Themen. Dies geschieht in Form von Überschriften. Erst im nun folgenden dritten Schritt erfolgt ein thematischer Vergleich mit Aussagen aus anderen Interviews und das Zusammenfassen der jeweiligen Themenbereiche der einzelnen Interviews, sofern möglich. Das Ziel ist auch hier also eine Reduktion der Texte auf die relevanten Passagen. Im vierten Schritt werden die verdichteten Textpassagen nun in eine allgemeingültige wissenschaftliche Sprache umgeschrieben und Begriffe der Experten durch Fachtermini ersetzt. Dies erfolgt unter der Zuhilfenahme anderer empirischer Studien und sonstiger Wissensbestände. Zu guter Letzt wird untersucht, inwiefern diese Texte bereits bestehende Theorien stützen oder widerlegen.

2.4. Forschungsdaten und deren Management

Im letzten Grundlagenkapitel befassen wir uns mit Forschungsdaten und der Verwaltung dieser. Dazu klären wir zunächst, was Forschungsdaten eigentlich sind, worin Forschungsdatenmanagement besteht und welche Anforderungen daran gestellt werden.

2.4.1. Forschungsdaten und deren Relevanz

Das *Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg (MWK BW)* veröffentlichte in einem *„Fachkonzept zur Weiterentwicklung der wissenschaftlichen Infrastruktur in Baden-Württemberg“* 2014 folgende Definition des Begriffs der Forschungsdaten:

Definition 2.7. Forschungsdaten sind digitale Daten, die durch wissenschaftliche Quellenforschungen, Experimente, Messungen, Simulationen, Erhebungen oder Befragungen entstehen. Sie umfassen strukturierte Massendaten sowie unstrukturierte Daten unterschiedlichsten Formats und Inhalts, so auch Prozessdaten wie Bearbeitungsschritte, in einem Forschungsvorhaben entwickelte Algorithmen und Programme oder digitalisierte Schriften, Bilder oder Filme [Min14]. □

Das *MWK BW* bezieht sich dabei vor dem Hintergrund der *E-Science* vor allem auf digitale Forschungsdaten. Ganz allgemein können mit Forschungsdaten jedoch auch handschriftliche Texte (man denke an die Philosophie) oder auch Artefakte (Archäologie) und sogar Lebewesen (Zoologie, Biologie) sein [Heu20]. In den meisten Fällen sind allerdings tatsächlich digitale Daten gemeint.

2.4.2. Forschungsdatenmanagement

Damit diese Forschungsdaten kurz- und langfristig verwendet werden können, müssen sie entsprechend verwaltet werden. Dieser Vorgang wird Forschungsdatenmanagement genannt und von *Simukovic*, *Kindling* und *Schrimbacher* wie folgt definiert:

Definition 2.8. Forschungsdatenmanagement bezeichnet „alle Aktivitäten, die mit der Aufbereitung, Speicherung, Archivierung und Veröffentlichung von Forschungsdaten verbunden sind“ [SKS13]. □

Laut der *Deutschen Forschungsgemeinschaft (DFG)*, welche 2015 ihre *Leitlinien zum Umgang mit Forschungsdaten* veröffentlichte, sollen Forschungsdaten möglichst früh öffentlich bereitgestellt werden. „Forschungsdaten“ meint hierbei entweder die erhobenen Rohdaten oder aber Daten, die bereits ausreichend strukturiert worden sind, um von Außenstehenden weiterverarbeitet werden zu können.

Besonders hervorzuheben sind die *FAIR-Prinzipien*, welche *Wilkinson et al.* 2016 erstmals vorstellten [WD+16]. Sie besagen, dass Forschungsdaten **auffindbar (findable)**, **zugänglich (accessible)**, **interoperabel (interoperable)** und **wiederverwendbar (re-usable)**[3] sein sollten. Dazu wurden 15 Anforderungen formuliert, welche Daten erfüllen müssen, um als „fair“ zu gelten[4], welche nachfolgend ausführlicher erläutert werden.

Daten sollen **auffindbar** sein. Dazu gehört, dass jedes Datum persistent und eindeutig über ein dafür vorgesehenes Attribut identifizierbar ist, beispielsweise über eine ID. Persistente Aufbewahrung bedeutet, dass die Informationen langfristig verfügbar bleiben – bei Links ist das beispielsweise oft nicht der Fall, bei PDF-Kopien von Webseiten hingegen schon. Außerdem sollten zu den Daten reichlich Metadaten existieren, sodass diese verständlich bleiben. Im Idealfall ist es außenstehenden Personen dann möglich, die Daten nur anhand ihrer Metadaten direkt zu finden. Für gewöhnlich sind Daten und Metadaten zwei verschiedene Dateien, weshalb es wichtig ist, dass die Identifikatoren der Daten in den Metadaten als Verweise auf die eigentlichen Daten vorhanden sind. Zu guter Letzt sollten sowohl die Daten als auch ihre Metadaten einfach zu durchsuchen sein. Dies bedeutet gewöhnlicherweise eine Indizierung [HSS19].

[3] Auch der Begriff „rekonstruierbar“ kann an dieser Stele verwendet werden.
[4] `https://www.force11.org/group/fairgroup/fairprinciples`

Um die **Zugänglichkeit** zu gewährleisten, sollten Daten mittels ihres Identifikators über ein standardisiertes Kommunikationsprotokoll abgerufen werden können. Für gewöhnlich werden sie online zur Verfügung gestellt; die Protokolle wären dann HTTP(S) oder FTP bzw. TCP/IP. Aber auch andere Protokolle sind denkbar, solange diese frei und universell implementierbar sind. HTTP(S) und FTP sind es, das Exchange-Protokoll von Microsoft ist es jedoch beispielsweise nicht. Außerdem muss das Protokoll eine Möglichkeit zu Authentifizierung und Autorisierung bereitstellen, wo immer es notwendig ist. Das letzte Prinzip zur Zugänglichkeit besagt, dass Metadaten auch dann noch verfügbar sein sollten, wenn die Originaldaten es nicht mehr sind, beispielsweise weil eine URL nicht mehr funktioniert. Grund dafür ist, dass Metadaten in einem solchen Fall dazu genutzt werden könnten, Personen oder Organisationen ausfindig zu machen, die mit diesen Daten zu tun hatten.

Das dritte Prinzip, die **Interoperabilität** der Daten, setzt sich aus drei Unterprinzipien zusammen. Zunächst einmal sollten die (Meta-)Daten eine formale Sprache nutzen, die auch allgemein dazu geeignet ist, Wissen zu repräsentieren, damit diese nicht nur von Menschen, sondern auch von Computern interpretiert werden können. Als Beispiel sei das *Dublin-Core*-Datenformat angegeben. Außerdem müssen alle Daten ein Vokabular verwenden, das die FAIR-Prinzipien einhält – das heißt, dass das Vokabular dokumentiert und auflösbar sein muss, indem eindeutige und persistente Bezeichnungen bzw. Identifikatoren genutzt werden. Zu guter Letzt sollten die eigentlichen Daten sowie die Metadaten qualitative Verweise auf andere (Meta-)Daten besitzen. Qualitativ bedeutet, dass klar ersichtlich ist, inwiefern zwei Datensätze miteinander verbunden sind – ein „basiert auf“ ist beispielsweise aussagekräftiger als ein einfaches „siehe“.

Zu guter Letzt sollten Daten auch **wiederverwendbar** sein. Dazu gehört, dass sie mit Schlag- bzw. Stichworten markiert werden, damit andere Personen schnell entscheiden können, ob die Daten für sie relevant bzw. nützlich sind oder nicht. Weiterhin muss die Lizenz klar sein: Unter welchen Bedingungen dürfen die Daten verwendet, bearbeitet, veröffentlicht werden? Als Beispiel seien hier die *Creative-Commons-Lizenzen* angeführt, welche diverse Fragen wie die Weitergabe, Namensnennung, kommerzielle Nutzung etc. einfach mittels einer Zeichenkette beantworten, welche zudem auch maschinell lesbar ist. Provenance-Informationen sollten außerdem vorhanden sein, damit nachvollziehbar ist, woher die Daten stammen, inwiefern sie verarbeitet wurden, wer sie erstellt oder gesammelt hat und ob sie zuvor schon veröffentlicht wurden. Der letzte Punkt, den es zur Wiederverwendbarkeit zu beachten gilt, ist das Einhalten gewisser Community-abhängiger Standards wie Leitfäden.

3. Aktueller Stand der Forschung

In diesem Kapitel beschäftigen wir uns mit dem aktuellen Stand der Forschung. Zunächst definieren wir den Begriff der Big Data, um anschließend sowohl Provenance als auch Privacy im Kontext der Big Data betrachten zu können. Am Ende schauen wir uns dann die Kombination aus Provenance und Privacy an und werden feststellen, dass es bislang kaum relevante Veröffentlichungen dazu gibt, die das Zusammenspiel der zwei Bereiche untersuchen.

Was genau „Big Data“ ist bzw. sind, ist unterschiedlich definiert. Diverse Definitionen stützen sich dabei jedoch auf einige „V-Begriffe“, in der Regel auf drei: **Volume**, **Velocity** und **Variety** [TN16]. *Heuer et al.* definieren den Begriff noch mithilfe eines vierten „Veracity“-Begriffs [HSS18]. Treffen alle vier der folgenden Eigenschaften zu, so wird die Datenmenge als Big Data bezeichnet:

- **Volume**: Die Datenbanken beinhalten hiesige Datenmengen bis hin zu hunderten Exabyte – ein Exabyte entspricht 1.000 Petabyte, was wiederum einer Million Terabyte entspricht.
- **Velocity**: Die Daten werden sehr schnell, meist sogar in Echtzeit, zu der Datenbank hinzugefügt; man spricht dann von Stromdaten. Für die Reaktion verbleiben dann nur wenige Millisekunden.
- **Variety**: In konventionellen Datenbanken liegen die Daten oft strukturiert vor, beispielsweise relational oder in strukturierten Dateiformaten wie XML oder JSON. Big-Data-Stromdaten können jedoch in den verschiedensten Formaten anfallen oder sogar gänzlich unstrukturiert sein, was zum Beispiel bei Audio- und Videoaufnahmen der Fall ist.
- **Veracity**: Stromdaten können auch ungenau sein. Beispielsweise können Daten fehlen, ein Rauschen beinhalten, approximiert sein oder durch diverse Heterogenitäten inkonsistent und/oder mehrdeutig sein.

Durch Big Data entstehen hinsichtlich des Datenschutzes diverse neue Herausforderungen. *Torra* und *Navarro-Arribas* haben 2016 drei wesentliche Probleme ermittelt [TN16]: Zum einen gehen durch Big Data Kontrolle bzw. Nachvollziehbarkeit und Transparenz verloren. Mehr und mehr Organisationen, Institutionen und Unternehmen sammeln (personenbezogene) Daten, weil es einerseits nie einfacher, andererseits nie profitabler war. Dadurch wird es schwieriger, wenn nicht schier unmöglich, einen Überblick über die erhobenen Daten an den verschiedensten Stellen zu erhalten. Ein weiteres Problem ergibt sich, wenn mehrere große Datensätze miteinander verknüpft werden. Bereits in Abschnitt 2.2 haben wir gesehen, dass das Verbinden zweier eigentlich harmloser Datensätze sensible Informationen offenlegen kann. Auch in Kapitel 4 werden wir das Problem noch einmal verdeutlichen. Durch Big Data wird es einerseits einfacher, Datensätze miteinander zu verknüpfen, andererseits auch gleichzeitig gefährlicher, beides aufgrund der wachsenden Datenmengen, welche viel mehr Informationen – auch personenbezogene – beinhalten als „gewöhnliche“ Datensätze. **Data Mining** bezeichnet die Wissenschaft, unter der Anwendung mathematisch-statistischer Methoden aus Datensätzen nützliche Informationen zu extrahieren, zu verbinden und Zusammenhänge herzustellen. Handelt es sich dabei um einen Datensatz, der die vier V erfüllt (siehe oben), so spricht man von **Big Data Analytics**. Das dritte der angesprochenen Probleme handelt genau davon: Das Schlussfolgern aus und Wiederverwenden von riesigen Datensätzen. *Kosinski et al.* konnten bereits 2013 nur anhand von Facebook-Daten mit relativ hohen Genauigkeiten bestimmen, ob eine US-amerikanische Person männlich oder weiblich ist (93%), der Demokratischen oder

Republikanischen Partei nahesteht (85%), christlich oder muslimisch (82%) sowie schwul (88%) oder lesbisch (75%) ist ([KSG13], via [Noc18]). Die Möglichkeit, nur anhand dieser Daten auf schützenswerte Attribute wie die sexuelle Orientierung oder Parteizugehörigkeit einer Person schließen zu können – beides darf laut *DSGVO* nicht ohne explizite Einwilligung verarbeitet werden[1] –, ist äußerst bedenklich.

3.1. Provenance/Privacy und Big Data

Da nun klar ist, was mit Big Data gemeint und wieso dieses Thema allgemein relevant ist, können wir uns sowohl Provenance als auch Privacy im Zusammenhang mit Big Data ansehen.

Big Data Provenance: State-Of-The-Art Analysis and Emerging Research Challenges *Alfredo Cuzzocrea* untersuchte 2016 in seinem Paper *„Big Data Provenance: State-Of-The-Art Analysis and Emerging Research Challenges“* den aktuellen Stand von Data Provenance im Zusammenhang mit Big-Data-Anwendungen [Cuz16]. In dieser Veröffentlichung stellt *Cuzzocrea* fest, dass bisherige Data-Provenance-Modelle, welche Annotationen (wie bspw. Zeugenbasen und Provenance-Polynome) nutzen, für Big-Data-Umgebungen aufgrund der riesigen Datenmengen nicht geeignet sind. Andere Aspekte wie die Vertraulichkeit, Datensicherheit und Privatheit der Daten sind ebenfalls kaum erforscht. *Cuzzocrea* benennt außerdem einige Herausforderungen, die es zu meistern gilt. So stellt er die Frage, *wann* die Provenance-Daten berechnet werden sollten – nur bei Bedarf (*„lazy provenance model“*) oder nach jeder Veränderung des Datenbestands (*„eagerly provenance model“*) – und welche Datenmodelle dafür genutzt werden sollten, da sich die internen Strukturen je nach Anwendung massiv voneinander unterscheiden können. Weitere von ihm benannte Probleme sind der Support von Nutzer-Annotationen in Provenance-Systemen, die Implementation flexibler Anfragesysteme für Data Provenance und das Fehlen von Anwendungen, die Provenance-Daten interaktiv visualisieren können. Allerdings benennt der Autor auch das Kombinieren von Provenance mit Datenschutz und -sicherheit als eine Herausforderung der nächsten Jahre.

Big Data Privacy and Anonymization Ebenfalls 2016 veröffentlichten *Vicenç Torra* und *Guillermo Navarro-Arribas* das Paper *„Big Data Privacy and Anonymization“* und stellten darin ihre Überlegungen zu Anonymisierungstechniken und Datenschutz im Big-Data-Kontext vor [TN16]. Auch die zunehmende Relevanz von Data Provenance wird erkannt. Die beiden Autoren sehen die Lösung für Privacy in Big Data in diversen Techniken zur Anonymisierung, sprechen synonym jedoch auch von Maskierungen. Die Grundidee ist bekannt: durch ein Reduzieren der Qualität des Datensatzes kann die Privatsphäre einzelner Individuen gesichert werden. Dabei ergeben sich drei wesentliche Forschungsfragen:

1. Wie kann die Qualität der Daten reduziert werden, bzw. welche Methoden existieren?
2. Wie kann der Verlust der Informationen minimiert werden? (Wie wird garantiert, dass nicht mehr Daten entfernt werden als notwendig?)
3. Wie kann gemessen werden, ob der Datenschutz letztendlich tatsächlich eingehalten wird bzw. in welchem Maße?

[1] Art. 9 DSGVO: „Die Verarbeitung personenbezogener Daten, aus denen die [...] ethnische Herkunft, politische Meinungen, religiöse oder weltanschauliche Überzeugungen oder die Gewerkschaftszugehörigkeit hervorgehen, sowie die Verarbeitung von [...] Daten zum Sexualleben oder der sexuellen Orientierung einer natürlichen Person ist untersagt.“

Anonymisierungsmethoden werden in zwei Kategorien unterteilt: Störende und nicht-störende Methoden. **Störende Methoden** sind Methoden, die den Daten einen gewissen Fehler beifügen. Das kann beispielsweise ein Rauschen sein, wie es bei der Differential Privacy zum Einsatz kommt, aber auch Mikro-Aggregationen oder Permutationen können damit gemeint sein. **Nicht-störende Methoden** sind hingegen Methoden, die zwar ebenfalls die Genauigkeit der Daten reduzieren, dabei aber keine Fehler hinzufügen. Dies ist zum Beispiel bei der Generalisierung und Unterdrückung von Tupeln (siehe Unterabschnitt 2.2.2) der Fall: Gewisse Daten werden zwar vereinfacht oder entfernt, womit der Informationsgehalt schwindet, es werden jedoch keine fehlerhaften Informationen hinzugefügt oder bestehende Daten so verändert, dass sie hinterher fehlerhaft wären. Zu diesen zwei Funktionsklassen gilt es zu forschen.

Forschung ist laut den Autoren ebenfalls bei der Messung des Informationsverlustes nötig. Wendet man oben genannte störende und nicht-störende Methoden an, verringert sich der Informationsgehalt des Datensatzes. Diese Veränderung, also den „Abstand“ zwischen altem und neuem Datensatz, gilt es effizient zu bestimmen. Abstrakt lässt sich das Problem mit folgender Funktion beschreiben:

Definition 3.1. Seien f eine Funktion, die den Informationsgehalt eines Datensatzes bestimmt, a eine Anonymisierungsfunktion (s.o.) und X ein Datensatz. Der **Informationsverlust** kann dann mit

$$IV_f(X, a(X)) = \text{divergence}(f(X), f(a(X)))$$

ausgedrückt werden, wobei $\text{divergence}(x, y)$ eine Funktion ist, die die inhaltliche Distanz zweier Datensätze misst ($\forall x : \text{divergence}(x, x) = 0$) (vgl. [TN16]). □

Das dritte Problem liegt im Ermitteln des tatsächlichen Datenschutzes. Für gewöhnlich besteht auch nach einer Anonymisierung ein gewisses Restrisiko, dass Teile der Daten wieder de-anonymisiert werden können. Dieses Risiko gilt es zu berechnen. Insgesamt werden also diverse Probleme, die bei Big Data im Zusammenhang mit (Data) Provenance und Privacy entstehen, erkannt, konkrete Lösungen gibt es bislang allerdings nicht.

3.2. Provenance und Privacy

Mit Blick auf die vorherigen Kapitel stellt sich nun die Frage, inwieweit Provenance und Privacy kombiniert werden können beziehungsweise inwiefern Data Provenance den Datenschutz verletzt oder der Datenschutz das Generieren gewisser Provenance-Informationen erschwert.

On Provenance and Privacy 2011 veröffentlichten *Susan B. Davidson, Sanjeev Khanna, Sudeepa Roy et al.* den Artikel *„On Provenance and Privacy“* [DKR+11]. Ausgangspunkt dieser Veröffentlichung ist die gestiegene Bedeutung von Provenance in wissenschaftlichen Arbeitsabläufen – also Workflow Provenance – und dem Schutz gewisser Daten. Damit sind nicht nur personenbezogene, sondern auch Modul- und Strukturdaten gemeint, also Informationen darüber, welche Software zum Einsatz kam und inwiefern mehrere Anwendungen miteinander verknüpft wurden. Dazu wird zwischen „data privacy“, „module privacy“ und „structural privacy“ unterschieden. **Data Privacy** hat nach den Autorinnen und Autoren die Aufgabe, die Daten zu schützen, die zwischen je zwei Modulen übertragen werden. **Module Privacy** schützt hingegen die Funktionalität der einzelnen Module, das heißt, es soll nicht möglich sein, von einer Eingabe in ein Modul auf die (ungefähre) Ausgabe dieses Moduls schließen zu können. **Structural Privacy** schützt zuletzt die Zusammensetzung des Arbeitsablaufs. Dazu gehört auch, dass einigen Nutzern möglicherweise Teile des entstehenden Provenance-Graphen vorenthalten werden, weil er oder

sie nicht berechtigt ist, diese Informationen einzusehen. Interessant für die Bachelorarbeit ist der Teil der Data Privacy. Im Artikel wird jedoch vorausgesetzt, dass die Forschungsdaten bei Workflow-Provenance-Anwendungen nicht aggregiert werden und Techniken des Datenschutzes in statistischen und relationalen Datenbanken deshalb nicht benötigt werden.

A roadmap for privacy-enhanced secure data provenance Im 2014 veröffentlichten Artikel *„A roadmap for privacy-enhanced secure data provenance“* von *Elisa Bertino, Gabriel Ghinita et al.* [BGK+14] stellen die Autorinnen und Autoren fest, dass Data Provenance seit Beginn des Jahrhunderts in der Forschung zwar immer bedeutsamer wurde, diverse Aspekte des Datenschutzes und der Datensicherheit dabei jedoch bislang keine große Rolle spielten. Deshalb entwickeln sie ein abstraktes Modell, welches „privacy-enhanced, secure provenance“ gewährleisten soll. Dabei wird zwischen privaten und nicht-privaten Daten sowie privater und nicht-privater Provenance unterschieden und eine Matrix aus den vier möglichen Kombinationen gebildet. Für jede dieser Kombinationen werden dann einige Ansätze vorgestellt. Dabei werden sogar Anonymisierungstechniken wie k-Anonymität, l-Diversität, t-Closeness und auch *Differential Privacy* angeschnitten. Leider nutzen die Autorinnen und Autoren jedoch eine Provenance-Notation, welche auf *gerichteten, azyklischen Graphen (DAG; engl. directed acyclic graphs)* basiert und vernachlässigen Data Provenance im Zusammenhang mit relationalen Datenbanken.

ProvChain: A Blockchain-Based Data Provenance Architecture in Cloud Environment with Enhanced Privacy and Availability *Liang et al.* stellten 2017 mit *ProvChain* eine cloudbasierte Architektur auf Blockchain-Basis vor, welche Privacy als Datensicherheit versteht: In einer Blockchain werden Provenance-Informationen darüber gespeichert, welcher Nutzer wann welche Operation(en) auf Objekte in einer Cloud ausführten, um kollaboratives Arbeiten nachvollziehbar zu machen [LST+17]. Das Problem sensibler Informationen in Data Provenance wurde zwar aufgegriffen, allerdings liegt der Fokus in dieser Veröffentlichung darauf, den Zugang zu diesen Daten zu beschränken und das Abrufen dieser transparent und integer zu speichern. Es ist also gar nicht gewollt, Daten der Öffentlichkeit zur Verfügung zu stellen. Somit erübrigen sich auch die Bestrebungen nach Datenschutz im Sinne der Privatheit einzelner Individuen.

Provenance and Privacy *Vicenç Torra et al.* untersuchten ebenfalls 2017 im Artikel *„Provenance and Privacy“* das Zusammenspiel von Data Provenance und Privacy [TNSM17]. Zunächst begründen sie die Relevanz des Themas mit zwei stark debattierten EU-Gesetzen: Zum Einen wurde seit 2013 das *„Recht auf Vergessenwerden“*, welches besagt, dass Nutzerinnen und Nutzer die Möglichkeit haben müssen, ihre Daten bei Konzernen löschen zu lassen, diskutiert. Zum Anderen wurde 2016 auch der Entwurf der *Europäischen Datenschutzgrundverordnung (DSGVO)* verabschiedet, welche unter Anderem beinhaltet, dass Personen Auskunft über ihre personenbezogenen Daten erhalten müssen. Bezogen auf Data Provenance ist dies nach *Torra et al.* jedoch keinesfalls trivial: Daten liegen oft in aggregierter Form vor (bspw. bei der Bildung von Durchschnitten oder einer anderen Generalisierung; man denke auch an Data-Mining-Techniken). Löscht man nun jedoch einzelne Datensätze, kann dies ebendiese Daten ungültig machen, sofern die entsprechenden Datensätze an der Konstruktion der aggregierten Daten beteiligt waren. Data Provenance, welche den Datenschutz betroffener Personen bewahrt, wurde seitdem also umso relevanter.

Der Privacy-Begriff wird dabei zunächst als Schutz der Daten vor unautorisierten Zugriffen verstanden: Wurden Provenance-Metadaten erhoben, so beinhalten diese möglicherweise sensible Informationen, welche – wenn überhaupt – nur für einen kleinen Personenkreis einsehbar sein dürfen. Es bedarf folglich einer

Möglichkeit, bestimmte Personen zu autorisieren und allen anderen den Zugang zu den Daten zu untersagen. Zudem muss die **Integrität** der Daten gewährleistet werden. Dies bedeutet, dass sie nicht manipuliert werden dürfen, während sie in irgendeiner Weise – beispielsweise innerhalb eines verteilten Systems – übertragen werden. Später wird der Privacy-Begriff jedoch auch auf das klassische Verständnis von Datenschutz ausgeweitet. Demnach soll es nicht möglich sein, gewisse Personen anhand der Provenance-Daten zu reindentifizieren bzw. Rückschlüsse zu ermöglichen, welche die Privatheit der Person verletzen würden.

Weiterhin stellen *Torra et al.* in dem Paper diverse Provenance-Anwendungen kurz vor: *RAMP (Reduce And Map Provenance)* ist eine Erweiterung für *Apache Hadoop*, welche für Ergebnisse von *Map-Reduce*-Berechnungen Provenance-Daten erzeugen. Einen ähnlichen Weg geht *HadoopProv*. Auch für *Apache Spark* gibt es mit *Titian* eine Erweiterung für **why-** und **where**-Provenance. Allerdings stellen die Autorinnen und Autoren fest, dass es insgesamt eher wenig Software für Data Provenance in Big-Data-Anwendungen gibt. Zudem beachtet keine davon Privacy-Aspekte. Als Lösungsansatz werden vor allem k-Anonymität und Differential Privacy angeführt; vor allem letzteres wird eine große Bedeutung zugeschrieben. Im nächsten Abschnitt werden wir uns deshalb noch genauer mit diesem Prinzip auseinandersetzen.

3.3. Differential Privacy

In den letzten zwei Jahrzehnten gewann neben der k-Anonymität und der l-Diversität auch eine weitere Datenschutzmaßnahme zunehmend an Bedeutung: die **Differential Privacy**, welche von *Cynthia Dwork* in [Dwo06] etabliert wurde und heute unter anderem von Unternehmen wie *Apple* verwendet wird, um statistische Daten(banken) zu veröffentlichen, ohne dabei den Schutz einzelner Personen innerhalb dieser Daten(banken) zu gefährden. Die grundlegende Idee dahinter ist, dass eine Funktion $\mathcal{K}$ existiert, welche Anfragen auf einen Datensatz ausführt und deren Ergebnis – möglicherweise modifiziert – liefert. Dabei gilt grundsätzlich, dass ein neuer Eintrag in einem Datensatz das Ergebnis der Funktion nicht verändern darf: Ist eine einzelne Zeile bzw. ein einzelnes Tupel für ein Ergebnis also entbehrlich, so ist der Datenschutz gewährleistet. Eine individuelle Person (oder auch kleinere Gruppe) fällt also innerhalb dieser Daten nicht auf. Der Vorteil besteht hier auch darin, dass Angriffe durch Hintergrundwissen – mehr dazu in Kapitel 4 – somit ausgeschlossen werden können [PS17].

Um Differential Privacy zu ermöglichen, wird ein zufälliges Rauschen zum Anfrageergebnis hinzugefügt. Dadurch wird das Ergebnis zwar ungenauer, beinhaltet aber immer noch ausreichend Informationen, ohne dabei sensible (persönliche) Daten zu offenbaren.

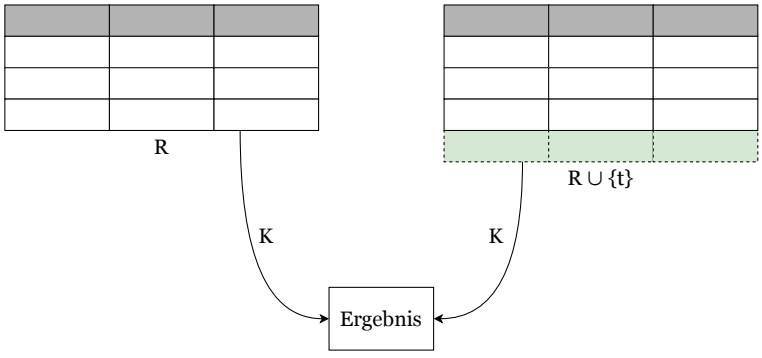


Abbildung 3.1.: Differential Privacy ist genau dann erfüllt, wenn ein neues Tupel das Ergebnis der Funktion nicht (wesentlich) beeinflusst

Zusätzlich zur klassischen Definition der Differential Privacy, welche besagt, dass das neue Ergebnis äquivalent zum vorherigen Ergebnis sein muss, gibt es diverse Erweiterungen dieser Definition, welche das Prinzip abschwächen. Eine davon stellt die ϵ-Differential-Privacy dar.

Definition 3.2. Eine randomisierte Funktion $\mathcal{K}$ bietet **ϵ-Differential-Privacy**, wenn für je zwei Datensätze D_1 und D_2, welche sich in höchstens einem Element unterscheiden, sowie für alle $S \subseteq W(\mathcal{K})$ gilt [Dwo06]:

$$P[\mathcal{K}(D_1) \in S] \leq e^{\epsilon} \cdot P[\mathcal{K}(D_2) \in S]$$

□

Diese Definition bezieht sich allgemein auf zwei Datensätze. Um sie ein wenig zu konkretisieren, definieren wir sie an dieser Stelle für Relationen und Tupel wie folgt:

Definition 3.3. Eine randomisierte Anfragefunktion $\mathcal{K}$ bietet ϵ-Differential-Privacy, wenn für eine Relation R und ein neues Tupel t sowie für alle $S \subseteq W(\mathcal{K})$ gilt:

$$P[\mathcal{K}(R) \in S] \leq e^{\epsilon} \cdot P[\mathcal{K}(R \cup \{t\}) \in S]$$

□

Die Funktion $\mathcal{K}$ muss also mit einer gewissen (hohen bis sehr hohen) Wahrscheinlichkeit sowohl für R als auch für $R \cup \{t\}$ denselben Wert annehmen. Das heißt im Umkehrschluss, dass sich der Funktionswert – also das Ergebnis – nur marginal verändern darf, wenn ein neues Tupel zu der Relation hinzugefügt oder ein bestehendes entfernt wird. Weiterhin gilt, dass der Datenschutz umso besser garantiert werden kann, je kleiner ϵ ist (geht ϵ gegen 0, so geht der Faktor e^{ϵ} gegen 1 und die obige Ungleichung wird zur Gleichung). Allerdings steigt mit zunehmendem ϵ auch der zu erwartende Informationsgehalt, sodass auch bei der Differential Privacy zwischen dem Datenschutz und der Menge an Informationen abgewogen werden muss. In gewissen Situationen kann es auch von Interesse sein, anstelle einer einzelnen Person eine Gruppe von Personen schützen zu wollen. Auch dies lässt sich quantifizieren, indem in der Definition der ϵ-Differential-Privacy das ϵ mit der Anzahl der Personen c multipliziert wird [Dwo06]. Daraus folgt das Offensichtliche: Je größer eine Gruppe ist, desto schwieriger ist es, ihre Daten in einer statistischen Datenbank zu schützen.

Oftmals, so auch in [Dwo08] von *Cynthia Dwork*, wird als Störfunktion die *Laplace*-Verteilung verwendet (siehe Abbildung 3.2 und Definition 3.4). μ nimmt dabei den tatsächlichen Wert an. Für σ gilt: je kleiner der Wert, desto stärker konzentriert sich die Verteilung um μ.

Definition 3.4. Die ***Laplace*-Verteilung** $\mathrm{Lap}_{\sigma;\mu}$ ist eine stetige Wahrscheinlichkeitsverteilung mit der Dichte

$$f_{\sigma;\mu}(x) = \frac{1}{2\sigma} \cdot e^{\frac{|x-\mu|}{\sigma}}$$

und den Parametern σ und μ, wobei σ die Skalierung und μ die Lage der Funktion bestimmt (an der Stelle $x = \mu$ ist $f_{\sigma;\mu}(x)$ maximal). □

Wir definieren nun noch nach *Cynthia Dwork* einen weiteren Begriff: Die Sensitivität einer Funktion f. Diese gibt die inhaltliche Distanz zweier Datensätze an:

Definition 3.5. Für $f : \mathcal{D} \to \mathbb{R}^k$ ist die **Sensitivität** von f definiert als

$$\Delta f = \max_{D_1, D_2} \|f(D_1) - f_(D_2)\|_1$$

für alle D_1, D_2, die sich in höchstens einem Element unterscheiden. k gibt hierbei die Anzahl der zurückgegebenen Werte an [Dwo08]. □

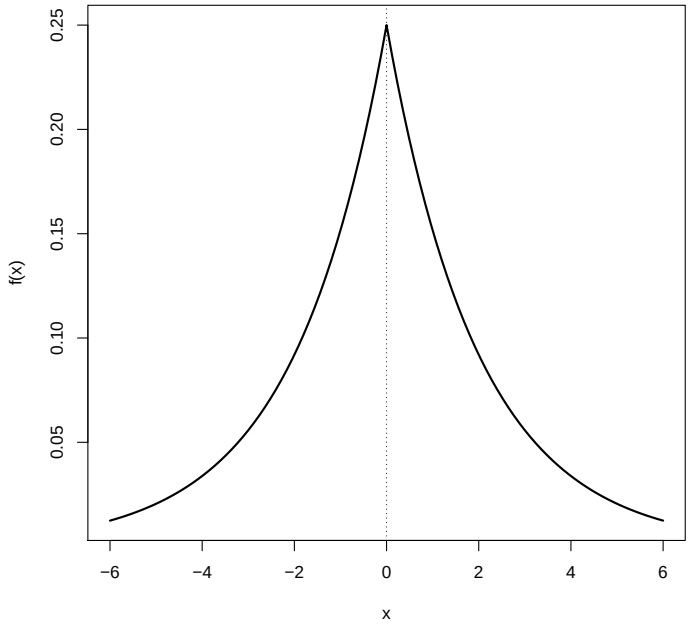


Abbildung 3.2.: Graph der *Laplace*-Verteilung mit $\sigma = 2$ und $\mu = 0$

Eine recht simple Definition von $f(R)$ bezüglich relationaler Datenbanken wäre beispielsweise die Anzahl aller Tupel einer Relation R. Die Distanz Δf wäre dann 1, da sich die Datensätze um nur (genau) ein Tupel unterscheiden. Der Einfachheit halber sei an dieser Stelle angenommen, dass $k = 1$ gilt, also f jeweils nur einen reellen Wert liefert. Dieser absolute Abstand, also Δf, dient als unser μ der *Laplace*-Verteilung; ϵ^{-1} dient als σ. Eine Anfragefunktion $\mathcal{K}(x)$ liefert nun nicht den tatsächlichen Wert $f(x)$, sondern

$$\mathcal{K}(x) = f(x) + (\mathrm{Lap}_{\epsilon^{-1};\Delta f})^k = f(x) + \mathrm{Lap}_{\epsilon^{-1};\Delta f},$$

wobei $\mathrm{Lap}_{\epsilon^{-1};\Delta f}$ den Wert der Dichtefunktion an einer zufällig gewählten Stelle dieser Funktion beschreibt. Entsprechend Definition 3.4 gilt nun, dass der Datenschutz umso besser gewährleistet wird, je kleiner $\epsilon = \sigma^{-1}$ ist. Der Wert von ϵ ist dabei meist öffentlich bekannt, da er als Gütemaß dient, wohingegen die zufällig gewählten x-Werte unbekannt sind, da sie andernfalls den originalen Wert offenbaren können, sofern die zugrundeliegende Verteilung bekannt ist. Dies kann durchaus der Fall sein, da auch sie etwas über die Qualität des Rauschens aussagt.

Im nächsten Schritt werden wir uns in Kapitel 4 mit einer eigenen Kombination von Provenance und Privacy beschäftigen. Die Idee, Privacy und Provenance zu kombinieren, ist ein noch sehr junges Forschungsthema, sodass bisher wenig Literatur zu diesem Thema vorliegt. Wir werden uns daher verschiedene Szenarien ansehen und überprüfen, ob und wenn ja, inwiefern **where**-, **why**- und **how**-Provenance in relationalen Datenbanken gegen ausgewählte Datenschutzaspekte verstoßen und wie mögliche Lösungsansätze aussehen könnten.

4. Provenance und Privacy

In diesem Kapitel setzen wir uns mit der Frage auseinander, inwieweit Data Provenance und Privacy miteinander einhergehen, ob gewisse Provenance-Techniken und -Anfragen möglicherweise den Datenschutz verletzen und falls ja, welche Lösungsansätze es dafür geben kann. Dazu betrachten wir zunächst in Abschnitt 4.1 die Invertierbarkeit von **where**-, **why**- und **how**-Provenance, bevor wir uns in Abschnitt 4.2 mit möglichen Problemen in Kombination mit dem Datenschutz beschäftigen. Im letzten Schritt, in Abschnitt 4.3, schauen wir uns an, welche Lösungen es für die ermittelten Probleme geben könnte.

4.1. Zur Invertierbarkeit von where, why und how

Data Provenance kann, das haben wir bereits in Abschnitt 2.1 gelernt, verschiedene Ausprägungen annehmen, die jeweils unterschiedliche Provenance-Informationen erzeugen. Bei der **where**-Provenance erhalten wir pro Tupel eine Liste aller beteiligen Tupel – die **Zeugenliste** – beziehungsweise eine Liste aller beteiligten Relationen. Bei der **why**-Provenance erhalten wir eine Menge von Zeugenmengen, die **Zeugenbasis**. Eine Zeugenmenge besteht dabei aus den Ursprungstupeln, die an einem Tupel der Ergebnisrelation beteiligt sind. Bei der **how**-Provenance erhalten wir zu guter Letzt ein **Provenance-Polynom**, welches auch auf die exakte Berechnungsvorschrift des Tupels schließen lässt.

4.1.1. Invertierbarkeit von where

Zunächst sehen wir uns an, welche Daten mittels **where**-Provenance wiederherstellbar sind. Wir betrachten dazu folgende Anfrage:

```
SELECT MatNr, AVG(Note) AS Durchschnitt
FROM PRUEFUNG
WHERE Semester = 'SS 20'
GROUP BY MatNr
HAVING COUNT(*) > 1
```

Anfrage 4.1: Die Durchschnittsnote je Student/-in (Matrikelnummer) im Sommersemester 2020, die mehr als eine Prüfung absolvierten

	MatNr	Durchschnitt	
R_1	10002	1.9	$\{P_{15}, P_{24}, P_{25}, P_{26}\}$
R_2	10003	1.5	$\{P_{16}, P_{27}\}$
R_3	10005	2.0	$\{P_{17}, P_{28}, P_{29}\}$

Tabelle 4.1.: Ergebnis von Anfrage 4.1

Bei der **where**-Provenance wissen wir einerseits, dass die Daten aus der Relation `PRUEFUNG` stammen, andererseits wissen wir auch, dass die Tupel $\{P_{15}, P_{16}, P_{17}, P_{24}, P_{25}, P_{26}, P_{27}, P_{28}, P_{29}\}$ am Ergebnis beteiligt

waren. Letzteres kennen wir natürlich auch. Über die Zusammensetzung der Ergebnistupel wissen wir allerdings nichts und auch die genaue Anzahl an Ausgangstupeln ist uns nicht bekannt. Aus den uns gegebenen Informationen lässt sich exemplarisch also nur ableiten, dass eine unbekannte Anzahl an Tupeln – ausgegangen wird von drei – zum Ergebnis 1.9 führte (siehe Tabelle 4.2). Das Semester lässt sich hierbei direkt aus der Anfrage ablesen; η_i repräsentiert den i-ten unbekannten Wert.

MatNr	ModNr	Semester	Note
10002	η_1	SS 20	1.9
...	...	...	...
10003	η_2	SS 20	1.5
...	...	...	...
10005	η_3	SS 20	2.0
...	...	...	...

Tabelle 4.2.: Die **where**-Provenance lässt nicht auf die konkrete Anzahl der Tupel der `PRUEFUNG`-Relation schließen, die in das Ergebnis eingeflossen sind.

4.1.2. Invertierbarkeit von why

Bei der **why**-Provenance erhalten wir hingegen schon einige weitere Informationen. Anstelle einer einfachen Zeugenliste wird hier nämlich eine Zeugenbasis (für die gesamte Relation) gespeichert, welche die Zeugen je Tupel (Zeugenmengen) beinhaltet. Aus dieser Zeugenbasis (siehe unten) lässt sich zunächst auf die Anzahl der Tupel der Ergebnisrelation schließen.

$$\text{Zeugenbasis: } B = \{\{\{P_{15}, P_{24}, P_{25}, P_{26}\}\}_{R_1}, \{\{P_{16}, P_{27}\}\}_{R_2}, \{\{P_{17}, P_{28}, P_{29}\}\}_{R_3}\}$$

MatNr	ModNr	Semester	Note
10002	η_1	SS 20	1.9
10002	η_2	SS 20	1.9
10002	η_3	SS 20	1.9
10002	η_4	SS 20	1.9
10003	η_5	SS 20	1.5
10003	η_6	SS 20	1.5
10005	η_7	SS 20	2.0
10005	η_8	SS 20	2.0
10005	η_9	SS 20	2.0

Tabelle 4.3.: Die **why**-Provenance lässt auf die Anzahl der Tupel der `PRUEFUNG`-Relation schließen, die in das Ergebnis eingeflossen sind.

Von nun an wissen wir also auch konkret, welche Originaltupel an welchen Ergebnistupeln beteiligt waren. Das erste Tupel setzt sich aus Werten von P_{15}, P_{24}, P_{25} und P_{26} zusammen, das zweite aus Werten von P_{16} und P_{27} und das dritte aus P_{17}, P_{28} und P_{29}. Wir können an dieser Stelle zwar noch nicht auf das Endergebnis schließen, da wir nicht wissen, wie diese Daten verarbeitet wurden (auch wenn das Bilden der arithmetischen Mittels in dem Fall recht trivial ist), wir können aber bereits Rückschlüsse auf die Zusammensetzung der Originaldaten ziehen. Im nächsten Schritt werden wir darauf genauer eingehen.

4.1.3. Invertierbarkeit von how

Bei der **how**-Provenance – zur Erinnerung: „Wie wurde das Ergebnis berechnet?“ – merken wir uns schließlich nicht nur, welche Daten wie am Ergebnis beteiligt sind, sondern auch, wie dieses konkret

berechnet wird. Dazu werden sogenannte **Provenance-Polynome** gespeichert. Die Polynome für eine solche Anfrage bestehen – das haben wir in Abschnitt 2.1 bereits gesehen – aus dem Quotienten der Summe und der Anzahl aller Werte, da `AVG(X)` $= \frac{\texttt{SUM(X)}}{\texttt{COUNT(X)}}$ gilt. Wir berechnen also im Falle von R_1 zunächst das Polynom für die Summe, welches $(P_{15} \odot 1.3) \oplus (P_{24} \odot 3.3) \oplus (P_{25} \odot 2.0) \oplus (P_{26} \odot 1.0)$. lautet. Das Polynom für die Anzahl lautet $P_{15} \oplus P_{24} \oplus P_{25} \oplus P_{26}$. Das vollständige Provenance-Polynom für das Ergebnistupel R_1 sieht entsprechend wie folgt aus:

$$p_1 = \frac{(P_{15} \odot 1.3) \oplus (P_{24} \odot 3.3) \oplus (P_{25} \odot 2.0) \oplus (P_{26} \odot 1.0)}{P_{15} \oplus P_{24} \oplus P_{25} \oplus P_{26}}.$$

Analog dazu lautet das Polynom für R_2

$$p_2 = \frac{(P_{16} \odot 1.3) \oplus (P_{27} \odot 1.7)}{P_{16} \oplus P_{27}}$$

und das für R_3

$$p_3 = \frac{(P_{17} \odot 1.0) \oplus (P_{28} \odot 2.0) \oplus (P_{29} \odot 3.0)}{P_{17} \oplus P_{28} \oplus P_{29}}.$$

Aus diesen Polynomen lassen sich schrittweise die Zeugenmengen und -basen und somit entsprechend auch eine Zeugenbasis für die gesamte Relation bestimmen: Der „`Note`“-Wert für das Tupel P_{15} lautet laut Polynom p_1 1.3, der Wert für P_{24} lautet 3.3, der für P_{25} lautet 2.0 und so weiter. Daraus lässt sich auch ohne weitere Informationen die Liste aller Noten rekonstruieren:

MatNr	ModNr	Semester	Note
10002	η_1	SS 20	1.3
10002	η_2	SS 20	3.3
10002	η_3	SS 20	2.0
10002	η_4	SS 20	1.0
10003	η_5	SS 20	1.3
10003	η_6	SS 20	1.7
10005	η_7	SS 20	1.0
10005	η_8	SS 20	2.0
10005	η_9	SS 20	3.0

Tabelle 4.4.: Die **how**-Provenance offenbart auch die konkreten Noten der `PRUEFUNG`-Relation.

4.2. Datenschutzprobleme bei where, why und how

Da wir nun wissen, welche Daten in welcher Dimension aus der **where**-, **why**- sowie **how**-Provenance rekonstruierbar sind, müssen wir uns die Frage nach dem Datenschutz stellen.

Datenschutzaspekte sind bei der **where**-Provenance zunächst vernachlässigbar, ganz einfach, weil keine schützenswerten Daten existieren. Die Durchschnittsnote ist ohnehin bekannt und unbedenklich. Auch eine Rekonstruktion der Originalrelation auf Basis der Zeugenliste führt uns nicht erneut zum Ergebnis oder zu einem der Zwischenergebnisse, da wir nicht wissen, wie die einzelnen Tupel miteinander zusammenhängen. Allerdings muss man sich ferner die Frage stellen, welche Daten im Sinne der Reproduzierbarkeit veröffentlicht werden müssen, beispielsweise in der Wissenschaft. Die **where**-Provenance liefert wahlweise die Menge aller am Ergebnis beteiligten Relationen oder aber die Menge aller beteiligten Tupel. Je nach Anfrage und Zusammensetzung der Datenbank kann dies für die Nachvollziehbarkeit oder sogar Reproduzierbarkeit von Anfrageergebnissen genügen. In einem solchen Fall genügt es, die Menge aller beteiligten Tupel zu veröffentlichen. Sollte die Nachvollziehbarkeit jedoch nicht gegeben sein,

beispielsweise weil ein natürlicher Verbund zweier Relationen mittels einer dritten, „mittleren“ Tabelle erfolgte, welche von der **where**-Provenance – wir sahen das in Unterabschnitt 2.1.1 – nicht erfasst wird, so müsste man in der Konsequenz alle Relationen veröffentlichen, die auf dem Rechenweg verwendet wurden. Da die **where**-Provenance diese jedoch nicht bestimmen kann, müsste man letztendlich die gesamte Datenbank preisgeben. Aus Sicht des Datenschutzes stellt dies natürlich den Worst Case dar. Ähnlich verhält es sich bei der relationenorientierten **where**-Provenance: Reichen die Informationen der Provenance-Anfrage nicht aus, um das Ergebnis nachvollziehen zu können, so müssen darüber hinaus weitere Daten veröffentlicht werden, um diese zu gewährleisten. Doch auch wenn diese gegeben ist, stellt hier natürlich bereits das Veröffentlichen einer gesamten Relation ein Problem dar, da – sofern nicht die gesamte Relation am Ergebnis beteiligt ist – (wesentlich) mehr Tupel veröffentlicht werden, als es nötig ist.

Bei der **why**-Provenance erhalten wir, siehe Abschnitt 4.1, auch die Anzahl der Tupel, die zur Ergebnisrelation gehören. An dieser Stelle ist es unter Einhaltung bestimmter Nebenbedingungen bereits möglich, den Datenschutz zu verletzen, nämlich genau dann, wenn

1. die Liste der Studierenden, die die Prüfung schrieben, in derselben Reihenfolge veröffentlicht wird, die auch vor der Aggregation der Werte genutzt wurde und
2. alle Noten, die eine Person erbracht hat, identisch sind: Sei N_P die Menge aller Noten, die von einer Person P erbracht wurden. Gilt dann $\forall n_i, n_j \in N_P : n_i = n_j$, so lautet der Durchschnitt ebenfalls n_i.

Dieser Fall ist allerdings grenzwertig und erfordert zusätzliches Wissen über die Streuung der Noten einer Person. Ist allerdings, beispielsweise aus statistischen Gründen, ein Streuungsmaß angegeben – zum Beispiel die Varianz – und ist dieses gleich 0, so ist es möglich, der jeweiligen Person all ihre erbrachten, identischen Noten zuzuordnen. Allerdings ist davon auszugehen, dass das Eintreffen dieses Falles in der Realität äußerst unwahrscheinlich ist.

Bei der **how**-Provenance ist es dank des Provenance-Polynoms möglich, die einzelnen Noten zu rekonstruieren. Betrachten wir Tabelle 4.4, so sehen wir, dass wir bereits ohne weitere Informationen die Matrikelnummern und die einzelnen Noten erhalten können. Allein letzteres kann bereits bedenklich sein, nämlich genau dann, wenn (1) die Liste aller Studierenden, die zu dieser Anfrage passen, (2) in derselben Reihenfolge veröffentlicht wird. Eine bijektive Zuordnung jedes Namen zu einer Note ist aufgrund der Gruppierung zwar nicht möglich, da ohne Matrikelnummer nicht bekannt ist, wo die Noten von Sarah Sonnenschein aufhören und die von Max Müller beginnen, für die erste Zeile ist dies jedoch zweifelsfrei möglich, was auch aus Tabelle 4.5 hervorgeht. Berücksichtigt man nun auch noch, dass wir ebenfalls die Matrikelnummern rekonstruieren können, so ist sogar eine Zuordnung aller Noten möglich, da wir ebenfalls wissen, wo die Noten von Sarah Sonnenschein aufhören und die von Max Müller beginnen. Gleiches gilt für den Übergang von Max Müller zu Ulrike Gebauer. Natürlich ist die Matrikelnummer selbst bereits personengebunden und erlaubt Rückschlüsse auf den Namen und Vornamen der dahinterstehenden Person, doch selbst, wenn diese bijektiv durch andere Identifikatoren ersetzt werden, ist diese Zuordnung möglich. Tabelle 4.6 illustriert dieses Problem.

Name	**Vorname**
Sonnenschein	Sarah
Müller	Max
Gebauer	Ulrike

Name	**Vorname**	**Note**
Sonnenschein	Sarah	1.3
?	?	3.3
?	?	2.0
?	?	1.0
?	?	1.3
?	?	1.7
?	?	1.0
?	?	2.0
?	?	3.0

Tabelle 4.5.: Unter den genannten Bedingungen ist eine Verknüpfung (nur) der ersten Zeilen möglich.

Name	**Vorname**	**MatNr-Pseudonym**	**Note**
Sonnenschein	Sarah	Pseudonym A	1.3
		Pseudonym A	3.3
		Pseudonym A	2.0
		Pseudonym A	1.0
Müller	Max	Pseudonym B	1.3
		Pseudonym B	1.7
Gebauer	Ulrike	Pseudonym C	1.0
		Pseudonym C	2.0
		Pseudonym C	3.0

Tabelle 4.6.: Unter dem Vorhandensein *irgendeines* Identifikators ist eine konkrete Zuordnung *aller* Noten zu der entsprechenden Person möglich.

Noch kritischer ist dies natürlich, wenn auch ohne der Anwesenheit von Identifikatoren eine (ein-)eindeutige Zuordnung möglich ist. Um dies zu erläutern, ziehen wir deshalb eine neue Anfrage heran, welche nur den Durchschnitt aller Noten einer Prüfung eines bestimmten Semesters ermittelt und jegliche andere Attribute dabei ignoriert:

```
SELECT AVG(Note) AS Durchschnitt
FROM PRUEFUNG
WHERE Semester = 'SS 20'
AND ModNr = 8
```

Anfrage 4.2: Die Durchschnittsnote aller Studierenden im Sommersemester 2020, die die Prüfung „Datenbanken“ absolvierten

	Note
P_{15}	1.3
P_{16}	1.3
P_{18}	3.0
P_{21}	1.0

Durchschnitt
1.65

Tabelle 4.7.: Ergebnis von Anfrage 4.2 inklusive Zwischenschritt vor der Bildung des Aggregats

Das Anfrageergebnis ist in Tabelle 4.7 zu sehen. Das entsprechende Provenance-Polynom lautet nun wie folgt:

$$p = \frac{(P_{15} \odot 1.3) \oplus (P_{16} \odot 1.3) \oplus (P_{18} \odot 3.0) \oplus (P_{21} \odot 1.0)}{P_{15} \oplus P_{16} \oplus P_{18} \oplus P_{21}}.$$

Daraus lässt sich problemlos die Liste aller Noten ableiten, die auch als Zwischenschritt vor der Aggregation in Tabelle 4.7 dargestellt ist.

Gehen wir nun davon aus, dass die Liste aller Studierenden, die diese Prüfung in diesem Semester schrieben, anderweitig veröffentlicht wird. Unter der Bedingung, dass die Reihenfolge identisch ist, lässt sich diese (`Name`, `Vorname`)-Tabelle dann problemlos neben die Tabelle von Anfrage 4.2 legen und ermöglicht eine 1:1-Zuordnung (siehe Tabelle 4.8). Man spricht dann von einem **Unsorted-Matching-Angriff** [PS17]. Im nachfolgenden Unterkapitel werden wir sehen, wie wir unter anderem dieses Problem lösen können.

Name	Vorname
Sonnenschein	Sarah
Müller	Max
Bach	Franziska
Müller	Mira

Note	
1.3	P_{15}
1.3	P_{16}
3.0	P_{18}
1.0	P_{21}

Tabelle 4.8.: Die Liste der betroffenen Studierenden, die anderweitig veröffentlicht wird (links) und die rekonstruierte Notenliste (rechts)

4.3. Mögliche Lösungsansätze

Nachdem wir uns nun angesehen haben, welche Probleme bei den verschiedenen Provenance-Anfragen auftreten können, stellt sich die Frage, was dagegen unternommen werden kann. Diese gilt es in diesem Abschnitt zu diskutieren und erste Lösungsansätze vorzustellen. Zunächst greifen wir dabei das Prinzip der Differential Privacy auf, welches wir in Abschnitt 3.3 kennenlernten, und untersuchen es hinsichtlich der Nützlichkeit als Problemlösung. Anschließend betrachten wir die Idee der intensionalen Provenance-Antworten und ihren Beitrag zum Datenschutz, bevor wir im letzten Schritt über diverse weitere Privacy-Ansätze diskutieren, darunter das Generalisieren und Unterdrücken von Tupeln sowie das Permutieren einzelner Provenance-Polynome.

4.3.1. Differential Privacy

Differential Privacy ist auf den ersten Blick ein vielversprechender Ansatz, da die Grundidee – etwas über die Grundgesamtheit, aber nicht über einzelne Entitäten zu lernen – den Datenschutz per Definition wahrt. In Kombination mit Data Provenance treten dabei jedoch diverse Probleme auf, welche wir uns in diesem Unterabschnitt ansehen werden. Gehen wir dazu einmal davon aus, dass Anfrage 4.3 an unsere Datenbank gestellt wird:

```
SELECT MatNr, AVG(Note) AS Durchschnitt
FROM PRUEFUNG
GROUP BY MatNr
ORDER BY MatNr ASC
LIMIT 5
```

Anfrage 4.3: Die Durchschnittsnoten sowie die Matrikelnummern der ersten fünf Studierenden

	MatNr	Durchschnitt	
R_1	10001	1.3	$\{P_1\}$
R_2	10002	1.92	$\{P_2, P_{15}, P_{24}, P_{25}, P_{26}\}$
R_3	10003	1.675	$\{P_3, P_{16}, P_{23}, P_{27}\}$
R_4	10004	1.3	$\{P_4\}$
R_5	10005	1.925	$\{P_5, P_{17}, P_{28}, P_{29}\}$

Tabelle 4.9.: Ergebnis von Anfrage 4.3

Uns interessiert also die Durchschnittsnote je Person, hier jeweils repräsentiert durch ihre Matrikelnummer. Da die einzelnen Provenance-Polynome recht umfangreich sind, wird an dieser Stelle auf die genaue Darstellung verzichtet. Wir beschränken uns hier auf die Zeugenlisten (siehe Tabelle 4.9). Prinzipiell gibt es nun zwei Möglichkeiten, um Differential Privacy anzuwenden: (1) Den Endergebnissen oder aber (2) den Rohdaten kann ein Rauschen beigefügt werden.

Fall 1: Wir verrauschen die Ergebnisse, in unserem Beispiel also die Werte der Spalte „Durchschnitt“. Gehen wir der Einfachheit halber einmal davon aus, dass unsere Anfragefunktion $\mathcal{K}(x)$ – die genaue Definition der Funktion sei an dieser Stelle egal – folgende Werte liefert:

tatsächlicher Wert	1.3	1.92	1.675	1.3	1.925
verrauschter Wert	1.5	2.0	1.5	1.2	2.0
absolute Differenz	0.2	0.08	0.175	0.1	0.075

Wir sehen diese auch in Tabelle 4.10. Sie ersetzen an dieser Stelle die konkreten Noten der „Durchschnitt“-Spalte:

	MatNr	**Durchschnitt**	
R_1	10001	1.5	$\{P_1\}$
R_2	10002	2.0	$\{P_2, P_{15}, P_{24}, P_{25}, P_{26}\}$
R_3	10003	1.5	$\{P_3, P_{16}, P_{23}, P_{27}\}$
R_4	10004	1.2	$\{P_4\}$
R_5	10005	2.0	$\{P_5, P_{17}, P_{28}, P_{29}\}$

Tabelle 4.10.: Ergebnis von Anfrage 4.3, modifiziert durch eine Anfragefunktion $\mathcal{K}$

Wir kennen nun nicht mehr die genauen, sondern nur noch die ungefähren Noten. Aus Sicht der Differential Privacy ist der Datenschutz verbessert worden. Allerdings verlieren wir dadurch die Möglichkeit, das Ergebnis zu validieren. Wir betrachten dazu exemplarisch das Provenance-Polynom p für das Tupel R_3 und die berechnete Teilrelation `PRUEFUNG`. Diese Teilrelation, dargestellt als Tabelle 4.11, hat nun die Aufgabe, das Ergebnis von 1.5 zu erklären, wird dem aber nicht gerecht: Berechnen wir den Durchschnitt manuell, so erhalten wir natürlich 1.675. Ohne weitere Informationen – beispielsweise das konkrete Rauschen – können wir dieses Ergebnis somit nicht validieren; das Speichern zusätzlicher Informationen würde allerdings das Prinzip der Differential Privacy aufheben. Differential Privacy kann also nicht auf die Endergebnisse angewandt werden, ohne die **how**-Provenance zunichte zu machen.

$$p = \frac{(P_3 \odot 2.7) \oplus (P_{16} \odot 1.3) \oplus (P_{23} \odot 1.0) \oplus (P_{27} \odot 3.0)}{P_3 \oplus P_{16} \oplus P_{23} \oplus P_{27}}$$

	MatNr	**ModNr**	**Semester**	**Note**
	...	...	...	...
P_3	10003	η_1	η_5	2.7
P_{16}	10003	η_2	η_6	1.3
P_{23}	10003	η_3	η_7	1.0
P_{27}	10003	η_4	η_8	3.0
	...	...	...	...

Tabelle 4.11.: Rekonstruktion der `PRUEFUNG`-Relation nach Anfrage 4.3

Fall 2: Wir verrauschen die Rohdaten. Aus Gründen der Übersichtlichkeit berechnen wir an dieser Stelle nur noch die Durchschnittsnote des Studenten mit der Matrikelnummer 10003 (siehe Anfrage 4.4 und Tabelle 4.12), verrauschen in einem Zwischenschritt jedoch die Werte:

```
SELECT AVG(Note) AS Durchschnitt
FROM PRUEFUNG
WHERE MatNr = 10003
```

Anfrage 4.4: Die Durchschnittsnote des Studenten mit der Matrikelnummer 10003

	MatNr	ModNr	Semester	Note
	...	...	...	...
P_3	10003	15	WS 19/20	2.7
P_{16}	10003	8	SS 20	1.3
P_{23}	10003	8	WS 20/21	1.0
P_{27}	10003	7	SS 20	1.7
	...	...	...	...

$\mathcal{K}$(Note)	abs. Differenz
...	...
2.55	0.15
1.5	0.2
0.87	0.13
1.85	0.15
...	...

Durchschnitt	
1.6925	$\{P_3, P_{16}, P_{23}, P_{27}\}$

Tabelle 4.12.: Berechnung und Ergebnis von Anfrage 4.4

Das Provenance-Polynom beinhaltet nun die verrauschten Werte. Allerdings verlieren wir in diesem Fall die Möglichkeit, die Originaldaten zu rekonstruieren: schließen wir vom Polynom auf die „originalen“ Tupel, so erhalten wir offensichtlich die verrauschten Werte als Werte der Spalte „Noten“. Ein Rückschluss von 2.55 auf 2.7, von 1.5 auf 1.3 und so weiter ist ohne weiteres nicht möglich, wie wir in Tabelle 4.13 und dem zugehörigen Polynom p sehen. In unserem Beispiel wären die Daten noch nicht einmal plausibel, da $0.87 < 1$ und somit keine zulässige Note ist. Auch hier ist der Gedanke naheliegend, dass wir uns weitere Informationen merken, beispielsweise die jeweiligen Differenzen (inkl. Vorzeichen) zwischen einer Note x und der verrauschten Note $\mathcal{K}(x)$, allerdings würde auch dies – wie bereits im ersten Fall – das Prinzip der Differential Privacy ad absurdum führen. Auch das Anwenden der Differential Privacy auf die Rohdaten macht somit die **how**-Provenance unbrauchbar.

$$p = \frac{(P_3 \odot 2.55) \oplus (P_{16} \odot 1.5) \oplus (P_{23} \odot 0.87) \oplus (P_{27} \odot 1.85)}{P_3 \oplus P_{16} \oplus P_{23} \oplus P_{27}}$$

	MatNr	ModNr	Semester	Note
	...	...	...	...
P_3	10003	η_1	η_5	2.55
P_{16}	10003	η_2	η_6	1.5
P_{23}	10003	η_3	η_7	0.87
P_{27}	10003	η_4	η_8	1.85
	...	...	...	...

Tabelle 4.13.: Das Provenance-Polynom p führt uns nicht zu den originalen Werten

4.3.2. Intensionale Provenance-Antworten

Ein weiterer Ansatz zur Einhaltung des Datenschutzes stellen **intensionale Provenance-Antworten** dar. Während extensionale Antworten grundsätzlich die beteiligten Rohdaten liefern, liefern intensionale Antworten nur *„Information[en] über die Daten in der Datenbank und nicht die Daten selbst“* [Sva16]. Die Wahrscheinlichkeit, den Schutz einer einzelnen Person zu verletzen, ist offensichtlich geringer, wenn wir nur allgemeine Informationen über einen Datensatz statt konkrete Teile dieses Datensatzes erhalten, weshalb es intensionale Provenance-Antworten wert sind, diskutiert zu werden.

Eine Möglichkeit, eine Provenance-Anfrage intensional zu beantworten, haben wir in Unterabschnitt 2.2.2 bereits kennengelernt: die **Generalisierung**. Indem die Informationen der Rohdaten generalisiert, also in unschärfere Gruppen kategorisiert werden, erfahren wir etwas *über* die Daten, aber nicht, wie die Daten konkret aussehen. Dabei können **Konzepthierarchien** zum Einsatz kommen. Eine solche Hierarchie sehen wir in Abbildung 4.1: Die Rohdaten, in dem Fall Jahreszahlen, werden in einem ersten Schritt zu Fünf-Jahres-Intervallen zusammengefasst. Sollte dies bereits genügen, um eine hinreichende k-Anonymität zu gewährleisten, sind wir an dieser Stelle fertig. Falls nicht, gehen wir über zur nächsten Stufe und beschränken uns auf die Jahrzehnte. Sollte auch dies nicht genügen, so nutzen wir die **any**-Wurzel und entfernen die Werte in der Konsequenz vollständig (bzw. ersetzen jeden Wert durch „any“). Zwischen Stufe 1 und 2 sind jedoch noch weitere Zwischenschritte denkbar; so können wir beispielsweise auch noch 20- und 50-Jahres-Intervalle nutzen, bevor wir an die Wurzel der Hierarchie gelangen.

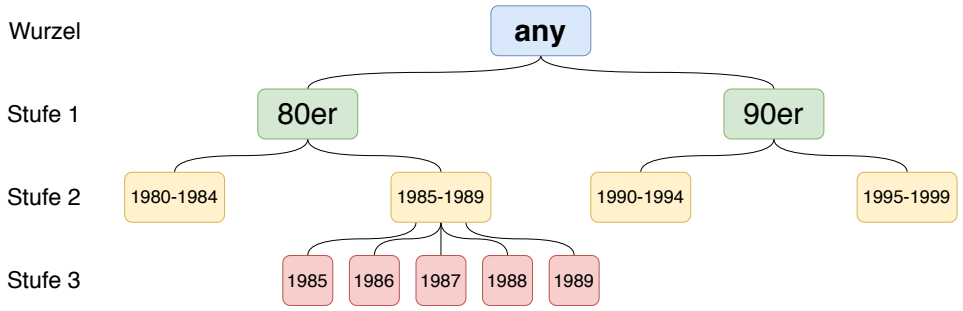


Abbildung 4.1.: Beispiel einer Konzepthierarchie für Jahreszahlen. Aus Gründen der Übersichtlichkeit sind nur fünf der 20 Rohdaten dargestellt.

Bei numerischen Werten funktioniert eine solche Unterteilung ad-hoc. Schwierig wird es bei allen anderen Datentypen, da die Generalisierung in diesen Fällen meist inhaltlich erfolgen muss. Hier müssen wir eine solche Konzepthierarchie vorweg manuell festlegen oder von gewissen *Ontologien* Gebrauch machen. An dieser Stelle sei auf die Bachelorarbeit *„Intensional Answers for Provenance Queries in Big Data Analytics“* von *Jan Svacina* verwiesen, welcher sich mit genau diesem Problem auseinandersetzte. Neben der Generalisierung können jedoch auch textuelle Beschreibungen oder sonstige Erklärungen als Antwort auf eine intensionale Provenance-Anfrage dienen. Generell muss jedoch darauf geachtet werden, dass die Antworten dennoch detailliert genug sind, um die wissenschaftliche Nachvollziehbarkeit bzw. Reproduzierbarkeit oder sogar Rekonstruierbarkeit gewährleisten zu können. Bei intensionalen Antworten ist dies zunächst nicht der Fall: Auf einer aggregierten Ebene kann nur noch die Plausibilität eines Ergebnisses überprüft werden, indem geprüft wird, ob jeweils innerhalb eines angegebenen Wertebereichs (beispielsweise „90er“) ein Wert existiert, der zu diesem Ergebnis führt. Das Testen aller möglichen Kombinationen ist jedoch entsprechend aufwändig.

4.3.3. Weitere Lösungsansätze

Generalisieren und Unterdrücken Eine weitere Möglichkeit, um Privacy-Aspekte zu bewahren, kann das **Generalisieren** bzw. **Unterdrücken von Tupeln** sein. Als nützliche Maßstäbe dienen hier die bereits bekannte k-Anonymität sowie die l-Diversität (siehe Abschnitt 2.2). Wir sahen das bereits im vorherigen Unterabschnitt, konzentrieren uns an dieser Stelle jedoch wieder auf extensionale Antworten. Die Grundidee hierbei ist, dass Tupel der Ergebnisrelation so weit verallgemeinert oder vollständig entfernt werden, bis die Relation die zwei genannten Kriterien erfüllt. Erst danach werden die Provenance-Informationen berechnet. Auf den ersten Blick lassen sich etwaige Datenschutzprobleme damit lösen: Dadurch, dass als Grundlage ein Datensatz mit geringerem Informationsgehalt dient, können im Zweifel auch weniger sensitive Informationen offenbart werden. Werden k-Anonymität und l-Diversität erfüllt, so verringert sich die Wahrscheinlichkeit, dass einzelne Personen dieses Datensatzes identifiziert werden können, immens. Allerdings sehen wir uns hierbei zunächst mit demselben Problem wie bei der Differential Privacy in Abschnitt 4.3.1 konfrontiert: Generalisieren und unterdrücken wir Tupel der Originaldaten und arbeiten mit diesen modifizierten Daten weiter, so erhalten wir möglicherweise ein Ergebnis, welches uns – bei der Rückverfolgung dessen – zu anderen, nämlich den originalen, Daten führt. Anders als bei der Differential Privacy, wo die Daten jedes Mal neu verrauscht werden, können wir hier jedoch prinzipiell mit einer k-anonymen und l-diversen Kopie des Datensatzes arbeiten. Diese würde persistent, also dauerhaft, gespeichert werden. Die Idee ist somit, sowohl die Datenbankanfrage als auch die Provenance-Berechnungen nicht auf der tatsächlichen, sondern auf einer zweiten, anonymisierten Relation auszuführen. Gleichzeitig entsteht dadurch der Vorteil, dass wir die am Ergebnis beteiligten Tupel bedenkenlos veröffentlichen können, da sie dann bereits anonymisiert vorliegen.

Permutieren Wir sahen in Kapitel 4.2, dass die Reihenfolge der Teilpolynome möglicherweise der der Ausgangsrelation(en) entsprechen kann, was dazu führt, dass die ursprüngliche Reihenfolge direkt aus dem gesamten Polynom abgelesen und ein sogenannter **Unsorted-Matching-Angriff** ermöglicht werden kann. Um dies zu vermeiden, können die einzelnen Polynome innerhalb des Provenance-Polynoms permutiert werden, idealerweise fixpunktfrei (alle Polynome erhalten eine neue Position). Betrachten wir erneut Anfrage 4.2 sowie Tabelle 4.7. Diesmal stellen wir das Provenance-Polynom jedoch nicht gemäß der Reihenfolge des Zwischenschritts auf, sondern wählen eine zufällige (siehe p'). Eine richtige Zuordnung ist nun allein anhand des Provenance-Polynoms nicht mehr möglich, wie wir in Tabelle 4.14 sehen. Eine solche fixpunktfreie Permutation ist allerdings nur dann möglich, wenn die Anzahl voneinander verschiedener Werte entsprechend groß ist. Gäbe es in diesem Beispiel eine weitere Person, welche ebenfalls eine 1.3 erhielt, so ließen sich natürlich immer noch alle Polynome permutieren, (mindestens) eine Person hätte hinterher aber den gleichen Wert wie vor der Permutation – nämlich eben die Note 1.3. Ihr Datenschutz wäre dann nicht gewährleistet.

$$p' = \frac{(P_{21} \odot 1.0) \oplus (P_{18} \odot 3.0) \oplus (P_{16} \odot 1.3) \oplus (P_{15} \odot 1.3)}{P_{21} \oplus P_{18} \oplus P_{16} \oplus P_{15}}.$$

Name	Vorname
Sonnenschein	Sarah
Müller	Max
Bach	Franziska
Müller	Mira

Note	
1.0	P_{21}
3.0	P_{18}
1.3	P_{16}
1.3	P_{15}

Tabelle 4.14.: Die Liste der betroffenen Studierenden, die anderweitig veröffentlicht wird (links) und die rekonstruierte Notenliste (rechts), allerdings in falscher Reihenfolge

Wir beenden dieses Kapitel mit dem Zwischenfazit, dass das Prinzip der Differential Privacy nicht mit den Prinzipien der Data Provenance einhergeht. Mögliche Lösungen hingegen können intensionale Provenance-Antworten, Methoden zur Generalisierung und Unterdrückung von Tupeln der originalen Relationen sowie das Permutieren einzelner Teilpolynome innerhalb der gesamten Provenance-Polynome sein. Im nächsten Kapitel werden wir 20 Expertinnen und Experten bezüglich ihrer Auffassung von Provenance und Privacy sowie den Umgang mit ihren Forschungsdaten befragen.

5. Das Experteninterview

In diesem Kapitel beschäftigen wir uns mit der Befragung der Expertinnen und Experten. Zunächst klären wir in Abschnitt 5.1, wieso die Wahl der Datenerhebung auf das Leitfadeninterview und nicht auf den standardisierten Fragebogen fällt. Im nächsten Schritt wird dann in Abschnitt 5.2 der Fragenkatalog vorgestellt und dabei jede Frage einzeln begründet. Im letzten Schritt erfolgt in Abschnitt 5.3 dann die Auswertung der erhobenen Daten.

5.1. Vorbereitung und Durchführung

Zunächst ist die Frage nach der Art der Befragung zu klären. An dieser Stelle entscheiden wir uns für ein Leitfadeninterview, also eine qualitative Befragung. Die Gründe dafür liegen in der Art der Daten, die gesammelt werden sollen: quantitative Daten, egal ob nominal, ordinal oder kardinal, sind eher zweitrangig; viel eher stehen persönliche Definitionen und Interpretationen im Vordergrund, die mit einem Fragebogen nur bedingt erfasst werden können. Zudem können die Antworten zwischen je zwei Personen so unterschiedlich sein, dass Anschlussfragen, die vorher mangels Informationen gar nicht absehbar waren, notwendig werden. Diese können in einem standardisierten Fragebogen natürlich nicht gestellt werden.

Zu Beginn wurden Interviews mit vier Personen vereinbart. Die Stichprobe sollte dann nach und nach auf zehn Personen erweitert werden, wobei deren Fachbereiche je nach den vorherigen Ergebnissen ausgewählt wurden. Nachdem dann Interviews mit zehn Expertinnen und Experten durchgeführt wurden, entschieden wir uns dazu, die Stichprobe auf insgesamt 20 Teilnehmer/-innen zu erweitern, um bei ungefähr gleicher Gewichtung weitere Personen aus anderen, nichtwissenschaftlichen Bereichen befragen zu können. Als **Gatekeeper** fungierte die Betreuerin dieser Bachelorarbeit, *Tanja Auge (M. Sc.)*.

Das Interview ist in drei grobe Bereiche unterteilt: Wir beginnen es mit einigen persönlichen Fragen (Fragen 1 bis 6). Diese dienen der späteren Klassifikation der befragten Personen in einzelne Gruppen (je nach Merkmal; Schnittmengen sind dabei möglich). Anschließend gibt es einen Frageblock zu den Themen Provenance und Privacy (Fragen 7 bis 10). Hierbei wird zunächst nach allgemeinen Definitionen dieser zwei sehr weitläufigen und bewusst vage formulierten Begriffe gefragt, bevor es um das Verständnis von Datenschutz der einzelnen Personen geht. Der dritte und letzte Frageblock (Fragen 11 bis 15) befasst sich dann mit dem Thema Forschungsdaten. Hierzu zählen das Forschungsdatenmanagement sowie der Umgang mit und die Veröffentlichung von Forschungsdaten. Zum Schluss wird mit Frage 16 noch einmal nach weiteren Anmerkungen und Gedanken gefragt, die die befragte Person eventuell noch zum Interview hinzufügen möchte.

Genau wie Mayer es in [May13] beschreibt, beginnt unser Experteninterview zunächst mit einem **Pretest**. Dazu werden zwei Personen aus dem Informatik-Forschungsbereich probeweise interviewt. Dieser Test offenbart keine größeren Diskrepanzen zwischen den Fragen und den erwarteten Antworten, weshalb ausschließlich Frage 15 leicht abgewandelt wird.

5.2. Der Fragenkatalog

Wie bereits im vorherigen Abschnitt erwähnt, lässt sich der Fragenkatalog in drei Teile untergliedern. Mit den ersten sechs Fragen wollen wir die persönliche Situation der jeweiligen Teilnehmerin oder des jeweiligen Teilnehmers klären. Dazu befragen wir sie/ihn unter anderem nach ihrem/seinem Forschungsbereich, dem Status an der Universität bzw. dem Beruf und der Art und Dimension der anfallenden Daten. Im zweiten Teil stellen wir dann Fragen zu den Begriffen „Provenance“ und „Privacy“ und wollen herausfinden, was die Personen damit konkret assoziieren. Außerdem wollen wir das Verständnis von Datenschutz der jeweiligen Person untersuchen und präsentieren ihr dafür auch einen problematischen Datensatz. Zu guter Letzt wollen wir dann das Verhältnis der Teilnehmerin oder des Teilnehmers zu Forschungsdaten und den Umgang mit diesen ermitteln.

1) Woran forschst du? (In welchem Bereich?)
Die Frage dient der Einordnung der Antworten in eine bestimmte Wissenschaft oder ein Teilgebiet einer solchen.

2) Welchen Status hast du an der Universität, bzw. welchen Beruf?
Diese Frage dient der Klassifikation der Erfahrung in der Forschung. Studierende haben üblicherweise weniger Erfahrung als Professorinnen und Professoren. Ist die Person außerhalb der Universität beruflich tätig, ist der Umgang mit Daten eventuell auch ein ganz anderer als universitätsintern. Zumindest in Bezug auf die Datenspeicherung erwarten wir hier deutliche Unterschiede.

3) Was für Daten fallen bei der Forschung an?
Daten können in den unterschiedlichsten Formen anfallen. Aufgezeichnete Interviews sind ebenso Daten wie physikalische Messwerte, der Umgang damit kann jedoch vollkommen unterschiedlich sein.

4) Welche Datenmengen fallen an?
Je nach Dimension (Kilo-, Mega-, Giga-, Terabyte, ...) ist der Umgang mit den Daten wahrscheinlich ein anderer. Einige Kilobyte bis wenige Megabyte dürften noch recht einfach zu verwalten sein, während Datenmengen, die unter den Begriff der „Big Data“ fallen – also Terabyte oder mehr –, unmöglich zu überschauen sind.

5) Wie lange werden die Daten gespeichert?
Werden Primärdaten nur für eine kurze Zeit gespeichert, entfallen große Aspekte des Forschungsdatenmanagements für ebendiese Daten.

6) Wer übernimmt die Speicherung?
Dies ist auch eine Frage nach der Verantwortung. Übernimmt die Person die Speicherung selbst, behält sie die volle Kontrolle über die Daten, ist aber auch für alles selbst verantwortlich. Externe Anbieter knüpfen die Speicherung möglicherweise an gewisse Rahmenbedingungen, garantieren dafür aber auch die Verwaltung der Daten.

7) Was verstehst du unter Provenance? Was verstehst du unter Privacy?
Diese zwei Fragen dienen der allgemeinen Findung einer Definition der beiden Begriffe und sind deshalb auch absichtlich vage gehalten. So ist weder von „Data Provenance“ noch von „Datenschutz“ im Speziellen die Rede.

8) Kannst du dir auch Situationen vorstellen, in denen „Privacy“ Daten meint, die nicht personengebunden sind? Falls ja, welche?
Die Frage dient als mögliche Anschlussfrage an die erfragte Privacy-Definition. Erwartungsgemäß werden viele Personen zunächst an personengebundene Daten denken, weshalb diese Frage in solchen Situationen versucht, die Definition auszuweiten.

Beispiel
Um zu erforschen, wie gut das Verständnis von Datenschutz der befragten Person ist, wird ihr ein Auszug eines Beispiel-Datensatzes präsentiert.

PLZ	Geburtsdatum	Geschlecht	Diagnose
18059	06.03.1998	männlich	Influenza
18055	21.09.1995	weiblich	Depressionen
18106	29.02.1994	männlich	Herzinfarkt
...	...	...	...

Tabelle 5.1.: Auszug eines fiktiven medizinischen Datensatzes

9) Kann ich diesen Datensatz so veröffentlichen? / Diesen Datensatz kann ich ja so veröffentlichen, richtig?
Der Datensatz ist aus Sicht des Datenschutzes höchst problematisch: Zwar fehlen Vor- und Nachnamen, aber es besteht keine k-Anonymität und somit auch keine l-Diversität. Zudem war es um die Jahrtausendwende herum möglich, 87% aller in den USA lebenden Personen nur anhand der Kombination aus Postleitzahl (ZIP-Code), Geburtsdatum und Geschlecht eindeutig zu identifizieren, wie *Latanya Sweeney* 2000 herausfand [Swe00]. Noch dazu ist eine medizinische Diagnose ein höchst privates und daher sehr schützenswertes Attribut. Diese Frage ist bewusst als „Fangfrage“ formuliert, um die Antwort nicht versehentlich in eine bestimmte Richtung zu lenken. Die Person sollte die Probleme des Datensatzes erkennen und an dieser Stelle widersprechen.

10) Was könnte ich denn tun, um diesen Datensatz dennoch veröffentlichen zu können?
Dies ist eine Anschlussfrage an die vorherige, falls die Antwort „nein“ lautet. Hiermit soll untersucht werden, ob und inwiefern die Experten mit Techniken der Anonymisierung wie beispielsweise der Generalisierung und Unterdrückung von Tupeln vertraut sind.

11) Was sind Forschungsdaten?
Der Begriff der Forschungsdaten kann sehr weit gefasst werden. Auch können diese je nach wissenschaftlichem Bereich vollkommen unterschiedlich aussehen.

12) Worin besteht Forschungsdatenmanagement?
Diese Frage hat das Ziel, das Verständnis des Forschungsdatenmanagements zu untersuchen. Üblicherweise zählen dazu die langfristige und nachhaltige Speicherung sowie die Aufbereitung der Forschungsdaten, damit diese auch von außenstehenden bzw. an dem Forschungsprojekt unbeteiligten Personen verstanden und ggf. verwendet werden können.

13) Unter welchen Bedingungen und in welchem Umfang würdest du einer anderen Person Einsicht in deine Forschungsdaten gewähren, nachdem du über diese publiziert hast?
Das eine Extrem stellt das vollständige Veröffentlichen sämtlicher (erhobener) Forschungsdaten dar. Das andere Extrem ist das Zurückhalten aller Daten. Es ist davon auszugehen, dass sich die befragten Personen für einen Mittelweg entscheiden werden/würden – diesen Mittelweg gilt es mit dieser Frage zu ermitteln.

14) Besteht ein Interesse daran, Forschungsdaten nicht zu veröffentlichen, falls ja, wann?
Die Frage dient als Anschlussfrage an die vorherige. Das Veröffentlichen der Forschungsdaten kann im Widerspruch zu bestimmten Interessen stehen, welche an dieser Stelle ermittelt werden sollen.

15) Steht das für dich im Widerspruch zur Idee von Open Science? Wie stehst du dazu?
Die Idee hinter Open Science ist, dass sowohl wissenschaftliche Publikationen als auch die zugrundeliegenden Forschungsdaten frei zugänglich gemacht werden – das heißt kostenlos und ohne Beschränkung. Werden Daten jedoch zurückgehalten, so widerspricht dies der Open-Science-Idee. Hier ist zu ermitteln, inwieweit dies aus Sicht der Teilnehmer/-innen problematisch ist – oder eben nicht.

16) Hast du selbst Fragen oder Anmerkungen, die du an dieser Stelle noch loswerden möchtest?
Für den Fall, dass die Person noch weitere Gedanken hat, die durch keine der vorherigen Fragen abgedeckt wurden, hat sie an dieser Stelle die Gelegenheit, sie frei zu äußern.

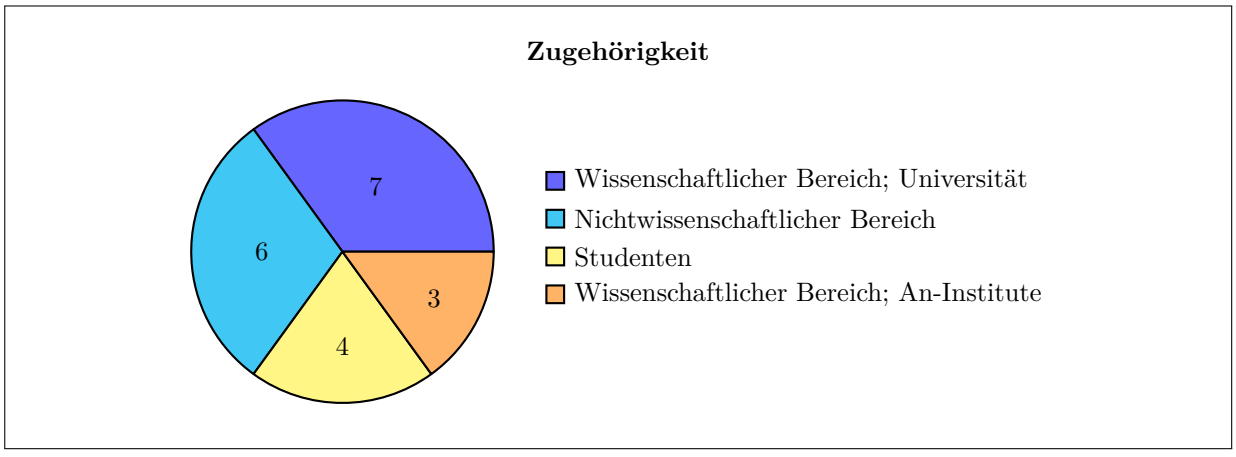
5.3. Die Auswertung der Interviews

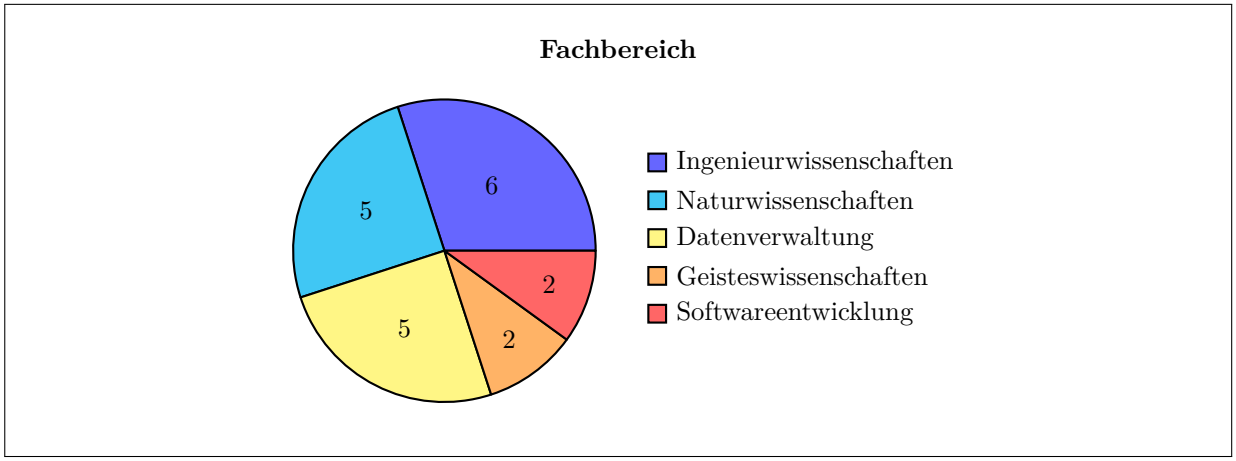
Zunächst betrachten wir einmal detailliert, aus welchen Gruppen die befragten Personen stammen. Insgesamt haben wir 20 Interviews geführt, wovon die Hälfte mit Wissenschaftlerinnen und Wissenschaftlern und die andere Hälfte mit Studenten und Personen außerhalb der Wissenschaft geführt wurden. Ein Viertel der Befragten sind hierbei im Bereich der Datenverwaltung angesiedelt, zwei sind in der Softwareentwicklung tätig. Die Übrigen lassen sich einer konkreten Wissenschaft zuordnen.

Sofern einzelne Aussagen zitiert werden, wird die Quelle durch die Angabe eines „IPxx“-Kürzels erkenntlich gemacht. „IP“ steht dabei für „Interviewpartner/-in“, wobei jede/-r Interviewpartner/-in vorweg eine eindeutige Nummer erhielt. In einem Fall wurde ein Interview mit zwei Personen geführt (wovon eine Person eine eher nebensächliche Rolle einnahm); an dieser Stelle ist von „IP10.1“ für Person 1 und „IP10.2“ für Person 2 die Rede.

Einige Antwortmengen lassen sich gut kategorisieren, insbesondere die der ersten sechs Fragen, und eignen sich deshalb gut für grafische Darstellungen. Wann immer eine ordinale Skalierung vorliegt, beispielsweise bei einer Zeit- oder Volumenangabe, wird die Verteilung der Antworten durch Balkendiagramme dargestellt. Liegt hingegen eine nominale Klassifikation vor, so werden Kreisdiagramme verwendet.

Art der Teilnehmer/-innen nach...



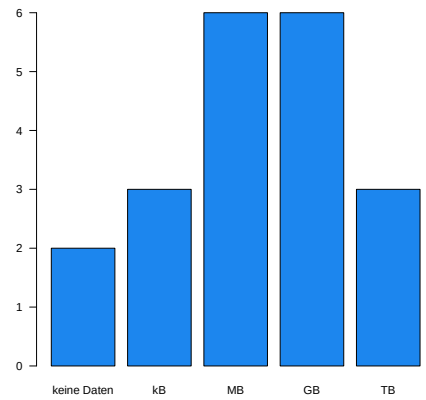


Was für Daten fallen an?

- keine (2)
- Datenbanken (2)
- Publikationen (2)
- Paketdaten/Netzwerk-Logs (2)
- Personendaten (2)
- Simulations- und Modelldaten (2)
- Forschungsdaten anderer bzw. des Instituts (4)
- Sensordaten/Messwerte/Zeitreihen (7)
- Interviews und deren Transkripte
- Excel-Tabellen
- Quellcode
- annotierte PDF-Dateien
- aggregierte Ergebnisse
- Merkmale von Tieren
- Filmmaterial

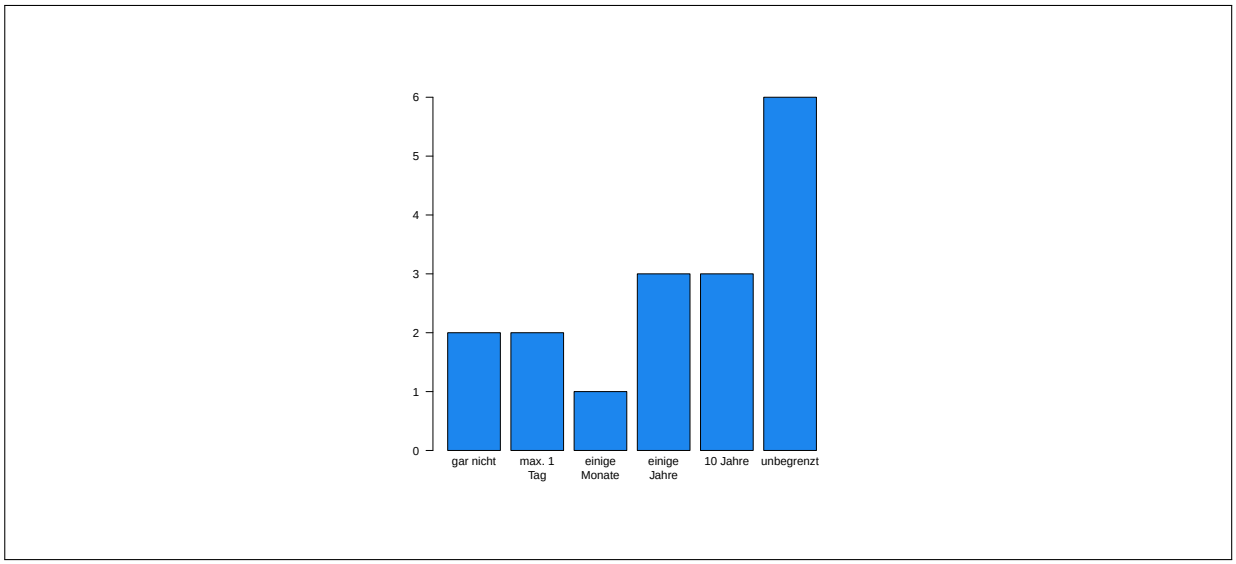
Hat eine Person mehrere Arten von Daten genannt, so haben wir diese auch einzeln erfasst, weshalb unsere Liste auch mehr als 20 Antworten umfasst. Ähnliche Antworten werden zusammengefasst; die Anzahl, wie oft diese Antwort vorkommt, steht jeweils dahinter. Fehlt diese Angabe, so kam sie nur ein Mal vor. Auffällig ist, dass viele der befragten Personen mit unstrukturierten Daten arbeiten. Strukturierte Daten hingegen, wie etwa (relationale) Datenbanken, sind kaum vertreten.

Welche Datenmengen fallen an?



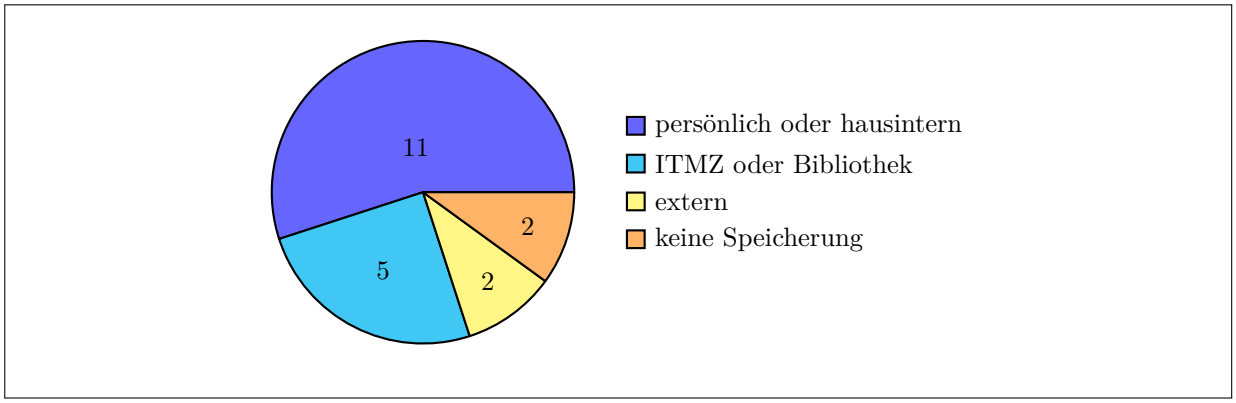
Bei den Datenmengen fällt auf, dass vor allem mit Mega- und Gigabyte an Daten gearbeitet wird. Wurde hier ein Intervall genannt, beispielsweise „Kilobyte bis Megabyte“, so wird die obere Grenze als Antwort gewertet, in diesem Beispiel also Megabyte.

Wie lange werden diese Daten gespeichert?



An dieser Stelle fällt klar auf, dass viele Personen oder deren Einrichtungen ihre Daten für eine unbegrenzte Zeit speichern. Ebenfalls oft – angesichts der kleinen Stichprobe – werden 10 Jahre genannt. Dies ist der Zeitraum, der von diversen Forschungsorganisationen, beispielsweise der *Deutschen Forschungsgemeinschaft*, vorgegeben wird. Drei Antworten sind in der Grafik nicht zu sehen: eine Person speichert die Daten je nach Unternehmen beliebig lange und kann keinen konkreten Zeitraum nennen, eine Person speichert die Daten bis zur Veröffentlichung einer Publikation und eine weitere Person gibt gar keinen Zeitraum an.

Wer übernimmt die Speicherung?



Bei der Frage, wer die Speicherung der Daten übernimmt, gibt es eine klare Tendenz zur persönlichen oder hausinternen Speicherung; mehr als die Hälfte der Befragten speichern ihre Daten selbst oder an der jeweiligen Einrichtung. Ein Viertel der Befragten nutzt die Angebote der Universitätsbibliothek oder des IT- und Medienzentrums der Universität. Lediglich zwei der 20 Befragten greifen auf externe Anbieter zurück; in einem Fall wird die Speicherung von dem Institut übernommen, welches die Rohdaten zur Verfügung stellt, im anderen Fall übernimmt das jeweils involvierte Unternehmen die Speicherung. „ITMZ oder Bibliothek“ trifft als Kategorie nur dann zu, wenn eine der beiden Einrichtungen mit der Speicherung beauftragt werden, nicht aber für Personen, die dort arbeiten und für die Speicherung verantwortlich sind.

An dieser Stelle ist die persönliche Situation der befragten Person klar. Als nächstes folgt deshalb der Fragenblock zu den Themen Provenance und Privacy.

Was verstehst du unter dem Begriff „Provenance“?

Die Antworten auf die Frage nach dem Provenance-Begriff fallen sehr vielfältig aus. Ein Viertel der Befragten gibt an, den Begriff noch nie gehört zu haben oder sich gar nichts darunter vorstellen zu können. Sieben der 20 Personen assoziieren ihn mit der „Herkunft von Daten“, vier Personen mit der „Herkunft von Dingen im Allgemeinen“. Als Beispiele werden hier Bücher, Kunstwerke oder aber Lebensmittel genannt. Die restlichen Antworten fallen sehr individuell aus. So beschreibt Provenance:

- in welchem Kontext Daten erhoben wurden;
- zu wissen, welche Schritte Daten oder andere Dinge durchlaufen sind;
- die Dokumentation des Weges, den etwas genommen hat;
- nachvollziehen zu können, wie eine datenbasierte Entscheidung zustande gekommen ist oder wie genau ein Wert generiert wurde;
- wie man Daten sammelt; Akquisition von Daten;
- Vertraulichkeit, Vertrauensbasis von Informationen (als Beispiel wurden Nachrichtenquellen genannt);
- wie Ergebnisse zustande kamen.

In einigen Fällen gaben die befragten Personen mehrere Antworten. Diese werden einzeln gezählt, weshalb die Anzahl aller Antworten wiederum größer als die Anzahl aller Befragten ist. Zudem wurden ähnliche Antworten entsprechend gruppiert. Selbiges gilt auch für die nächste Frage.

Was verstehst du unter dem Begriff „Privacy“?

14 Befragte und somit 70% assoziieren den Begriff mit dem Datenschutz (10) oder dem Schutz der Privatsphäre eines Menschen (4). Vier Personen verstehen unter „Privacy“ den Schutz jeglicher Daten. Zwei Befragte verstehen darunter das Verhindern von Rückschlüssen auf einzelne Menschen, eine Person assoziierte den Begriff mit Datenminimalität und -sparsamkeit sowie informationeller Selbstbestimmung (konkreter: dass Menschen Einsicht in auf sie bezogene Daten erhalten). Zwei Antworten stechen besonders hervor, da sie den Privacy-Begriff überhaupt nicht mit Personendaten verbinden:

> *„Die Möglichkeit, dafür zu sorgen, dass etwas nicht öffentliches Wissen wird.“*
> — IP08

> *„Wie mit den Daten umgegangen wird, wer darauf Zugriff hat. Ob überhaupt jemand darauf Zugriff hat. Was für Vorkehrungen es gibt, dass niemand unbefugten Zugriff kriegt. Wie lange sie gespeichert werden, bei wem auch immer.“*
> — IP13

Insgesamt gibt es somit eine klare Tendenz zu personenbezogenen Daten und dem Schutz von Personen, eindeutig ist dies allerdings nicht. Auf die Anschlussfrage, ob der Privacy-Begriff auch bei Daten ohne Personenbezug verwendet werden kann, antworten fünf Befragte mit Nein, zwei waren sich nicht sicher. Zwei Personen beziehen sich jedoch auf Daten, aus denen ein Patent entstehen kann, eine bezog sich auf Firmeninteressen und vertraglichen Vereinbarungen, die schützenswert seien, eine bezog sich auf Militär- und Regierungsdaten. Eine weitere Person denkt an allgemein schützenswerte Daten und weichte damit die zuvor genannte Privacy-Definition wieder auf. Zudem unterscheidet eine befragte Person auch zwischen personenbezogenen und persönlichen Daten und antwortete mit Daten, die zwar keinen Personenbezug

haben, aber dennoch persönlich seien. Als Beispiel wird das unveröffentlichte Manuskript eines Romans genannt. Auch hier gibt es eine besonders auffällige Antwort:

> *„Das würde ich jetzt nicht als Privacy bezeichnen, aber schon als Datenschutz.“*
> — IP20

Der bewusst vage formulierte „Privacy“-Begriff ist also unpassend, der „Datenschutz“-Begriff trifft jedoch zu. Privacy und Datenschutz sind aus Sicht der befragten Person somit zwei verschiedene Dinge.

Beispieldatensatz

Bis auf eine Person erkennen alle Teilnehmer/-innen (und somit 95%) die Probleme, die mit diesem Datensatz einhergehen. Die Frage, ob die 87%, die Latanya Sweeney in [Swe00] ermittelte, überraschen, beantworten zehn Befragte und somit die Hälfte mit Ja, die andere Hälfte antwortet mit Nein. Daraus lässt sich keine Erkenntnis ableiten. Anders sieht dies bei den Möglichkeiten zur Manipulation des Datensatzes aus:

Was könnte man tun, um den Datensatz dennoch veröffentlichen zu können?

- Generalisieren (16)
- Spalten entfernen (10)
- Kodieren (4)
- Slicing (1)
- Masking (1)

Als mögliche Anonymisierungstechnik wird neben dem Weglassen ganzer Spalten (10) in nahezu allen Fällen (16) die **Generalisierung** empfohlen. Sehr häufig wird vorgeschlagen, das Geburtsdatum in größere Kategorien zu unterteilen, indem man nur das Jahr oder alternativ das Alter der Person angibt. Eine einzelne Person nennt auch **Slicing** – die Werte nicht-korrelierender Spalten werden permutiert, korrelierende Spalten bleiben unverändert –, eine andere nennt **Masking** – einzelne, zu eindeutige Attribute werden maskiert – als mögliche Technik. Die Generalisierung scheint als Datenschutz-Methode also verbreitet zu sein.

Was sind Forschungsdaten?

- Alles, was im Forschungsprozess entsteht (10)
- Alles, was im Forschungsprozess aggregiert/verknüpft wird (4)
- Alles, was für Forschung genutzt werden kann (3)
- Alles, was als Grundlage für Untersuchungen/Erkenntnisgewinn dient (3)
- Alles, was für Forschungsprozesse relevant ist – auch Literatur; alle Informationsquellen (2)

Tendenziell werden unter Forschungsdaten alle Daten verstanden, die während der Forschung neu geschaffen werden (10). Ebenfalls insgesamt 10 Personen sehen aber auch bereits vorhandene Daten als Forschungsdaten, wenn sie für die Forschung herangezogen werden. Zwei Personen gehen noch einen Schritt weiter und verstehen darunter auch die verwendeten Informationsquellen, einschließlich der Literatur, die für eine etwaige Recherche genutzt wurde. Gelegentlich wird gefragt, ob Aggregate der Forschungsdaten wiederum Forschungsdaten seien. Zwei Antworten sind dabei besonders hervorzuheben, da sie die Relevanz von Data Provenance unterstreichen:

> *„Es ist auf jeden Fall ein wertvolleres Forschungsdatum, wenn es Provenance hat.“*
> — IP08

> *„Sofern der Prozess erkennbar ist, wie die Daten von A nach B gekommen sind und Sie die Möglichkeit haben, auch wieder von B nach A zu kommen, ist es ja einfach nur eine Modifizierung des Kerndatensatzes. Also ja.“*
> — IP10.2

Worin besteht Forschungsdatenmanagement?

Die Antworten auf die Frage, worin Forschungsdatenmanagement besteht, fallen vollkommen unterschiedlich aus. Eine Klassifikation der Antworten ist hier kaum möglich. Sehr oft wird ein Bezug zum Schutz personenbezogener Daten hergestellt, was jedoch eine Auswirkung der vorherigen Fragen sein könnte. Weiterhin wird oft das sichere Speichern dieser Daten genannt. Gemeint ist hier zum einen eine verlustfreie Aufbewahrung, ermöglicht durch Backups und anderen Methoden zur Datenkonservierung, zum anderen das Absichern der Daten nach außen, sodass unbefugte Personen keinen Zugang dazu erhalten können. Auch das Strukturieren von Daten und das Hinzufügen von Metadaten wird oft genannt, in einigen Fällen auch im Zusammenhang mit der Veröffentlichung der Forschungsdaten für andere Forscher/-innen. Mehrere Antworten bezogen sich – möglicherweise unwissentlich – auf Provenance:

> *„Es ist wichtig, dass man über die Herkunft der Daten Bescheid weiß. Dass man weiß: Wie wurden die Daten erhoben? Wo? Und zu welchem Zweck? Und vielleicht auch sogar, welche Schwierigkeiten? Wie schwer ist es, diese Forschungsdaten zu erheben? Tatsächlich, damit man weiß, welchen Wert haben die?“*
> — IP20

Weitere, ausführliche Antworten:

> *„Was will ich mit den Forschungsdaten machen? Will ich die Forschungsdaten vielleicht noch in einem anderen Projekt verwenden? Sollen die frei verfügbar gemacht werden, auch für andere Forscher? Dann muss ich mir Gedanken machen, wie ich die dokumentiere. In welcher Form? Muss ich eine Software dazu packen, weil das vielleicht sehr spezielle Messwerte aus irgendeinem komischen Gerät sind? Weil die in einem proprietären Format sind? Und dieses Gedanken machen, um all diese Schritte vorher, über die Erhebung, die Dokumentation, über die Ablage, die Speicherung bis hin zur Einhaltung von Gesetzen – dass ich bestimmte Daten, wenn ich mit Textkorpora arbeite, hinterher löschen muss; also auch wie vernichte ich Daten, ohne den ganzen Forschungsprozess nicht mehr nachvollziehbar zu machen –, all das gehört für mich zum Forschungsdatenmanagement.“*
> — IP03

> *„Wie kann ich das durchführen? Wie kann ich die [Daten] vernünftig ablegen, sodass ich eben auch in einem halben Jahr noch weiß: Was bedeutet diese Datei? Was bedeuten diese Spalten, die Werte und das alles in einer Datei? Wie war der Ablauf dieses Forschungsprozesses? [...] Welche Skripte habe ich nacheinander aufgerufen? Welche Daten sind wo rein geflossen? Wo wurde vielleicht manuell interagiert oder ähnliches?“*
> — IP17

Unter welchen Bedingungen und in welchem Umfang würdest du einer anderen Person Einsicht in deine Forschungsdaten gewähren, nachdem du über diese publiziert hast?

Zu diesem Thema wurden alle Personen befragt, unabhängig davon, ob sie selbst in der Wissenschaft tätig waren/sind und publizierten/publizieren oder nicht. Sollte letzteres der Fall sein, wurde die Person gebeten, sich in die Rolle eines Wissenschaftlers/einer Wissenschaftlerin zu versetzen.

Neun Befragte würden Daten unter der Bedingung veröffentlichen, dass der Datenschutz gewährt wird. Drei würden die Entscheidung vom jeweiligen individuellen Fall abhängig machen. So lautet eine Antwort beispielsweise, dass die Daten innerhalb des Lehrstuhls jederzeit geteilt würden, Anfragen von außerhalb jedoch einzeln betrachtet werden. Zwei Personen würden die Daten nach einer Publikation oder nach dem Projektende veröffentlichen, da andernfalls das Risiko bestünde, dass andere Personen die Daten als Grundlage für eigene Publikationen nutzen würden. Zwei Befragte würden nur das absolute Minimum veröffentlichen – also nur die für das Ergebnis relevanten Daten – und zwei Personen würden ihre Rohdaten gar nicht veröffentlichen (eine Person schließt es vollständig aus, die andere lässt sich die Option dennoch offen). Die Veröffentlichung der Daten wird oft mit der öffentlichen Finanzierung der Personen/der Projekte begründet; viele Befragte sehen sich in der Pflicht, ihre Daten öffentlich bereitzustellen. Einige Befragte sprechen von der Konkurrenz innerhalb der Wissenschaft und begründen restriktivere Herangehensweisen damit. Interessant ist auch die folgende Antwort einer befragten Person:

> *„Und da hat man halt diesen Zwiespalt zwischen: Welche Daten veröffentliche ich? Welche nicht? Was ziehe ich noch heran? Zum Beispiel habe ich eben gesagt, dass ich Zitate veröffentliche und es gibt exorbitant gute Zitate, die ich veröffentliche, aber es fehlt dann ja der ganze Kontext drumherum. Und das ist auch sehr, sehr fraglich, denn: Hat diejenige das jetzt tatsächlich so gesagt? Und in welchem Kontext? Denn das steht ja im Artikel nicht. Ich weiß gar nicht, ob es darauf so eine präzise, konkrete Antwort gibt. Also im Zweifel, auch im Rahmen dieser ganzen gesellschaftsrelevanten Fake-News-Debatte, veröffentliche ich dann die Daten nicht. Also wenn ich mir dessen nicht sicher bin, dass ich meine Ergebnisse auch nachvollziehbar [und] transparent gestalten kann, dann würde ich es halt nicht veröffentlichen.“*
> — IP11

Besteht allgemein ein Interesse daran, Forschungsdaten nicht zu veröffentlichen? Falls ja, wann?

- eigene Reputation (6)
- Datenschutz (5)
- externe Interessen (5)
- eigene wirtschaftliche Interessen (4)
- politische Gründe (4)
- finanzieller Aufwand (3)
- Patentierungen (2)

...

Wichtig bei dieser Frage und dem Verstehen der Antworten ist, dass die beschriebenen Interessen und Gründe nicht unbedingt die der Forscher/-innen widerspiegeln, sondern genereller Natur sind. Häufig werden eigene Interessen genannt, insgesamt 12 Mal (eigene Reputation, wirtschaftliche Interessen und Patentierungen). Auch externe Interessen (5), beispielsweise von an Kooperationen beteiligten Unternehmen, sowie der Datenschutz (5) werden oft genannt. Politische Interessen (4), darunter zwei Mal die öffentliche Sicherheit, werden ebenfalls mehrfach genannt, genau wie der finanzielle Aufwand (3), der eventuell betrieben wurde, um die Forschung zu betreiben. Zu den weiteren Antworten, jeweils einzeln genannt, gehören ein großer persönlicher Einsatz, die ungeklärte Frage, wem bestimmte Daten gehören, ein Widerspruch zu den eigenen Hypothesen, kommerzielle und militärische Daten und die Befürchtung,

dass sich andere mit diesen Daten profilieren, weil sie die Quelle nicht angeben. Auch die Gefahr, dass in der Zukunft erlassene Gesetze bestimmte Daten oder deren Publikation illegalisieren, wird als Möglichkeit genannt.

Wie stehst du zur Idee von Open Science? Ist es okay, Daten/Wissen vorzuenthalten?

Die Antworten fallen hier erwartungsgemäß individuell und differenziert aus, es lassen sich dennoch einige Tendenzen feststellen: Die Mehrheit der befragten Personen ist sich darüber einig, dass Daten, die öffentlich finanziert erhoben wurden, auch öffentlich zugänglich sein sollten. Argumentiert wird dabei mit dem gesellschaftlichen Nutzen von Forschung sowie mit dem Recht der Gesellschaft, an die von ihr finanzierten Daten zu gelangen. Nichtsdestotrotz, das war eine weitere Antwort, sollte diese Entscheidung jeder Forscherin und jedem Forscher frei stehen; die Freiheit der Forschung sollte nicht angetastet werden. In manchen Situationen seien Embargo-Fristen jedoch ein legitimer Kompromiss. In solchen Fällen würden die Daten für einen gewissen Zeitraum zurückgehalten und anschließend veröffentlicht werden.

> *„Wir haben zum Beispiel Projekte, in denen Daten erhoben werden und die werden über Steuergelder finanziert. Das heißt, man hat die Verpflichtung, irgendwann die Ergebnisse aus den Projekten herauszugeben. Aber trotzdem denke ich, müssen die vorher geschützt werden. Wir haben tatsächlich auch eine Policy, dass die Daten so lange geschützt werden, wie die Forscher damit arbeiten und ihre Publikationen erstellen. Dass nicht jemand anderes vorher Publikationen mit diesen Daten erstellen kann. So lange dürfen wir sie hier vorhalten und noch nicht der Öffentlichkeit preisgeben.“*
> — IP20

Uneinigkeit besteht darin, in welchem Umfang und ab wann das Veröffentlichen von Daten geschehen sollte. Manche Befragten beschränken sich auf die Ergebnisse, andere würden die Daten erst nach einer Publikation oder dem Ende eines Projektes veröffentlichen. Sollten die Daten im Sinne des Datenschutzes schützenswert sein, dann sei eine vollständige Zurückhaltung gerechtfertigt. Der Personenschutz wäre dann wichtiger als die Reproduzierbarkeit.

> *„Also Forschung dient der Menschheit, sollte der Menschheit dienen. Ob das jetzt explizit die Codes selber betrifft, okay, kann man drüber diskutieren, aber die Ergebnisse auf jeden Fall.“*
> — IP13

> *„Wenn irgendjemand eine Möglichkeit gefunden hat, mit Haushaltsmitteln eine Atombombe zu bauen, ist es vielleicht nicht wahnsinnig sinnvoll, das auf 9GAG hochzuladen. Das ist dann natürlich auch wieder eine moralische Sache. [...] Solange niemand gefährdet ist oder gefährdet wird, sollte meiner Meinung nach auch alles so öffentlich sein, wie es nur überhaupt geht.“*
> — IP16

Wir sehen, dass der Provenance-Begriff oftmals nicht geläufig, aber zumindest recht eindeutig ist. Beim Begriff der Privacy ist dies anders: Dieser ist weitgehend bekannt, allerdings keinesfalls eindeutig. Zwar gehen die ersten Assoziationen meist in Richtung „Datenschutz“ und „Privatheit“, allerdings können sich viele der befragten Personen den Begriff auch in anderen Kontexten vorstellen, beispielsweise beim Schutz jeglicher Art von Daten, nicht nur bei solchen mit Personenbezug. Doch nicht nur der Privacy-Begriff ist uneindeutig, auch der „Datenschutz“-Begriff ist es. Anders sieht dies bei Forschungsdaten und ihrem Management aus; die Antworten fallen oft ähnlich aus. Lediglich bei der Frage, ob Daten neu geschaffen worden sein müssen, um als Forschungsdaten zu gelten, gibt es Diskrepanzen, ebenso bei der Frage nach der Veröffentlichung der eigenen Daten. Im letzten Kapitel dieser Bachelorarbeit werden wir ein ausführlicheres Fazit ziehen und einen Ausblick auf weitere Gedanken geben.

6. Fazit und Ausblick

Zum Abschluss der Bachelorarbeit *„Provenance und Privacy in ProSA“* wollen wir zunächst in Abschnitt 6.1 ein Fazit ziehen. Dabei gehen wir sowohl auf den ersten Teil – die Kombination von Data Provenance mit Privacy – als auch auf den zweiten Teil – die qualitative Befragung von 20 Expertinnen und Experten – ein, bevor wir uns abschließend in Abschnitt 6.2 einigen weiterführenden Gedanken widmen, die es nicht (mehr) in diese Arbeit geschafft haben.

6.1. Fazit

Beginnen wir mit einem Fazit zu Provenance und Privacy. Wir haben gesehen, dass die **where**-Provenance zwar auf den ersten Blick hinsichtlich des Datenschutzes unbedenklich, allerdings auch nicht für die wissenschaftliche Reproduzierbarkeit geeignet ist, wodurch in der Konsequenz möglicherweise der Worst Case eintritt und der Datenschutz überhaupt nicht gewährleistet werden kann. Anders sieht dies bei der **why**-Provenance aus: Auch hier fehlt zwar noch die Berechnungsvorschrift, wodurch zwar keine Reproduzierbarkeit/Rekonstruierbarkeit gegeben ist, aber zumindest die Nachvollziehbarkeit kann erfüllt werden. Datenschutzprobleme treten hier nur in Grenzfällen auf, deren Auftreten in der Praxis zwar unwahrscheinlich, jedoch nicht unmöglich ist. Schwieriger ist es hingegen bei der **how**-Provenance: Aufgrund der großen Menge an Informationen, die die Provenance-Antworten liefern, eignen sie sich einerseits sehr zur Sicherung der wissenschaftlichen Reproduzier- bzw. Rekonstruierbarkeit, können aber auch sensible Informationen über einzelne Personen offenbaren. In jedem Fall – sowohl bei der **why**- als auch bei der **how**-Provenance – ist dazu jedoch externes Wissen erforderlich. Genau dieser Fall wird beim Datenschutz allerdings häufig unterschätzt, wie das Beispiel des *Gouverneurs von Massachusetts* zeigt [Swe02b].

Aus unter anderem diesem Grund haben wir im zweiten Teil dieser Bachelorarbeit 20 Personen interviewt, um ihr Verständnis von Datenschutz, aber auch ihre Vorstellung der Provenance- und Privacy-Begriffe sowie den Umgang mit den eigenen (Forschungs-)Daten zu erforschen. Dabei fanden wir heraus, dass „Privacy“ zwar oft mit Datenschutz verbunden wird, allerdings keinesfalls auf diese Bedeutung beschränkt ist. So können beispielsweise auch sensitive Firmen- und Geschäftsdaten, politische Informationen oder aber die eigenen Forschungsdaten im Sinne des Privacy-Begriffes als schützenswert angesehen werden. Auch der Datenschutz-Begriff muss hierbei nicht zwangsläufig mit personenbezogenen Daten in Verbindung gebracht werden:

> *„Das würde ich jetzt nicht als Privacy bezeichnen, aber schon als Datenschutz.“*
> — IP20

Abschließend können wir sagen, dass die Vorstellung der Privacy im Sinne personenbezogener Daten zu kurz greift. Bezüglich Privacy bei Data Provenance halten wir fest, dass es insbesondere bei der **how**-Provenance zu Problemen kommen kann, sofern externes Wissen miteinbezogen wird. Das Permutieren und Generalisieren bzw. Unterdrücken von Tupeln sowie intensionale Antworten anstelle von extensionalen sind hierbei erste Lösungsansätze, wohingegen die auf dem ersten Blick vielversprechende Differential Privacy per Definition ungeeignet ist.

6.2. Ausblick

An dieser Stelle wollen wir diese Bachelorarbeit noch mit einigen weiterführenden Überlegungen abschließen, welche es nicht in die Arbeit geschafft haben, da sie den Umfang dieser überschritten hätten. Es folgt eine Auflistung offener Fragen und Anregungen. Einige davon ergaben sich aufgrund der Abweichung von der ursprünglichen Aufgabenstellung, andere ergaben sich während der geführten Interviews, der Bearbeitung der Themenstellung und der Literaturrecherche.

- Es wurden verschiedene Ansätze genannt, um einen wirksamen Datenschutz bei Data-Provenance-Anfragen zu gewährleisten. Diese sollten in das Programm *ProSA* implementiert werden.
  - Ein erster Schritt besteht darin, das Projekt *ChaTEAU* um den Provenance-Aspekt zu erweitern. Sobald die **how**-Provenance implementiert ist, sollte eine Permutation der Polynome recht einfach umsetzbar sein.
  - Schwieriger wird eine Implementation zur Generalisierung und Unterdrückung bzw. zum Herstellen der k-Anonymität und ggf. l-Diversität für gegebene k und l. Das Paper [Swe02a] kann hier als Einstieg dienen.
  - Ebenfalls sinnvoll wäre die Implementation intensionaler zusätzlich zu extensionalen Antworten. Hier sei auf die Bachelorarbeit *„Intensional Answers for Provenance Queries in Big Data Analytics“* von *Jan Svacina* verwiesen, in welcher sich der Autor bereits mit diesem Thema beschäftigt hat [Sva16].
- Mehrfach wurde Bezug auf die 87% genommen, die *Latanya Sweeney* in [Swe00] für die Bevölkerung der USA ermittelte. Einige der befragten Personen in Kapitel 5 wollten wissen, wie sich dies in Deutschland widerspiegelt. Eine solche Recherche wäre sicherlich interessant, allerdings auch schwieriger durchzuführen, da die Datenschutzbestimmungen in Deutschland für gewöhnlich restriktiver sind als in den USA.

Literaturverzeichnis

[AH19] Auge, Tanja ; Heuer, Andreas: ProSA - Using the CHASE for Provenance Management. In: *ADBIS* Bd. 11695, Springer, 2019 (Lecture Notes in Computer Science), S. 357–372

[Aug17] Auge, Tanja: *Umsetzung von Provenance-Anfragen in Big-Data-Analytics-Umgebungen.* Masterarbeit, Universität Rostock, Lehrstuhl für Datenbank- und Informationssysteme, 2017

[BGK+14] Bertino, Elisa ; Ghinita, Gabriel ; Kantarcioglu, Murat u. a.: A roadmap for privacy-enhanced secure data provenance. In: *J. Intell. Inf. Syst.* 43 (2014), Nr. 3, S. 481–501

[BW15] Bleymüller, Josef ; Weissbach, Rafael: *Statistik für Wirtschaftswissenschaftler.* 17., überarbeitete Auflage. München : Verlag Franz Vahlen, 2015

[CCT09] Cheney, James ; Chiticariu, Laura ; Tan, Wang-chiew: Provenance in Databases: Why, How, and Where. In: *Foundations and Trends in Databases* (2009), Januar, S. 379–474

[Cuz16] Cuzzocrea, Alfredo: Big Data Provenance: State-Of-The-Art Analysis and Emerging Research Challenges. In: *EDBT/ICDT Workshops* Bd. 1558, CEUR-WS.org, 2016 (CEUR Workshop Proceedings)

[DF08] Davidson, Susan B. ; Freire, Juliana: Provenance and scientific workflows: challenges and opportunities. In: *SIGMOD Conference*, ACM, 2008, S. 1345–1350

[DKR+11] Davidson, Susan B. ; Khanna, Sanjeev ; Roy, Sudeepa u. a.: On Provenance and Privacy. In: *ICDT*, ACM, 2011, S. 3–10

[Dwo06] Dwork, Cynthia: Differential Privacy. In: *ICALP (2)* Bd. 4052, Springer, 2006 (Lecture Notes in Computer Science), S. 1–12

[Dwo08] Dwork, Cynthia: Differential Privacy: A Survey of Results. In: *TAMC* Bd. 4978, Springer, 2008 (Lecture Notes in Computer Science), S. 1–19

[Gru19] Grunert, Hannes: *Tafelübung 7.* Übungsunterlagen „Informationssysteme und -dienste / Data Science“, 2019

[GT17] Green, Todd J. ; Tannen, Val: The Semiring Framework for Database Provenance. In: *Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems - PODS '17.* Chicago, Illinois, USA : ACM Press, 2017, S. 93–99

[HDB17] Herschel, Melanie ; Diestelkämper, Ralf ; Ben Lahmar, Houssem: A survey on provenance: What for? What form? What from? In: *VLDB J.* 26 (2017), Nr. 6, S. 881–906

[Heu20] Heuer, Andreas: Research Data Management. In: *it - Information Technology* 61 (2020), Nr. 2. `http://dx.doi.org/10.1515/itit-2020-0002`. – DOI 10.1515/itit–2020–0002

[HSS18] Heuer, Andreas ; Saake, Gunter ; Sattler, Kai-Uwe: *Datenbanken - Konzepte und Sprachen, 6. Auflage.* MITP, 2018

[HSS19] Heuer, Andreas ; Saake, Gunter ; Sattler, Kai-Uwe: *Datenbanken - Implementierungstechniken, 4. Auflage.* MITP, 2019

[KSG13] Kosinski, Michal ; Stillwell, David ; Graepel, Thore: Private traits and attributes are predictable from digital records of human behavior. In: *Proceedings of the National Academy of Sciences* 110 (2013), Nr. 15, S. 5802–5805

[LST+17] Liang, Xueping ; Shetty, Sachin ; Tosh, Deepak K. u. a.: ProvChain: A Blockchain-based Data Provenance Architecture in Cloud Environment with Enhanced Privacy and Availability. In: *CCGrid*, IEEE Computer Society / ACM, 2017, S. 468–477

[May13] Mayer, Horst O.: *Interview und schriftliche Befragung: Grundlagen und Methoden empirischer Sozialforschung.* 6., überarbeitete Auflage. München : Oldenbourg Verlag, 2013

[Min14] Ministerium für Wissenschaft, Forschung und Kunst: *E-Science: Wissenschaft unter neuen Rahmenbedingungen, Fachkonzept zur Weiterentwicklung der wissenschaftlichen Infrastruktur in Baden-Württemberg.* 2014

[MW04] Meyerson, Adam ; Williams, Ryan: On the Complexity of Optimal K-Anonymity. In: *PODS*, ACM, 2004, S. 223–228

[Noc18] Nocun, Katharina: *Die Daten, die ich rief: Wie wir unsere Freiheit an Großkonzerne verkaufen.* Originalausgabe. Köln : Lübbe, 2018

[PS17] Petrlic, Ronald ; Sorge, Christoph: *Datenschutz: Einführung in technischen Datenschutz, Datenschutzrecht und angewandte Kryptographie.* Wiesbaden : Springer Vieweg, 2017. – ISBN 978–3–658–16838–4

[Sam01] Samarati, Pierangela: Protecting Respondents' Identities in Microdata Release. In: *IEEE Trans. Knowl. Data Eng.* 13 (2001), Nr. 6, S. 1010–1027

[SKS13] Simukovic, Elena ; Kindling, Maxi ; Schirmbacher, Peter: *Umfrage zum Umgang mit digitalen Forschungsdaten an der Humboldt-Universität zu Berlin*

[Sva16] Svacina, Jan: *Intensional Answers for Provenance Queries in Big Data Analytics.* Bachelorarbeit, Universität Rostock, Lehrstuhl für Datenbank- und Informationssysteme, 2016

[Swe00] Sweeney, Latanya: Simple Demographics Often Identify People Uniquely, Carnegie Mellon University, 2000

[Swe02a] Sweeney, Latanya: Achieving k-Anonymity Privacy Protection Using Generalization and Suppression. In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10 (2002), Nr. 5, S. 571–588

[Swe02b] Sweeney, Latanya: k-Anonymity: A Model for Protecting Privacy. In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10 (2002), Nr. 5, S. 557–570

[TN16] Torra, Vicenç ; Navarro-Arribas, Guillermo: Big Data Privacy and Anonymization. In: *Privacy and Identity Management* Bd. 498, 2016 (IFIP Advances in Information and Communication Technology), S. 15–26

[TNSM17] Torra, Vicenç ; Navarro-Arribas, Guillermo ; Sanchez-Charles, David ; Muntés-Mulero, Victor: Provenance and Privacy. In: *MDAI* Bd. 10571, Springer, 2017 (Lecture Notes in Computer Science), S. 3–11

[WD+16] Wilkinson, Mark D. ; Dumontier, Michel u. a.: The FAIR Guiding Principles for scientific data management and stewardship. In: *Scientific Data* 3 (2016), S. 160018. `http://dx.doi.org/10.1038/sdata.2016.18`. – DOI 10.1038/sdata.2016.18

Tabellenverzeichnis

Anfragenverzeichnis

Abbildungsverzeichnis

A. Anhang: Aufbau des Datenträgers

Auf dem Datenträger, der zusätzlich zu dieser Bachelorarbeit eingereicht wurde, befinden sich ergänzende Materialien, die der Sicherung der Nachvollziehbarkeit dienen. An dieser Stelle werden die einzelnen Verzeichnisse und Dateien erläutert, die sich auf diesem Datenträger befinden.

Literatur Im Ordner *„Literatur“* befinden sich alle genutzten Quellen dieser Bachelorarbeit. Die Dateinamen entsprechen den Kürzeln im Literaturverzeichnis; so liegt beispielsweise [TN16] als Datei `Literatur/TN16.pdf` vor. Nicht enthalten sind die Bücher [BW15], [HSS18] und [HSS19], da diese digital nicht verfügbar waren bzw. sind.

Es sei an dieser Stelle angemerkt, dass diese Quellen lediglich der wissenschaftlichen Nachvollziehbarkeit dienen und im Sinne des Urheberrechts nicht öffentlich zur Verfügung gestellt werden dürfen. Insbesondere gilt dies für die Werke [May13] und [PS17], welche durch Lizenzverträge der *Universität Rostock* durch die *Universitätsbibliothek Rostock* bereitgestellt wurden, sowie für [Noc18], welches privat bereitgestellt wurde und darüber hinaus mit einem eindeutigen Wasserzeichen versehen ist. Alle Dateien liegen im `PDF`-Format vor.

SQLite Browser und Beispieldatenbank Im Ordner *„SQLite Browser“* befinden sich der Quellcode des gleichnamigen Programms in der Version 3.11.2 (sowohl `zip`- als auch `tar`-und-`gzip`-komprimiert) sowie ausführbare Versionen für Windows (sowohl in einer 32-Bit- als auch 64-Bit-Version; beides in Form eines `zip`-komprimierten Archivs) und macOS (als `.dmg`-Datei). Um die Integrität dieser Dateien zu bestätigen, sind an dieser Stelle die `SHA-256`-Prüfsummen der ausführbaren Dateien angegeben, welche auch unter `https://github.com/sqlitebrowser/sqlitebrowser/releases/tag/v3.11.2` zu finden sind und damit verglichen werden können:

Version	**`SHA-256`-Checksum**
Windows – 32 Bit	bdfcd05bf1890a3336a1091c6e9740d582167494d0010da061f9effab2243b9e
Windows – 64 Bit	c6117e9d75bde6e0a6cbf51ee2356daa0ce41ca2dd3a6f3d1c221a36104531a0
macOS	022536d420dca87285864a4a948b699d01430721b511722bcf9c8713ab946776

Der *SQLite Browser* wurde genutzt, um die Beispieldatenbank zu verwalten und SQL-Anfragen auf diese auszuführen. Die entsprechende Datenbank befindet sich in Form der Datei `database.db` im Hauptverzeichnis und kann mit dem *SQLite Browser* geöffnet werden. Anschließend können die einzelnen Relationen dieser Datenbank betrachtet und SQL-Anfragen an die Datenbank gestellt werden.

Transkribierte Interviews Die Transkripte der geführten Interviews liegen in anonymisierter Form im Ordner *„Transkripte“* vor. Jede/-r Teilnehmer/-in erhielt dabei eine eindeutige ID *„IPxy“*, wobei xy eine fortlaufende zweistellige Zahl kennzeichnet; die Dateinamen entsprechen dieser ID. Alle Dateien liegen im `PDF`-Format vor. Alle Gespräche sind in Form von Stichpunktsätzen wortgetreu nach bestem Wissen und Gewissen dokumentiert worden. Die Antworten des Fragestellers sind dabei schwarz, die Antworten

der befragten Person(en) sind blau gefärbt. In einigen Fällen erfolgten Fragen oder Anmerkungen der Betreuerin dieser Bachelorarbeit; diese sind rot hervorgehoben.

Bachelorarbeit Sowohl die digitale Version dieser Bachelorarbeit als auch die Quelldaten befinden sich im Ordner *„Bachelorarbeit“*. Die Bachelorarbeit liegt dabei als `PDF`-Datei vor; die Quelldaten liegen als `TeX`-Dateien in Unterordner *„LaTeX“* vor. Auch alle weiteren, von `pdfTeX` erzeugten Dateien, darunter in den Formaten `.aux` und `.log`, befinden sich in diesem Verzeichnis.

Selbstständigkeitserklärung

Hiermit erkläre ich, dass ich die vorliegende Bachelorarbeit selbstständig und nur unter Verwendung der angegebenen Literatur und Hilfsmittel angefertigt habe. Die aus fremden Quellen direkt oder indirekt übernommen Stellen sind als solche kenntlich gemacht. Die Arbeit wurde bisher keiner anderen Prüfungsbehörde in gleicher oder ähnlicher Form vorgelegt und auch nicht veröffentlicht.

Rostock, den 18. Februar 2020 _____