

Der Umgang mit Forschungsdaten im Fach Psychologie: Konkretisierung der DFG-Leitlinien

Felix Schönbrodt, Mario Gollwitzer & Andrea Abele-Brehm*

Im Auftrag des DGPs Vorstands (verabschiedet am 17.9.2016; aktuelle Version: 9.11.2016)

Die vorliegenden Empfehlungen sollen – als einer von mehreren Bausteinen – zur Qualitätssicherung der psychologischen Forschung beitragen. Sie sind getragen von der Idee einer offenen und transparenten Wissenschaft, in der publizierte Befunde nachvollziehbar sind und Daten, die im Kontext publizierter wissenschaftlicher Arbeiten und drittmittelgeförderter Forschungsprojekte erhoben wurden, anderen Forscherinnen und Forschern zur Nachnutzung zur Verfügung stehen.

Schon seit langem gibt es die Forderung, Forschungsdaten öffentlich verfügbar zu machen¹. So fordert die Deutsche Forschungsgemeinschaft (DFG) in den "Empfehlungen zur gesicherten Aufbewahrung und Bereitstellung digitaler Forschungsprimärdaten" (2009), dass Daten, die aus öffentlich-rechtlichen Mitteln finanziert wurden, nach Abschluss des Projekts öffentlich zugänglich und frei verfügbar gemacht werden sollen². Daran anschließend hat sich auch die Allianz der Wissenschaftsorganisationen 2010 für die langfristige Sicherung und den grundsätzlich offenen Zugang zu Forschungsdaten ausgesprochen³. Die DFG hat im September 2015 neue Leitlinien zum Umgang mit Forschungsdaten veröffentlicht, in denen sie diese Ziele bekräftigt und die Fachgesellschaften dazu auffordert, ihren Umgang mit Forschungsdaten zu reflektieren und angemessene Regularien zur disziplinspezifischen Nutzung und Bereitstellung von Forschungsdaten zu entwickeln⁴. Die *Deutsche Gesellschaft für Psychologie* (DGPs) schließt sich den Zielen der DFG und der Allianz der Wissenschaftsorganisationen an und präzisiert hier die Erwartungen der DFG für das Fach Psychologie.

Hierbei ist ein sorgfältiges Abwägen von Rechten, Kosten und Nutzen aus der Perspektive (1) der an den Studien beteiligten Teilnehmerinnen und Teilnehmer, (2) der Forscherinnen und Forscher, die die Originaldaten erhoben haben, (3) der Öffentlichkeit (einschließlich potenzieller Inhaberinnen und Inhaber relevanter Urheberrechte) und (4) der wissenschaftlichen Gemeinschaft (einschließlich potenzieller Nachnutzerinnen und Nachnutzer erhobener Daten) erforderlich. Das Interesse der *scientific community* an möglichst umfassender Datennutzung muss mit dem Interesse der individuellen Forscherinnen und Forscher an der Verwertung der von ihnen erhobenen Daten sowie mit dem Interesse der Studienteilnehmerinnen und -teilnehmer an einem ethisch verantwortungsvollen Umgang mit ihren Daten in Beziehung gesetzt werden.

In diesem Zusammenhang kann nicht genügend betont werden, dass dem Erheben von Originaldaten in der Psychologie nach wie vor eine zentrale Bedeutung zukommt. Daten sind eine *conditio sine qua non* für jede empirische Wissenschaft. Wer Daten generiert und diese offen zur Verfügung stellt, leistet einen wichtigen Beitrag zur Wissenschaft, der entsprechend gewürdigt werden sollte. Forscherinnen und Forscher, die Originaldaten erheben, dürfen daher gegenüber Forscherinnen und Forschern, die die Daten anderer nachnutzen, keinen karrieretechnischen Nachteil haben (bspw. weil letztere im gleichen Zeitraum eine größere Anzahl an Publikationen vorweisen können).

Entsprechend betont auch die DFG, dass "das Engagement und die Bemühungen von Wissenschaftlern und Wissenschaftlerinnen um die Verfügbarmachung von Forschungsdaten [...] bei der Würdigung von wissenschaftlichen Qualifikationen und Leistungen zukünftig stärker berücksichtigt

* Wir danken Malte Elson, Johannes Breuer und Zoe Magraw-Mickelson für die englische Übersetzung des Dokuments (siehe http://www.dgps.de/fileadmin/documents/Empfehlungen/Data_Management_eng.pdf). Viele Aspekte dieser Empfehlungen sind durch zahlreiche konstruktive Beiträge während eines ausgiebigen Diskussionsprozesses mit der Mitgliedschaft der DGPs entstanden. Es würde zu weit führen, alle Beteiligten einzeln hier aufzuführen; wir möchten jedoch allen ausdrücklich für ihr Engagement danken.

1 Für einen Überblick bis 2012 siehe auch Fahrenberg, J. (2012). Open Access – nur Texte oder auch Primärdaten? *Working Paper Series des Rates für Sozial- und Wirtschaftsdaten (RatSWD)*. Nr. 200/2012. http://www.jochen-fahrenberg.de/fileadmin/openaccess/Open_Access_Primaerdaten.pdf; siehe auch American Psychological Association (2015). *Data Sharing: Principles and Considerations for Policy Development*. URL: <https://www.apa.org/science/leadership/bsa/data-sharing-report.pdf>

2 http://dfg.de/download/pdf/foerderung/programme/lis/ua_inf_empfehlungen_200901.pdf

3 <http://www.allianzinitiative.de/de/handlungsfelder/forschungsdaten/grundsaeetze.html>

4 http://www.dfg.de/download/pdf/foerderung/antragstellung/forschungsdaten/richtlinien_forschungsdaten.pdf

werden [sollten]^{5 6}. Gleichzeitig gilt, dass eine sinnvolle Datennachnutzung erwünscht ist und zu wichtigen und validen wissenschaftlichen Erkenntnissen führen kann.

Die vorliegende Konkretisierung der DFG-Leitlinien verfolgt das Ziel, eine Balance zwischen diesen unterschiedlichen Interessen zu realisieren. Sie

- thematisiert die Bedeutung eines nachhaltigen Managements von Forschungsdaten,
- definiert, was „Primärdaten“ sind und wie sie gespeichert werden sollen,
- definiert Standards und eventuelle Einschränkungen bei der Veröffentlichung von Forschungsdaten und
- definiert die Rechte und Pflichten von „Datenbereitstellenden“ und „Datennachnutzenden“.

Die DGPs regt an, dass alle Drittmittelgeber diese Empfehlungen zum Datenmanagement bei der Vergabe von Forschungsgeldern und bei der Begutachtung von Projektabschlussberichten berücksichtigen.⁷

Die vorliegenden Empfehlungen sollen nach fünf Jahren evaluiert und gegebenenfalls überarbeitet werden.

Die Empfehlungen liegen ebenfalls in englischer Sprache vor (http://www.dgps.de/fileadmin/documents/Empfehlungen/Data_Management_eng.pdf).

Die Herausgeberinnen und Herausgeber der Organzeitschriften der DGPs werden gebeten, auf die Umsetzung der vorliegenden Empfehlungen bei Manuskripteinreichungen zu achten.

Die DGPs wirkt international im Rahmen ihrer Möglichkeiten darauf hin, eine Diskussion über eine mögliche Harmonisierung unterschiedlicher Leitlinien bzw. Empfehlungen im hier vorgelegten Sinn anzustoßen.

Die DGPs empfiehlt ihren Mitgliedern, bei der Besetzung von Professuren und der Evaluation der wissenschaftlichen Leistungen der Bewerberinnen und Bewerber auch darauf zu achten, inwiefern sie Kriterien der Transparenz – im Sinne der vorliegenden Empfehlungen – in ihrer Forschung berücksichtigen.

1. Forschungsdatenmanagement

Ziele eines nachhaltigen Datenmanagements in der Psychologie bestehen unter anderen in der

- a) Qualitätssicherung (Gewährleistung der langfristigen Nachprüfbarkeit von wissenschaftlichen Ergebnissen, auch im Hinblick darauf, dass erhobene Daten zukünftig mit anderen und ggf. besseren Methoden analysiert werden können als dies zum Zeitpunkt der Erhebung der Fall war)⁸, der

5 http://www.dfg.de/foerderung/antragstellung_begutachtung_entscheidung/antragstellende/antragstellung/nachnutzung_forschungsdaten/

6 Eine Möglichkeit, die Leistungen datengenerierender Forschung stärker zu würdigen, ist die Schaffung einer neuen Publikationskategorie, die „Datenbereitstellung“. Artikel, die auf Datennachnutzungen beruhen und bei denen die Personen, die die Daten erhoben haben, nicht als Koautorinnen bzw. Koautoren genannt sind, sollten in der Nähe des Titels die Autorinnen bzw. Autoren der Primärdaten und die exakte Zitation des Repositoriums, auf dem diese zu finden sind, nennen. Datenbereitstellende können in ihren Publikationslisten eine neue Rubrik einführen (z.B. „Nachnutzung der Daten von *** in Publikation****“), in der die durch die Datennachnutzung resultierenden Publikationen aufgeführt sind.

7 Siehe auch: Hartig, K. & Soßna, V. (2016). *Forschungsdatenmanagement in DFG-Anträgen: Was kann, was soll, was muss beschrieben werden?* Institutionelles Repositorium der Leibniz Universität Hannover. DOI: <http://dx.doi.org/10.15488/262>

8 Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2015). The prevalence of statistical reporting errors in psychology (1985-2013). *Behavior Research Methods* [Online]. doi:10.3758/s13428-015-0664-2

- b) Optimierung des Erkenntnisgewinns (Nutzung von Daten für Re- und Metaanalysen, Analysen „einmaliger“ Datensätze)⁹ sowie der
- c) Maximierung des Kosten-Nutzen-Verhältnisses (optimale Ausnutzung bereits erhobener Daten; Vermeidung redundanter Belastungen bei Mensch und Tier).

Ein offener, langfristiger und kostenloser Zugang zu Forschungsdaten trägt dazu bei, diese Ziele zu erreichen. Forscherinnen und Forscher, die Originaldaten erheben (wir bezeichnen sie im Weiteren als „Datenbereitstellende“), müssen Sorge dafür tragen, dass die von ihnen bereitgestellten Daten nachhaltig nutzbar sind (siehe hierzu Abschn. 7.2). Forscherinnen und Forscher, die Originaldaten für Nachanalysen nutzen (sog. „Nachnutzende“), sind zur Einhaltung bestimmter Standards verpflichtet (siehe hierzu Abschn. 7.3).

Bei der Veröffentlichung von Forschungsdaten gilt es, datenschutzrechtliche, urheberrechtliche und forschungsethische Aspekte zu berücksichtigen (siehe hierzu Abschn. 5). Diese Aspekte können ggf. Einschränkungen in Bezug auf die Veröffentlichung von Primärdaten implizieren.

2. Primärdaten

Da in der Folge häufig von „Primärdaten“ die Rede ist, soll zunächst definiert werden, was damit gemeint ist. Eine sinnvolle Unterscheidung ist diesbezüglich die zwischen Rohdaten und Primärdaten. *Rohdaten* sind die Ursprungsaufzeichnungen, z.B. Kreuze auf einem Papierfragebogen, Zeichnungen oder auch Audio- oder Videoaufnahmen. Mit *Primärdaten* ist die erste Übertragung der Rohdaten in ein digitales Format gemeint, also z.B. Code „1“ für eine Ja-Antwort usw.

Somit sind Primärdaten in der Psychologie vollkommen unbearbeitete (d.h. untransformierte, nicht-aggregierte etc.) quantitative und qualitative Daten, zum Beispiel

- bei Experimenten alle manipulierten und gemessenen Variablen für jeden Experimentaldurchgang jeder Person;
- bei Fragebögen die Antworten jeder Person auf jedem Item;
- bei Freitext-Eingaben der Originalwortlaut (unter Berücksichtigung des Datenschutzes, s.u.);
- digitalisierte Videoaufnahmen (Anmerkung: da diese im Allgemeinen jedoch nicht hinreichend anonymisierbar sind, können sie nicht im öffentlichen Repositorium gespeichert werden. Stattdessen können Kodierungen des beobachteten Verhaltens in Repositorien gespeichert werden);
- Downloads oder Screenshots von Inhalten sozialer Medien (z.B. Facebook-Profile oder Twitter-Nachrichten);
- bei (neuro)physiologischen Daten (wie EEG- oder fMRT-Daten) verlustfrei umgewandelte Daten in einem standardisierten Rohdatenformat (z.B. EDF, DICOM, oder NIFTI), die nicht aggregiert sind und nicht nur auf wenige „regions of interest“ beschränkt sind¹⁰.

„Primärdaten“ beinhalten auch die Daten jener Personen, die von der Datenanalyse ausgeschlossen wurden (außer, der Ausschluss beruht darauf, dass die teilnehmende Person während oder nach der Datenerhebung ihre Einwilligung zurückgezogen hat).

3. Art der Speicherung

Die Primärdaten sollen in digitaler Form¹¹ auf einem vertrauenswürdigen Repositorium bereitgestellt werden. Wichtige Qualitätsmerkmale eines vertrauenswürdigen Repositoriums sind¹²:

9 http://www.allianzinitiative.de/fileadmin/user_upload/redakteur/Grundsätze_Forschungsdaten_2010.pdf

10 Für MRI-Daten siehe z.B. die „Best Practices in Data Analysis and Sharing in Neuroimaging using MRI“ der Organization for Human Brain Mapping: <http://www.humanbrainmapping.org/files/2016/COBIDASreport.pdf>

11 Es wird davon ausgegangen, dass der größte Anteil der in der Psychologie erhobenen Daten bereits digital vorliegt oder digitalisierbar ist. Bei nicht-digitalisierbaren Daten wird empfohlen, diese direkt an der Institution aufzubewahren. Dabei muss dafür Sorge getragen werden, dass diese Daten auch bei einem Affiliationswechsel oder bei einem Ausscheiden aus dem akademischen Betrieb erhalten bleiben.

12 Siehe z.B. auch Data Seal of Approval: <http://datasealofapproval.org/en/>

- Die wirtschaftliche bzw. ideologische **Unabhängigkeit** und wissenschaftliche **Professionalität** der bereitstellenden Institution;
- Die **Persistenz** der Daten: Es muss gewährleistet sein, dass die Daten über einen langen Zeitraum (mindestens 10 Jahre, im Idealfall jedoch deutlich länger) gesichert sind, und es muss geklärt sein, was im Falle einer Auflösung des Repositoriums mit den Daten geschieht;
- Die **Zugänglichkeit** der Daten: Es muss möglich sein, die Daten öffentlich und kostenlos abzurufen; es muss aber auch möglich sein, Zugangsbeschränkungen (im Sinne von „Scientific Use Files“) zu definieren (zur Diskussion einer eventuellen Zugangsbeschränkung siehe Abschn. 5);
- Die **Identifizierbarkeit** der Daten: Es muss ein persistenter Identifikator (z.B. eine persistente URL oder, falls möglich, ein DOI) vergeben werden;
- Die **Klärung der Rechte** an den Daten: Mit dem Speichern der Daten darf kein Abtreten der ausschließlichen Nutzungsrechte an Dritte verbunden sein (das einfache Nutzungsrecht, d.h. das Recht zur Archivierung und Vervielfältigung muss dem Betreiber des Repositoriums übertragen werden, damit der Betrieb des Repositoriums überhaupt möglich ist);
- Die Möglichkeit, Dateien bzw. Datensätze **sowohl öffentlich als auch nicht-öffentlich** abzuspeichern.

Daher ist ein vertrauenswürdige öffentliches Repositorium (z.B. PsychData des ZPID¹³, datorium bei GESIS¹⁴, oder gut ausgebaute universitäre Repositorien) einem Zeitschriftenrepositorium vorzuziehen. Eine Bereitstellung auf privaten oder persönlichen universitären Webseiten wird hingegen nicht empfohlen.

Bei der Wahl des Repositoriums sind Einschränkungen, die sich aus forschungsethischen Richtlinien ergeben (z.B. ein evtl. Verbot, Daten auf einem ausländischen oder außereuropäischen Server zu speichern), zu beachten.

Die Institution, die das Repositorium bereitstellt, sollte Beratung und Unterstützung bei der Speicherung der Primärdaten zur Verfügung stellen.

4. Kosten der Datenarchivierung

Daten mit hohen Qualitätsstandards aufzubereiten und bereitzustellen bedeutet meist zusätzlichen Ressourcenaufwand. Daher kann und sollte für die Aufbereitung und Archivierung von Datensätzen bereits im Rahmen der Drittmittel-Anträge finanzielle Unterstützung in Form von Personal- und auch Sachmitteln beantragt werden. Im Falle sehr großer Datenmengen (z.B. EEG oder fMRI) gibt es Repositorien, die sich auf Datenmengen im Gigabytebereich und größer spezialisiert haben. Die Nutzung dieser Repositorien verursacht Kosten, die ebenfalls bereits bei der Antragstellung berücksichtigt werden sollten.

5. Datenschutz und Urheberrecht

Datenschutz- und urheberrechtliche Einschränkungen müssen bereits in der Planungsphase einer Untersuchung berücksichtigt werden¹⁵. So muss zum Beispiel durch eine geeignete Anonymisierung bzw. Pseudonymisierung sichergestellt werden, dass nicht aus der Kombination verschiedener erhobener Merkmale – auch solcher, die in unterschiedlichen Studien mit den gleichen Teilnehmerinnen und Teilnehmern (z.B. Studierende im 1. Fachsemester Psychologie an der Universität XY) erhoben wurden – Personen identifizierbar sind¹⁶. Datenschutzrechtliche Erwägungen spielen jedoch nicht nur auf der Ebene einzelner Personen, sondern auch auf relevanten Aggregatebenen eine Rolle: so muss speziell im Falle sensibler Fragestellungen (z.B. illegales

13 Dehnhard, I., Weichselgartner, E. & Krampen, G. (2013). Researcher's willingness to submit data for data sharing: A case study on a data archive for psychology. *Data Science Journal*, 12, 172-180.

14 <https://datorium.gesis.org>

15 Siehe z.B. Hrynaszkiewicz, I., Norton, M. L., Vickers, A. J., & Altman, D. G. (2010). Preparing raw clinical data for publication: Guidance for journal editors, authors, and peer reviewers. *BMJ*, 340, c181. doi:10.1136/bmj.c181; oder: <http://theodi.github.io/ukan-course/#0.1>

16 Gola, P. & Schomerus, R. (2010). *Kommentar Bundesdatenschutzgesetz*, 10. Auflage. München: C.H. Beck.

Verhalten, Suizidraten etc.) besonders darauf geachtet werden, inwiefern einzelne Orte (Schulen, Firmen etc.) in den Daten bzw. durch eine Zusammenführung von Datensätzen eindeutig identifiziert werden können.

Bereits bei der Rekrutierung von Teilnehmerinnen und Teilnehmern sind die einschlägigen Gesetze und Verordnungen (zum Datenschutz und zum Recht auf informationelle Selbstbestimmung) zu beachten. Dies gilt insbesondere auch dann, wenn Kinder untersucht werden, die noch nicht selbst ihre Einwilligung geben können. So müssen Teilnehmerinnen und Teilnehmer einer Studie darauf aufmerksam gemacht werden, dass ihre anonymisierten Daten ggf. für eine Nachnutzung durch Dritte zur Verfügung gestellt werden und dass Zweck, Art und Umfang dieser Nachnutzung zum gegenwärtigen Zeitpunkt noch nicht absehbar sind. Eine explizite Einwilligung zur Datennachnutzung ist in jedem Fall einzuholen, wenn diese Daten nicht vollständig anonymisiert werden können¹⁷. Wenn Daten vollständig anonymisiert sind, z.B. Fragebogendaten oder Daten aus Experimentalfreihen, muss eine entsprechende Einwilligung nicht notwendigerweise eingeholt werden, da keine individuellen Zuordnungen mehr möglich sind. In begründeten Zweifelsfällen sollte die lokale Ethikkommission bzw. die Ethikkommission der DGPs konsultiert werden¹⁸.

Einwilligungserklärungen und Ethikanträge sollten daher so angepasst werden, dass sie kompatibel mit den vorliegenden Empfehlungen sind. Außerdem werden Ethikkommissionen aufgefordert, ihre Richtlinien daraufhin zu überprüfen, ob bestimmte Vorschriften zu restriktiv in Bezug auf Forschungstransparenz sind, ohne dem Datenschutz wirklich dienlich zu sein (zum Beispiel ist ein Löschzwang von komplett anonymisierten Daten unnötig). Vorschläge für entsprechende Formulierungen finden sich in den Anhängen B (Einwilligungserklärung) und C (Leitlinien Ethikkommission).

Wenn rechtliche Bedenken gegen die Veröffentlichung von Daten vorliegen, sind diese unbedingt zu beachten. Verweigern einzelne Teilnehmerinnen und Teilnehmer darüber hinaus ihre Zustimmung zu einer eventuellen Nachnutzung ihrer nicht vollständig anonymisierten Daten, dürfen die Daten dieser Teilnehmerinnen und Teilnehmer nicht veröffentlicht werden. Wenn Daten nicht veröffentlicht werden können, sollten die entsprechenden Gründe an geeigneter Stelle (z.B. in einer Fußnote in der Publikation) benannt werden¹⁹. Solche Bedenken sollten jedoch umgekehrt nicht als Rechtfertigung dienen, Forschungsdaten nicht zu veröffentlichen, obwohl es rechtlich und ethisch unproblematisch wäre. Darüber hinaus soll bei bestehenden rechtlichen Einschränkungen dargelegt werden, welche Arten von aggregierten Daten oder anonymisierbaren Teildatensätzen dennoch veröffentlicht werden können.

„Scientific use files“ (SUF) bieten eine abgestufte Einschränkung des Nutzerkreises bei datenschutzrechtlichen Bedenken. SUF setzen eine faktische Anonymisierung voraus (d.h. eine Deanonymisierung ist nur mit einem unverhältnismäßig hohen Aufwand an Zeit, Kosten und Arbeitskraft möglich). Solche Datensätze werden auf Antrag nur Forscherinnen und Forschern von wissenschaftlichen Einrichtungen zu wissenschaftlichen Zwecken zur Verfügung gestellt. Diese Datensätze dürfen üblicherweise die Räume der nachnutzenden Institution nicht verlassen, Nachnutzende müssen eine Geheimhaltungs- und Nichtweitergabeerklärung unterzeichnen und die Rohdaten müssen nach einer definierten Frist gelöscht werden. Insbesondere bei besonders schützenswerten Daten wie etwa Selbstauskünfte über die ethnische Herkunft, politische, religiöse oder philosophische Überzeugungen sowie Daten über Gesundheit oder Sexualleben sollte man das Identifizierungsrisiko genau prüfen und die Verwendung eines SUF in Betracht ziehen. Die Anwendung einer SUF-Lizenz kann in Einzelfällen auch dann angemessen sein, wenn begründbar anzunehmen ist, dass Datensätze für missbräuchliche Zwecke genutzt werden²⁰.

Der Entscheidungsbaum in Abbildung 1 verdeutlicht die wichtigsten Aspekte bei der Verarbeitung und Bereitstellung von anonymen und personenbezogenen Daten. Diese Abbildung dient zur Illustration

17 Metschke, R., & Wellbrock, R. (2002). *Datenschutz in Wissenschaft und Forschung*. Berlin: Berliner Beauftragter für Datenschutz und Informationsfreiheit. <https://datenschutz-berlin.de/attachments/47/Materialien28.pdf?1166527077>

18 Deutsche Gesellschaft für Psychologie (Hrg.) (2016). *Ethisches Handeln in der Psychologischen Forschung*. Göttingen: Hogrefe, in Druck.

19 Taichman, D. B., Backus, J., Baethge, C., Bauchner, H., de Leeuw, P. W., Drazen, J. M., Fletcher, J., et al. (2016). Sharing clinical trial data: A proposal from the international committee of medical journal editors. *New England Journal of Medicine*, 374, 384–386. doi:10.1056/NEJMe1515172

20 Lewandowsky, S., & Bishop, D. (2016). Research integrity: Don't let transparency damage science. *Nature*, 529(7587), 459–461. <http://doi.org/10.1038/529459a>

von typischen Situationen. Sie bildet bei weitem nicht alle denkbaren Fälle ab, wie personenbezogene Daten zu handhaben sind; insbesondere bei der Nachnutzung personenbezogener Daten gilt es viele Aspekte zu berücksichtigen. Hierbei sei auf den umfassenden Leitfaden von Metschke und Wellbrock (2002) verwiesen (siehe Fußnote 17).

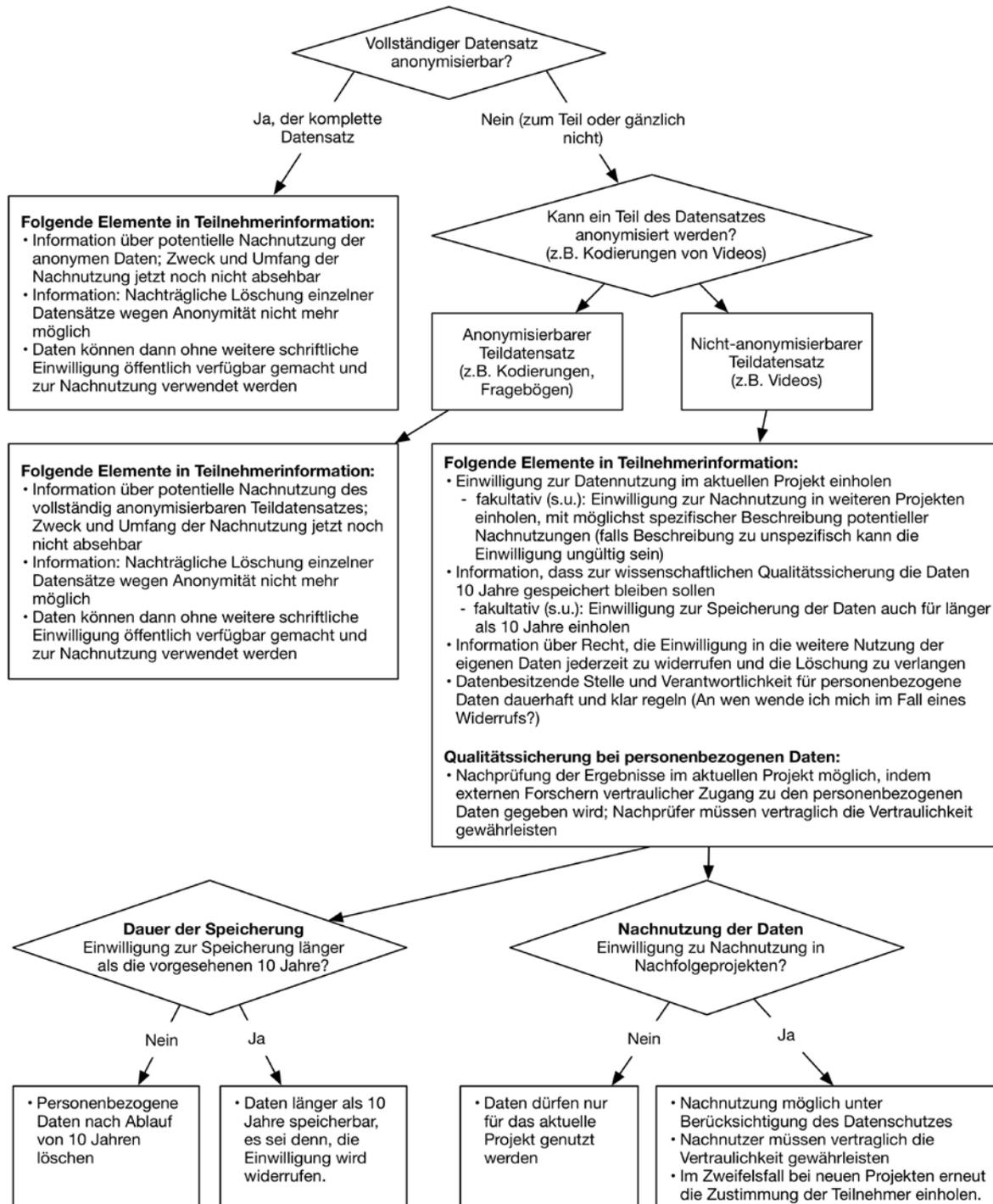


Abbildung 1: Vereinfachte Darstellung der wichtigsten Aspekte bei der Verarbeitung und Bereitstellung von anonymen und personenbezogenen Daten

Daten, die im Internet nicht frei zugänglich gemacht werden können, da sie nicht vollständig anonymisierbar sind, wie zum Beispiel bestimmte Bild- und Tondateien, Videoaufzeichnungen von Personen, Interviewtranskripte (einschl. klinische Interviews) sollten nicht-öffentlich archiviert und mindestens 10 Jahre aufbewahrt werden.

Wenn eine Anonymisierung bzw. eine Pseudonymisierung der Daten vorgenommen wird, ist dies nachvollziehbar zu dokumentieren (z.B. durch die Bildung von Kategorien, Löschung von Variablen etc.). Ein DGPs-Leitfaden zum Thema Forschungsethik wird demnächst veröffentlicht. Zudem gibt es bereits jetzt einschlägige Leitfäden²¹.

In Bezug auf Forschungstransparenz ist es wünschenswert, alle verwendeten Instrumente in der Originalsprache bereitzustellen, allerdings nur dann, wenn es sich um Instrumente handelt, bei denen durch die Zugänglichmachung deren Nutzungsverwertung nicht beeinträchtigt wird. Insbesondere bei kommerziellen Testverfahren, Trainingsmanualen und Testverfahren, die für die Anwendung bestimmt sind, müssen die fachlich gebotenen und urheberrechtlichen Einschränkungen berücksichtigt werden²². Dies gilt auch für Softwareapplikationen oder ähnliche Produkte, für die eine Patentanmeldung bzw. eine kommerzielle Nutzung vorgesehen ist. Falls eine Publikation auf bestehenden Daten beruht und ein Datennutzungsvertrag geschlossen wurde (wie z.B. bei SOEP oder NEPS²³), muss dieser ebenso berücksichtigt werden. Bei Kooperationen mit ausländischen Universitäten müssen darüber hinaus die dort geltenden Gesetze beachtet werden.

6. Zeitpunkt und Umfang der Datenveröffentlichung

In Bezug auf den Zeitpunkt und Umfang der Datenveröffentlichung unterscheiden wir zwischen zwei Typen.

Datenveröffentlichung Typ 1: Veröffentlichung von Daten, die Bestandteil einer Publikation sind

Mit Erscheinen einer Publikation soll die Person oder Gruppe, die die Daten erhoben hat (die „Datenbereitstellenden“), alle Primärdaten bereitstellen, die zur Reproduktion der publizierten Ergebnisse notwendig sind. Diese Art der Datenbereitstellung bezieht sich nicht nur auf Daten, die im Rahmen von DFG-Projekten erhoben wurden, sondern generell auf Daten, die publizierten Artikeln bzw. Berichten zugrunde liegen. Verantwortlich für die langfristige technische und inhaltliche Nachvollziehbarkeit des Datensatzes ist – sofern nicht anders geregelt – der Erstautor bzw. die Erstautorin, bzw. der korrespondierende Autor oder die korrespondierende Autorin.

In der Publikation sollten im Regelfall alle anderen Variablen genannt werden, die im Rahmen der bzw. den in der Publikation berichteten Studie(n) ebenfalls erhoben, in der Publikation aber nicht berücksichtigt wurden (vgl. "standard reviewer disclosure request", <https://osf.io/hadz3/>). Diese anderen Variablen werden später erst als Daten bereitgestellt, wenn sie entweder selbst Teil einer Publikation sind, oder wenn das drittmittelgeförderte Forschungsprojekt abgeschlossen ist und der Gesamtdatensatz bereitgestellt wird (s.u., *Veröffentlichung Typ 2*). Im Fall von umfangreichen Survey-Studien, in denen die Anzahl der erhobenen Variablen (z.B. Fragebogen-Items) sehr groß ist, reichen eine zusammenfassende Übersicht über die erhobenen Konstrukte bzw. Themenbereiche und ein Link zu einem Dokument mit detaillierteren Informationen aus.

Diese Datenveröffentlichung Typ 1 gilt auch bei Forschung aus laufenden universitären Mitteln (z.B. auch bei Veröffentlichungen, die aus Abschlussarbeiten entstanden sind), bei Forschung in Unternehmen sowie bei nicht-öffentlichen Drittmittelgebern. Es muss vor Datenerhebung abgeklärt werden, welche Art von bzw. welcher Teil der Primärdaten zu Zwecken der Reproduzierbarkeit und Nachnutzung bereitgestellt werden dürfen und welche(r) nicht. Im Regelfall basieren Publikationen nur auf solchen Daten, die im Sinne der Datenveröffentlichung Typ 1 bereitgestellt werden können. Ausnahmen (bspw. Sondervereinbarungen mit dem Auftraggeber bzgl. der Datenveröffentlichung) müssen in der Publikation begründet werden.

Datenveröffentlichung Typ 2: Veröffentlichung nach Projektabschluss

Gemäß den Empfehlungen der DFG soll der Gesamtdatensatz eines Projekts "unmittelbar nach Abschluss der Forschungen oder nach wenigen Monaten der Öffentlichkeit frei zur Verfügung gestellt werden"²⁴. Das beinhaltet auch alle relevanten Daten des Projekts, die zu diesem Zeitpunkt noch nicht

21 <http://www.psychdata.de/downloads/PsychData-Handbuch.pdf>

22 Für eine Zugänglichmachung kommerziell publizierter Testverfahren bedarf es immer der Zustimmung des Urhebers, bzw. des betreffenden Verlages.

23 SOEP = Sozio-ökonomisches Panel, <http://www.diw.de/de/soep>; NEPS = Nationales Bildungspanel, www.neps-data.de

24 siehe "Empfehlungen zur gesicherten Aufbewahrung und Bereitstellung digitaler Forschungsprimärdaten" der DFG (Januar 2009; http://dfg.de/download/pdf/foerderung/programme/lis/ua_inf_empfehlungen_200901.pdf)

Bestandteil einer Publikation sind. Dazu gehört auch das zum Verständnis der Daten nötige Material (insbesondere Auswertungsskripte, Codebücher; wenn möglich, auch Untersuchungs- bzw. Stimulusmaterial). Bei Simulationsstudien sollten sowohl der datengenerierende Code als auch die simulierten Daten bereitgestellt werden (außer wenn etwa die Menge der simulierten Daten die Speicherkapazität aktueller Repositorien übersteigen würde). Kann das zur Replikation eines Forschungsergebnisses notwendige Material nicht bereitgestellt werden, sollte dies begründet werden, z.B. in einer README-Datei im Repository.

Welche Daten als relevant anzusehen sind, liegt im Ermessen der Projektverantwortlichen. Irrelevant sind beispielsweise Daten, die auf fehlerhaft programmierten Experimenten beruhen oder im Rahmen von hochgradig explorativen Pilotstudien erhoben wurden. Um dem Problem des Publikationsbias entgegenzuwirken, müssen jedoch Daten von korrekt durchgeführten Studien, welche nicht das erwartete Ergebnis gebracht haben („Null-Befunde“), offen bereitgestellt werden; es dürfen auf keinen Fall nicht-hypothesenkonforme Ergebnisse unterdrückt werden. Bei der Begutachtung von Abschlussberichten sollte darauf geachtet werden, ob die Ergebnisse und Primärdaten der im Antrag angekündigten Studien auch vorhanden sind.

Der Zeitpunkt, zu dem ein Projekt als abgeschlossen gilt, kann von der Komplexität des Projekts und von weiteren Faktoren abhängig sein; bei drittmittelgeförderten Projekten gilt jedoch im Regelfall, dass der Abschlussbericht den Projektabschluss darstellt. Die Datenveröffentlichung soll so zeitnah wie möglich nach Abschluss des entsprechenden Forschungsprojekts erfolgen.

Im Falle eines Embargos (siehe 7.1 und 7.2) kann die Speicherung der Daten im Repository zum Zeitpunkt des Projektabschlusses zunächst nicht-öffentlich erfolgen. Das bedeutet, die Dateien werden zeitnah mit dem Projektabschluss hochgeladen und erhalten bereits einen persistenten Identifikator, stehen aber der Öffentlichkeit noch nicht zur freien Verfügung. Nach dem Ablauf der Embargoperiode werden die entsprechenden Dateien dann öffentlich verfügbar gemacht.

Die Veröffentlichung Typ 2 bezieht sich insbesondere auf Projekte von öffentlichen (bzw. falls möglich auch nicht-öffentlichen) Drittmittelgebern, die einen definierten Erhebungsumfang und einen definierten Projektabschluss haben. Bei kontinuierlich laufenden Projekten (z.B. aus universitären Mitteln) ist ein "Projektabschluss" möglicherweise schwer zu definieren; hier gilt jedoch auf jeden Fall die Regelung zu "Veröffentlichung Typ 1".

7. Rechte und Pflichten der Erst- sowie der Nachnutzenden

Die Idee einer offenen Wissenschaft, die Verpflichtung zur Veröffentlichung von Forschungsdaten und die Möglichkeit der Nachnutzung veröffentlichter Forschungsdaten stellt sowohl Datenbereitstellende als auch Nachnutzende vor Herausforderungen. Beide Parteien haben konkrete Rechte, sind dadurch aber auch mit konkreten Pflichten konfrontiert.

7.1. Rechte der Datenbereitstellenden

Forscherinnen und Forscher, die Primärdaten produzieren, d.h. Datenbereitstellende, haben das Recht der ersten Nutzung dieser Daten. Falls mehrere Forscherinnen und Forscher an einem Projekt beteiligt sind, sollten sie die Rechte der Erstnutzung innerhalb ihrer Gruppe vorab regeln.

Datenbereitstellende können ein Embargo für die Datennachnutzung definieren. Das bedeutet, dass die bisher noch nicht in Publikationen verwendeten Daten zwar – wie von der DFG gefordert – auf einem Repository gespeichert werden, allerdings für eine gewisse Karenzzeit nicht für Dritte zugänglich sind (z.B. durch einen Passwortschutz oder ein Verzeichnis im Repository, das nicht der Öffentlichkeit zugänglich ist). Somit können sie noch nicht von Dritten als Grundlage für eigene Analysen genutzt werden.

Datenbereitstellende haben ferner das Recht, zu erfahren, wer ihre Daten für welchen Zweck nutzt. Sie haben das Recht, vor einer Veröffentlichung der Nachanalyse ihrer Daten – insbesondere dann, wenn die Nachanalyse die Originalbefunde nicht reproduzieren konnte – von den Nachnutzenden informiert zu werden (siehe Abschn. 7.3: Pflichten der Nachnutzenden). Manche Repositorien bieten die technische Möglichkeit, die Datenbereitstellenden darüber zu informieren, wann und von wem ein Datensatz heruntergeladen wurde. Diese *Information* über den Download von Daten sollte im Allgemeinen jedoch nicht mit einer *Zugangsbeschränkung* für bestimmte Personengruppen verbunden sein (außer in begründeten Ausnahmefällen; siehe Punkt 5 unter „Scientific use files“). Das heißt, durch die Wahl eines geeigneten Repositoriums kann über einen Download der Daten informiert

werden. Unabhängig davon *muss* über jegliche veröffentlichte Nachnutzung (sei es in Publikationen, Präsentationen oder Blogbeiträgen) durch die Nachnutzenden informiert werden.

7.2. Pflichten der Datenbereitstellenden

Datenbereitstellende sind verpflichtet, die Veröffentlichung ihrer Daten so zu gestalten, dass eine Nachnutzung sinnvoll möglich ist. Dazu gehört, dass (a) Einwilligungen zur Datennutzung von den Forschungsteilnehmenden vorliegen und dass (b) alle Daten sowie die entsprechenden Metadaten, welche den Datensatz als Ganzes beschreiben, sorgfältig und verständlich beschrieben sind. Um Forschenden das entsprechende Datenmanagement zu erleichtern, werden zurzeit geeignete Hilfsmittel entwickelt (z.B. DataWiz des ZPID).

Datenbereitstellende haben das Recht, ein Embargo zu definieren (siehe Abschn. 7.1); in diesem Falle sind Datenbereitstellende jedoch verpflichtet, dieses Embargo gemeinsam mit der Speicherung der zunächst nicht frei verfügbaren Primärdaten bekannt zu machen und explizit das Ende der Embargoperiode zu benennen. Dazu liegt z.B. im Repositorium eine öffentlich zugängliche Datei, die das Embargo beschreibt, das Ende benennt und die erhobenen Daten beschreibt (z.B. durch einen Verweis auf das Codebuch).

Nach Ablauf der Embargoperiode werden die Daten öffentlich verfügbar gemacht und stehen in der Regel uneingeschränkt und offen für Nachnutzung frei, auch wenn die Datenbereitstellenden die Daten selbst noch nicht für Publikationen genutzt haben. Da die Erfahrung lehrt, dass mit Abschluss eines Forschungsprojekts nicht alle geplanten Publikationen bereits realisiert sind, sieht die DGPs im Regelfall ein Embargo von nicht länger als 5 Jahren nach Projektabschluss als angemessen an. Längere Embargoperioden müssen begründet werden (z.B. in der Datei im Repositorium, die auch das Ende des Embargos benennt).

In aller Regel sollten Daten, die als Teil einer Publikation veröffentlicht wurden (Datenveröffentlichung Typ 1), nicht mit einem Embargo belegt werden können. In Ausnahmefällen (z.B. bei einer extrem aufwändigen Datenerhebung und wenn bestimmte weiterführende Forschungsfragen zu diesem Datensatz bereits angelegt sind) können auch diese Daten mit einem Embargo belegt werden. Dieses Embargo sollte jedoch deutlich kürzer sein als das für Datenveröffentlichung Typ 2 definierte Embargo. Darüber hinaus muss auch hier sichergestellt sein, dass die Daten ab dem Zeitpunkt der Veröffentlichung der Publikation auf Anfrage hin für eine Reproduktion der berichteten Ergebnisse zur Verfügung stehen.

7.3. Pflichten der Nachnutzenden

Um die Vorteile einer Datennachnutzung zu gewährleisten und gleichzeitig ihre Gefahren zu minimieren, sind Transparenz, Vertrauen und Kooperationsbereitschaft zwischen allen Beteiligten essentiell. Nachnutzende sollten daher Kontakt zu den Datenbereitstellenden aufnehmen, um eine möglichst valide Nutzung der Daten zu ermöglichen und Missverständnisse zu vermeiden. Bei jeder Art der Datennachnutzung gilt der Grundsatz, dass sie immer mit dem Ziel des größtmöglichen Erkenntnisgewinns in Bezug auf eine Forschungsfrage stattfinden soll. Entsprechend darf eine Nachnutzung von Daten nicht durch das Ziel motiviert sein, die Reputation der Datenbereitstellenden zu beschädigen. Umgekehrt dürfen aber Datenbereitstellende nicht verhindern, dass in Folge einer Reanalyse Ergebnisse veröffentlicht werden, die den Originalergebnissen widersprechen oder Fehler in den jeweiligen Originalarbeiten aufdecken.

Insbesondere wenn Nachnutzende beabsichtigen, ihre Reanalysen öffentlich zu machen (z.B. in Präsentationen, Publikationen oder Blogbeiträgen), müssen die Datenbereitstellenden darüber informiert werden, (1) dass und mit welchem Ziel die Datennachnutzung erfolgt, (2) welche Ergebnisse bei dieser Nachnutzung erzielt wurden und (3) wo die Ergebnisse der Datennachnutzung veröffentlicht werden sollen.

In jedem Fall gilt: Wer Daten anderer für eigene Forschung nutzt, muss diese Daten adäquat zitieren²⁵. Dafür ist es hilfreich, wenn den Daten im Repositorium ein Zitationshinweis (mit Angabe eines persistenten Identifikators) beiliegt. Eine Publikation, die auf Datennachnutzung beruht, sollte keine eigene Variante des Datensatzes bereitstellen, sondern immer auf den persistenten Identifikator des

25 Data Citation Synthesis Group (2014). *Joint declaration of data citation principles*.
<https://www.force11.org/datacitation>

Originaldatensatzes verweisen, auch wenn der Datensatz in Folge der Reanalyse verändert wurde (z.B. durch Bildung neuer Variablen). Eventuelle Datentransformationen und Neuberechnungen von Variablen sind durch reproduzierbare Analyseskripte zu dokumentieren.

Nachnutzende sind außerdem dazu verpflichtet, die Daten so zu analysieren, dass Rechte der Teilnehmerinnen und Teilnehmer der Originalstudie nicht verletzt werden. Die Nachnutzung von Daten unterliegt den gleichen datenschutz- und urheberrechtlichen Bestimmungen wie deren Erstnutzung. Hierfür sind die Nachnutzenden entsprechend verantwortlich.

Im Falle einer Datennachnutzung gelten selbstverständlich die gleichen Anforderungen an Transparenz und wissenschaftliche Sorgfalt wie bei der Erstnutzung. Für eine wissenschaftliche Neuinterpretation der Daten müssen die wissenschaftlichen Standards zum Zeitpunkt der Nachnutzung gelten. Bei der *Bewertung* von Originalanalysen im Kontext einer Reanalyse müssen diejenigen Standards angelegt werden, die zur Zeit der Originalanalysen galten.

Das Angebot einer Koautorenschaft. Die Frage, ob bzw. unter welchen Umständen die Datenbereitstellenden als Koautorinnen bzw. -autoren einer Publikation, die aus der Reanalyse entsteht, geführt werden müssen, kann nicht abstrakt geregelt werden, sondern muss im Einzelfall entschieden werden. So wie es jetzt bereits eine Abwägungsfrage ist, welche Beteiligten an einem Projekt einen substantiellen Beitrag geliefert haben, der eine Koautorenschaft rechtfertigt, muss sich diese Frage auch bei der Datennachnutzung stellen. Allerdings erscheint uns folgende Kategorisierung sinnvoll, die einige exemplarische (aber nicht erschöpfende) Beispiele für drei Kategorien der Datennachnutzung aufzeigt:

- *Einfache Datennachnutzung:* z.B. Effektstärken für Meta-Analysen aus den Daten extrahieren; Mittelwerte oder Verteilungen von Variablen berechnen. Bei dieser Art der Nachnutzung wird üblicherweise keine Koautorenschaft angeboten. Reanalysen, welche ausschließlich versuchen, die Originalbefunde zu reproduzieren (und zum Beispiel in Blogs berichtet werden), zählen im Allgemeinen zu dieser einfachen Nachnutzung.
- *Weiterführende Datennachnutzung:* hier unterscheiden wir zwei Unterkategorien:
 - *Weiterführende Fragestellungen, welche die Forschungsfrage der Originalpublikation ergänzen bzw. erweitern:* z.B. ein vorhandener Datensatz wird reanalysiert; dabei zeigt sich, dass der zentrale publizierte Effekt von einer weiteren erhobenen Variablen moderiert wird; in einer Folgestudie wird dieser Moderationseffekt konfirmatorisch geprüft, somit wird die Theorie der Originalautoren konzeptuell weiterentwickelt. In solchen Fällen soll den Datenbereitstellenden eine Koautorenschaft angeboten werden.
 - *Orthogonale Fragestellungen, welche die Daten zu einer anderen Fragestellung als die der Originalpublikation nutzen:* z.B. eine Forscherin entwickelt ein neuartiges Maß der Reliabilität und nutzt eine Vielzahl von bereitgestellten Fragebogendaten verschiedener Forscherinnen und Forscher (die auf inhaltliche Fragen fokussiert waren), um ihr neues Reliabilitätsmaß darauf anzuwenden. Bei dieser Art der Nachnutzung wird den Datenbereitstellenden nicht notwendigerweise eine Koautorenschaft angeboten.

Wenn unklar ist, ob es sich um eine einfache oder eine weiterführende Nachnutzung im oben beschriebenen Sinne handelt, sollte im Zweifelsfall immer der Kontakt zu den Originalautorinnen bzw. -autoren gesucht werden.

Eine Koautorenschaft sollte außerdem immer angeboten werden, wenn über das reine Bereitstellen von Daten ein darüber hinaus gehender Beitrag der Datenbereitstellenden zur Datennachnutzung vorliegt. Ein solcher Beitrag kann zum Beispiel schon dann vorliegen, wenn wichtige Hinweise und Unterstützung bei der Reanalyse der Daten gegeben werden.

Wenn eine Datennachnutzung über die oben skizzierten Szenarien „einfacher Datennachnutzung“ hinausgeht, halten wir es für empfehlenswert, dass Datenbereitstellende und Nachnutzende vorab eine Vereinbarung über die Datennachnutzung treffen, in der unter anderem die Frage von Koautorenschaften geregelt ist. In einer solchen Vereinbarung könnten ggf. auch Fragen des Datenschutzes angesprochen werden; zudem kann geregelt werden, in welcher Form die Datenbereitstellenden die Ergebnisse der Datennachnutzung kommentieren können (falls sie nicht ohnehin Koautor(in) der entsprechenden Publikation sind). Eine solche Vereinbarung darf jedoch nicht dazu genutzt werden, bestimmte Forscher(gruppen) selektiv von der Nachnutzung auszuschließen.

Wurde von den Datenbereitstellenden ein Embargo definiert, ist dies zu respektieren. Wenn klar gegen ein valides Embargo (oder eine andere vertraglich getroffene Regelung) verstoßen wurde, sollten Zeitschriften von einer Publikation absehen bzw. bei Bekanntwerden einer solchen Verletzung die Publikation zurückziehen.

Anhang E illustriert die verschiedenen Datenveröffentlichungstypen und Möglichkeiten der Datennachnutzung an einem Beispiel.

Die hier dargestellten Regeln der Datennachnutzung können auch dem Repository beigelegt werden (z.B. in einer Readme-Datei), damit potentielle Datennachnutzende auch wirklich mit ihnen vertraut werden, insbesondere wenn sie die vorliegenden Empfehlungen vorher nicht kannten.

Anhang A: Konkretisierungsbeispiele

Die DGPs ist sich bewusst, dass die offene Bereitstellung von Forschungsdaten neue Anforderungen stellt, die bisher oft nicht im Forschungsprozess abgebildet wurden. Daher möchte sie hier Empfehlungen geben, wie die Bereitstellung dergestalt ablaufen könnte, dass sie hohen Qualitätsstandards der Forschungstransparenz entspricht. Darüber hinaus bietet die DGPs Workshops an, die einen praktischen Einstieg in das Thema ermöglichen.

1. Dauer der Aufbewahrung: Die DFG schreibt vor, die Primärdaten *mindestens* 10 Jahre aufzubewahren. Diese Regelung leitet sich aus der Guten Wissenschaftlichen Praxis ab, die insbesondere auf die Vermeidung von wissenschaftlichem Fehlverhalten abzielt und sicherstellen soll, dass in einem Verdachtsfall die Forschungsergebnisse nachvollzogen und geprüft werden können. Das impliziert nicht, dass die Daten nach diesem Zeitraum gelöscht werden *sollten*. Da es, außer bei Studien mit extrem großen Datenmengen, kein Argument für die Löschung anonymisierter Daten gibt, befürwortet die DGPs im Normalfall eine unbegrenzte Sicherung ohne zeitgebundenen "Löschmechanismus" oder „Löschzwang“.
2. Die Speicherung in einem nicht-proprietären Datenformat (wie z.B. csv-Dateien oder Text-Dateien) ist einem proprietären Dateiformat (.mat-Datei, SPSS-Datei, SAS-Datei) vorzuziehen, damit die Transparenz nicht auf Besitzer(innen) ggf. teurer Spezialsoftware beschränkt wird. Entsprechend sollen die Daten so aufbereitet sein, dass sie ohne Spezialsoftware auslesbar und nutzbar sind. Zum Beispiel würde ein binärer Output aus einer unbekanntem Experimentalsoftware dem Ziel der Forschungstransparenz nicht genügen; hier sind die Daten soweit aufzubereiten, dass sie in einem nachvollziehbaren Datenformat vorliegen. Wenn das nicht möglich ist, sollte klar dokumentiert werden, mit welcher Software die Daten auslesbar sind.
3. Neben der technischen Zugänglichkeit muss auch die inhaltliche Verständlichkeit der Daten gewährleistet sein. Alle Variablen müssen in einem digitalen Codebuch dokumentiert sein²⁶. Es muss klar sein, welche manipulierte bzw. gemessene Variable in der Publikation zu welcher Variablen im Datensatz gehört. Idealerweise liegen den Daten auch Analyseskripte bei (z.B. R-Skripte oder SPSS-Syntax), die die publizierten Ergebnisse reproduzieren.
4. Der Speicherort sollte einen persistenten Identifikator erhalten (z.B. eine persistente URL oder, falls möglich, einen DOI). Das erlaubt eine einheitliche Zitation der Daten. Der Speicherort soll in der Publikation, die auf diesen Daten beruht, genannt werden, so dass Forschungsdaten auch gefunden werden können.
5. Neben der Dokumentation auf Variablenebene (Codebuch) ist auch eine Dokumentation auf Studienebene und auf Datenobjektebene (Dateien, Versionen, etc.) erforderlich. Das kann zum Beispiel durch eine README-Datei im Repository gelöst werden, die einen Überblick über die archivierten Dateien gibt, und gegebenenfalls Hinweise zur Reproduktion gibt (z.B., Welche Software ist notwendig? In welcher Reihenfolge sind die Skripte auszuführen? Wie soll der Datensatz zitiert werden?).
6. Es wird empfohlen, neben den Daten auch Instrumente, Software und Materialien bereitzustellen²⁷.

26 Datenmanagement und Data Sharing in der Psychologie. Einführung und Manual. Herausgegeben vom Forschungsdatenzentrum PsychData des Leibniz-Zentrums für Psychologische Information und Dokumentation (ZPID), Trier, 15. Mai 2013.

27 Für eine Zugänglichmachung kommerziell publizierter Testverfahren bedarf es immer der Zustimmung des Urhebers, bzw. des betreffenden Verlages.

Anhang B:**Mögliche Formulierung "Einwilligungserklärung" (für den Fall, dass von vornherein nur anonyme Daten anfallen)***Freiwilligkeit:*

Die Teilnahme an der Studie ist freiwillig. Sie können jederzeit und ohne Angabe von Gründen Ihre Einwilligung zur Teilnahme an dieser Studie widerrufen, ohne dass Ihnen daraus Nachteile entstehen. Auch wenn Sie die Studie vorzeitig abbrechen, haben Sie Anspruch auf eine entsprechende Vergütung oder die entsprechende Anzahl Versuchspersonenstunden für den bis dahin erbrachten Zeitaufwand. Sie können Ihre Einwilligung zur Speicherung der Daten bis zum Ende der Datenerhebung widerrufen, ohne dass Ihnen daraus Nachteile entstehen.

Datenschutz:

Da keine personenbezogenen Daten erhoben werden, ist nach Abschluss der Datenerhebung prinzipiell keine Zuordnung mehr zwischen den Daten im Datensatz und Ihrer Person möglich – der Datensatz ist anonym. Entsprechend ist nach Abschluss dieser Datenerhebung auch keine gezielte Löschung Ihres persönlichen Datensatzes möglich, da wir diesen nicht zuordnen können.

Verwendung der anonymisierten Daten

Die Ergebnisse und Daten dieser Studie werden als wissenschaftliche Publikation veröffentlicht. Dies geschieht in anonymisierter Form, d. h. ohne dass die Daten einer spezifischen Person zugeordnet werden können. Die vollständig anonymisierten Daten dieser Studie werden als offene Daten im Internet in einem Datenarchiv namens _____ zugänglich gemacht. Damit folgt diese Studie den Empfehlungen der Deutschen Forschungsgemeinschaft (DFG) und der Deutschen Gesellschaft für Psychologie (DGPs) zur Qualitätssicherung in der Forschung.

Hiermit versichere ich, dass ich die oben beschriebenen Teilnehmerinformationen verstanden habe und mit den genannten Teilnahmebedingungen einverstanden bin.

Anhang C:

Auszug aus den Leitlinien Ethikkommission (in Druck)

Information, Aufklärung, Einverständnis, Bundesdatenschutzgesetz
--

1. Teilnehmerinformation

Entspricht die hier vorgelegte schriftliche Teilnehmerinformation der finalen Version?

Enthält die Einverständniserklärung einen Zusatz mit einer ausdrücklichen Einwilligung des Probanden, falls geplant ist, die Studie zu einem späteren Zeitpunkt mit einer erneuten Kontaktaufnahme/Datenerhebung fortzuführen?

Falls ja, ist dargelegt, wie in diesem Fall der Schutz der personenbezogenen Daten gewährleistet ist?

Ist für die Beteiligung von Minderjährigen ein der Altersgruppe entsprechender Informationsbogen erstellt?

2. Datenschutz

Ist der Abschnitt Datenschutz in die Teilnehmerinformation integriert und deutlich hervorgehoben?

Ist der Abschnitt Datenschutz ausführlich genug und laienverständlich geschrieben?

3. Pseudonymisierung

Ist die Art und Weise der Pseudonymisierung verständlich beschrieben worden und stellt sie den Datenschutz gemäß den gesetzlichen Vorgaben sicher (Bundesdatenschutzgesetz §3, Abs. 6)?

Abschlusscheck

Erfüllt die Teilnehmerinformation die erforderlichen Kriterien?

Erfüllt die Einverständniserklärung die erforderlichen Kriterien?

Sind Bild- und/oder Tonaufnahmen geplant? Falls ja, sind hierfür gesonderte Einwilligungserklärungen einzuholen.

Liegen Teilnehmerinformation und Einverständniserklärung als voneinander getrennte Bogen vor?

Sind eventuell spezielle Teilnehmerinformationen (z. B. für EEG-, MRT-, TMS-Studien) nötig?

Werden im Falle von Untersuchungen an Minderjährigen und anderen vulnerablen Gruppen besondere Schutzmaßnahmen vorgenommen?

Wie werden im Falle von Onlinestudien die allgemeinen ethischen Prinzipien eingehalten (insbesondere Gewährleistung der Freiwilligkeit der Teilnahme, Anonymität, Datenschutz)?

Weitere Details können den Ethik-Richtlinien der DPGs www.dgps.de/dgps/aufgaben/ethikrl2004.pdf (3a-e; 6,9) sowie den DPGs-Richtlinien zur Antragstellung (Punkt C) entnommen werden:
www.dgps.de/kommissionen/ethik/hinweise_zur_antragstellung.pdf

Anhang D:

Auszug aus dem Antragsformular der Ethikkommission der Fakultät für Psychologie und Pädagogik der Ludwig-Maximilians-Universität München (Stand: 14.3.2016)²⁸

4) Angaben zum Datenschutz

4.1 Welche personenbezogenen Daten (z. B. Name, Email-Adresse, Wohnort, weitere personenbezogene Daten) werden erhoben?

4.2 Sind Video- oder Tonaufnahmen oder andere Verhaltensregistrierungen vorgesehen?

4.3 Wie wird die Anonymisierung oder Pseudonymisierung der erhobenen Daten gesichert?

4.4a Wann werden die gespeicherten Daten gelöscht?

Hinweis für Antragsteller: Die personenbezogenen Daten (z.B. Erhebung von Name, Emailadresse, weitere personenbezogene Daten) sind, sobald sie nicht mehr zur Rekrutierung der Teilnehmerinnen und Teilnehmer oder für Nachfragen benötigt werden, zu löschen. Es bietet sich an, einen entsprechenden Abschnitt in die Datenschutzerklärung und Einwilligungserklärung einzufügen. Beispielsweise: „Die Löschung Ihrer personenbezogenen Daten erfolgt gemäß der Grundsätze der Forschung am Menschen der Deutschen Forschungsgesellschaft (DFG). Die personenbezogenen Daten werden gelöscht, sobald sie nicht mehr zur Rekrutierung der Teilnehmerinnen und Teilnehmer oder für Nachfragen benötigt werden“.

Beachten: Die erfolgte Löschung personenbezogener Daten ist zu protokollieren und auf Anforderung nachzuweisen.

Demgegenüber müssen vollständig anonymisierte Rohdaten nicht gelöscht werden und sollten gemäß der „Leitlinien zum Umgang mit Forschungsdaten“ der DFG in offen zugängliche wissenschaftliche Repositorien/Datenbanken überführt werden. Nur dann ist eine Replizierbarkeit der Ergebnisse für weiterführende Forschungen gewährleistet. Siehe hierzu

http://www.dfg.de/download/pdf/foerderung/antragstellung/forschungsdaten/richtlinien_forschungsdaten.pdf, Punkt 2 und 3. Über die mögliche Veröffentlichung der vollständig anonymisierten Daten sollten aber die Teilnehmerinnen und Teilnehmer in der Probandeninformation informiert werden.

4.4b Im Antrag ist auszuführen, wie die Pseudonymisierung oder Anonymisierung sowie Art und Zeit der Löschung der personenbezogenen Daten durch verantwortliche Personen (welche?) durchgeführt wird.

4.5 Können Studienteilnehmer jederzeit die Löschung ihrer Daten verlangen?

28 <http://www.fak11.lmu.de/forschung/ethikkommission/>

Anhang E: Datenveröffentlichungstypen und Datennachnutzung in einem Beispiel

Die hier vorliegende Konkretisierung der DFG-Leitlinie soll im Folgenden an einem fiktiven Beispiel verdeutlicht werden, welches alle angesprochenen Szenarien der Datenbereitstellung und – nachnutzung vereint. In der Realität werden selten all diese Szenarien so gebündelt in einem Projekt auftreten. Das Beispiel dient vielmehr der komprimierten Veranschaulichung.

Ein großangelegtes Forschungsprojekt plant einen Längsschnitt über 10 Jahre, bei dem eine Vielzahl von Variablen gemessen wird. Schon bei der Beantragung bei der DFG wurde ein Absatz zum Datenmanagement aufgenommen und es wurden Personalmittel beantragt, um die Daten angemessen dokumentieren und für eine Veröffentlichung aufbereiten zu können. Die Daten werden vor Veröffentlichung nach dem HIPAA Safe Harbor Standard und entlang der Richtlinien von Hrynaszkiwicz et al. (2010) (<http://www.bmj.com/content/340/bmj.c181>) anonymisiert (das bedeutet, dass z.B. Namen, Email-Adressen, Fotos oder spezifische geographische Daten entfernt werden).

Zwei Jahre nach Projektbeginn ist die erste Erhebungswelle aufbereitet und die beteiligte Forscherin A publiziert einen Artikel auf Basis von drei Variablen (Lebenszufriedenheit, Neurotizismus und Anzahl der Freunde). Die Primärdaten dieser drei Variablen werden zusammen mit der Publikation veröffentlicht (Datenveröffentlichung Typ 1). Darüber hinaus verweist die Autorin im Artikel auf eine öffentlich zugängliche Variablenliste der anderen Variablen dieser Erhebung.

- Forscher V macht eine Reanalyse dieser drei Variablen und entdeckt mithilfe eines moderneren nicht-parametrischen Analyseverfahrens, dass der Effekt im Mittel sogar deutlich stärker ist, als ursprünglich publiziert. Er informiert A über diese Ergebnisse, berichtet die Reanalyse in einem Blogeintrag und auf PubPeer und zitiert dabei die Primärdaten und den Originalartikel. Forscherin A verlinkt im Repositorium zu dieser Reanalyse.

Der im Projekt beteiligte Forscher B publiziert in einem weiteren Artikel mittels einer explorativen Analyse einen recht überraschenden Befund zur ersten Erhebungswelle (basierend auf anderen Variablen). Er macht die Daten nach Datenveröffentlichungstyp 1 öffentlich.

- Forscherin W reanalysiert die Daten und findet einen Fehler in der Analyse. Sie informiert Forscher B über ihre Erkenntnisse; dieser bestätigt den Fehler. B und W entscheiden, gemeinsam einen Kommentar an die Zeitschrift zu schicken, was zur Folge hat, dass der Originalartikel in Bezug auf die fehlerhafte Analyse korrigiert wird.

Forscher C veröffentlicht eine weitere Publikation inklusive Daten zur Welle 1 des Längsschnittprojekts.

- Forscher X kann diese Primärdaten im Rahmen einer Metaanalyse verwenden (da es ein within-subject-Design ist, konnte der korrekte Standardfehler des Effektstärkenschätzers nur mit Kenntnis der Primärdaten ermittelt werden). Der Datensatz wird in der Publikation zitiert; eine Koautorenschaft wird nicht angeboten, da das bei Metaanalysen nicht üblich ist.
- Forscher Y kann die Primärdaten nutzen, um in einer Bayesianischen Analyse (eines anderen Datensatzes) seine priors zu kalibrieren, indem er sich auf die nun bekannte Verteilung der bestehenden Studie beziehen kann. Er zitiert den Datensatz, bietet aber keine Koautorenschaft an.
- Forscherin Z prüft explorativ eine neue Fragestellung an dem veröffentlichten Datensatz und möchte diese Analyse als "Studie 1" berichten (in Studie 2 wird die Hypothese präregistriert und confirmatorisch an einem neuen Datensatz geprüft). Sie bietet Forscher C eine Koautorenschaft an; dieser lehnt jedoch ab, da er aus Zeitmangel im Moment keinen substantiellen Beitrag zum Manuskript liefern kann.
- Forscherin Q macht eine Reanalyse und kommt zu dem Schluss, dass der Effekt nicht mehr vorhanden ist, wenn man "Depressive Symptome" für "Neurotizismus" kontrolliert. Sie informiert Forscher C über diese Reanalyse und reicht einen Kommentar bei der Zeitschrift ein. Forscher C hält diese Analyse nicht für sinnvoll und veröffentlicht seinerseits eine Replik auf die Reanalyse. Auf diese Weise entsteht ein wissenschaftlicher Diskurs.

Analog werden mit den Daten der weiteren Erhebungswellen Publikationen angefertigt.

In der dritten Welle werden zusätzlich autobiographische Erinnerungen erhoben und nach Inhaltskategorien kodiert. Den Teilnehmerinnen und Teilnehmern wird in der Einwilligungserklärung mitgeteilt, dass die konkreten Texte nicht offen zur Verfügung gestellt werden, da eventuell ein Personenbezug hergestellt werden könnte. Es wird nachgefragt, ob die Daten zu Zwecken der

wissenschaftlichen Qualitätssicherung auch länger als die vorgesehenen 10 Jahre gespeichert werden dürfen. Es wird die Einwilligung eingeholt, diese autobiographischen Erinnerungen in der aktuellen Studie zu nutzen. Es wird gefragt, ob die Daten darüber hinaus in weiterführenden Forschungsprojekten (die kurz beschrieben werden) genutzt werden dürfen. Es wird ein Kontakt benannt, bei dem der weiteren Datenverarbeitung widersprochen und eine Löschung der personenbezogenen Daten verlangt werden kann.

Forscherin E publiziert zu den autobiographischen Erinnerungen; allerdings können die Texte aus Datenschutzgründen nicht mitveröffentlicht werden. Forscherin E begründet dies in der Publikation und in einer Readme-Datei im Repositorium. Allerdings werden die Häufigkeiten der Inhaltskategorien pro Person veröffentlicht (welche anonym sind), so dass die berichteten Analysen reproduzierbar sind.

- Forscher U hat eine Idee, wie die Texte der Erinnerungen mit einem alternativen Kodierschema ausgewertet werden könnten. Er erfragt bei Forscherin E die Primärdaten und sichert schriftlich die Vertraulichkeit zu. Die Texte derjenigen Teilnehmer und Teilnehmerinnen, die einer Nachnutzung zugestimmt haben, werden Forscher U dafür bereitgestellt. Da Forscherin E substantiell bei der Aufbereitung der Texte geholfen hat, um die Analysen von Forscher U zu ermöglichen, bietet Forscher U ihr eine Koautorenschaft an, die sie auch annimmt.

Nach Abschluss des Gesamtprojekts (d.h. drei Erhebungswellen und 10 Jahre nach Projektbeginn) entscheiden die Projektleiter und Projektleiterinnen, ein Embargo von 5 Jahren in Anspruch zu nehmen, um die bis dahin noch nicht verwerteten Daten exklusiv nutzen zu können. In diesem Zeitraum entstehen weitere Publikationen. Die Datenveröffentlichung gemäß Typ 1 erfolgt jeweils selektiv für diejenigen Primärdaten aus dem Gesamtdatensatz, die zur Reproduktion der jeweils berichteten Ergebnisse notwendig sind.

Nach Ablauf der Embargoperiode wird dann der anonyme Gesamtdatensatz veröffentlicht, der nun zum allergrößten Teil aus bereits in Publikationen verwerteten Daten besteht. Im Repositorium wird genau definiert, wie der Datensatz zu zitieren ist. Der Datensatz ist nun ein öffentliches Gut. Zehn Jahre nach Projektabschluss werden die personenbezogenen Daten derjenigen Teilnehmer und Teilnehmerinnen gelöscht, die einer längerfristigen Speicherung nicht zugestimmt haben. Im Laufe der nächsten Jahre werden zahlreiche Publikationen von Drittforschern und -forscherinnen basierend auf dem Datensatz veröffentlicht, was zu einer großen Anzahl an Zitationen für die Datenbereitstellenden führt. In einigen Fällen wurden auch weitere Koautorenschaften angeboten. In jedem Fall haben die Datenbereitstellenden mehr Zitationen und mehr Koautorenschaften, als es im bisherigen Modell einer "closed science" der Fall gewesen wäre.