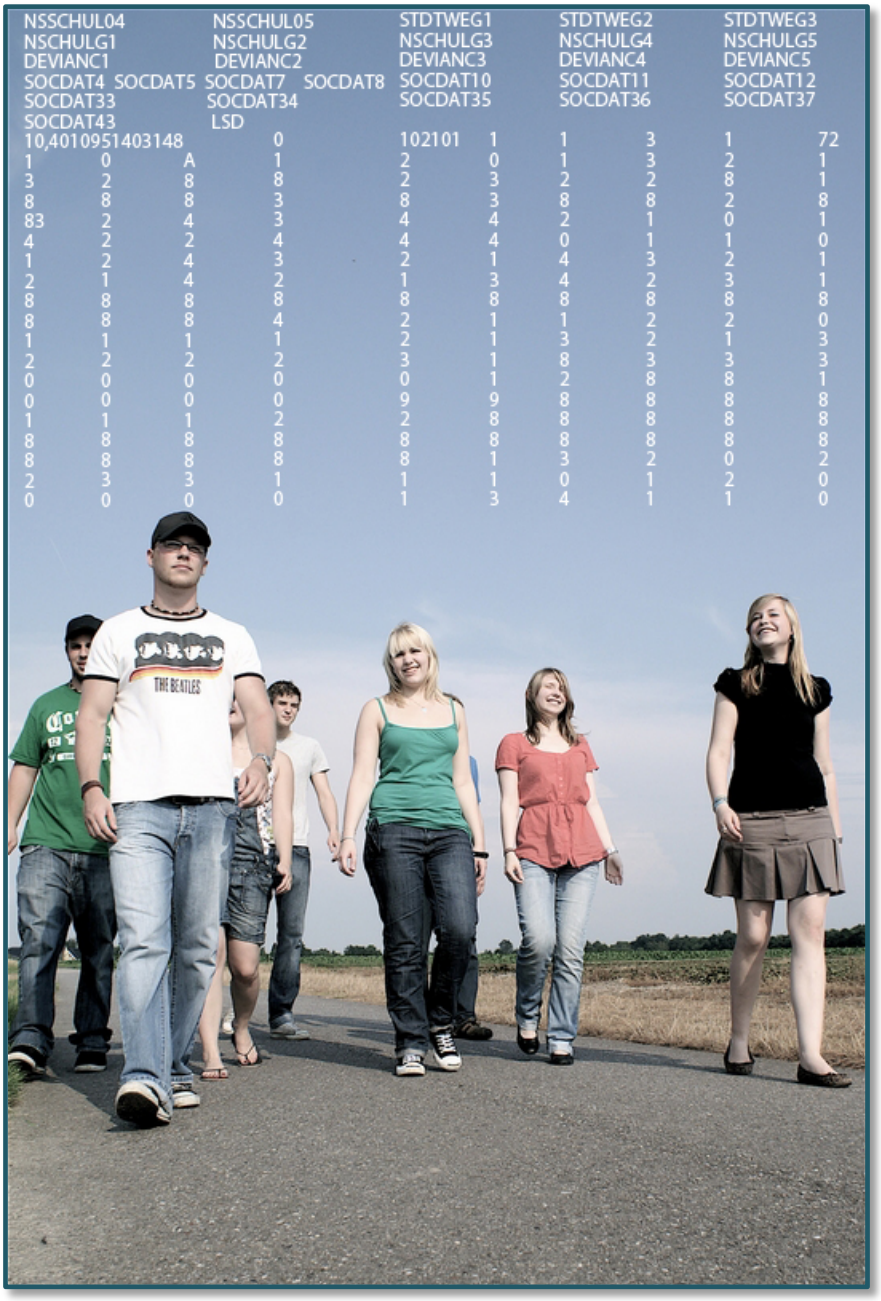
Datenmanagement und Data Sharing in der Psychologie

Einführung und Manual

Herausgegeben vom Forschungsdatenzentrum PsychData des Leibniz-Zentrums für Psychologische Information und Dokumentation (ZPID), Trier, 15. Mai 2013



Psychologie Information
ZPID - Leibniz Institut

Leibniz-Zentrum für Psychologische Information und Dokumentation (ZPID) · Universitätsring 15 · 54296 Trier
Fon: +(49)-(0)651-201-2877 · Fax: +(49)-(0)651-201-2071 · E-Mail: info@zpid.de · Internet: www.zpid.de

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1. Einleitung

Forschung beruht auf Daten, und deren Nutzung hat sich mit der „Digitalisierung der Welt“ maßgeblich gewandelt. Die zunehmende Geschwindigkeit und immer stärkere globale Vernetzung in der Informations- und Kommunikationstechnologie ermöglicht einerseits die Erhebung und Verarbeitung immer größerer Datenmengen, andererseits umfassende Kooperationen zwischen Forschergruppen auf der ganzen Welt.

Dies eröffnet die Gewinnung neuer Erkenntnisse durch Forschungsvorhaben und -methoden, die zuvor aufgrund ihrer Komplexität schlichtweg nicht denkbar gewesen wären. In diesem Zusammenhang ist sogar von der Etablierung eines neuen, „datengetriebenen“ Forschungsparadigmas die Rede, welches primär auf der Exploration automatisiert erhobener Datenmassen basiert (Hey, Transley & Tolle, 2009). Allerdings stehen Forschende angesichts der erhöhten Komplexität aber auch vor Problemen: Schlimmstenfalls ertrinkt man in der „Datenflut“ (Hey & Trefethen, 2003); zumindest aber ist ein erhöhter Aufwand zur Koordination der Beteiligten und zur Kontrolle der Daten nötig. Um das vorhandene Potential außerdem wirklich ausschöpfen zu können, müssen erhobene Daten zudem für interessierte Forscher überhaupt erst einmal zugänglich sein.

Das Bemühen, die komplexen Anforderungen im Umgang mit Forschungsdaten zu bewältigen, findet in dem zunehmend eigenständiger werdenden Tätigkeitsbereich des *Datenmanagements* [1], also der Gesamtheit der „Maßnahmen (…), die sicherstellen, dass (…) Forschungsdaten nutzbar sind“ (Ludwig & Enke, 2013, S. 13), seinen Ausdruck. Ebenfalls an Bedeutung gewinnt die Praxis des *Data Sharing:* das Bestreben, Daten interessierten Forschergruppen oder ganz der Öffentlichkeit zur Verfügung zu stellen. Für erfolgreiches Datenmanagement und Data Sharing in der Forschung ist einerseits eine angemessene technische und organisatorische Infrastruktur nötig, etwa die Berücksichtigung von damit verbundenem Aufwand in Förderplänen, die Verfügbarkeit adäquater Hard- und Software an den Forschungsstätten (inklusive spezialisierter Hilfsmittel, etwa zur Datendokumentation), die Einrichtung von Datenarchiven sowie das Vorhandensein allgemein anerkannter Standards und Richtlinien. Andererseits müssen bei den Forschenden selbst ein Bewusstsein für die Bedeutung von Datenmanagement und Data Sharing und grundlegende Kenntnisse und Fertigkeiten vorhanden sein. Die Verankerung in den Lehrplänen der Studiengänge sowie fortbildende Angebote, Einführungen und Manuale sind Mittel dies zu erreichen.

In diesem Sinne soll hier eine speziell an Forschende in der Psychologie gerichtete Einführung in den Themenbereich gegeben werden. Dabei soll insbesondere auch auf die Angebote von

[1] Da sich diese Einführung allein auf Forschungsdaten bezieht, sind hier mit „Daten“ immer „Forschungsdaten“ gemeint, und dementsprechend „Forschungsdatenmanagement“ mit „Datenmanagement“, „Research data sharing“ mit „Data sharing“, usw.

PsychData, dem Forschungsdatenarchiv des Leibniz-Zentrums für Psychologische Information und Dokumentation (ZPID) eingegangen werden. PsychData stellt ein speziell auf die psychologische Forschung ausgerichtetes Forschungsdatenzentrum dar, dass sich die Archivierung und Bereitstellung von Daten zur Aufgabe gemacht hat.

Der Fokus der Einführung liegt, gemäß der derzeitigen Ausrichtung der psychologischen Mainstream-Forschung, auf dem quantitativen Forschungsparadigma. Ein Großteil der Konzepte und Vorgehensweisen lässt sich aber (gegebenenfalls leicht angepasst) auch auf qualitative Ansätze übertragen. In Kapitel 2 werden Grundbegrifflichkeiten und Bedeutsamkeit von Datenmanagement und Data Sharing über das bereits Gesagte hinaus vertiefend dargestellt. Dabei wird auch auf die besonderen Bedingungen in der Psychologie eingegangen. Kapitel 3 soll einen Überblick über wichtige Teilaspekte und konkrete Tipps zu Datenmanagement und Data Sharing in den verschiedenen Abschnitten des Forschungsprozesses geben. In Kapitel 4 werden schließlich die Angebote von PsychData detailliert dargestellt. Leser, die primär an der Nutzung dieser Dienste interessiert sind, oder aber mit Datenmanagement und Data Sharing bereits vertraut sind und sich einen Überblick über die Angebote verschaffen wollen, können direkt zu diesem Kapitel übergehen.

1.1 Weiterführende Ressourcen

- Das von Büttner et al. (2011) herausgegebene *Handbuch Forschungsdatenmanagement* gibt eine (nicht fächerspezifische) ausführliche Einführung in die vielfältigen Aspekte des Datenmanagements (und auch des Data Sharing).
- Auf dem *Informationsportal Forschungsdaten* [2] sind unter anderem grundlegende Erläuterungen zum Forschungsdatenmanagement, Einführungen und Manuale, Richtlinien und Erklärungen von Forschungsförderern und Wissenschaftsorganisationen sowie fachspezifische Informationsangebote zusammengestellt.
- Im *Archive and Data Management Training Center* [3] des Leibniz-Instituts für Sozialwissenschaften (GESIS) finden sich an Sozialwissenschaftler gerichtete Informationen zu wichtigen Aspekten des Datenmanagements und der Langzeitarchivierung.
- Die Website des *Digital Curation Centre (DCC)* [4] enthält zahlreiche auch für Forscher außerhalb von Großbritannien nützliche Ressourcen, z.B. Einführungen oder eine Zusammenstellung von Datenmanagement-Tools.

[2] http://www.forschungsdaten.org/ [26.04.2013]
[3] http://www.gesis.org/archive-and-data-management-training-and-information-centre/training-center-home/ [26.04.2013]
[4] http://www.dcc.ac.uk/ [26.04.2013]

2. Datenmanagement und Data Sharing: Konzeptuelle Grundlagen

Dieses Kapitel dient der Klärung einer Reihe von Fragen, die sich stellen, wenn man sich erstmalig eingehend mit Datenmanagement und Data Sharing beschäftigt. Zunächst werden in Kapitel 2.1 grundlegende Begriffe erklärt und gegeneinander abgegrenzt. Kapitel 2.2 ordnet Datenmanagement/Data Sharing-Aktivitäten dann in den übergreifenden Kontext eines wissenschaftlichen Forschungsvorhabens ein und in Kapitel 2.3 werden Gründe genannt, warum es sich lohnt, zusätzlich Zeit und Aufwand in diese Aktivitäten zu investieren. Anschließend wird in Kapitel 2.4. der Frage nach dem Zusammenhang zwischen der Höhe dieses (zusätzlichen) Aufwands und den Merkmalen eines Forschungsvorhabens nachgegangen. Der letzte Abschnitt des Kapitels geht auf besondere Umstände ein, die beim Datenmanagement und Data Sharing in der Psychologie zu berücksichtigen sind.

2.1 Grundbegriffe

2.1.1. Daten, Dokumentation und Metadaten

Gemäß dem Duden (1976)sind *Daten* „durch Beobachtungen, Messungen, statistische Erhebungen u. a. gewonnene (Zahlen-)Werte“, also die Symbole, die in einem Erhebungsprozess zur Repräsentation des interessierenden Phänomens erzeugt werden. Für sich genommen sind diese Symbole wenig aussagekräftig. Um Daten sinnvoll interpretieren zu können, benötigt man zusätzliche Informationen etwa zu Erhebungskontext, Erhebungsgerät, erhebender Person, Zeitpunkt der Erhebung usw.

Die *Dokumentation*, also die mehr oder weniger dauerhafte Hinterlegung solcher zusätzlicher Informationen ist ein alltäglicher Bestandteil der Arbeit eines jeden Forschenden. Idealerweise sollte die Dokumentation so vollständig und verständlich sein, dass die Daten auch für nicht direkt an der Erhebung beteiligte Personen interpretierbar sind; allein schon deshalb, weil nicht alle Mitglieder einer Forschergruppe mit allen Aspekten der Erhebung vertraut sind. Auch kann es innerhalb einer Forschergruppe zu Personalwechseln und damit zur „Abwanderung“ von Hintergrundwissen kommen. Aber auch Personen außerhalb der eigenen Gruppe sollte eine Interpretation möglich sein (ohne dass am Telefon alles stundenlang geklärt werden muss), wenn z.B. ein externer Auswerter oder Projektpartner die Daten begutachten soll oder die Daten im Sinne des Data Sharing-Gedankens anderen Forschern zur Verfügung gestellt werden sollen.

Um die Interpretierbarkeit (nicht nur durch Menschen, sondern auch an der Datenverarbeitung beteiligte Maschinen) sicherzustellen, kann die Art und Weise der Dokumentation stärker formalisiert werden. Dies ist der Hintergrundgedanke des Konzepts der *Metadaten*, also „Daten oder Informationen, die in strukturierter Form (…) Forschungsdaten dokumentieren“ (Jensen, Katsanidou & Zenk-Möltgen, 2011, S. 83). Metadaten lassen sich in etwa analog zu Variablenwerten eines Versuchsteilnehmers verstehen, nur dass sie sich auf einen Forschungsdatensatz oder eine Variable im Datensatz

als „Beobachtungseinheit“ beziehen. Beispielsweise könnte ein (fiktives) Schema für Metadaten zu einem Datensatz die Angaben „Projektfördernummer“, „Erhebungsleiter“, „Erhebungsbeginn“ und „Erhebungsende“ enthalten, wobei die „Variablenwerte“ bei den ersten beiden Angaben eine unspezifische Zeichenkette, bei den letzten beiden eine Datumsangabe sein müssten. Mittlerweile wurden außerdem Standards für die Vergabe von Metadaten zu Forschungsdaten entwickelt. Mehr dazu findet sich in Kapitel 3.2.2.

Daten sind durch Beobachtungen, Messungen, statistische Erhebungen u.a. gewonnene (Zahlen-)Werte (Duden, 1976).

Mit Dokumentation bezeichnet man zusätzliche Informationen, die notwendig sind, um Daten langfristig sinnvoll interpretieren zu können.

Metadaten sind Daten oder Informationen, die in strukturierter Form (…) Forschungsdaten dokumentieren (Jensen et al., 2011).

2.1.2. Rohdaten, Primärdaten, abgeleitete Daten, Sekundärdaten

Im Verlauf der Datenerhebung und -analyse machen die Daten verschiedene Transformationsschritte durch. Im Kontext des Datenmanagements und Data Sharing werden für die dabei entstehenden verschiedenen Datenformen verschiedene Begriffe gebraucht, wobei die Nutzung allerdings nicht immer einheitlich ist. Mit *Rohdaten* sind in der Regel die Daten in Form der Erstaufzeichnung gemeint, etwa eine Videoaufzeichnung, eine Tonbandaufzeichnung, die angekreuzten Kästchen auf einem Fragebogen oder der Fallbericht des untersuchenden Arztes.

Diese Daten werden heutzutage in aller Regel anschließend in eine digitale, stärker strukturierte Form übertragen, in der Psychologie meist eine Datenmatrix, bei der Beobachtungseinheiten durch Zeilen und Variablen durch Spalten repräsentiert sind. Die digitale, strukturierte Form der Daten ist es, die in aller Regel mit dem Begriff „Daten“ bzw. „Forschungsdaten“ gemeint ist – so auch in dem vorliegenden Text. Als Abgrenzung zwischen dieser Datenform und Rohdaten wurde bei PsychData lange Zeit der Begriff *Primärdaten* verwendet. Je nach Art der Rohdaten kann die Transformation zu Primärdaten verschieden aufwändig sein: In einem Reaktionszeitexperiment können die Daten bereits direkt in Matrixform in einer Computerdatei erfasst werden; die Kodierung von Videoaufzeichnungen erfordert dagegen einen großen zusätzlichen Arbeitsaufwand. Eng mit diesem Transformationsaufwand zusammen hängt außerdem die Tatsache, dass es bereits bei der Umwandlung von Roh- in Primärdaten zu einem mehr oder weniger großen Informationsverlust kommt, weshalb die Rohdaten soweit wie möglich ebenfalls aufbewahrt werden sollten.

Im weiteren Verlauf der statistischen Analyse basierend auf den Primärdaten werden meist weitere Variablen zum Analysedatensatz hinzugefügt, wie z.B. rekodierte Fragebogenitems, Summenscores, Indizes, oder dichotomisierte Variablen. Diese *abgeleiteten Daten* sind für die eigene Analyse von

großer Bedeutung, für eine eventuelle Nachnutzung dagegen unter Umständen weniger, da die Berechnungsvorschrift zusammen mit den Primärdaten zur Reproduktion ausreichend sein sollte.

Zum Begriff der Primärdaten ist anzumerken, dass er in der Literatur in zum Teil stark unterschiedlicher Bedeutung verwendet wird, weshalb man im Zweifelsfall genau prüfen sollte, was damit eigentlich gemeint ist. Neben der Bedeutung im Sinne von (auch für die Nachnutzung) aufbereiteten Daten meinen manche damit den durch ein Messgerät produzierten Datenstrom, andere alle Daten, die als Grundlage einer wissenschaftlichen Publikation dienen (vgl. Klump, 2011, Fußnote 1). Schließlich können mit „Primärdaten" auch von einem Forscher selbst erhobene Daten gemeint sein, in Abgrenzung von *Sekundärdaten*, also bereits vorhandenen Daten, die durch den Forscher genutzt werden (etwa Zensusdaten oder nachgenutzte Daten anderer Forschender). Deswegen wird bei PsychData seit 2012 nicht mehr der Begriff „Primärdaten" verwendet, sondern – etwas allgemeiner – von „Forschungsdaten" gesprochen. Mit „Forschungsdaten" sind bei PsychData sowohl Primärdaten (im oben beschriebenen Sinn) als auch abgeleitete Daten gemeint[5].

Im Verlauf der Datenerhebung und -analyse machen Daten verschiedene Transformationsschritte durch. Die resultierenden Daten können als Rohdaten, Primärdaten oder abgeleitete Daten bezeichnet werden.

Ursprungsaufzeichnungen stellen sogenannte *Rohdaten* dar.
Mit *Primärdaten* bezeichnet man die erste Übertragung von Rohdaten in eine digitale, stärker strukturierte Form.
***Abgeleitete Variablen* werden durch Analyse- und Berechnungsschritte gebildet, die an den Primärdaten durchgeführt werden.**

2.1.3. Datenmanagement, Data Preservation, Data Sharing, Data Curation

Anknüpfend an die in der Einleitung gegebene Begriffsbestimmung lässt sich Datenmanagement verstehen als die Gesamtheit der gezielten Maßnahmen, um die bestmögliche Nutzung erhobener Daten sicherzustellen. Die Maßnahmen umfassen unter anderem die Sicherstellung der Qualität der Erhebungsinstrumente und –prozesse, die Prüfung erhobener Daten auf Integrität und Konsistenz sowie die Erstellung von hochwertiger Dokumentation und Metadaten. Einzelne Datenmanagement-Maßnahmen werden in Kapitel 3 ausführlich dargestellt.

Eng mit Datenmanagement verbunden sind Maßnahmen zur langfristigen Erhaltung der Daten (*Data Preservation*). Darunter fallen beispielsweise die Datenspeicherung in voraussichtlich langfristig unterstützten und gepflegten Dateiformaten, die regelmäßige Anfertigung und sichere Aufbewahrung von Sicherheitskopien, die Prüfung der Datenintegrität auf physikalischer Ebene (*bitstream preservation*) und Systeme zur Kontrolle von Datenzugang und -manipulation. Diese Aufgaben werden oft von spezialisierten *Datenarchiven*, wie z.B. PsychData, wahrgenommen. Bei diesem *Data*

[5] Bei älteren Studien, die bei PsychData archiviert wurden, findet sich noch standardmäßig die Unterscheidung in „Primärdaten" und „abgeleitete Daten".

Archiving übergibt die Forschergruppe ihren Datensatz an das Archiv, welches ihn in seinen Datenbestand einpflegt (*Ingest*).

Datenmanagement und Data Preservation garantieren die langfristige Nutzbarkeit der Daten und sind damit Grundlage des bereits erwähnten Data Sharing – der Zugänglichmachung des Datensatzes für die Öffentlichkeit oder interessierte Forschergruppen. Wie weitgehend die Zugangsmöglichkeiten durch Dritte sein sollen (z.B.: Wer soll Zugriff haben? Ab wann? Auf welche Teile des Datensatzes?), lässt sich in einer Vereinbarung zwischen Archiv und Datengeber regeln. Data Sharing erfordert über eine Qualitätssicherung der Daten hinaus auch die korrekte Handhabung rechtlicher Bestimmungen, insbesondere im Umgang mit personenbezogenen Daten. Datenschutzbestimmungen sind hier von Anfang an zu berücksichtigen. Je nach verwendeten Erhebungsprozeduren oder Bestimmungszwecken der Daten sind auch urheberrechtliche Regelungen oder verwandte Schutzrechte zu beachten.

Über die Aufbewahrung, Erhaltung und Zugänglichmachung hinaus stehen Datenarchive, die sich oft auch als kulturelle Gedächtnisorganisation verstehen, langfristig gesehen angesichts einer immer weiter wachsenden Datensammlung außerdem vor der Herausforderung, eine Selektion der erhaltenswerten Daten zu treffen. Dies erfordert zum einen eine Auswahl, welche Datensätze überhaupt für eine Archivierung in Frage kommen, zum anderen eine periodische Prüfung des Datenbestandes auf fortbestehende Erhaltungswürdigkeit. Die gesamten Aktivitäten eines Datenarchivs, von der Auswahl über die Aufnahme, Erhaltung, Zugänglichmachung bis hin zur eventuellen Entfernung der Daten, werden auch als „Datenkuratierung“ (*Data Curation*) bezeichnet (Pennock, 2007). „Kuratierung“ ist hier durchaus im aus dem Kulturbetrieb bekannten Sinne zu verstehen, da Forschungsdaten auch als Kulturgut betrachtet werden (vgl. Kommission Zukunft der Informationsinfrastruktur, 2011, Kap. III.1).

Die hier vorgestellten Begrifflichkeiten überschneiden sich zum Teil beträchtlich. Insbesondere „Datenmanagement“ kann sich auf alle hier vorgestellten Aspekte des Umgangs mit Daten beziehen, also auf die Handhabung durch Forscher oder durch Archive, auf Data Sharing, Data Curation, usw. Andererseits impliziert „Data Curation“ immer auch ein aktives Management der Daten. Da sich das vorliegende Manual primär an Forschende richtet, sind hier, sofern nicht anders angegeben, mit „Datenmanagement“ die Aktivitäten „auf Forscherseite“ gemeint. Kapitel 3 stellt diese Aktivitäten, inklusive der Vorbereitungen für anschließendes Data Sharing, detailliert dar.

Trotz der engen Beziehung zwischen Datenmanagement, Data Preservation, und Data Sharing müssen diese nicht zwangsläufig miteinander einhergehen. In der Privatwirtschaft wird oft ein beträchtlicher Aufwand für Datenmanagement betrieben (zu Verwaltungszwecken, aber z.B. auch für Data Mining und Konsumentenforschung). Doch gerade weil diese Daten aufwändig bearbeitet wurden, achten die Unternehmen darauf, sie anderen *nicht* zugänglich zu machen. In der klinischen Forschung bestehen aufgrund der gesundheitspolitischen Relevanz strengste Vorschriften für die

Durchführung und Dokumentation von Medikamentenstudien und es hat sich die eigenständige Profession des „Clinical Data Manager“ herausgebildet (vgl. McFadden, 2007). Unabhängig davon ist die Zurückhaltung von Forschungsergebnissen und -daten durch pharmazeutische Unternehmen eine mittlerweile wohlbekannte Problematik (Goldacre, 2012; Rising, Bacchetti & Bero, 2008). Auch viele akademische Forscher in der klinischen Forschung halten ihre Daten zurück, oft aufgrund der Befürchtung, von Nachnutzern der Daten übervorteilt zu werden (Vickers, 2006). Eine ähnliche Problematik existiert auch in der Psychologie (s. Kapitel 2.5).

Unter Datenmanagement versteht man alle Maßnahmen, die die bestmögliche Nutzung erhobener Daten sicherstellen sollen.

Mit *Data Preservation* sind Vorgehensweisen zur langfristigen Erhaltung von Forschungsdaten gemeint.
***Data Sharing* bezeichnet die Zugänglichmachung von Daten für die Öffentlichkeit und/oder die wissenschaftliche Forschungsgemeinschaft.**
***Data Curation* impliziert aktives Datenmanagement und beinhaltet alle Aktivitäten eines Datenarchivs von der Auswahl der Daten bis zu deren Zugänglichmachung.**

2.2 Forschungsdaten im Kontext wissenschaftlicher Forschung

Forschungsaktivität lässt sich anhand eines Zyklenmodells beschreiben: Basierend auf früheren Forschungsergebnissen wird eine (ungelöste) Fragestellung formuliert, ein Forschungsprojekt initiiert (oft beginnend mit einem Fördermittelantrag) und geplant, Daten erhoben und ausgewertet, und basierend auf den Ergebnissen eine Publikation erstellt, die wiederum Ausgangspunkt für weitere Fragestellungen sein kann (s. z.B. JISC, n.d.).

Solche Modelle sind traditionellerweise auf die wissenschaftliche Publikation als zentrales Ergebnis der Forschungsarbeit ausgerichtet. Mit der zunehmenden Bedeutung von Forschungsdatensätzen als Objekt von eigenständiger Bedeutung, unabhängig von darauf basierenden Publikationen, wurden derartige Lebenszyklusmodelle auch für Forschungsdaten postuliert (DCC, n.d.; ICPSR, 2012, S. 8; Ludwig & Enke, 2013, S. 15). Der Datenlebenszyklus nimmt seinen Anfang quasi aus einem Forschungslebenszyklus heraus mit der Erstellung eines Datenmanagementplans (vgl. Kapitel 3.1) und der Erhebung, Bearbeitung und Analyse des Datensatzes. Werden die Daten im Verlauf aber sorgfältig gepflegt, anschließend archiviert und anderen zugänglich gemacht, entwickeln sie ein „Eigenleben“. So können sie ihrerseits andere Forschungsprojekte anregen, beziehungsweise Vorhaben beschleunigen oder überhaupt erst ermöglichen, da die aufwändige Datenerhebungsphase ganz oder teilweise entfällt (Abbildung 1). Sind erst einmal genügend nachnutzbare Datensätze vorhanden, kann ein Forschungsprojekt auch allein auf der Integration verschiedener nachgenutzter Datensätze entwickelt werden. Damit erfolgt letztendlich eine Verschiebung von einer publikationszentrierten hin zu einer datenzentrierten Perspektive.

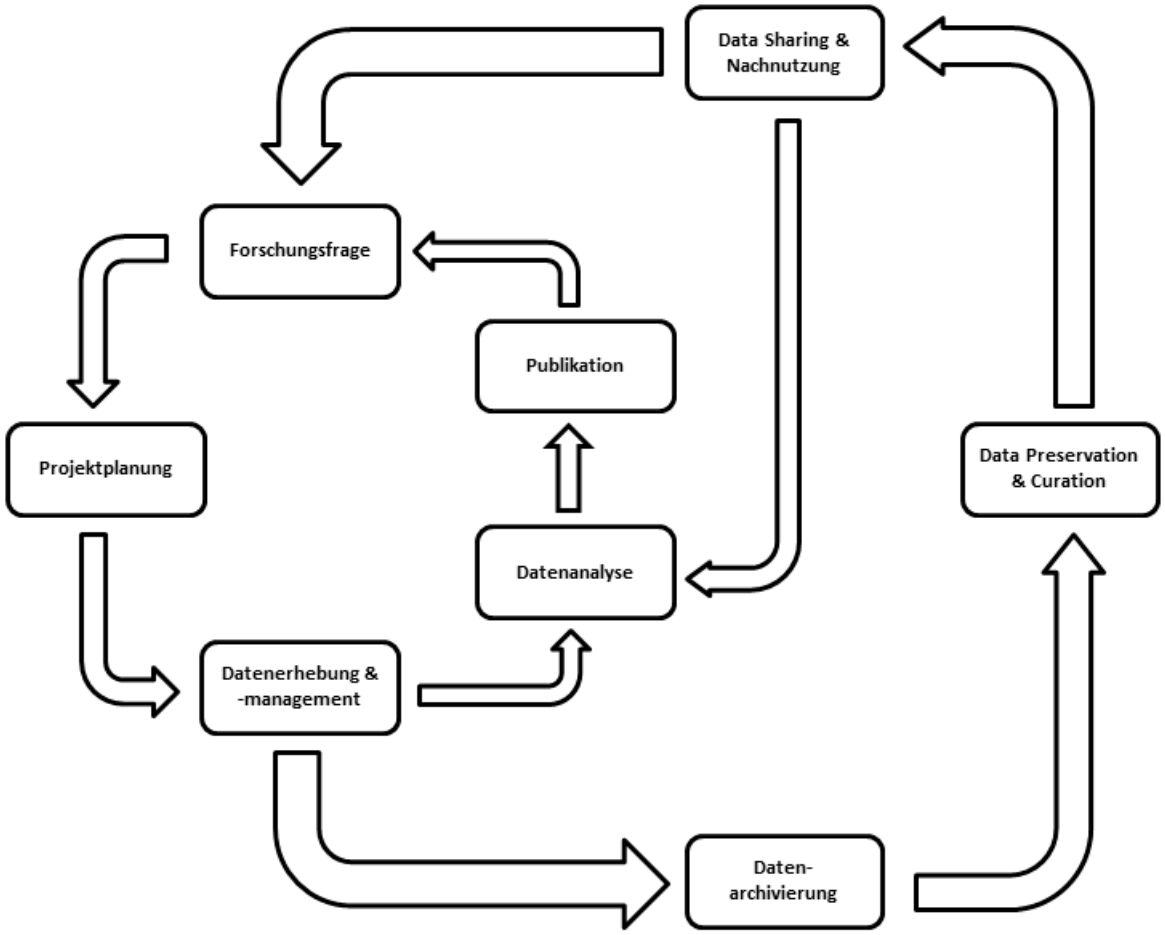


Abbildung 1. Zusammenhang der Lebenszyklen von Forschungsprojekt und Datensatz

2.3 Gründe für Datenmanagement und Data Sharing

Datenmanagement und Data Sharing im eigenen Forschungsprojekt zu berücksichtigen, ist zwangsläufig mit Mehraufwand verbunden. Es gibt jedoch eine ganze Reihe von Gründen, die für diesen zusätzlichen Aufwand sprechen (siehe z.B. Weichselgartner, Günther & Dehnhard, 2011a):

2.3.1. Erkenntnisgewinn

Vorhandene Forschungsdaten lassen sich unter der Perspektive neuer Fragestellungen, unter Verwendung neuer Analysemethoden oder im Lichte neuer Erkenntnisse reanalysieren und so zu neuen Befunden führen. Dies ist von ganz besonderer Bedeutung im Fall von Daten, die unter einmaligen, nicht replizierbaren Umständen erhoben wurden, beispielsweise nach Naturkatastrophen oder einschneidenden gesellschaftlichen Ereignissen.

Neue, anhand von Einzelstudien nicht zu erlangende Erkenntnisse lassen sich auch durch den Vergleich von Datensätzen, die zu verschiedenen Zeitpunkten oder unter unterschiedlichen Umständen erhoben wurden, gewinnen. So ließen sich beispielsweise in der Psychologie historische Verschiebungen in Testnormen (man denke an den Flynn-Effekt) oder die Übertragbarkeit (scheinbarer) psychologischer Gesetzmäßigkeiten in andere Kontexte untersuchen. Eine umfassende Verfügbarkeit von Datensätzen würde die Durchführung von Metaanalysen ermöglichen, die aufgrund der höheren Auflösung der Daten präziser wären als die bislang üblichen publikationsbasierten Metaanalysen.

Schließlich kann ein Erkenntnisgewinn mittelbar durch die Anregung von Kooperationen und Gedankenaustausch zwischen datengebenden und –nehmenden Forschergruppen entstehen.

2.3.2. Nutzung in der Lehre

Aus realer Forschungsarbeit stammende Datensätze stellen ein wertvolles Hilfsmittel bei der Vermittlung wissenschaftlicher Fertigkeiten für Hochschullehre sowie Fort- und Weiterbildung dar. Gerade zur Vermittlung komplexer statistischer Verfahren wie Mehrebenenanalysen sind die oft umfangreichen nachnutzbaren Datensätze hilfreich.

2.3.3. Bewältigung von Komplexität in umfangreichen Forschungsvorhaben

Das Gelingen von Forschungsvorhaben, bei denen große, komplex strukturierte Datenmengen anfallen (z.B. Längsschnitterhebungen), setzt ein strukturiertes, planvolles Vorgehen voraus. Datenmanagement-Aktivitäten sind ein wesentlicher Bestandteil davon. Die Erstellung eines Datenmanagementplans (s. Kapitel 3.1) erfordert es, die unterschiedlichen Anforderungen in den verschiedenen Phasen des Forschungsvorhabens von Anfang an zu bedenken und hilft bei der Aufdeckung von bislang nicht bedachten Problemen. Auch die Auseinandersetzung mit Data Sharing kann eine Rolle spielen, da man sich Gedanken über eine langfristig nutzbare und verständliche Form von Daten und Dokumentation machen muss, was auch der eigenen Forschergruppe (z.B. nach Personalwechseln) zugute kommt.

2.3.4. Einhaltung guter wissenschaftlicher Praxis und ethischer Prinzipien

Der Bedeutungsgewinn von Datenmanagement und Data Sharing geht auch zurück auf das Bemühen um die Sicherstellung „guter wissenschaftlicher Praxis“, die am Ideal der vorurteilsfreien und für Kritik offenen Wissenschaft orientiert ist, welches z.B. klassisch von Merton (1973) formuliert wurde. Um die Einhaltung guter wissenschaftlicher Praxis hat sich, angestoßen durch die Aufdeckung von Datenfälschung und anderen Fällen wissenschaftlichen Fehlverhaltens, eine nach wie vor anhaltende Debatte entsponnen. So geht die Verabschiedung von den „Vorschlägen zur Sicherung guter wissenschaftlicher Praxis“ durch die Deutsche For-

schungsgemeinschaft (DFG, 1998) maßgeblich auf einen prominenten Fall von Datenfälschung durch deutsche Krebsforscher (Bartholomäus & Schnabel, 1997) zurück.

Auch in der Geschichte der Psychologie wurden Fälle von Fehlverhalten bekannt, erst kürzlich beispielsweise die umfangreiche Datenfälschung durch einen niederländischen Sozialpsychologen (Enserink, 2012). In letzterem Fall bemängelte die Untersuchungskommission nicht nur gezielte Fälschung, sondern auch ein Arbeitsumfeld, in dem das Bemühen um methodische Sorgfalt vernachlässigt wurde. Auch grobe Fahrlässigkeit im Umgang mit Daten kann also wissenschaftliches Fehlverhalten darstellen (vgl. Freedland & Carney, 1992). Psychologische Fachgesellschaften wie die American Psychological Association (APA) und die Deutsche Gesellschaft für Psychologie (DGPs) haben seit einiger Zeit eigene Verhaltenskodizes für Forschende in der Psychologie verabschiedet, gemäß derer Forschungsdaten interessierten Kollegen zur Verfügung zu stellen sind (APA, 2010; DGPs, 2004).

Die Forderung nach Zugang zu auf transparente Weise erhobenen, verständlichen Daten erschwert wissenschaftliches Fehlverhalten und fördert damit letztlich das Vertrauen in die Qualität der Forschung. Auch in der großen Mehrheit der Fälle, in denen kein Fehlverhalten vorliegt, können durch wechselseitige Kritik Schwächen in Studiendesign und -methoden entdeckt und behoben werden. Außerdem wird es möglich zu prüfen, ob sich wichtige Befunde replizieren lassen (King, 1995).

Qualitativ hochwertige und (offen) zugängliche Daten zu erzeugen, erhält angesichts des für die Erhebung von Forschungsdaten nicht selten beträchtlichen finanziellen Aufwands eine zusätzliche ethische Komponente. Weil die meisten Forschungsgelder aus öffentlicher Hand stammen, sind ein verantwortungsvoller Umgang mit den Daten und eine möglichst umfangreiche Nutzung im Sinne der Gesellschaft wünschenswert. Eine besondere ethische Verpflichtung zur sorgfältigen Nutzung der Daten besteht, wenn diese Daten an Menschen oder Tieren erhoben wurden, da eine Erhebung mit Risiken für die Versuchsteilnehmer verbunden sein kann (und bei Versuchstieren in der Regel mit dem Tod endet).

2.3.5. Richtlinien von Förderorganisationen

Aufgrund der zuvor diskutierten Argumente für Datenmanagement und Data Sharing enthalten die Förderrichtlinien vieler wichtiger Drittmittelgeber wie DFG, National Institutes of Health (NIH), National Science Foundation (NSF) oder Wellcome Trust mittlerweile Vorgaben zum Umgang mit Forschungsdaten (Pampel & Bertelmann, 2011). Beispielsweise fordert die DFG seit 2010 die Erstellung eines Datenmanagementplans als Teil von Projektanträgen (DFG, 2012). Die NIH erwarten, dass Anträge über Fördersummen ab 500,000 $ einen Plan zum Data Sharing oder eine Begründung dafür, weshalb Data Sharing nicht möglich ist, enthalten (NIH, 2003). Damit sind Datenmanagement und Data Sharing nicht mehr nur aus rein „idealisti-

schen“, sondern auch aus materiellen Gründen für Forschende von Bedeutung. In den Kapiteln 3.1 und 3.2.1 wird näher auf die *Data Policies* von Förderorganisationen sowie Richtlinien und Vereinbarungen von Fachgesellschaften und Forschungspolitik eingegangen.

2.3.6. Anerkennung durch die Forschungsgemeinschaft

Neben diesem finanziellen Anreiz spielt als „extrinsischer“ Motivator auch die Anerkennung des Verdienstes, hochwertige Daten als Grundbaustein der Forschung beigetragen zu haben, eine Rolle. Das zentrale Mittel zum Ausdruck einer solchen Anerkennung ist die Zitation. Daher wurden Initiativen gestartet mit dem Ziel, Datensätze als eigenständige Forschungsleistung leichter auffindbar und „zitierbar“ zu machen. Beispielsweise bemüht sich die *Research Data Alliance* [6] um die Entwicklung und Harmonisierung von Standards, damit u.a. die Suche nach Forschungsdaten erleichtert wird. Es existieren außerdem bereits mehrere Online-Verzeichnisse von Datenrepositorien zum Auffinden relevanter Archive bzw. Datensätze (s. Kapitel 3.2.1).

Als Standard für die Zitation elektronischer Datensätze scheint sich der *Digital Object Identifier* (DOI) zu etablieren, der auch zur Zitation elektronischer Zeitschriftenartikel sehr populär ist. Der DOI ist ein sogenannter *Persistent Identifier*, ein abstrakter „Name“ für ein bestimmtes Objekt (hier: ein Datensatz), der durch eine Registrierungsagentur gepflegt wird, so dass ein Datensatz langfristig, z.B. auch nach einem „Ortswechsel“ (etwa ein Umzug auf einen anderen Webhost), auffindbar und die Zitation damit nachvollziehbar bleibt. Weitere Informationen zu Persistent Identifiers und DOIs finden sich in Kapitel 3.2.3.

Der zusätzliche Aufwand für Datenmanagement und Data Sharing lässt sich rechtfertigen durch folgende Gründe:

- **Erkenntnisgewinn,**
- **Einsatz von Daten in der wissenschaftlichen Lehre,**
- **Förderung eines strukturierten Vorgehens bei komplexen umfangreichen Forschungsvorhaben,**
- **Einhaltung guter wissenschaftlicher Praxis und ethischer Prinzipien,**
- **Einhaltung der Richtlinien von Förderorganisationen,**
- **Verstärkte Anerkennung durch die Forschungsgemeinschaft.**

2.4 Wie viel Aufwand sollte in Datenmanagement und Data Sharing investiert werden?

Unabhängig von den zahlreichen Argumenten, die für Datenmanagement und Data Sharing sprechen, bleibt zunächst unklar, wie groß der Aufwand sein sollte, den man als Forschergruppe darin investieren sollte. Genau genommen geht es nämlich nicht so sehr um die Frage,

[6] http://rd-alliance.org/ [01.05.13]

ob man Datenmanagement und Data Sharing betreiben sollte, sondern *in welchem Umfang*. Ein Datenmanagementplan etwa kann unterschiedlich detailliert ausgearbeitet sein und unterschiedlich aufwändige Prozeduren vorschreiben. Data Sharing wiederum kann sich auf die Aufbewahrung innerhalb des eigenen Instituts zur Weitergabe auf Nachfrage beschränken, oder aber die Übergabe an ein Archiv zur Nachnutzung umfassen. Hier ist eine Abwägung zwischen dem nötigen Aufwand und dem erwarteten Nutzen (auf die Forschergemeinde und die Gesellschaft als Ganzes bezogen, nicht bloß auf die eigene Gruppe) angebracht.

Einen ersten, eindeutigen Maßstab liefern gesetzliche Anforderungen, beispielsweise zur Qualitätssicherung der Forschung und Dokumentation in der klinischen Forschung. Jedoch sind solche harten gesetzlichen Mindeststandards an die Daten- und Dokumentationsqualität nur in bestimmten Forschungsbereichen relevant, und auch beim Großteil der psychologischen Forschung dürften sie (abgesehen von Bestimmungen zum Datenschutz) eher selten eine Rolle spielen. Richtlinien von Förderorganisationen spielen wie erwähnt eine zunehmende Rolle und können weitere eindeutige Vorgaben darstellen. Auch die Forschungsinstitutionen selbst können über eigene Qualitätsrichtlinien verfügen. In der Regel sind diese Vorgaben aber bewusst nicht sehr spezifisch gehalten, um die Flexibilität und Entscheidungsfreiheit in der Forschung nicht übermäßig einzuschränken.

Als ungefährer Maßstab zur Bestimmung des angemessenen Aufwands kann dann das zu erwartende „Nachnutzungspotential“ dienen: Gibt es andere Forschungsbereiche oder Forschergruppen, für deren Fragestellungen die Daten von Interesse sein könnten? Diese Frage ist natürlich im Voraus schwer zu beantworten, vor allem auf lange Sicht. Auch Datenarchive beschäftigen sich mit diesem Aspekt, wenn sie eine Auswahl treffen müssen, ob ein Datensatz erhaltungswürdig ist. Von daher bietet es sich an, bei einem potentiell geeigneten Datenarchiv anzufragen, ob die zu erhebenden Daten archivierungswürdig sind. Auswahlkriterien können etwa die Reproduzierbarkeit bzw. die historische „Einmaligkeit“ von Daten, der Umfang und die Repräsentativität einer Erhebung oder der von einer Längsschnittstudie untersuchte Zeitraum sein (für eine ausführlichere Diskussion des Themas siehe Kapitel 3.2.3).

Im Allgemeinen ist die Arbeit, die in Datenmanagement investiert werden sollte, maßgeblich abhängig vom voraussichtlichen Nutzungsszenario, also insbesondere auch vom Ausmaß des geplanten Data Sharing. Denn je größer der Nutzerkreis, desto geringer ist der „kleinste gemeinsame Nenner“ an geteiltem Wissen über die Daten, für den keine explizite Dokumentation nötig ist. Dies lässt sich am *Data Curation Continuum*-Modell von Treloar & Harboe-Ree (2008) verdeutlichen (Abbildung 2). Das Modell unterscheidet zwischen den Domänen „privater“ Forschung, kollaborativer Forschung, und der (Forschungs-)Öffentlichkeit. Beginnend in der privaten Domäne, erfordert die Migration in die kollaborative Domäne eine erweiterte Ausstattung der Daten mit Metadaten sowie die Einrichtung einer entsprechenden Nutzungsumge-

bung. Die Weitergabe in die Public Domain erfordert die Auswahl eines geeigneten Archivs, eine noch weiter gehende Ausstattung mit Metadaten, die Zuteilung eines Persistent Identifier, usw.

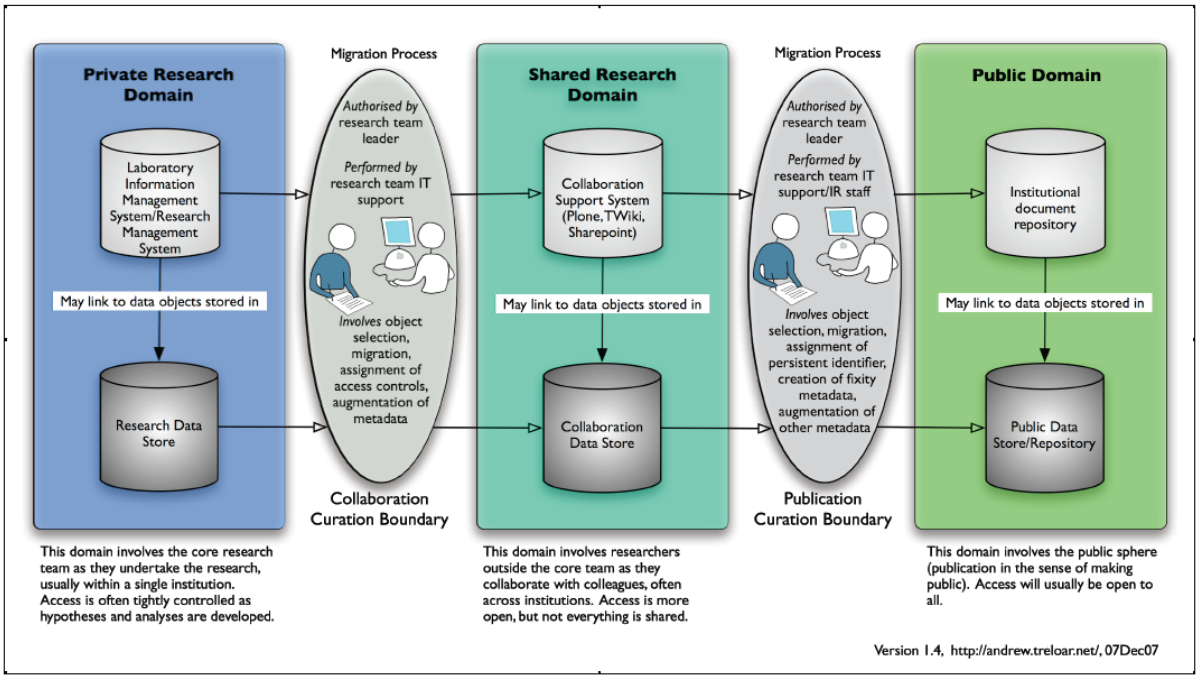


Abbildung 2. Data Curation Continuum (aus Treloar & Harboe-Ree, 2008)

In welchem Umfang Datenmanagement betrieben wird, sollte sich am zu erwartenden Nachnutzungspotential und am Umfang des potentiellen Nachnutzerkreises orientieren.

2.5 Besonderheiten in der Psychologie

An verschiedenen Stellen dieses Kapitels wurde bereits angedeutet, dass Datenmanagement und Data Sharing in verschiedenen Forschungsdisziplinen unterschiedlich verbreitet sind. Dies zeigte sich unter anderem in einer Umfrage unter Forschenden aus verschiedenen Disziplinen (Key Perspectives, 2010). Während etwa in Astronomie und Genforschung die Bereitschaft zum Data Sharing relativ hoch war, hielt sie sich bei den Sozial- und Gesundheitswissenschaften in Grenzen. Auch im Datenmanagement, etwa der Ausstattung der Daten mit Metadaten, zeigten sich Unterschiede. Darüber hinaus gibt es Forschungsbereiche, in denen Daten zwar gut dokumentiert und gepflegt, aber nicht geteilt werden (s. Kapitel 2.1). Ursachen für solche Unterschiede liegen vermutlich in Unterschieden in der Art (Key Perspectives, 2010) und Menge (Vlaeminck, 2008) der üblicherweise erhobenen Daten, in den verschiedenen, historisch gewachsenen „Forschungskulturen“ (Becher & Trowler, 2001) und in den jeweils vorherrschenden Anreizstrukturen (Weichselgartner et al., 2011a). Ein Bewusstsein für diese fächerspezifischen Umstände erleichtert es, angemessene Datenmanagement- und Data Sharing-Strategien zu wählen. In diesem

Abschnitt soll dementsprechend auf einige Besonderheiten bezüglich Datenmanagement und Data Sharing in der Psychologie eingegangen werden.

2.5.1. Die „Datenkultur" in der Psychologie

Breckler (2009) kommentierte das Verhältnis der Psychologie zum Data Sharing folgendermaßen: „The data culture of psychology is one of limited sharing, and then only among a select few with demonstrated competence and legitimate need. The federal rules right now support this culture, and psychologists cling tightly to it."

Dies zeigte sich beispielsweise auch beim Versuch von Wicherts, Borsboom, Kats und Molenaar (2006), für eine Reanalyse Datensätze, auf denen Publikationen in APA-Zeitschriften basierten, durch Anschreiben der Autoren zu erlangen. Sie erhielten auf diese Weise lediglich ein Viertel der angefragten Datensätze, obwohl die Ethikrichtlinien der APA die Weitergabe der Daten anmahnen. Auch in der Psychologie in Deutschland ist die Bereitschaft zum Data Sharing noch relativ gering (vgl. Weichselgartner, 2011a, S. 12).

Gründe dafür liegen in einer Reihe verschiedener negativer Anreize, die aber nicht allein spezifisch für die Psychologie sind (ähnliche Probleme existieren z.B. in der klinischen Forschung). Dies sind etwa ein erhöhter Aufwand (insbesondere da in der Psychologie Datenmanagement- und Data Sharing-Infrastrukturen noch wenig ausgeprägt sind), die potentielle Preisgabe von Vorteilen im wissenschaftlichen Wettbewerb durch die Veröffentlichung des eigenen, aufwändig erhobenen Datensatzes, die Gefahr einer Reputationsschädigung durch die Aufdeckung methodischer Unzulänglichkeiten oder die Einschränkung wirtschaftlicher Verwertungsmöglichkeiten (Weichselgartner et al., 2011a).

Darüber hinaus können die typische Form psychologischer Forschungsprojekte und -daten sowie spezifische rechtliche Risiken eine Rolle spielen, wie im Folgenden ausgeführt wird.

2.5.2. Art und Umfang psychologischer Forschungsdaten

In der Psychologie existiert sowohl eine qualitative als auch eine quantitative Forschungstradition. Darüber hinaus gibt es in der psychologischen Forschung eine Vielzahl verschiedener Forschungsparadigmen mit je eigenen Datenerhebungsmethoden. Dementsprechend wird aus einer großen Vielfalt von Datenquellen geschöpft, von Tagebüchern sowie Video- und Tonaufzeichnungen über strukturierte, teilstrukturierte oder unstrukturierte Interviews und Fragebögen zu Laborwerten und Computeraufzeichnungen über Verhaltensparameter wie Reaktionszeiten oder psychophysiologische Parameter. Eine vollständige Nachvollziehbarkeit des Erhebungsprozesses erfordert oft eine umfassende und damit arbeitsintensive Dokumentation, z.B. über Kodierungsvorschriften oder verwendete Kategorisierungssysteme (vgl. Weichselgartner, 2011b).

Anders als in den Sozial-, Wirtschafts- oder Bildungswissenschaften, in denen oft Daten aus großen nationalen Panelstudien verwendet werden, sind Forschungsprojekte in der Psychologie außerdem in der Regel eher auf kleine Stichproben ausgelegt, wobei aufwändige und umfangreiche Testverfahren zum Einsatz kommen (z.B. Intelligenztests). Da die in Panelstudien verwendeten Indikatoren in der Regel möglichst knapp gehalten sind (oft nur ein Item pro Konstrukt), um die Probanden nicht über die Gebühr zu belasten, bestehen in der Psychologie oft Vorbehalte gegenüber der methodischen Qualität der Befunde (Weichselgartner, 2011a).

Diese Faktoren führen zu einer stark „individuellen" Form psychologischer Forschungsprojekte, so dass die Nützlichkeit der entstehenden Datensätze für eventuelle Nachnutzer manchmal in Frage gestellt wird. Eine Abwägung, ob die Archivierung eines Datensatzes überhaupt sinnvoll ist, ist also durchaus angebracht. Gleichwohl lassen sich ohne weiteres auch Forschungsdaten in der Psychologie identifizieren, die zweifellos ein großes Nachnutzungspotential haben, etwa umfangreiche Normdatensätze zu Testverfahren, die z.B. für die Konstruktion von Kurzformen oder zur Untersuchung der historischen Verschiebung von Normwerten genutzt werden könnten oder die Daten der über Jahrzehnte angelegten längsschnittlichen Studien aus der Entwicklungs- und Persönlichkeitspsychologie. Wie das Beispiel der Beziehungs- und Familienpanelstudie *pairfam* [7] zeigt, lassen sich außerdem auch psychologische Fragestellungen erfolgreich in Panelstudien integrieren.

2.5.3. Rechtliche Aspekte im Umgang mit psychologischen Forschungsdaten

Schließlich sind besondere rechtliche Rahmenbedingungen im Umgang mit psychologischen Forschungsdaten zu berücksichtigen. Diese spielen bereits in der Erhebung und Analyse der Daten eine Rolle, bekommen aber im Falle einer Archivierung und Zugänglichmachung für Dritte ein besonderes Gewicht. Psychologische Forschungsdaten haben oft Personenbezug. Außerdem sind sie meist besonders sensibler Natur, wie etwa diagnostische Informationen im Fall der klinischen Psychologie, Angaben zur Arbeitszufriedenheit in der Arbeits- und Organisationspsychologie oder zu Schulleistungen in der pädagogischen Psychologie. Die Handhabung personenbezogener Daten ist in Deutschland insbesondere im Bundesdatenschutzgesetz und den Datenschutzgesetzen der Länder geregelt. Ein sachgemäßer Umgang erfordert insbesondere eine angemessene Anonymisierung der Daten sowie die informierte Einwilligung der Probanden. Eine solche Einwilligung kann sich auch auf die Weitergabe der Daten an Dritte erstrecken (und muss es auch, wenn Data Sharing stattfinden soll). Mehr zur Thematik findet sich in den Kapiteln 3.2.2 und 3.2.3.

Außerdem können urheberrechtliche Regelungen eine Rolle spielen, in Deutschland maßgeblich geregelt durch das Urheberrechtsgesetz. Betroffen sein können zum Beispiel die bei der Daten-

[7] http://www.pairfam.de/ [01.05.2013]

erhebung verwendeten Instrumente, insbesondere psychologische Tests, deren Nutzungsrechte oft bei Testverlagen liegen. Eine adäquate Datendokumentation sollte jedoch möglichst den Wortlaut der verwendeten Fragebogenitems sowie Informationen zur Kodierung der Antworten umfassen. Wird eine Archivierung und Zugänglichmachung solcher Dokumentation oder Daten angestrebt, sollte man diesen Umstand frühzeitig mit den (derzeitigen oder zukünftigen) Rechteinhabern abklären. Diese Problematik wird in Kapitel 3.2.3 eingehender behandelt.

In der Psychologie gibt es (bisher) keine Data-Sharing-Kultur. Als Gründe dafür kommen ein erhöhter Dokumentationsaufwand, die potentielle Preisgabe von Vorteilen im wissenschaftlichen Wettbewerb, Bedenken vor einer möglichen Aufdeckung von methodischen Unzulänglichkeiten oder die Einschränkung wirtschaftlicher Verwertungsmöglichkeiten in Betracht.

Der erhöhte Dokumentationsaufwand ergibt sich dadurch, dass in der Psychologie häufig recht komplexe Erhebungsmethoden (bei eher kleinen Stichprobenumfängen) eingesetzt werden.

Psychologische Forschungsdaten haben meist Personenbezug, weswegen Datenschutzbelange zu berücksichtigen sind. Empfehlenswert ist es, vor der Erhebung eine informierte Einwilligung der Probanden in die wissenschaftliche Nachnutzung durch Dritte einzuholen.

3. Leitfaden zum Datenmanagement in der Psychologie

Jede Forscherin und jeder Forscher wird bereits in irgendeiner Form Vorgehen und Ergebnisse bei der eigenen Forschung planen und dokumentieren. Dabei bewegt sich jeder Einzelne in dem Spannungsfeld zwischen „Ordnung ist das halbe Leben“ und „Das Genie überschaut das Chaos“. Ziel dieses Kapitels ist es, Forschenden aus der Psychologie, die die Balance ein wenig in Richtung „Ordnung“ verschieben wollen, einen strukturierten Überblick zu wichtigen Themenbereichen und konkreten Vorgehensmöglichkeiten beim Datenmanagement und Data Sharing zu geben. Vorausplanendes Handeln nimmt dabei eine Schlüsselrolle ein, weshalb zu Beginn des Forschungsprojekts ein Datenmanagementplan ausformuliert werden sollte (Kapitel 3.1). In diesem sollten alle in den einzelnen Phasen des Projekts für Datenmanagement und Data Sharing relevanten Tätigkeiten, soweit dies a priori möglich ist, beschrieben werden. In Kapitel 3.2 wird auf diese Tätigkeiten genauer eingegangen.

3.1 Datenmanagementpläne

In der Planungsphase eines Forschungsvorhabens sollten möglichst früh Überlegungen zu Datenmanagement und Data Sharing mit einfließen und in einem strukturierten Dokument, dem *Datenmanagementplan*, festgehalten werden. Sofern das Vorhaben über Drittmittel finanziert werden soll, sind die Antragsrichtlinien der potentiellen Förderer ein wichtiger Ausgangspunkt, da diese Vorgaben zur Anfertigung eines Datenmanagement- und Data Sharing-Plans enthalten

können. Tabelle 1 listet solche Vorgaben für einige Förderorganisationen auf. Da der Umgang mit Forschungsdaten ein sehr aktuelles Thema der Forschungspolitik und -förderung ist, können sich diese Vorgaben natürlich in naher Zukunft auch ändern.

Neben diesen handfesten Vorgaben existieren auch eine Reihe „weicherer" Empfehlungen von Wissenschaftsorganisationen und -politik sowie Richtlinien von Zeitschriften zum Umgang mit Forschungsdaten und relevante gesetzgeberische Bestimmungen. Diese beziehen sich in der Regel allerdings nicht spezifisch auf Datenmanagement*pläne* und werden daher in Kapitel 3.2.1 behandelt. Prinzipiell können natürlich auch einzelne Universitäten, Forschungsinstitute oder Fakultäten eigene Richtlinien bezüglich Datenmanagement- oder Data-Sharing-Plänen haben. Auch wenn dies zumindest in der Psychologie in Deutschland derzeit noch unüblich ist, ist es sinnvoll, sich im Zweifelsfall zu erkundigen, ob derartige Richtlinien existieren.

Tabelle 1. Vorgaben verschiedener Forschungsförderer zu Datenmanagement- und Data-Sharing-Plänen in Förderanträgen (Stand: Mai 2013)[8]

Organisation	**Vorgaben**
DFG (2012)	„Wenn aus Projektmitteln systematisch (Mess-)Daten erhoben werden, die für die Nachnutzung geeignet sind, legen Sie bitte dar, welche Maßnahmen ergriffen wurden bzw. während der Laufzeit des Projektes getroffen werden, um die Daten nachhaltig zu sichern und ggf. für eine erneute Nutzung bereit zu stellen. Bitte berücksichtigen Sie dabei auch -sofern vorhanden- die in Ihrer Fachdisziplin existierenden Standards und die Angebote bestehender Datenrepositorien."
NIH (2003)	„Investigators seeking $500,000 or more in direct costs in any year should include a description of how final research data will be shared, or explain why data sharing is not possible (…) in the form of a brief paragraph. The precise content of the data-sharing plan will

[8] Es existieren außerdem einige, online verfügbare Übersichten über datenbezogene Anforderungen von Forschungsförderern:

- Informationsportal Forschungsdaten: http://www.forschungsdaten.org/wissenswertes/uber-foerderer/ [02.05.2013] Übersicht über Anforderungen deutscher und wichtiger internationaler Förderer
- Oxford University Administration and Services: http://www.admin.ox.ac.uk/rdm/managedata/funderpolicy/ [02.05.2013] Links zu den Policies von britischen, amerikanischen und EU-Förderern
- Digital Curation Centre: http://www.dcc.ac.uk/resources/data-management-plans/funders-requirements [02.05.2013] Detaillierte Aufstellung der Anforderungen britischer Förderer

	vary, depending on the data being collected and how the investigator is planning to share the data. (…) Applicants (…) may wish to describe briefly the expected schedule for data sharing, the format of the final dataset, the documentation to be provided, whether or not any analytic tools also will be provided, whether or not a data-sharing agreement will be required and, if so, a brief description of such an agreement.“
NSF (n.d.)	„Proposals must include a supplementary document of no more than two pages labeled ‘Data Management Plan’. This supplement should describe how the proposal will conform to NSF policy on the dissemination and sharing of research results (…), and may include: 1. the types of data, samples, physical collections, software, curriculum materials, and other materials to be produced in the course of the project; 2. the standards to be used for data and metadata format and content (where existing standards are absent or deemed inadequate, this should be documented along with any proposed solutions or remedies); 3. policies for access and sharing including provisions for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights or requirements; 4. policies and provisions for re-use, re-distribution, and the production of derivatives; and 5. plans for archiving data, samples, and other research products, and for preservation of access to them.“
ESRC (2013)	„ESRC grant applicants who plan to generate data are responsible for preparing and submitting data management and sharing plans for their research projects as an integral part of the application (…). It is expected that an outline data management and sharing plan will include the following points: • an explanation of the existing data sources that will be used by the research project with references • an analysis of the gaps identified between the currently available

	and required data for the research • information on the data that will be produced by the research project, including: data volume; data type, eg. qualitative or quantitative data; data quality, formats, standards documentation and metadata; methodologies for data collection • planned quality assurance and backup procedures [security/storage] • plans for management and archiving of collected data • expected difficulties in data sharing, along with causes and possible measures to overcome these difficulties • explicit mention of consent, confidentiality, anonymisation and other ethical considerations • copyright and intellectual property ownership of the data • responsibilities for data management and curation within research teams at all participating institutions.“

Wie man an der Tabelle erkennt, sind die Vorgaben je nach Förderer unterschiedlich strikt und differieren in ihrer Ausführlichkeit. Die zitierte DFG-Richtlinie macht beispielsweise nahezu keine Vorgaben zur Gestaltung des Plans. Natürlich sollte man einen Datenmanagement-Plan nicht lediglich unter dem Gesichtspunkt der Förderrichtlinien betrachten, sondern auch als ein Hilfsmittel zur Strukturierung und vorausschauenden Gestaltung des eigenen Forschungsprojekts. Die systematische Beschäftigung mit den zu erhebenden Daten kann helfen, potentielle Probleme frühzeitig zu entdecken und zu entschärfen. Eine sorgfältig geplante Erhebung und Dokumentation erleichtert zudem wesentlich die spätere Datenanalyse und Verfassung von Manuskripten sowie die möglicherweise geplante Übergabe der Daten an ein Datenarchiv. Der eigentliche Zeitgewinn eines vorausschauenden Datenmanagements zeigt sich daher oft erst zum Abschluss einer Studie, da die mühsame (und manchmal gar nicht mehr mögliche) Rekonstruktion und Zusammenstellung benötigter Informationen entfällt. Im Rahmen der Zusammenstellung eines Förderantrages oder einer Projektbeschreibung müssen außerdem ohnehin vorausplanende Überlegungen getroffen werden, so dass der Datenmanagementplan als Teildokument dieser Gesamtplanung verstanden werden sollte.

Ein Datenmanagementplan besteht üblicherweise aus Unterabschnitten, in denen jeweils ein bestimmter Aspekt mehr oder weniger in sich geschlossen abgehandelt wird, wobei aber zwischen den Aspekten viele Abhängigkeiten bestehen. Beispielsweise sind strengere Sicherheits-

mechanismen nötig, wenn die Daten sensible personenbezogene Informationen enthalten, und die Art und Menge der erhobenen Daten bestimmen mit, welche Dateiformate geeignet sind.

Für eine bessere Übersichtlichkeit erfolgt eine Grobgliederung in die Tätigkeiten „vor", „während" und „nach" der „eigentlichen" Forschungsarbeit, also der Datenerhebung und -analyse. Wegen der bestehenden inhaltlichen Abhängigkeiten sollte man bei der Gestaltung der einzelnen Aktivitäten versuchen, die anderen Aktivitäten im Hinterkopf zu behalten.

Mögliche Unterabschnitte eines Datenmanagementplans (die nicht alle zwangsläufig enthalten sein müssen) behandeln die folgenden Themen, die in Kapitel 3.2 eingehend besprochen werden:

VOR DER DATENERHEBUNG UND -ANALYSE

- **Hintergrund des Forschungsprojekts**: Was ist die Motivation, die Zielsetzung, das Design der Studie? Was ist der Zusammenhang zu vorhergehender Forschung? Dies kann mehr oder weniger unmittelbar aus bereits bestehenden Beschreibungen des Forschungsvorhabens übernommen werden
- **Zusammenstellung relevanter Richtlinien, Empfehlungen, gesetzlicher Bestimmungen und Lizenzverträgen**: Welche für den Umgang mit den entstehenden Forschungsdaten relevanten Richtlinien und Empfehlungen sind zu berücksichtigen? Ist für die Datenerhebung die Lizenzierung von Testverfahren, Software oder Ähnlichem nötig? Falls bestehende Daten nachgenutzt werden, fallen dafür Nutzungsgebühren oder sonstige Aufwände an?
- **Klärung von Rollen und Verantwortlichkeiten**: Wer hat die Leitung des Forschungsprojekts inne? Wer trägt die Verantwortung für das Management der Daten als Ganzes? Wer führt welche Teilaufgaben durch? Welche Akteure außerhalb des Forschungsteams sind an Datenmanagement und Data Sharing beteiligt?
- **Recherche und Prüfung bereits existierender Daten**: Gibt es bereits (zugängliche) Daten, die für die Untersuchung der Fragestellung relevant sind? Wenn ja, ist die Erhebung zusätzlicher, neuer Daten oder die Integration von Sekundärdatensätzen aus verschiedenen Quellen nötig?
- **Planung des Datenerhebungsprozesses**: Was für Daten sollen erhoben werden, und wie? Wie groß wird die Stichprobe voraussichtlich sein? Gibt es bei der Datenerhebung logistische Probleme zu berücksichtigen (z.B. aufgrund von Transport, einer sehr großen Daten- oder Materialmenge)? Falls Daten aus unterschiedlichen Quellen erhoben werden, wie werden sie integriert?
- **Kosten-Budget für Datenmanagement und Data Sharing**: Welche Kosten fallen voraussichtlich speziell für Datenmanagement und Data Sharing an (z.B. durch zusätzlichen

Dokumentationsaufwand, Lizenzgebühren, Zusammenstellung der relevanten Materialien für eine Archivierung, eventuelle Gebühren für die Archivierung und Langzeiterhaltung der Daten)?

- **Gesetzliche Regelungen zum Datenschutz vor der Datenerhebung - Formulierung der informierten Einwilligung**: Welche Richtlinien und gesetzlichen Regelungen sind zu beachten? Welche Maßnahmen zur Sicherstellung werden ergriffen (z.B. Antrag an Ethikkommission, Informierte Einwilligung, Anonymisierung)? Wie wirken sich diese auf das Datenmanagement und auf Data Sharing aus?

WÄHREND DER DATENERHEBUNG UND ANALYSE

- **Erstellung der Daten- und Studiendokumentation**: Welche Informationen zu Studien- und Variablenmerkmalen sowie Datenerhebungs- und Datentransformationsprozeduren sollen in welcher Form erfasst werden? Was sind neben den Metadaten weitere für die Nutzung der Daten wichtige Dokumente? Welche Metadatenstandards sollen verwendet werden? Stehen kontrollierte Vokabularien zur Beschreibung der Daten zur Verfügung?
- **Art der Datenstruktur**: Wie sind die erhobenen Daten und Metadaten strukturiert? In welchen Dateiformaten werden sie gespeichert (dies kann sich zwischen Projektphase und anschließender Archivierung unterscheiden)? In welchem Format sind weitere für die Interpretation der Daten wichtige Dokumente gespeichert?
- **Organisation der Dateien, Versionierung**: Nach welcher Systematik werden die entstehenden Dateien benannt? Gibt es ein Versionskontrollsystem? Wie hängen die erzeugten Dateien miteinander zusammen? Gibt es eine „Master"-Datei, aus der abgeleitete Dateien erzeugt werden? Wie werden abgeleitete Datensätze oder anderweitig sich auf Quelldaten beziehende Dateien auf Ursprungsdateien bezogen?
- **Qualitätssicherung von Daten und Metadaten**: Werden die Erhebungsinstrumente und -prozeduren pilotgetestet? Wie wird das Problem fehlender Daten gehandhabt? Gibt es eine automatische Fehlerkontrolle und Konsistenzprüfung während der Dateneingabe bzw. -erhebung? Findet eine unabhängige Mehrfacheingabe der Daten statt? Welche Prüfungen werden im Anschluss an die Dateneingabe und nach Datentransformationen durchgeführt? Gibt es Prüfprozeduren für Metadaten (z.B. XML-Validierung)?
- **Sicherheitskopien, Datenintegrität, Zugriffskontrolle und Aufbewahrung**: In welchen Abständen werden Sicherheitskopien angefertigt? Wo werden diese aufbewahrt (welches Speichermedium, welche Räumlichkeiten)? Wer soll (während der Projektphase) Zugang zu den Daten haben, mit welchen Zugriffsrechten? Welche Schutzmaßnahmen werden implementiert (Zugang zu Räumlichkeiten, Passwörter, Verschlüsselung, Virenschutz- und Firewallprogramme)?

NACH DER DATENERHEBUNG UND -ANALYSE

- **Langzeiterhaltung und Archivierung**: Wie lange sollen die Daten mindestens vorgehalten werden? Ist Archivierungswürdigkeit der Daten gegeben? Wo sollen sie gegebenenfalls archiviert werden? Wie sollen Daten (z.B. zum Zweck der Anonymisierung) zerstört werden? Besteht die Möglichkeit einer Abwicklung des Datenarchivs, und was wird in diesem Fall zur Sicherung der Daten unternommen?
- **Berücksichtigung rechtlicher Bestimmungen zu Datenschutz und Urheberrecht**: Welche Maßnahmen zur Sicherstellung werden ergriffen, besonders bezüglich der Anonymisierung der Daten? Welche Daten müssen anonymisiert werden und in welchem Umfang? Werden die entstehenden Daten voraussichtlich urheberrechtlich oder anderweitig geschützt sein, und wenn ja, wie sollen Nutzungsrechte beim Data Sharing übertragen werden? Bei multinationalen Projekten: Ist allen relevanten gesetzlichen Regelungen der involvierten Jurisdiktionen Rechnung getragen?
- **Zugänglichmachung der Daten (Data Sharing)**: Wer sind potentielle Nachnutzer der Daten? Soll es eine Sperrfrist für den Zugang geben? Wie wird der Zugang geregelt (vertragliche Regelungen; Zugangsmodalitäten)? Soll der Zugang auf bestimmte Nutzergruppen beschränkt sein, oder gibt es verschiedene Versionen des Datensatzes für verschiedene Gruppen? Wie werden die Daten durch Dritte auffindbar gemacht (Persistent Identifiers)?

Wie detailliert ein Datenmanagement-Plan sein sollte, hängt auch von dem Umfang des geplanten Forschungsvorhabens ab. Für komplexe Vorhaben und solche, bei denen eine spätere Freigabe der Daten zur Nachnutzung geplant ist, sollte im Allgemeinen ein detaillierterer Datenmanagement- (und Data Sharing-) Plan angefertigt werden (vgl. Kapitel 2.4).

Einige weitere wichtige **Leitprinzipien beim Verfassen eines Datenmanagementplans** sind (Jones, 2011):

- **Konsultation und Kollaboration**: Eine Beratung durch Akteure, die mit Datenmanagement und Data Sharing vertraut sind, kann Orientierung geben. Universitätsbibliotheken und Datenarchive sind gute Ansprechpartner. Institutionelle Ethikkommissionen können Auskunft zu den nötigen Datenschutzmaßnahmen geben. Falls urheberrechtliche Bedenken bestehen, sollte eine Absprache mit Verlagen und anderen potentiellen Rechteinhabern stattfinden, bei komplizierter Rechtslage gegebenenfalls auch eine Rechtsberatung. Rechenzentren können Informationen zu vorhandener IT-Infrastruktur, zur sicheren Datenübertragung und zur Datensicherung geben.
- **Nutzung bestehender Ressourcen**: Vorhandene Infrastrukturen zu Datenmanagement und Data Sharing zu nutzen, kann viel Arbeit ersparen und erhöht in der Regel die Quali-

tät der Maßnahmen. Darunter fallen z.B. IT-Dienste des lokalen Rechenzentrums, die Dienstleistungen disziplinspezifischer Datenarchive, Software- oder Online-Tools zur Erstellung eines Datenmanagementplans und für das laufende Datenmanagement (deren Inhalte überschneiden sich relativ stark) sowie im Rahmen früherer Projekte im Institut gesammelte Erfahrungen, Dokumentvorlagen, usw.

- **Begründung der gewählten Maßnahmen**: Vorgaben zu Inhalten eines Datenmanagementplans, sofern sie überhaupt bestehen, sind meist bewusst unspezifisch gehalten. Dies gibt Entscheidungsspielraum bei der Anpassung des Datenmanagementplans an den Kontext des Forschungsvorhabens. Soweit möglich sollte eine Begründung für die gewählten Maßnahmen angegeben werden.
- **Umsetzbarkeit des Plans**: Die vorgesehenen Maßnahmen sollten im Verhältnis zu den vorhandenen Kapazitäten stehen. Gerade der für die Dokumentation während der Studie nötige Aufwand ist nicht zu unterschätzen. Eine Diskussion der bestehenden Arbeitsbelastung und der potentiellen Rollen und Verantwortlichkeiten für die verschiedenen Datenmanagement- und Data Sharing-Tätigkeiten im Forschungsteam (z.B. Pflege der Metadaten, Kontrolle der Datenkonsistenz, Anfertigung von Backups, Korrespondenz mit IT-Abteilung, Datenarchiv etc.) hilft bei der Feststellung, was machbar ist und was nicht.

Auch wenn es empfehlenswert ist, Datenmanagement- und Data-Sharing-Maßnahmen möglichst frühzeitig zu bestimmen, lässt sich selbstverständlich nicht jede Eventualität vorab planen, oder es stellt sich im Lauf des Projekts heraus, das bestimmte Aspekte des Plans in der vorgesehenen Form nicht umsetzbar sind. Daher bietet es sich an, ein für Daten- und Dokumentationsdateien vorgesehenes Versionierungssystem (Kapitel 3.2.2) auch auf den Datenmanagementplan selbst anzuwenden, d.h. eine „Logdatei“ zu führen, in der Änderungen am Plan und ggf. eine Begründung dafür festgehalten werden.

Ein Datenmanagementplan sollte sich an den Vorgaben des jeweiligen Forschungsförderers oder der Forschungseinrichtung orientieren. Er dient aber vor allem auch der Strukturierung und vorausschauenden Gestaltung des eigenen Forschungsprojekts.

Datenmanagementpläne bestehen aus verschiedenen Unterabschnitten, die einen bestimmten Aspekt des Datenmanagements behandeln. Diese lassen sich der Übersichtlichkeit halber den verschiedenen zeitlichen Phasen der Datenerhebung („vor“; „während“; „nach“) zuordnen.

3.1.1. Weiterführende Ressourcen

Zum Datenmanagement existieren bereits vergleichbare Leitfäden, von denen jedoch keiner spezifisch für die Psychologie ist. Da die Prinzipien eines rationalen Umgangs mit Daten trotz aller Disziplinunterschiede aber ähnlich sind, gibt es naturgemäß große inhaltliche Überschnei-

dungen zwischen allen solchen **Leitfäden;** auch der vorliegende Leitfaden baut maßgeblich auf den im Folgenden genannten auf:

- Der 45 Seiten starke Leitfaden des amerikanischen Inter-University Consortium for Political and Social Research (ICPSR) ist auf die Sozialwissenschaften zugeschnitten und damit auch für Psychologen von Interesse (ICPSR, 2012).
- Der vom *UK Data Archive* herausgegebene Leitfaden (van den Eynden, Corti, Woollard, Bishop & Horton, 2011; 40 S.) ist nicht fachspezifisch, enthält aber zahlreiche Fallbeispiele aus verschiedenen Fachbereichen.
- Vom Projekt *WissGrid* stammt der erste deutschsprachige (fächerunspezifische) Leitfaden, der auch eine ausführliche Checkliste für Datenmanagement enthält (Ludwig & Enke, 2013; 122 S.)
- Eine kurze Einführung inklusive eines Glossars und zahlreicher Links zu weiterführenden Informationen stammt vom *Data Observation Network for Earth (DataONE)* (Strasser, Cook, Michener & Budden, n.d., 11 S.).
- Eine webbasierte Einführung für Sozialwissenschaftler wird von der GESIS (n.d.) angeboten.
- Die Datenbibliothek *EDINA* der Universität Edinburgh bietet den Online-Trainingskurs *MANTRA* an, der eine multimedial aufbereitete Einführung sowie Übungsaufgaben für Datenmanagement in verschiedenen Statistiksoftwarepaketen (inklusive Trainingsdatensätzen) umfasst (EDINA, n.d.).

Weitere ähnliche Online-Leitfäden werden von vielen Universitätsbibliotheken angeboten (Curtin University Library, 2013; MIT Libraries, n.d.; Oxford University Administration and Services, 2012; University of Edinburgh Information Services, 2013).

Eine Reihe von **Checklisten** hilft, bei der Erstellung eines Datenmanagementplans keinen wichtigen Aspekt zu übersehen:

- Ludwig und Enke (2013, S. 83 ff).: Deutschsprachige Checkliste für Datenmanagement, relativ ausführlich und nach Tätigkeitsbereich gegliedert, nicht spezifisch für bestimmte Fächer oder Förderrichtlinien.
- Checkliste der Curtin University Library [9]: Englischsprachig, nach Tätigkeitsbereich gegliedert, nicht spezifisch für bestimmte Förderrichtlinien.
- Donnelly und Jones (2011): Englischsprachige Checkliste des Digital Curation Centre, sehr detaillierte Gliederung mit Erläuterungen und Ratschlägen zu möglichen Inhalten, nicht fachspezifisch, allerdings fokussiert auf die Anforderungen britischer Förderein-

[9] http://libguides.library.curtin.edu.au/research-data-management [02.05.2013], Link „Planning checklist“

richtungen; von Plant (2011) stammt ein auf diese Checkliste abgestimmter Musterdatenmanagementplan für ein psychologisches Forschungsvorhaben.

- ICPSR (2012, S. 15 ff): Kompakte englischsprachige Checkliste für Sozialwissenschaften, mit kurzem Beispiel-Datenmanagementplan.

Die folgenden **Webtools** können bei der Datenmanagementplanung behilflich sein (bei allen ist eine kostenlose Registrierung erforderlich). Bislang existiert allerdings kein dediziertes Datenmanagementplanungstool für den deutschsprachigen Bereich.

- *DMPonline* [10]: Das Planungstool vom Digital Curation Centre basiert auf den Items der oben genannten Checkliste. Es enthält Vorlagen gemäß der Anforderungen britischer Fördereinrichtungen und Universitäten sowie der NSF, aber auch eine generische Vorlage.
- *DMPTool* [11]: Von der California Digital Library angebotenes Planungstool. In der Funktionsweise ähnlich DMPonline, allerdings mit Vorlagen für US-Forschungsförderer (inkl. fächerspezifischer Anforderungen der NSF).
- *MyPsychData*[12]: Das vom ZPID angebotene, in Kapitel 4.2.1 im Detail dargestellte Webtool für die psychologische Forschung ist nicht primär für die Erstellung eines Datenmanagementplans gedacht, sondern zur Dokumentation wichtiger Studien- und Datenmerkmale. Diese hängen aber naturgemäß eng mit den Inhalten eines Datenmanagementplans zusammen, daher ist das Tool auch zur Ausgestaltung dieses Plans hilfreich.

Es existieren bereits verschiedene, online verfügbare Leitfäden, hilfreiche Checklisten und Webtools, die das Datenmanagement erleichtern.

3.2 Datenmanagement-Aspekte in einzelnen Phasen des Forschungsprozesses

Dieses Unterkapitel soll dabei helfen, die zuvor aufgezählten Aspekte eines Datenmanagement- und Data-Sharing-Plans mit Inhalt zu füllen.

3.2.1 Vor der Datenerhebung und -analyse

3.2.1.1. Hintergrund des Forschungsprojekts

Rahmenbedingungen für Datenmanagement-Aktivitäten sind durch bereits feststehende oder übergeordnete Merkmale des eigenen Forschungsvorhabens gegeben, wie z.B. Zielsetzung(en), geplanter Zeitrahmen, Kooperationspartner, aber auch Hintergrundwissen aus zur untersuchten Fragestellung gesammelter Literatur und aus eigenen früheren Forschungsarbeiten. Oft besteht

[10] http://dmponline.dcc.ac.uk/ [02.05.2013]
[11] http://dmp.cdlib.org/ [02.05.2013]
[12] http://mypsychdata.zpid.de/ [15.07.2013]

hier bereits eine Dokumentensammlung (Literatursammlung, Exposés oder sonstige Beschreibungen des Vorhabens, Dokumentenvorlagen, Kostenabrechnungen, …), auf denen die Dokumente zu Datenmanagement und Data Sharing aufbauen können.

Bereits existierende Dokumentensammlungen zum Forschungshintergrund (z. B. Literatur, Antragsskizzen, etc.) können als Grundlage für den Datenmanagementplan verwendet werden.

3.2.1.2. Zusammenstellung relevanter Richtlinien, Empfehlungen, gesetzlicher Bestimmungen, Lizenz- und Nutzungsverträge

Es bietet sich an, als Teil dieser Dokumentensammlung zum Forschungsvorhaben vorab rechtliche Bestimmungen und Policy-Dokumente, die voraussichtlich relevant für Datenmanagement und Data Sharing sind, zusammenzustellen und laufend zu erweitern, um den Überblick über bindende Vorgaben zu behalten. Übersieht man solche Vorgaben, kann dies später zu mehr oder weniger unangenehmen Störungen des Forschungsvorhabens führen (je nach „Härtegrad" der verletzten Vorgabe). Beispielsweise kann es passieren, dass die informierte Einwilligung so formuliert ist, dass sie nicht zur späteren Datenweitergabe im Rahmen der Nachnutzung autorisiert oder der Lizenzvertrag zur Nutzung eines Testverfahrens gestattet nicht, bestimmte Details zur Anwendung des Verfahrens weiterzugeben, die für das Verständnis der Daten durch Nachnutzer wichtig wären.

Es existieren verschiedene **Arten von Dokumenten**, die in so eine Sammlung aufgenommen werden können (vgl. Pampel & Bertelmann, 2011; Spindler & Hillegeist, 2011):

- **Verbindliche Anforderungen durch Drittmittelgeber**: Die für Fördermittelnehmer der DFG verbindlichen „Vorschläge zur Sicherung guter wissenschaftlicher Praxis" fordern beispielsweise, „Primärdaten (…) auf haltbaren und gesicherten Trägern in der Institution, wo sie entstanden sind, für zehn Jahre" aufzubewahren (DFG, 1998). Auch die Datenmanagement- und Data-Sharing-Maßnahmen, zu denen man sich in einem Förderantrag verpflichtet hat, können hier aufgenommen werden.
- **Anforderungen der eigenen Institution**: Die oben genannten DFG-Richtlinien verpflichten geförderte Institutionen auch dazu, eigene Regelungen zu etablieren. Bei der Max-Planck-Gesellschaft sind Institute etwa verpflichtet, die Daten mindestens zehn Jahre zu erhalten und aufzubewahren und darüber hinaus „für berechtigte Interessenten" Zugang zu den Daten zu gewähren (Max-Planck-Gesellschaft, 2009).
- **Richtlinien von Fachverbänden**: Auch Fachverbände haben Regelwerke guter wissenschaftlicher Praxis und Ethikrichtlinien verabschiedet, die Aussagen zum Datenmanagement und Data Sharing enthalten können. Darunter fallen z.B. die „auf die Forschung bezogenen ethischen Richtlinien" der DGPs (2004, Punkt 14).

- **Editorial Policies von Zeitschriften**: Einige Zeitschriften stellen „datenbezogene" Bedingungen für die Publikation, etwa dass Daten auf Anfrage verfügbar gemacht werden oder in einem Archiv eingestellt werden. Dies gilt vor allem für *Open Access*-Zeitschriften oder Zeitschriften in Disziplinen, in denen eine starke Data-Sharing-Kultur existiert.
- **Empfehlungen von Wissenschaftsorganisationen und -politik**: Es existieren einige Empfehlungen, die zwar unverbindlich sind, aber als Richtschnur genutzt werden können: Zum Beispiel die „Empfehlungen zur gesicherten Aufbewahrung und Bereitstellung digitaler Forschungsprimärdaten" der DFG (2009) oder die „OECD principles and guidelines for access to research data from public funding" (OECD, 2007).
- **Gesetzliche Bestimmungen**: An erster Stelle stehen hier datenschutzrechtliche Bestimmungen, da in der Psychologie häufig personenbezogene Daten erhoben werden. Relevant sind in Deutschland das Bundesdatenschutzgesetz, das jeweilige Landesdatenschutzgesetz sowie das Sozialgesetzbuch X. Unter Umständen spielt auch das Urheberrecht eine Rolle, allerdings weniger bezüglich der erhobenen Daten, da diese selten unter Urheberschutz fallen, sondern häufiger bezüglich lizenzierter Testverfahren o.ä. (siehe unten). In der klinischen Forschung ist außerdem zu beachten, dass das Regelwerk zur „guten klinischen Praxis" (International Conference on Harmonisation, 1996), das strenge Anforderungen an das Datenmanagement stellt, durch das Arzneimittelgesetz und die GCP-Verordnung Gesetzescharakter hat.
- **Lizenzvereinbarungen**: Nutzungsbedingungen für Software, z.B. Experimentalsoftware, Software für Online-Umfragen oder Statistikprogramme sowie für psychologische Testverfahren sollten (auch) unter dem Gesichtspunkt der Eignung für die eigenen Datenmanagement- und Data-Sharing-Aktivitäten gesichtet werden. Dabei können Probleme entdeckt werden, die unter Umständen durch eine zusätzliche Vereinbarung mit dem Rechteinhaber ausgeräumt werden können. Hierunter fällt beispielsweise auch die Vereinbarung, Normdaten für ein zu entwickelndes Testverfahren, das später kommerziell verlegt wird, veröffentlichen zu dürfen.
- **Nutzungsbedingungen von Forschungsdatenarchiven**: Institutionelle Archive verfügen über eigene Regelungen darüber, welche Daten unter welchen Bedingungen aufgenommen werden, wie damit verfahren wird und unter welchen Bedingungen sie weiter gegeben werden können. Auch hier sind natürlich weitere Vereinbarungen zwischen Archiv und der eigenen Forschergruppe als potentiellem Datengeber denkbar.

Eine Zusammenstellung von Policies (z.B. von Forschungsförderern, Instituten, Fachverbänden, Zeitschriftenverlagen), rechtlichen Bestimmungen (z.B. von Wissenschaftsorganisationen, politischen Organen), Lizenzvereinbarungen und Regelungen von Datenarchiven wird für sinnvoll erachtet, damit eventuell existierende Verpflichtungen nicht übersehen werden.

3.2.1.3. Klärung von Rollen und Verantwortlichkeiten

Eine Vielzahl von Akteuren ist typischerweise an einem Forschungsvorhaben beteiligt und spielen eine (mehr oder weniger zentrale) Rolle für gelingendes Datenmanagement und Data Sharing. Eine Aufstellung der Beteiligten und ihrer Datenmanagement- und Data-Sharing-bezogenen Verantwortlichkeiten zu Beginn des Vorhabens hilft, diese Rollen zu erfüllen, insbesondere wenn Akteure aus verschiedenen Institutionen beteiligt sind (z.B. bei Forschungskooperationen). **Wichtige Akteure** sind zum Beispiel:

- Der/die **Gesamtleiter/-in des Forschungsvorhabens** (Gesamtverantwortung für das Vorhaben und damit auch für die Zuteilung und Koordination der Verantwortlichkeiten für Datenmanagement und Data Sharing);
- Gegebenenfalls ein/-e **Hauptverantwortliche/-r für Datenmanagement und Data Sharing** (z.B. verantwortlich für Qualitätssicherungs- und Dokumentationsprozeduren);
- **Wissenschaftliche Mitarbeiter** (Erhebung, Eingabe, Verarbeitung, Analyse von Daten);
- **Mitarbeiter der institutionellen IT-Infrastruktureinrichtungen** (Aufbewahrung und Sicherung der Daten, ggf. Einrichtung und Pflege von Datenbanken);
- **Nichtwissenschaftliche Mitarbeiter und Hilfskräfte** (z.B. Verwaltung und Pflege von Dokumenten, Durchführung von Erhebungen, Dateneingabe);
- **Mitarbeiter des Forschungsdatenarchivs** (Beratung zu Datenmanagement und Data Sharing, Unterstützung bei Vorbereitung der Übernahme der Daten, Langzeitarchivierung und Verfügbarmachung);
- **Sonstige externe Dienstleister** (z.B. Datentreuhänder bei sensiblen personenbezogenen Daten, externe Umfrage- und Interviewdienstleister, Hard- und Softwareanbieter).

Auf Grundlage einer Festlegung von Rollen und der damit verbundenen Tätigkeiten können Zugangs- und Zugriffsrechte zu Dateien und Räumlichkeiten sowie spezifische Schulungsmaßnahmen definiert werden, etwa eine Schulung für einen neuartigen Metadatenstandard, zur Dateneingabesoftware für nichtwissenschaftliche Mitarbeiter oder zu einer *virtuellen Forschungsumgebung* für wissenschaftliche Mitarbeiter.

Die eindeutige Festlegung von Rollen und Verantwortlichkeiten in einem Forschungsprojekt erleichtert das Datenmanagement, da u.a. Zugangs- und Zugriffsrechte sowie Schulungsmaßnahmen klar definiert werden können.

3.2.1.4. Suche nach und Prüfung von bereits existierenden Daten

Eine Durchsuchung der Bestände bereits vorhandener Daten kann helfen, den mit einer Datenerhebung verbundenen großen zeitlichen und finanziellen Aufwand zu vermeiden oder zu redu-

zieren. Dafür bieten sich **Verzeichnisse und Suchportale für Datenrepositorien** an, wie z.B. die folgenden (z.T. noch im Aufbau befindlich):

- **Open Access Directory – Data repositories** [13]: Nach Disziplinen gegliederte Auflistung von Repositorien, die (zum Teil) frei zugängliche Forschungsdaten enthalten;
- **re3data.org** [14]: Nicht disziplinspezifisches Suchportal, basierend auf einem detaillierten XML-Schema zur Beschreibung der Repositorien (z.B. abgedeckte Disziplinen, verantwortliche Institutionen, Datenpolicies des Repositoriums);
- **Databib** [15]: Nicht disziplinspezifisches Suchportal, Suche z.B. nach Disziplin und Sitzland;
- **DataCite – Suchoberfläche** [16]: Interface für die Suche im Metadatenbestand der DOI-Registrierungsagentur *DataCite* (s. Kapitel 3.2.3); nicht disziplinspezifisch, umfasst neben Datensätzen auch Ressourcen wie Software oder Audio- und Videodaten (Ressourcentyp über Suchoberfläche spezifizierbar);
- **PsychLinker - Datenarchive**[17]: Linksammlung von Datenarchiven, Datenzentren und einzelnen Datenanbietern, in denen Forschungsdaten zu psychologischen Forschungsthemen angeboten werden;
- **Katalog des Council of European Social Science Data Archives (CESSDA)** [18]: Verzeichnis von Datensätzen der CESSDA-Mitglieder (u.a. GESIS). Auf Sozialwissenschaften beschränkt; Datensätze anhand von mehrsprachigem Thesaurus recherchierbar**;**
- **PsychSpider – Kollektion „Forschungsdaten“** [19]: Psychologiespezifische Web-Suchmaschine, enthält eine gesondert absuchbare Kollektion zu Forschungsdaten (umfasst verschiedene psychologierelevante, im deutschsprachigen Raum angesiedelte Datenrepositorien, auf Datensatzebene durchsuchbar).

Das Vorgehen bei der Suche nach Datensätzen kann analog zur Recherche nach Publikationen zu früherer Forschung gesehen werden, wenn auch die Informations- und Suchinfrastruktur im Bereich der Forschungsdaten derzeit noch bei weitem nicht so gut ausgebaut ist wie bei der Forschungsliteratur. Ebenso, wie man einzelne Autoren wegen unpublizierter Forschungsberichte anschreiben kann, besteht eine weitere Recherchemöglichkeit (neben der Suche in Datenbanken und Verzeichnissen) darin, Autoren mit der Bitte um Verfügbarmachung potentiell relevanter, aber nicht publizierter Datensätze anzuschreiben. Dieses Vorgehen ist eher mühselig und es muss zumindest derzeit von einer eher geringen Bereitschaft ausgegangen werden Daten wei-

[13] http://oad.simmons.edu/oadwiki/Data_repositories [03.05.2013]
[14] http://www.re3data.org/ [03.05.2013]
[15] http://databib.org/ [03.05.2013]
[16] http://search.datacite.org/ [12.05.2013]
[17] http://www.zpid.de/redact/category.php?cat=8 [17.07.2013]
[18] http://www.cessda.org/accessing/catalogue/ [12.05.2013]
[19] http://www.zpid.de/PsychSpider.php [03.05.2013]

terzugeben (s. Kapitel 2.5). Diese Strategie erscheint also allenfalls von Interesse, wenn ein Datensatz fundamentale Bedeutung für das eigene Forschungsvorhaben hat.

Zum Zeitpunkt der Erstellung dieser Manualversion ist die Menge archivierter und nachnutzbarer Datensätze in der Psychologie noch relativ begrenzt, weshalb sich die Anzahl potentiell relevanter Datensätze vermutlich schnell eingrenzen lässt. Dennoch sollte ein gewisser Aufwand eingeplant werden, da solche *potentiell* relevanten Daten-sätze auf die tatsächliche Eignung zur Untersuchung der eigenen Forschungsfrage anhand der mitgelieferten Dokumentation genauer untersucht werden müssen, z.B. hinsichtlich Erhebungszeitraum, untersuchter Population, Stichprobengröße bzw. Teststärke, Effekt der interessierenden Stärke oder Operationalisierungen (gerade bei letzteren müssen erfahrungsgemäß oft eher Abstriche in dem Sinn gemacht werden, dass nicht punktgenau das interessierende Konstrukt erfasst wird).

Falls keine relevanten Daten gefunden werden, kann auf gewohnte Weise mit der Planung der Erhebung fortgefahren werden. Findet man dagegen Datensätze, die für die Untersuchung der eigenen Forschungsfrage geeignet sind, stellt sich die Frage, inwieweit diese Daten alleine zur Beantwortung der Frage ausreichend sind. Ist dies der Fall, kann direkt zur Datenanalyse übergegangen werden, die darüber hinaus durch die in der Regel hochwertige Qualität und Dokumentation archivierter Datensätze erleichtert ist.

Ist der Datensatz zu klein (beispielsweise weil nur eine spezifische Subgruppe von Interesse ist), kann unter Umständen (wenn sich zur nachgenutzten Erhebung mehr oder weniger vergleichbare Bedingungen herstellen lassen) eine eigene Nacherhebung durchgeführt werden, um auf die gewünschte Stichprobe zu kommen. Wurden relevante Variablen nicht erhoben, lässt sich der Datensatz eventuell durch eine Kombination mit anderen Datensätzen, z.B. mit Zensusdaten, entsprechend ergänzen. Die Herausforderung besteht hier natürlich darin, erst einmal einen geeigneten ergänzenden Datensatz zu finden, und die Einträge (*Records*) der beiden Datensätze dann miteinander zu verbinden. Ein derartiges Vorgehen kann z.B. sinnvoll zur Ergänzung von Daten auf Personenebene durch Informationen zu Kontext- oder aggregierten Variablen sein (etwa Daten zur Wohngegend). Selbstverständlich kann auf diese Weise auch ein neu erhobener Datensatz ergänzt werden. Der Prozess der Verbindung der Datensätze, die sogenannte *Record Linkage*, ist komplex und wird hier nicht weiter behandelt. Ressourcen finden sich z.B. auf der Website des *German Record Linkage Center* [20]. Weiterhin ist bei der Verbindung von Datensätzen mit personenbezogenen Daten stets auf datenschutzrechtliche Bestimmungen zu achten.

Für die Suche nach existierenden Datensätzen, die für das eigene Forschungsvorhaben verwendet werden können, bieten sich verschiedene, online verfügbare Verzeichnisse und Suchportale an oder – alternativ – eine direkte Anfrage bei dem Autor des interessierenden Datensatzes.

[20] http://record-linkage.de/ [03.05.2013]

Erweist sich ein Datensatz nach eingehender Prüfung als geeignet für die eigene Untersuchung, kann direkt zur Datenanalyse übergegangen werden. Andernfalls kann auch versucht werden, einen nur teilweise geeigneten Datensatz durch Kombination mit anderen Datensätzen zu ergänzen.

3.2.1.5. Charakterisierung der Daten und des Datenerhebungsprozesses

Die Erstellung eines kompakten Überblicks über die Erhebungsprozedur und die dabei entstehenden Daten hilft als Grundlage für weitere Datenmanagement- und Data-Sharing-Entscheidungen, zum Beispiel der Zuteilung von Arbeitsaufgaben, der Aufwands- und Kostenabschätzung sowie der Berücksichtigung von Datenschutzbestimmungen. In dem Überblick können auch eventuell vorgefundene nachnutzbare Sekundärdaten berücksichtigt werden.

So ein Überblick könnte zum Beispiel dergestalt strukturiert sein, dass der idealtypische Erhebungsablauf für eine Versuchsperson geschildert wird. Auf der obersten Gliederungsebene werden die einzelnen Erhebungstermine, die die Versuchsperson absolviert, sowie die beim jeweiligen Termin angewendeten Erhebungsprozeduren und –instru-mente aufgelistet (z.B. ein Interview, eine ärztliche Untersuchung, eine experimentelle Prozedur). Für jede solche Prozedur wird wiederum eine Kurzcharakterisierung (z.B. technische Details zum Randomisierungsverfahren, Name und Inhalt eines Fragebogens, wichtige Details zur Durchführungsmodalität), die dadurch erzeugten Daten und Dokumente (z.B.: Welche Dateitypen werden erzeugt? Welche nicht-digitalen Daten und Dokumente werden erzeugt) und sonstige Maßnahmen wie z.B. logistische Überlegungen (Welche Speicherkapazität benötigen die erzeugten Dateien, wie werden sie gespeichert und ggf. an einen zentralen Speicherort übermittelt? Wie werden nicht-digitale Materialien transportiert? Gibt es beim Transport bzw. der Datenübertragung Sicherheitsmaßnahmen zu beachten?) angegeben. In einem separaten Abschnitt werden dann weitere Schritte beschrieben, die nötig sind, um die erzeugten Daten in Primärdaten zu transformieren, so dass die Verbindung zu den konkreten Variablen, die später analysiert werden, hergestellt ist (z.B. Kodierung von Video- oder Audiodateien).

Die umgekehrte Herangehensweise bestünde darin, alle letztendlich interessierenden Variablen aufzulisten und für jede den Prozess ihrer Erzeugung zu beschreiben.

Ein Überblick über die Erhebungsprozedur und die entstehenden Daten kann entweder so strukturiert werden, dass von dem idealtypischen Erhebungsablauf her die einzelnen Schritte bis zur Beschreibung der erzeugten Daten dargestellt werden, oder dass umgekehrt ausgehend von den interessierenden Daten der Erhebungsprozess beschrieben wird.

3.2.1.6. Kostenabschätzung für Datenmanagement und Data Sharing

Vor allem bei Forschungsvorhaben größeren Umfangs ist eine Abschätzung des Aufwands für Datenmanagement- und Data-Sharing-bezogene Aktivitäten sinnvoll, um einen entsprechenden Budgetanteil zu reservieren. Dieser kann auch als Argumentationshilfe bei der Förderantragstellung verwendet werden. Denn in dem Maß, in dem Forschungsförderer Datenmanagement und Data Sharing verlangen, müssen sie selbstverständlich auch gewillt sein, (zusätzliche) Mittel dafür bereitzustellen. Bei kleineren Vorhaben kann eine Abschätzung sinnvoll sein, um zu bestimmen, welche Datenmanagement- und Data-Sharing-Aktivitäten im vorgegebenen Rahmen umsetzbar sind.

Meist ist so eine Abschätzung nur eine sehr „vage" Angelegenheit. Es bestehen zwei Vorgehensmöglichkeiten: Entweder können die Kosten für alle „datenbezogenen" Aktivitäten abgeschätzt werden, also die Kosten für die Erhebung und alle Datenmanagement und Data-Sharing-Aktivitäten zusammen, oder es werden die Kosten für letzteres über die ohnehin für die Datenerhebung, -verarbeitung und -analyse anfallenden Kosten hinaus geschätzt. Für letztere Herangehensweise hat das UK Data Archive (2012) eine Checkliste zusammengestellt. Ausgehend von dieser Liste seien hier einige **Faktoren, die bei der Erstellung eines Kostenplans eine Rolle spielen können**, genannt:

- **Planungs- und organisatorische Maßnahmen** (Erstellung des Datenmanagementplans, Absprache von Rollen und Verantwortlichkeiten, Erstellung des Kostenplans selbst);
- **Schulungs- und Beratungsmaßnahmen**, z.B. für Prozeduren zur Datensicherung, -prüfung und -dokumentation;
- **Gebühren für den Erwerb von Software, Dokumenten** und dergleichen, die spezifisch für Datenmanagement und Data Sharing benötigt werden;
- **Be- und Verarbeitung der entstehenden Daten** (über das auch ohne explizites Data Management und Data Sharing geplante Maß hinausgehend, z.B. zusätzliche Integritäts-, Konsistenzchecks, Versionierung und Datensicherung);
- **Erhöhter Dokumentationsaufwand** (z.B. Erzeugung von Metadatendateien);
- **Vorbereitung der Übergabe an ein Datenarchiv** (z.B. Dateiformatkonvertierungen, Anonymisierungsmaßnahmen, Zusammenstellung von Dokumenten);
- Bei der **Übergabe an ein Archiv** eventuell anfallende Gebühren (vom Archiv selbst, zum Erwerb von Nutzungsrechten, die eine Archivierung ermöglichen);
- **Nach der Archivierung anfallende Tätigkeiten** zur Betreuung des Datensatzes in Kooperation mit dem Archiv (laufende Korrespondenz mit dem Archiv, Besprechungen auf Anlass besonderer Ereignisse).

Diese Liste beruht auf der Annahme, dass die Daten an ein externes, nicht von der eigenen Forschungsgruppe betriebenes Archiv übergeben werden. Werden die Daten in einem eigenen Archiv (z.B. institutionseigenes Repositorium) aufbewahrt und zugänglich gemacht, sind natürlich auch die Kosten für den Betrieb des Archivs zu berücksichtigen (Organisation von Räumlichkeiten, Anschaffung und Betrieb einer geeigneten IT-Infrastruktur, Entwicklung von Arbeitsabläufen und eigener Softwarelösungen, Pflege und Zugänglichmachung des Datenbestands, Aufbau und Betrieb einer eigenen Website, …).

Ein Kostenplan für Datenmanagement und Data Sharing ist hilfreich bei der Budgetplanung und für eine Abschätzung des mit den verfügbaren Mitteln realisierbaren (zusätzlichen) Arbeitsaufwands. Dabei können entweder die Gesamtheit aller datenbezogenen Prozesse oder nur die Aufgaben bezüglich Datenmanagement und Data Sharing berücksichtigt werden.

Zu den zu berücksichtigenden Kosten gehören organisatorische Maßnahmen, Schulungs- und Beratungsmaßnahmen, anfallende Gebühren für Software, zusätzliche Tätigkeiten für die Aufbereitung, Dokumentation und für die Übergabe der Daten an ein Datenarchiv sowie laufende Betriebskosten bei eigener Archivierung.

3.2.1.7. Datenschutz vor der Datenerhebung: Formulierung der informierten Einwilligung (informed consent)

Gemäß den allgemein anerkannten Prinzipien ethischer Humanforschung (DGPs, 2004; World Medical Association, 2008) muss vor jedweder Datenerhebung am Menschen eine informierte Einwilligung der Probanden erfolgen. Die dem Probanden gegebenen Informationen sollten Angaben zur Datenerhebungsprozedur, ihrem Sinn und Zweck sowie zur im weiteren Verlauf erfolgenden Verarbeitung und Nutzung der Daten enthalten, inklusive eventueller Sekundärdaten, die mit den erhobenen Daten verknüpft werden sollen (in der Psychologie sind in begrenztem Maße Täuschungsmaßnahmen im Lauf der Datenerhebung gestattet, über die aber mit Abschluss der Datenerhebung aufzuklären ist). Die Einwilligung ist sowohl bezüglich der Teilnahme an der Datenerhebung als auch der Nutzung der Daten jederzeit revidierbar (letzteres zumindest solange Personenbezug besteht, die Daten also nicht anonymisiert wurden). Sie sollte auf jeden Fall in schriftlicher Form erfolgen, sicherheitshalber in mehrfacher Ausfertigung um sich gegen späteren Verlust oder Zerstörung einer Ausfertigung abzusichern (Spindler & Hillegeist, 2009).

Die Erstellung eines adäquat formulierten Einwilligungsformulars (und gegebenenfalls eines separaten Probandeninformationsblattes) ist also eine wichtige Grundlage dafür, was mit den Daten im weiteren Forschungsverlauf gemacht werden kann. Dies gilt insbesondere auch für die Möglichkeit, Daten langfristig aufzubewahren und sie anderen Forschern zur Verfügung zu stellen. Daher sollte bereits in der Probandeninformation auf ein solches Vorhaben hingewiesen werden und explizite Zustimmung eingeholt werden. Darüber hinaus sollte darauf geachtet

werden, dass Datenarchivierung und Data Sharing nicht implizit durch bestimmte Formulierungen verunmöglicht wird, etwa durch die Aussage, dass *alle* entstehenden Daten (also auch anonymisierte Daten) nach Ablauf des Forschungsprojekts vernichtet werden oder dass ausschließlich die erhebende Forschergruppe Zugang zu den Daten haben wird. Ein Einwilligungsformular kann auch die Möglichkeit zu einer „abgestuften" Einwilligung bieten, indem die Zustimmung zu verschiedenen Nutzungsarten der Daten separat abgefragt wird (siehe van den Eynden et al., 2011, S. 24 für ein Beispiel eines solchen Formulars, das allerdings auf die britische Gesetzgebung angepasst ist).

Das Bundesdatenschutzgesetz sieht vor, dass personenbezogene Daten gelöscht werden (d.h. die Daten anonymisiert werden), „sobald der Forschungszweck dies gestattet". Im Regelfall ist daher eine Anonymisierung erst nach Abschluss des laufenden Forschungsvorhabens (und der damit verbundenen Datenanalysen oder eventuell notwendigen Kontaktaufnahmen mit den Versuchspersonen) nötig. Deshalb wird auf Techniken zur Anonymisierung und Pseudonymisierung noch nicht hier, sondern erst in Kapitel 3.2.3 genauer eingegangen. Unter Umständen kann es aber bereits während des laufenden Projekts nötig sein, solche Maßnahmen zu ergreifen, etwa wenn äußerst sensible Daten erhoben werden sollen, die die Versuchsteilnehmer nur unter absoluter Anonymität preiszugeben bereit sind (siehe Dietz, Striegel, Franke, Lieb, Simon & Ulrich, 2013, für ein Beispiel, in dem sich ein dramatischer Unterschied in den Ergebnissen einer anonymisierten gegenüber einer nichtanonymisierten Erhebung zeigte).

Auch wenn es laut Gesetzestext nicht verboten wäre, anonymisierte Daten ohne eine vorherige Einwilligung des Probanden Dritten zugänglich zu machen (vgl. Metschke & Wellbrock, 2002, Kapitel 3), ist es gute Praxis, auf eine solche Verwendung hinzuweisen. Eine kurze Beschreibung von Mechanismen zur Gewährleistung der Anonymität kann Vertrauen schaffen, dass nicht doch durch die Hintertür wieder ein Personenbezug hergestellt wird. Eine Konsultation mit der institutionellen Ethikkommission (bei der in aller Regel ohnehin ein Antrag auf Prüfung des Forschungsvorhabens vorgelegt werden muss) kann helfen, eine angemessene Formulierung zu finden.

Mit der informierten Einwilligung werden den Probanden Angaben zur Datenerhebungsprozedur, dem Sinn und Zweck des Forschungsvorhabens sowie zur anschließenden Verarbeitung und Nutzung der Daten vorgelegt.

Es sollte darauf Wert gelegt werden, dass in der Formulierung die Möglichkeit zur anonymisierten Datenweitergabe (für wissenschaftliche Zwecke) ausdrücklich enthalten ist.

3.2.2 Während der Datenerhebung und -analyse

3.2.2.1. Daten- und Studiendokumentation anhand von Metadaten

In Kapitel 2.1 wurde bereits darauf hingewiesen, dass die erhobenen Datenwerte für sich genommen nur schwer oder gar nicht interpretierbar sind. Erst durch zusätzliche, beschreibende Informationen werden sie verständlich. So könnte etwa die Zeichenkette „1910“ für eine Jahreszahl, die Entfernung zweier Orte, eine Reaktionszeit aus einem Experiment oder eine Telefonnummer stehen. Die Dokumentation der zur Interpretation der Daten nötigen Informationen ist also ein unverzichtbarer Bestandteil der Forschungsarbeit. Idealerweise sollte die Dokumentation zeitnah zum Eintreten des dokumentierten Ereignisses bzw. zum Erhalt der zu dokumentierenden Information erfolgen; bei Informationen zu einer Variable beispielsweise, sobald man sich auf deren Erhebungsmodus, den Datentyp, die Messeinheit oder Ähnliches festgelegt hat, spätestens aber während der Datenerhebung, solange das notwendige Wissen noch präsent ist. Je mehr Zeit zwischen einem Ereignis und seiner Dokumentation verstreicht, desto umständlicher und unpräziser wird die Dokumentation im Allgemeinen.

Der Unterschied zwischen einer Studie mit und ohne Datenmanagement besteht nicht so sehr darin, *ob* eine Dokumentation stattfindet, sondern *wie*. Die Variablen einer SPSS-Datendatei und ihre Werte mit beschreibenden Labels zu versehen, ist in der Psychologie gängige Praxis. Dies allein würde aber nicht zur adäquaten Interpretation der Daten genügen – es ist darüber hinaus auch die Kenntnis von Details zur Erhebungsprozedur, zum Erhebungszeitraum, zum Hintergrund der Studie und vielen anderen Aspekten nötig. Hier besteht das Risiko einer lückenhaften Dokumentation, wenn man sich über Datenmanagement keine Gedanken macht, da den an der Erhebung unmittelbar Beteiligten die nötigen Informationen während der Erhebung ständig präsent sind und so kein unmittelbarer Anlass besteht, zusätzliche Arbeit für ihre Dokumentation aufzuwenden.

Um eine systematische Dokumentation aller Informationen zu erreichen, die mit der Studie nicht vertraute Forschende (oder man selbst fünf Jahre später) zum Verständnis benötigen, sollte man auf ein strukturiertes Schema der zu erfassenden Informationen zurückgreifen, also auf ein Metadatenschema. Dazu bestimmt man zunächst alle interessierenden Objekte, über die man beschreibende Metadaten sammeln möchte, z.B. Metadaten über einzelne Variablen, über den Datensatz insgesamt oder über die Studie als Ganzes.

Für jedes solche Objekt listet man dann auf, welche Merkmale man dokumentieren will. Für die Variablen könnten dies zum Beispiel die Merkmale „Variablenname“, „Variablenlabel“, „Wertebereich“ und „Erhebungsprozedur“ sein, für den Datensatz „Beobachtungseinheit“, „Anzahl Fälle“, „Anzahl Variablen“, „Datei, die den Datensatz enthält“, und für die Studie „Erhebungszeitraum“, „Beschreibung des Hintergrunds der Studie“, „Studienleiter“, etc. Für Zwecke der Quali-

tätssicherung (siehe unten) und der Maschinenlesbarkeit kann man außerdem definieren, welche Werte diese Metadatenelemente jeweils annehmen dürfen (Beispiel: Nur ganze Zahlen für „Anzahl Fälle“). Zwischen den Objekten „Studie“, „Datensatz“ und den „Variablen“-Objekten bestehen außerdem hierarchische Beziehungen: Der Datensatz wurde im Rahmen der Studie erhoben und ist ihr damit logisch untergeordnet. Die Variablen wiederum sind als Teil des Datensatzes diesem untergeordnet.

Erweitert man das oben skizzierte Schema noch um eine zusätzliche, ebenfalls der Studie untergeordnete Objektklasse „Sonstiges für die Studie wichtiges Dokument“, in der man beschreibende Informationen zu weiteren Dokumenten, wie etwa dem Datenmanagementplan, Programmkode für eine selbstprogrammierte Experimentalprozedur und dergleichen erfasst, hat man ein Schema, mit dem sich für die meisten quantitativen Studien alle relevanten Metadaten erfassen lassen. Ein vergleichbares Metadatenschema wird auch von PsychData verwendet (s. Kapitel 4). Im Weiteren geht es zunächst nur um die logische Form des abstrakten Metadatenschemas. Wie man eine konkrete Metadatendatei basierend auf dem Schema erstellen kann, wird am Ende des Abschnitts sowie im Abschnitt „Datenformate“ dieses Kapitels erörtert.

Im Folgenden werden einige oft verwendete **Metadatenkategorien auf Variablenebene für einen Datensatz aus der quantitativen Forschung** aufgezählt. Man beachte, dass abgesehen von einigen fundamentalen Angaben, die für die Verarbeitung durch den Computer unabdingbar sind (Variablenname, Datentyp), diese Angaben nicht zwingend für jede Variable gemacht werden müssen; oft ist das auch gar nicht sinnvoll, z.B. die Angabe von fehlenden Werten bei einer Identifikatorvariable, die einfach die Fälle durchnummeriert. Außerdem können die meisten der Metadatenkategorien je nach Bedarf auch anders „zugeschnitten“ werden, z.B. können Kategorien zusammengefasst werden (etwa: Angabe der Messeinheit im Variablenlabel) oder aufgespalten werden. Ein Teil der hier aufgeführten Merkmale ist vermutlich aus der Variablenansicht in SPSS bekannt.

- **Variablenname**: Ein eindeutiger (im Datensatz nur einmal vorkommender) Bezeichner für die Variable, der von den relevanten Statistikprogrammen „verstanden“ wird, also ihren syntaktischen Vorgaben entspricht (in der Regel: keine Leerzeichen und nur ASCII-Symbole enthalten, nicht mit einer Ziffer beginnend; bei älteren Programmen oft auch eine Begrenzung auf maximal acht Zeichen).
- **Datentyp**: Eine Vorschrift dazu, welche Werte die Variable annehmen darf. Alle von Computern verarbeiteten Daten haben einen solchen Typ, andernfalls könnte der Computer sie nicht interpretieren. Dadurch unterscheidet sich der Datentyp von der Angabe eines Wertebereichs, die primär inhaltlich begründet ist (beispielsweise, dass der systolische Blutdruck nicht den Wert 1000 annehmen kann). Geläufige Datentypen sind zum Beispiel Zeichenkette (beliebige Kombination von Text, Ziffern und Sonderzeichen),

Ganzzahliger Wert (oft auch: „Integer“), logischer / Wahrheitswert (WAHR oder FALSCH, oft auch „Boole’scher Wert“ genannt), Fließkommazahl, Datumsangabe (in der Regel unter Spezifikation einer bestimmten Formatierung, z.B. DD-MM-YYYY), und Aufzählungen (auch: „Enum“; entspricht einer Kategorialvariable, z.B. die Variable „Jahreszeit“, die nur die Werte „Frühling“, „Sommer“, „Herbst“ und „Winter“ zulässt).

- **Variablenlänge**: Die maximale (ggf. auch minimale) erlaubte Zeichenlänge des Variablenwerts.
- **Variablenlabel**: Eine knappe, möglichst aussagekräftige Beschreibung der Variablen. Das Label dient der Gewinnung eines schnellen Überblicks bzw. der Erinnerung der Bedeutung der Variablen und kann auch zur Beschriftung von Diagrammen oder Tabellen verwendet werden.
- **Messeinheit.**
- **Wertebereich**: Sofern nicht ohnehin durch die Angabe des Datentyps abgedeckt, dient dieses Merkmal der Spezifikation eines Intervalls oder einer Aufzählung erlaubter Werte. Bei Kategorialvariablen basierend auf einem komplexen Klassifikationssystem (z.B. einer standardisierten Kodierung der Berufstätigkeit) ist auch ein Verweis auf eine externe Datei sinnvoll.
- **Fehlende Werte**: Aufzählung von Variablenwerten, die vergeben werden, wenn kein Variablenwert für die Versuchsperson (bzw. Beobachtungseinheit) vorhanden ist. Je nach Art der Variable kann es sinnvoll sein, verschiedene solcher Werte zu definieren, abhängig davon, worauf die „Missingness“ zurückzuführen ist. In jedem Fall sollten Werte gewählt werden, die außerhalb des Bereichs gültiger Variablenwerte liegen (s. a. den Abschnitt zur Qualitätssicherung).
- **Wertelabels**: Eine kurze, aussagekräftige Beschreibung der Bedeutung der Werte einer kategorialen Variable; Labels zu fehlenden Werten können hier oder in einem eigenen Metadatenelement aufgeführt sein
- **Variablengruppe**: Ein Schlagwort zur Gruppierung inhaltlich verwandter Variablen, z.B. „Demographie“, „Persönlichkeitsmerkmale“, …
- **Datenerzeugender Prozesses**: Eine Beschreibung der Prozedur, anhand der die Primärdaten erzeugt wurden. Bei Fragebogendaten kann dies beispielsweise die Angabe des Fragebogens, der Itemnummer auf dem Fragebogen sowie der genaue Itemtext sein, bei einer abgeleiteten oder einer Gewichtungsvariable die Berechnungsvorschrift (oder, bei einem komplexen Algorithmus, ein Verweis auf eine externe Datei wie z.B. eine Programmkode-Datei), und bei einer aufwändig kodierten Kategorialvariable die Kodierungsinstruktionen mit Verweis auf die zugrundeliegende Rohdatendatei (z.B. ein Video). Wie man sieht ist dies eine relativ breite Merkmalsklasse, die gegebenenfalls auch in spezifischere Teilelemente aufgespalten werden kann.

- **Datenort**: Ein Verweis auf den genauen Ort, an dem man die Daten auffinden kann, etwa eine bestimmte Spalte einer Tabellendatei.
- **Editierungsschritte**: Eine Beschreibung der „Bearbeitungshistorie" der Daten, ausgehend von den Daten, die ursprünglich durch den datenerzeugenden Prozess generiert wurden. Hier können insbesondere die Editierungen im Rahmen von Qualitätskontrollen, Datenbereinigung und dergleichen vermerkt werden.
- **Details zu imputierten Werten**: Angaben dazu, ob die Variable imputierte Werte enthält, wie die Imputationsprozedur aussah, und ggf. ein Verweis auf die Ursprungsvariable mit fehlenden Werten.
- **Deskriptivstatistiken**: Beispielsweise eine Überblicksdarstellung absoluter und/oder relativer Häufigkeiten der Variablenwerte und deskriptivstatistische Kennwerte.

Neben den Metadaten auf Variablenebene sind insbesondere die auf die Studie als Ganzes bezogenen Metadaten wichtig für das Verständnis der Daten. Im Folgenden werden einige mögliche **Metadatenelemente auf Studienebene** aufgezählt. Eine Aufspaltung der Studienmetadaten in spezifische Subkategorien ist insbesondere unter dem Gesichtspunkt der Absuchbarkeit von Datenarchiven sinnvoll, da die Suchfunktion von Datenarchiven oft auf den Studienmetadaten basiert; beispielsweise kann so nach allen Studien gesucht werden, die eine bestimmte Mindest-Stichprobengröße haben. Aus Platzgründen sind die hier genannten Kategorien aber relativ breit gefasst.

- **Beteiligte**: Eine Aufzählung der an der Datenerhebung beteiligten Personen und Institutionen sowie ihrer Rollen bzw. des von ihnen geleisteten Beitrags
- **Titel**: Der „offizielle" Titel des Forschungsvorhabens;
- **Hintergrund der Studie**: Eine Beschreibung des theoretischen Hintergrunds (inklusive Verweisen auf Hintergrundliteratur), der Zielsetzung und Fragestellung, und ggf. der statistischen Hypothesen der Studie;
- **Stichprobenmerkmale**: Details zur Gewinnung der Stichprobe (Beschreibung der Zielpopulation, Ziehungsprozedur, Response Rate, Flowchart zur Illustration der Dropout-Raten, finale Stichprobengröße);
- **Gewichtungsverfahren**: Beschreibung der Prozedur zur Gewinnung von Gewichtungsfaktoren;
- **Erhebungszeitraum;**
- **Erhebungsort;**
- Beschreibung der **Datenerhebungsprozedur** (Terminierung und Ablauf der Erhebungssitzungen, angewendete Erhebungsinstrumente, Software und Hardware);
- Beschreibung der **Maßnahmen zur Sicherung der Datenqualität** (Pilotstudien, Konsistenzprüfungen, …);

- Angaben zu **nachgenutzten Sekundärdaten** (Quelle, Integration mit anderen Daten, …);
- **Fördermittel, Förderkennzeichen;**
- **Schlagworte zur inhaltlichen und methodischen Charakterisierung** der Studie („Klinische Psychologie", „Querschnittsstudie", …). Diese können frei gewählt sein oder aus einem kontrollierten Vokabular stammen (z.B. ZPID, 2011);
- **Aus der Studie hervorgegangene Datensätze**: Beschreibung der Datensätze und Verweis auf die entsprechenden Datensatz-Dateien. Dies ist vor allem bei Längsschnittstudien mit wiederholten Erhebungswellen relevant. Aber auch bei Studien, die auf einer Erhebung basieren, können verschiedene Teildatensätze oder verschiedene Versionen bestehen (z.B. aggregierte oder anonymisierte Versionen, Arbeitsversionen vs. nicht abänderbare Quelldaten);
- **Aus der Studie hervorgegangene Publikationen**: Ggf. mit Angabe, auf welchem Datensatz genau die Analysen, die in der Publikation berichtet werden, basieren;
- **Sonstige für die Dateninterpretation wichtige Dokumente**: Kurzbeschreibung und Verweis auf die entsprechenden Dateien (z.B. ausführliche Kodierinstruktionen, Lizenzbestimmungen für die Nachnutzung, Konkordanzlisten für einander entsprechende Items in verschiedenen Wellen einer Längsschnittstudie).

Bei Studien mit komplex strukturierten Daten oder anderen wichtigen Dokumenten, z.B. Softwarekode, sollten auch die einzelnen Datensätze bzw. Dokumente mit je eigenen Metadaten versehen werden (anstelle einer Beschreibung durch ein Metadatenelement auf Studienebene, wie in der Aufzählung oben). Bei qualitativen Studien ist es beispielsweise sinnvoll, ein auf die jeweilige Erhebungsform (Interview, Fokusgruppe, Tage-buch, …) eigens zugeschnittenes Metadatenschema zu verwenden, bei Interviewtranskripten etwa über Ort, Zeit, Setting, Interviewer und Angaben zum biographischen Hintergrund der Interviewten.

Die eigenständige Entwicklung eines Metadatenschemas hat den Vorteil hoher Flexibilität. Allerdings hat es den großen Nachteil, nicht *interoperabel* zu sein, d.h. nicht abgestimmt mit den von anderen Forschern und Archiven verwendeten Schemata, und damit letztendlich nur innerhalb der eigenen Gruppe verwendbar zu sein. Um Interoperabilität und damit letztendlich eine bessere Auffindbarkeit der zur Nachnutzung archivierten Datensätze zu ermöglichen (anhand der Recherche nach einheitlich dokumentierten, spezifischen Merkmalen der Datensätze über Suchportale), wurden mittlerweile verschiedene Metadatenstandards speziell für Forschungsdaten entwickelt (vgl. Jensen et al., 2011). In der Regel sind diese Standards spezifiziert über XML-Schemadateien oder Dokumenttypdefinitionen (DTD) (Yott, 2005).

Ein speziell für Daten aus den Sozialwissenschaften entwickelter internationaler Standard ist das Metadatenschema der *Data Documentation Initiative (DDI)* [21], das in mehreren Versionen existiert. Version 2 („DDI-Codebook“) ist geeignet zur Beschreibung weniger komplexer Studiendaten, etwa von Querschnittserhebungen und ähnelt in seiner Struktur dem hier skizzierten Schema (die Metadatenobjekte auf der höchsten Hierarchieebene sind das Metadatendokument selbst, die Studie als Ganzes, die Datendateien, der Datensatz und sonstige Materialien; zu jedem dieser Objekte sind jeweils eigene Metadatenelemente spezifiziert). Die aktuelle Version 3 („DDI-Lifecycle“) wurde erweitert, um eine bessere Beschreibung komplexer Datenstrukturen (z.B. hierarchisch organisierte Daten), miteinander in Beziehung stehender Studien (z.B. Längsschnittstudien), multinationaler Studien und weiterer Merkmale zu ermöglichen.

Auch Metadatenstandards, die nicht spezifisch für Forschungsdaten entwickelt wurden, können zum Teil für deren Beschreibung verwendet werden. Wenig aufwändig implementierbar und weit verbreitet ist das *Dublin Core Metadata Element Set* [22], das zur Beschreibung von Dokumenten und Ressourcen im Internet entwickelt wurde und für das eine XML-Spezifikation besteht. Es umfasst lediglich 15 Elemente, die grundlegende Eigenschaften wie Format, Gattung, Sprache, Titel und Urheber abdecken. Eine erweiterte Version beinhaltet 55 Elemente. Eine Beschreibung mit Dublin Core-Metadaten garantiert ein Mindestmaß an Interoperabilität und ist eine gute Alternative, wenn keine Kapazitäten für eine ausführliche Beschreibung mit Metadaten vorhanden sind.

Die Erstellung eines XML-Metadatendokuments wird durch die Verwendung eines der zahlreichen verfügbaren XML-Editor-Programme (darunter auch viele kostenfrei zugängliche) erleichtert. Diese können automatisch ein Grundgerüst des XML-Dokuments basierend auf einer Schemadatei oder einer DTD erstellen und insbesondere auch prüfen, ob ein Dokument konform mit der Schemadatei bzw. der DTD ist. Dennoch ist die Metadatenerstellung gemäß einem Metadatenstandard (oder auch einem selbst entwickelten Schema, das als XML-Datei implementiert werden soll) insbesondere für nicht mit Informatik und IT-Technologien Vertraute mit großem Aufwand verbunden, da eine Einarbeitung in XML und die Spezifikation des Standards selbst erforderlich ist. Wem das zuviel Aufwand ist, kann auf spezielle **Metadateneditierungssoftware** zurückgreifen:

- *Colectica for Microsoft Excel* [23] ist ein kostenfrei herunterladbares Add-In für Excel. Nach der Installation erscheint in Excel ein zusätzlicher Reiter, über den auf unkomplizierte Weise Metadaten zu einem in Excel geladenen Datensatz hinzugefügt und als DDI-XML-, PDF-, oder Word-Datei exportiert werden können.

[21] http://www.ddialliance.org/ [06.05.2013]
[22] http://dublincore.org/documents/dces/ [06.05.2013]
[23] http://www.colectica.com/software/colecticaforexcel [06.05.2013]

- *Nesstar Publisher* [24] ist ein ebenfalls kostenfrei verfügbares Programm zur Erstellung von DDI-XML. Das Programm ist auch zur Erzeugung detaillierter Metadaten geeignet, ist einfach bedienbar und benötigt keinerlei XML-Kenntnis.

Für die Psychologie existiert derzeit kein eigener Metadatenstandard. Das im ZPID für PsychData entwickelte psychologiespezifische Schema orientiert sich allerdings an den Spezifikationen von DDI und Dublin Core. *MyPsychData* ist ein webbasiertes Metadatenerzeugungstool, das eine einfache Studien- und Datendokumentation gemäß diesem Schema erlaubt. Es sind keine XML- oder sonstigen technischen Vorkenntnisse erforderlich (s. Kapitel 4.2.1).

Neben den Metadatenschemata selbst stellen *kontrollierte Vokabulare* einen weiteren wichtigen Aspekt der standardisierten Beschreibung von Forschungsdaten dar. Diese sollen helfen, mit einer begrenzten Anzahl von Begriffen die Facetten eines interessierenden Gegenstandsbereichs erschöpfend zu erfassen, möglichst ohne dass es Bedeutungsüberschneidungen zwischen Begriffen gibt. Die DDI entwickelt derzeit verschiedene solcher Vokabularien, die dann innerhalb der Metadatenelemente verwendet werden können, beispielsweise eine kontrollierte Liste möglicher Beobachtungseinheiten („Individual“, „Organization“, „Family“, „Geographic Unit“, …). Auch die aus der Literatursuche bekannten Thesaurussysteme (für die Psychologie z.B. Gallagher Tuleya, 2007; ZPID, 2011) sind kontrollierte Vokabulare, die sich zur thematischen und methodischen Beschreibung von Datensätzen verwenden lassen. Auf ähnliche Weise bietet sich auch die Verwendung bestehender Klassifikationssysteme an, etwa die ISO-Normbegriffe für geographische Einheiten (ISO 3166) und Sprachen (ISO 639-1:2002).

Wenn eine spätere Freigabe der Forschungsdaten zur Nachnutzung vorgesehen ist, sollte schließlich erwogen werden, Dokumentation und insbesondere Metadaten von vornherein in Englisch zu verfassen, da dies den potentiellen Nachnutzerkreis wesentlich vergrößert. Dies kann jedoch einen beträchtlichen zusätzlichen Arbeitsaufwand darstellen, beispielsweise müssen dann Fragebogenitems und Nutzungsanweisungen, die sonst einfach per Copy/Paste übernommen werden könnten, übersetzt werden. Eine Möglichkeit ist, eine Auswahl der wichtigsten Metadatenelemente und Dokumente zu treffen, die zweisprachig dokumentiert werden. Die Verwendung von englischsprachigen kontrollierten Vokabularien bei den Metadaten und von international standardisierten Klassifikationssystemen (z.B. von Schuldbildung oder Arbeitstätigkeiten) oder bereits im Englischen vorliegenden Erhebungsinstrumenten kann die Arbeitslast außerdem weiter verringern.

[24] http://www.nesstar.com/software/publisher.html [06.05.2013]

Empfehlenswert ist eine zeitnahe Dokumentation, die neben der Variablendokumentation auch eine Beschreibung des Studienhintergrunds und des Erhebungskontexts umfasst.

Um Interoperabilität zu gewährleisten ist es sinnvoll, ein bereits vorhandenes Metadatenschema zu verwenden, bspw. den für die Sozialwissenschaften entwickelten Standard DDI (Data Documentation Initiative) oder das weit verbreitete Dublin Core Metadata Element Set.
Als weitere Hilfen zur Dokumentation können Metadateneditierungssoftware zur Erstellung von XML-Metadatenelementen und kontrollierte Vokabulare verwendet werden.

Um die spätere Nachnutzbarkeit zusätzlich zu erhöhen, sollte eine englischsprachige Dokumentation erwogen werden.

3.2.2.2. Art der Datenstruktur (Datenorganistation)

Neben der Tatsache, dass die für die Dateninterpretation nötigen Informationen während der Erhebung ohnehin präsent sind, mag ein anderer Grund für eine lückenhafte Dokumentation darin liegen, dass SPSS als verbreitetste Statistiksoftware zwar eine bequeme Möglichkeit der Dokumentation von Variablen- und Wertelabels in einer Datei zusammen mit den Daten bietet, aber wenig dergleichen für darüber hinausgehende Informationen. Eine sorgfältige Dokumentation erfordert also die Pflege von mehr als bloß der Datendatei selbst.

Dies kann besonders während der Datenerhebung, -prüfung und -analyse, in der laufend Änderungen vorgenommen werden, schnell eine unübersichtliche Situation schaffen, in der die Aktualisierung bestimmter Dateien vergessen wird. Fällt dann später auf, dass die Informationen nicht mehr zusammenpassen, kann man sich schon nicht mehr erinnern, was wie geändert werden müsste. Weiter verkompliziert wird die Angelegenheit, wenn eine Datei von verschiedenen Bearbeitern aktualisiert wird, insbesondere wenn die Bearbeiter eigene Arbeitskopien anfertigen. Das Ergebnis kann Doppelarbeit, ein mühseliges Wiederzusammenfügen oder das Überschreiben einer zwischenzeitlich aktualisierten Datei sein. Eine gut durchdachte Festlegung auf ein System der Organisation von Daten, Dateien, und ihrer verschiedenen Versionen hilft, dieses Problem zumindest zu mildern.

Als Ausgangspunkt für Überlegungen zur Dateiorganisation kann zunächst die Struktur der Forschungsdaten unabhängig von einem bestimmten Dateiformat dienen. Daten können prinzipiell auf vielfältige Weise geordnet werden, je nachdem, wie es für den jeweiligen Arbeitskontext am vorteilhaftesten ist. Zur Übertragung von Daten im Netzwerk ist eine sequenzielle Repräsentation als eine lange „Datenzeile“ angemessen, für die Datenverarbeitung im Rahmen der Datenerhebung und –analyse dagegen kaum. Für diesen Zweck sinnvolle **Datenstrukturen** werden im Folgenden erläutert und in Abbildung 3 veranschaulicht:

- Die klassische und in den meisten Fällen angemessene Repräsentationsform ist die **rechteckige Datenmatrix**, in der Beobachtungseinheiten durch die Zeilen und Variablen durch die Spalten repräsentiert werden. Sie ist übersichtlich, erfasst alle Daten in einem Datensatz und lässt sich durch alle Statistik- und Tabellenkalkulationsprogramme darstellen. Allerdings hat dieses Format Schwierigkeiten damit, hierarchisch strukturierte Daten darzustellen, bei denen Beobachtungseinheiten auf verschiedenen Hierarchieebenen mit je eigenen Merkmalen existieren, beispielsweise bei Messwiederholungen an derselben Person: Die Person hat überdauernde Merkmale, aber es gibt auch Variablen, die für die einzelnen Messzeitpunkte spezifisch sind.

  Solange für jede Person dieselbe Anzahl an Messzeitpunkten vorliegt, lassen sich die Daten im gewohnten Format (manchmal auch „wide"-Format genannt) speichern: pro Merkmal und Messzeitpunkt eine Spalte. Falls aber unterschiedlich viele Messzeitpunkte vorliegen, ist man gezwungen, eine große Anzahl leerer Datenzellen zu erzeugen. Eine Lösung liegt in der Verwendung des Datensatzes im „long"-Format: Jede Zeile stellt hier eine Beobachtungseinheit der untersten Hierarchieebene dar, im genannten Beispiel also einen Messzeitpunkt bei einer bestimmten Person.

  Das long-Format erlaubt also die Repräsentation hierarchischer Daten in der klassischen Rechteckform und ist auf die übliche Weise in Statistik- und Tabellenkalkulationsprogrammen darstellbar. Manche Statistikprogramme verlangen sogar eine Repräsentation der Daten im long-Format zur Durchführung bestimmter Analysen (Mixed Effects-Modelle). Redundanterweise müssen aber die Merkmale der Beobachtungseinheit übergeordneter Ebenen in jeder Zeile wiederholt werden. Dies dürfte allerdings lediglich bei sehr großen Datenmengen ein Problem darstellen.
- Eine Möglichkeit, die oben erwähnte Redundanz zu vermeiden, bietet eine explizit **hierarchische Organisationsform** der Daten. Die Daten sind gemäß einer Baumstruktur organisiert: Ausgehend von einem „Wurzelobjekt" können die Daten in einem einzigen Datensatz in einer komplexen Aufgliederung in beliebig vielen Hierarchieebenen repräsentiert werden. Eine hierarchische Datenorganisation kann beispielsweise anhand einer XML-Datei erfolgen; diese bietet zusätzlich prinzipiell die Möglichkeit, Daten direkt zusammen mit den Metadaten in einer einzigen Datei zu speichern. Die Datenanalyse auf Grundlage hierarchischer Datendateien ist jedoch nicht so intuitiv wie bei „rechteckiger" Organisation und erfordert im Allgemeinen, dass die Daten anhand von Programmskripten in die Statistiksoftware importiert bzw. in die hierarchische Datei exportiert werden.

  Schließlich kann das Problem hierarchischer Datenstrukturen durch die Verwendung **relationaler Datenbankstrukturen** gelöst werden. Der Grundgedanke hierbei ist, dass Daten verschiedener Hierarchieebenen jeweils in einer eigenen rechteckigen Tabelle (die in diesem Kontext verwirrenderweise auch „Relation" genannt wird) gespeichert

werden und die Tabellen über bestimmte Schlüsselvariablen miteinander verknüpft werden. Im Fall einer messwiederholten Erhebung gäbe es etwa eine Tabelle zur Repräsentation von Individuen, und eine für individuelle Erhebungssitzungen, wobei die Verbindung durch den Identifikationskode der Versuchsperson hergestellt würde. Statistische Analysen können dann auf Grundlage einzelner Tabellen durchgeführt werden (z.B. die Person-Tabelle, wenn nur die Unterschiede zwischen überdauernden Persönlichkeitsmerkmalen von Interesse sind) oder es können durch die Verknüpfung von Tabellen neue Tabellen, die alle nötigen Informationen enthalten, erzeugt werden.

Mit relationalen Datenbankstrukturen lassen sich komplexe Zusammenhänge elegant repräsentieren. Allerdings erfordern sie einen erhöhten Verwaltungsaufwand, da die Tabellen in je eigenen Dateien gespeichert werden und zusätzlich eine Datenbankschemadatei nötig ist, welche die Zusammenhänge zwischen den Tabellen beschreibt. Außerdem ist auch die Verwendung einer Datenbankmanagementsystemsoftware erforderlich. Diese Programme basieren meist auf der relationalen Datenbanksprache *SQL*. Zur Generierung von Datentabellen zur statistischen Analyse ist also zusätzlich auch die Kenntnis von SQL nötig.

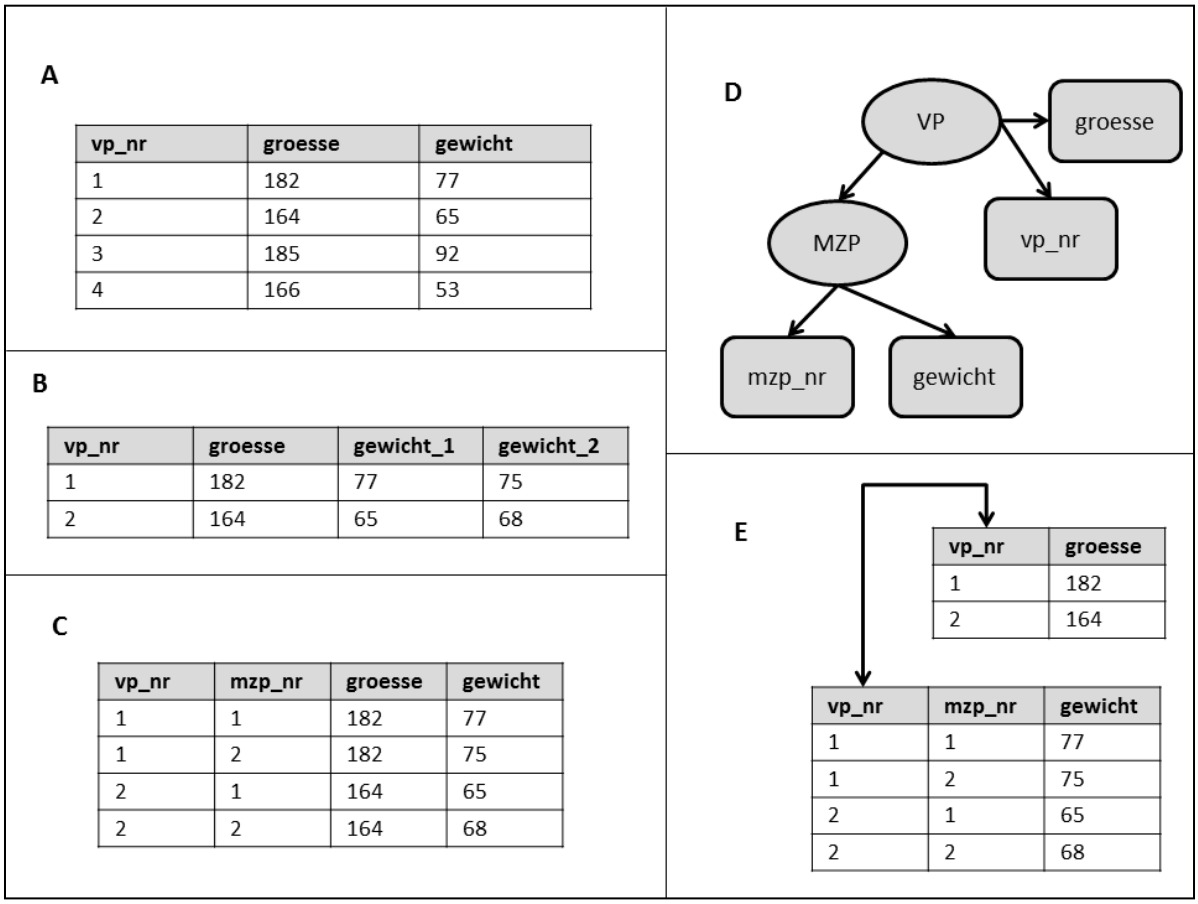


Abbildung 3. Verschiedene Möglichkeiten der Datenorganisation. A: Rechteckige Organisation nicht hierarchisch strukturierter Daten. B: Daten mit zwei Hierarchieebenen (Versuchsperson und Messzeitpunkt) im „wide"-Format. C: Daten mit zwei Hierarchieebenen im „long"-Format. D: Schema für hierarchische Datendatei. E: Daten in relationaler Datenbankstruktur.

Grundlage für die Organisation von Dateien ist die Struktur der Forschungsdaten. Forschungsdaten können in einer rechteckigen Form im „wide“- oder „long“-Format organisiert werden. Alternativ bieten sich hierarchische Organisationsformen an, z.B. in einer XML-Datei oder in einer relationalen Datenbank.

3.2.2.3. Organisation der Dateien und Versionierung

Dateiorganisation

Hat man sich auf eine Form der Datenstruktur festgelegt, die man während dem laufenden Forschungsvorhaben für Erfassung, Bearbeitung und Auswertung der Daten verwenden möchte, kann man darauf basierend ein System zur Organisation und Bearbeitung von Daten, Metadaten, und sonstigen Dokumenten während der Projektphase bestimmen. Dies umfasst z.B. die Bestimmung der Anzahl und des Formats nicht-statischer Dateien (die fortlaufend aktualisiert werden, insbesondere Datensätze und Metadatendateien, u.U. auch sonstige wichtige Dokumente wie Softwarekode oder Datenmanagementplan), Speicherorte der Dateien, ein System zu ihrer Benennung sowie Richtlinien zur Aktualisierung und Versionierung. Dieses Gesamtsystem dient primär dem kurzfristigen Datenmanagement während des laufenden Forschungsvorhabens. Die für Langzeitarchivierung und Data Sharing geeigneten Daten- und Dateiformate unterscheiden sich unter Umständen davon. Im Allgemeinen sollte die spätere Konversion in diese Formate aber keine größeren Probleme bereiten, solange im Rahmen des laufenden Forschungsprojekts eine adäquate Datenqualitätssicherung und Dokumentation bzw. Metadatensammlung implementiert ist. Orientieren kann man sich bei den Entscheidungen zur Erstellung einer sinnvollen Dateiorganisation an den folgenden Kriterien:

- **Integrierbarkeit in die Arbeitsabläufe**: An erster Stelle bietet sich zu diesem Zweck natürlich die Verwendung des nativen Formats der eigenen Statistiksoftware an, da diese neben der Analyse auch Möglichkeiten zur Dateneingabe, -be-arbeitung und -prüfung beinhalten. Zur Dateneingabe und -prüfung sind aber auch Datenbankmanagementsysteme und Tabellenkalkulationsprogramme verwendbar, letztere in begrenztem Umfang auch zur Analyse. Die ganzen Vorteile eines Datenbankmanagementsystems (z.B. automatisierte Datenprüfung, individualisierte Dateneingabemasken) lassen sich vor allem ausschöpfen, wenn genug Zeit und Kenntnisse für die Einrichtung und Pflege einer relationalen Datenbank vorhanden sind.
  Eine Alternative ist es, die Daten als *Plaintext*-Dateien, die lediglich aus den Daten, einer Kopfzeile mit Variablennamen, Zeilenumbrüchen sowie Trennsymbolen zur Trennung der Datenspalten (der sogenannte *Delimiter*, meist ein Komma oder Tabulator) bestehen, zu speichern (z.B. „comma separated value“-Format, csv), in das Statistikprogramm einzulesen und nach der Bearbeitung wieder als Textdatei abzuspeichern. Der Vorteil liegt

insbesondere in der Interoperabilität des Formats: Es wird von allen Statistik- und Tabellenkalkulationsprogrammen „verstanden“ und ist nicht proprietär. Dadurch kann es z.B. die kooperative Arbeit verschiedener Projektpartner erleichtern. Wird von verschiedenen Projektpartnern dieselbe Statistiksoftware auf unterschiedlichen Systemen (Windows, Mac OS, Linux, …) genutzt, bietet sich auch die Verwendung von „portablen“ Dateiformaten der Programme an, z.B. SPSS portable file oder SAS transport file.

Sowohl bei Plaintext- als auch bei den nativen Dateiformaten gilt allerdings, dass sie gar keine oder nur eine begrenzte Menge an Metadaten erfassen können. Ein möglicher „Workaround“ bei Statistikprogrammen wäre, zusätzliche Metadaten als Kommentare in den Syntax-/Skriptdateien zu erfassen.

- **Sparsamkeit**: Die Anzahl der zu pflegenden Dateien bzw. Speicherorte (und auch der analogen Dokumente!) sollte so gering wie möglich sein (aber natürlich so umfangreich wie nötig). Der oben genannte Workaround für Metadaten ist etwa nur sinnvoll bei einer begrenzten Menge von Metadaten. Bei einer größeren Menge sollte man die Metadaten in einen dedizierten Speicherort auslagern (anstatt Metadaten an zwei verschiedenen Stellen zu erfassen). Zur Erstellung der Metadaten kann dann ein eigenes Tabellenblatt in einer Spreadsheet-Arbeitsmappe, ein Texteditor oder ein Textverarbeitungsprogramm wie Word verwendet werden. Bequemer, da stärker vorstrukturiert, ist die Erstellung mit Metadatensoftware, z.B. MyPsychData oder dem Nesstar Publisher.

  Eine Auftrennung der Datendatei in verschiedene Teildatensätze sollte nur dann erfolgen, wenn das etwa aufgrund der Komplexität der Datenstruktur geboten ist.
- **Einfache Instandhaltung und Aktualisierung der Bezüge zwischen Datendateien, Metadatendateien und sonstigen Dokumenten**: Dazu sollten alle relevanten Dateien möglichst an einem zentralen Ort zusammen aufbewahrt werden, z.B. einem Projektordner, der für alle Projektbeteiligten zugänglich ist (z.B. über Netzwerklaufwerke, „Cloud“-Dienste). Eine readme-Datei in diesem Ordner kann das System der Datenaktualisierung erläutern (inklusive des Vorgehens bei der Versionierung der Dateien, siehe unten), wichtige nicht-digitale Dokumente auflisten und wichtige Bezüge erläutern (d.h. welche Dateien auf welche anderen Dateien verweisen und ggf. aktualisiert werden müssen, wenn diese anderen Dateien verändert werden). Steht ein Dokumentenscanner zur Verfügung, können „analoge“ Dokumente digitalisiert und dem Projektordner hinzugefügt werden.

Eine konsistente Systematik zur Benennung von Dateien und Ordnern hilft, die Dateiorganisation aufrechtzuerhalten. Dateinamen können z.B. Kürzel beinhalten, die das Forschungsprojekt, Bearbeiter, Dateityp, Versionsnummer, Dateistatus oder das Datum der letzten Änderung unmittelbar kenntlich machen. Die Kürzel sollten dann in der Projektordner-readme-Datei oder an

einem anderen Ort zugänglich sein. Eine andere Möglichkeit ist, eng in Zusammenhang stehenden Dateien Namen zu geben, die bis auf die Dateinamenerweiterung (z.B. „sav“ bei SPSS-Datendateien und „sps“ bei Syntaxdateien) identisch sind. Die Datei- und Ordnernamen sollten nicht zu lang sein (man sollte es also mit Kürzeln auch nicht übertreiben, vor allem nicht einfach immer weitere Kürzel anhängen – die gefürchtete „manuskript_v8_ueberarb_revHB_revGS_-FINAL_2013-05-08.docx“) und keine Sonderzeichen und Leerzeichen enthalten, da dies für Ärger mit manchen Betriebssystemen oder Anwendungen sorgen kann. Vor allem aber sollten die Namen der Dateien und Ordner eines einmal etablierten Dateinamenssystems nicht willkürlich geändert werden, da in anderen Dateien unter Umständen auf die Pfade dieser Dateien verwiesen wird und diese Verweise dann ins Leere führen würden.

Ein System zur Dateiorganisation sollte die Bestimmung der Anzahl und des Formats der nicht-statischen Dateien, deren Speicherorte, ein System zur Benennung der Dateien sowie ein Versionierungssystem beinhalten.

Die Benennung von Dateien sollte einheitlich erfolgen, mit nicht zu langen Namen ohne Sonder- oder Leerzeichen.

Versionierung

Bei nicht-statischen Dateien ist in der Regel eine Versionierung der Dateien empfehlenswert, das heißt, Änderungen nicht einfach in der alten Datei abzuspeichern, sondern als neue Kopie mit einer neuen Versionsnummer. Die alten Versionen können z.B. in einem Unterordner des Projektordners aufbewahrt werden. Versionierung erlaubt es, bei Aktualisierungen, die sich im Nachhinein als problematisch herausstellen, auf eine frühere Form der Datei zurückzugreifen. Außerdem kann der Entstehungsprozess der Datei, z.B. die Bearbeitungsschritte einer Datendatei bei der Qualitätskontrolle, nachvollzogen werden. Die Versionierung von Dateien sollte einer gewissen Systematik unterliegen. Es sollte zumindest festgelegt sein, bei welchen Änderungen eine neue Version der Datei erstellt werden muss. Je nachdem, wie komplex die Organisation des Arbeitsumfeldes ist (Anzahl der Datenbearbeiter, Speicherorte, Projektpartner, Dateien und Dokumente, …), können die folgenden **Versionierungsmaßnahmen** verwendet werden:

- Die Definition besonderer „**Meilensteine**“ bei der Erstellung der Dateien. Im Falle einer Datendatei kann dies z.B. die Eingabe aller erhobenen Daten, der Abschluss der Integration verschiedener Datenquellen, der Abschluss der Qualitäts- /Konsistenzprüfung des Datensatzes oder die Hinzufügung aller abgeleiteten Dateien sein. Bei einer Metadatendatei wäre als erster Meilenstein die Erstellung aller Metadaten, die bereits vor der Datenerhebung verfügbar sind, denkbar, danach Meilensteine parallel zu denen der Datendatei. Bei Erreichen eines solchen Meilensteins sollte eine gesonderte Meilenstein-Version der Datei („Master“-Datei) erstellt werden, die mit einem Schreibschutz versehen und in einem interoperablen Format (z.B. csv, xml) gespeichert wird. Eine zusätzli-

che Sicherungsmaßnahme ist die Generierung einer *Prüfsumme* bei Erstellung der Meilensteindatei, die zusammen mit ihr abgespeichert wird. Die Prüfsumme ist einfach eine Zahl oder Zeichenkette, die anhand eines Algorithmus aus dem Inhalt einer Datei erstellt wird. Ändert sich der Inhalt der Datei, ändert sich auch die Prüfsumme, die daraus entsteht. Durch Abgleich der ursprünglich angelegten Prüfsumme und einer neu generierten lässt sich prüfen, ob die Datei zwischenzeitlich verändert wurde. Ein verbreiteter Algorithmus zur Erstellung einer Prüfsumme ist der *MD5*-Algorithmus.

- Die Verwendung von **Sub-Versionen** (z.B. „Version 1.08"): Sub-Versionen kennzeichnen inkrementelle, z.B. innerhalb eines Arbeitstages vorgenommene Änderungen, Hauptversionen dagegen Meilensteine oder besonders wichtige Aktualisierungen.
- Eine **Vorschrift dazu, für welche anderen Dateien bei Erstellung einer neuen Version einer bestimmten Datei ebenfalls eine neue, aktualisierte Version erzeugt werden muss**. Dies kann z.B. in einer readme-Datei im Projektordner abgelegt sein.
- Die Festsetzung bestimmter **Termine, zu denen eine ausführliche Prüfung und gegebenenfalls Harmonisierung der Dateien vorgenommen wird** (z.B. bevor das Erreichen eines Meilensteins „abgesegnet" wird).
- Das **Führen eines separates *Changelog*** zu einer Datei, in dem aufgelistet ist, welche Änderungen von älteren zu neueren Versionen gemacht wurden
- Die Verwendung von **Programmfunktionen oder spezialisierter Software zur kollaborativen Bearbeitung von Dokumenten, zum Versionsmanagement und zur Synchronisierung** von Ordnerinhalten, oder die Verwendung eines Webspeichers oder „Cloud"-Dienstes (bei letzteren ist in besonderem Maß auf Datensicherheit und Datenschutz zu achten, z.B. durch Verwendung von VPN-Software, verschlüsselter Datenübertragung und einem vertrauenswürdigem Host; MyPsychData bietet die Möglichkeit einer webbasierten gemeinschaftlichen Bearbeitung von Daten und Metadaten, sofern diese anonymisiert wurden; siehe Kapitel 4.2.1).
- Die regelmäßige **Anfertigung von Sicherungskopien und Einrichtung eines Rechte- und Zugangskontrollsystems** (siehe den entsprechenden Unterabschnitt dieses Kapitels).

Die Pflege statischer Dokumente, die im Lauf des Forschungsprojekts unverändert bleiben (z.B. Förderrichtliniendokumente, Hintergrundliteratur, Fragebögen) ist naturgemäß weniger aufwändig, sollte aber nicht vergessen werden. Insbesondere sollten auch hier die Dateinamen nicht willkürlich geändert werden, bzw. auf die Instandhaltung der Bezüge geachtet werden (z.B. dass in einem Metadatendokument, in dem Informationen zu einem Fragebogen hinterlegt sind, der richtige Dateipfad zu diesem Fragebogen verzeichnet ist). Die Speicherung in bzw. Konver-

tierung zu nichtveränderlichen und durchsuchbaren Formaten wie PDF bietet sich an, um versehentlichen Änderungen vorzubeugen und schnell die gewünschte Information zu finden.

Als Versionierungsmaßnahmen können die Definition von „Meilensteinen“ und ihrer Terminierung, die Verwendung von Subversionen, das Führen eines *Changelog*, die regelmäßige Anfertigung von Sicherheitskopien oder die Verwendung spezieller Programme erfolgen.

Es ist sinnvoll, sich eine Übersicht über die aufeinander bezogenen Dateien zu erstellen sowie eine Vorschrift für deren regelmäßige Überprüfung und Aktualisierung.

Statische Dateien sollten in nicht veränderbaren Formaten gespeichert werden.

3.2.2.4. Qualitätssicherung von Daten und Metadaten

Mit Qualitätssicherung sind in diesem Kontext alle Maßnahmen gemeint, die helfen sollen, den Informationswert der generierten Daten und Metadaten (Validität, Integrität, Konsistenz) zu steigern. Realistisch betrachtet sind Fehler bei Daten- und Metadatenerhebung, -eingabe und -verarbeitung unvermeidbar, können aber durch Qualitätssicherungsmaßnahmen verringert oder besser entdeckt und korrigiert werden.

Vor der Datenerhebung

In diesem Sinne findet Qualitätssicherung bezüglich der Datenerhebung bereits beim **Design der Erhebungsprozedur und der Erhebungsinstrumente** statt. Je nach Forschungsbereich kann die Erhebungsprozedur durch ein Ablaufprotokoll mehr oder weniger stark a priori festgeschrieben werden und damit Störeinflüsse ausgeschaltet und die Objektivität erhöht werden. Fragebögen sollten möglichst so gestaltet werden, dass die Fragen, Instruktionen und Antwortoptionen für die Probanden nicht missverständlich oder verwirrend sind (z.B. keine doppelten Verneinungen, Abkürzungen; keine Umkehrung der Skalierungsrichtung von einem Itemblock zum nächsten). Messinstrumente sollten vor ihrer Nutzung kalibriert werden. Ein gutes Design des Versuchs und der Erhebungsinstrumente als grundlegendste Voraussetzung für die Qualitätssicherung der Daten bedarf aber einer eigenständigen Einführung und wird daher hier nicht weiter behandelt (siehe z.B. Bortz & Döring, 2002; Bühner, 2011; Moosbrugger & Kelava, 2012; Sedlmeier & Renkewitz, 2008). Auch die Qualität von Metadaten kann erhöht werden, indem a priori sorgfältig überlegt und in der Gruppe diskutiert wird, welche Metadatenelemente auf welche Weise erhoben werden sollen (bei Datenerhebungsprozessen durch den Computer, z.B. Online-Befragungen oder mit Experimentalsoftware, können bestimmte Metadaten auch automatisiert miterhoben werden).

Maßnahmen zur Qualitätssicherung der Daten (Validität, Integrität, Konsistenz) sollten vor, während und nach der Datenerhebung stattfinden.

Eine gründliche Planung des Versuchsdesigns und der Erhebung bildet eine wichtige Grundlage für die Qualitätssicherung der Daten.

Während der Datenerhebung und -eingabe

Eine **Datenqualitätssicherung während einer Erhebungssitzung** wird durch Verwendung computerisierter Erhebungsinstrumente und Hilfsmittel möglich, speziell bei Fragebogen- und Interviewmethoden. In den Sozialwissenschaften werden diese oft als „computer-assisted … interviewing" bezeichnet, z.B. „CATI" (telephone), „CAPI" (personal), „CAWI" (web), oder „CASI" (self) – Software. Die letzten beiden bezeichnen selbstausgefüllte Fragebögen. Direkt während der Bearbeitung des Fragebogens prüft die verwendete Erhebungssoftware, ob eine Eingabe innerhalb des erlaubten Wertebereichs liegt, füllt Informationen, die durch frühere Angaben impliziert werden, automatisch aus, springt bei Verzweigungen im Fragebogen an die richtige Stelle (z.B. Überspringen der Fragen zu Kindern nach Angabe der Kinderlosigkeit) und graut „unmögliche" Antwortoptionen aus.

Liegen die Daten der Probanden in einer nicht direkt speicherbaren Form vor, zum Beispiel bei einer Kartensortieraufgabe, sollte Arbeit in die Entwicklung einer durchdachten Erfassungsstrategie investiert werden. Ein strukturierter Prozess reduziert die Menge der sich bei der Datenerhebung einschleichenden Fehler. Im Falle einer Sortieraufgabe von Nationalitäten zum Beispiel könnte der Versuchsleiter bei jedem Durchlauf die Namen aller Nationalitäten aufschreiben. Besser wäre es aber, die Karten auf der Rückseite mit Zahlen oder den entsprechenden Länderkürzeln zu versehen und in ein vorgefertigtes Gruppierungsraster einzutragen.

Bei den Schritten zur **Vorbereitung der Dateneingabe** liegt eine erste Qualitätssicherungsmaßnahme in der guten Auswahl eines **Schemas zur Benennung der Variablen des Datensatzes.** Dies ist insbesondere bei einer großen Menge von Variablen relevant, wie sie z.B. bei umfangreichen Fragebögen oder bei Reaktionszeitexperimenten mit Hunderten von Trials entstehen, da sonst bei der Dateneingabe und -verarbeitung möglicherweise folgenschwere Verwechslungen entstehen könnten. Mögliche Namensschemata sind etwa:

- **Bezeichnung mit Laufnummer**: Die Variablen werden einfach ohne Berücksichtigung ihres Inhalts durchnummeriert, z.B. „V001", „V002", …. Dies ist oft die Voreinstellung in Statistikprogrammen. Die Beibehaltung dieses Systems ist im Allgemeinen aber nicht empfehlenswert, da die Namen nichtssagend und leicht verwechselbar sind und daher Eingabe- und andere Fehler im Lauf der Datenbearbeitung begünstigen.
- **Übernahme der Item-/Fragennummer**: Haben die Items des erhobenen Stimulusmaterials bereits irgendeine Form von Nummerierung oder Bezeichnung, kann diese ein-

fach für die Variablen übernommen werden, z.B. bei den Fragen eines Fragebogens. Dadurch ist eine einfache Zuordnung der Variablen zu der Erhebungsprozedur möglich, oft sagen die Namen aber nichts über den Inhalt aus und sind leicht verwechselbar (z.B. „Q_01a“, „Q_01b“, „Q_02“, …).

- **Mnemonische Bezeichner**: Um die Repräsentation für die Bearbeitenden zu erleichtern, können auch kurze, die Natur der Variable verdeutlichende Bezeichner gewählt werden. Diese sind jedoch nicht unbedingt für nachträglich hinzukommende Personen genauso nahe liegend wie für die ursprünglichen Namensgeber. Beschränkungen in der Länge der Variablennamen können es außerdem dazu führen, dass kein Platz für ausreichend deutliche Bezeichner ist, insbesondere bei Variablen, die einen verwandten Inhalt abbilden (z.B. „ZUFRIED_ALLG“, „ZUFRIED_FAM“, „GLUECK“, …).
- **Präfix-Stamm-Suffix-Kombinationen**: Hier wird jeder Variablengruppe ein Stamm zugewiesen, beispielsweise ZU für Zufriedenheit. Der einzelnen Variable wird dann als Präfix die jeweilige Spezifizierung vorgestellt, wenn es sich also um die Zufriedenheit mit der Familie handelt, wäre das FAM_ZU. Ein Suffix bietet sich besonders zur Kennzeichnung verschiedener Erhebungszeitpunkte oder für abgeleitete Variablen an (z.B.: „FAM_ZU_1“, „FAM_ZU_2“, „ARB_ZU_1“, …).

Die oben erwähnte Beschränkung der Variablennamenslänge gilt für die meisten modernen Statistikprogramme nicht mehr (zumindest nicht die sehr restriktive Beschränkung auf acht Zeichen). Dennoch ist es sinnvoll, eine gewisse „Disziplin“ bei der Namensgebung zu üben, da eine größere Menge sehr langer, sich zum Teil nur in wenigen Zeichen unterscheidender Namen schnell unübersichtlich werden kann. Bei einer späteren Veröffentlichung des Datensatzes ist außerdem oft eine Beschränkung auf acht Zeichen nötig, da die Archive in der Regel eine breite Nachnutzbarkeit durch verschiedenste Arten von Programmen anstreben. Damit diese Zeichenbegrenzung auch bei später hinzugefügten abgeleiteten Variablen eingehalten werden kann, sollte die Länge wichtiger Variablen außerdem auf sechs bis sieben Zeichen beschränkt sein (z.B. um den Zusatz „_Z“ für die z-standardisierte Version anhängen zu können).

Hat man Audio- oder Videodaten als Rohdaten erfasst, ist oft eine Transkription des Gesagten nötig. Transkription ist eine anspruchsvolle Aufgabe und sollte durch ein Transkriptionsmanual und eine vorstrukturierte Transkriptvorlagendatei, die beispielsweise schon Fragen nach auszufüllenden Hintergrundinformationen enthält, unterstützt werden. Im Transkriptionsmanual sollte zum Beispiel spezifiziert werden, welchen Detailgrad das Transkript erreichen soll, beziehungsweise welche Äußerungen wie festgehalten werden sollen. Bei hohem Detailgrad sind mehrere Transkriptionsdurchgänge („passes“), bei denen jeweils auf unterschiedliche Aspekte geachtet oder auf Fehler geprüft wird, eine Option. Weitere zu beachtende Aspekte sind z.B. die

Nutzung einheitlicher Bezeichner für die beteiligten Personen und die Markierung von Informationen, die gegebenenfalls einer Anonymisierung bedürfen.

Vor der Eingabe der Primärdaten ist außerdem womöglich eine Kodierung der Rohdaten, z.B. von offenen Antworten, nötig. Dies kann je nach Komplexität der Rohdaten eine einfache bis kognitiv komplexe, zeitaufwändige Tätigkeit sein. Leitlinien für die **Qualitätssicherung beim Kodieren** sind beispielsweise die Folgenden:

- Die **Kategorien sollten präzise definiert sein, einander wechselseitig ausschließen, und den interessierenden Phänomenbereich erschöpfend abdecken** (ein System nichtexklusiver Kategorien ist auch möglich, aber wesentlich komplexer, da für jede Kategorie einzeln geprüft werden muss, ob sie auf die Rohdaten zutrifft, und daher auch mehrere Variablen erzeugt werden müssen).
- Die **Originalinformation sollte, wann immer möglich, bewahrt werden**. Originaldaten können immer in Kategorien umgewandelt werden, der umgekehrte Schritt ist jedoch nicht möglich. Gerade unter dem Aspekt der Nachnutzung ist dies bedeutsam; Nachnutzer möchten womöglich ein eigenes, ihrer Fragestellung angemessenes Kodierungsschema verwenden. Will man etwa lediglich die Alterskategorien „Kind“, „Jugendlicher“, und „Erwachsener“ verwenden, sollte das genaue Alter dennoch in den Datensatz aufgenommen und die Kategorialvariable separat daraus abgeleitet werden. Auch Freitextantworten können, sofern nicht zu umfangreich, (gegebenenfalls anonymisiert) direkt in den Datensatz aufgenommen werden.
- **Fragen mit festgelegten Antwortmöglichkeiten sollten diese auch eins zu eins als Kodiermöglichkeiten abbilden**, um Verwirrung zu vermeiden. Wenn auf dem Fragebogen steht: „1 = ich stimme zu“, dann sollte auch der Kode „1“ „ich stimme zu“ abbilden. Eine Umpolung eines Items stellt einen Transformationsschritt dar, der dokumentiert werden sollte.
- Die **Arbeitsschritte der Kodierung und der Dateneingabe sollten voneinander getrennt sein**, um keine kognitiven Kapazitäten für „Task Switching“ zu verbrauchen. Dies gilt insbesondere für komplexe Kodieraufgaben.
- **Wenn eine algorithmische Kodierung möglich ist, sollte sie per Softwareskript erfolgen**, und das Skript als Teil der Dokumentation hinterlegt werden. Dadurch werden Kodierfehler vermieden, zumindest solange das Skript fehlerfrei ist. Aber auch wenn nicht, lässt sich dieser Fehler nachvollziehen, das Skript korrigieren und die Kodierung problemlos wiederholen.
- **Komplexe Kodieraufgaben sollten auf Grundlage eines Kodiermanuals durch zuvor trainierte Mitarbeiter vorgenommen werden**. Wenn auf Grundlage der Rohdaten mehrere Variablen kodiert werden sollen, sollten gegebenenfalls mehrere Kodierungs-

durchläufe erfolgen, in denen man die Aufmerksamkeit auf die jeweils relevanten Aspekte fokussiert. Eine Mehrfachkodierung durch unabhängige Rater und anschließende Konsistenzprüfung ist ein weiteres Mittel.

Je nach Erhebungsprozedur liegen die Daten direkt in digitaler Form vor, oder müssen zunächst manuell eingegeben oder eingescannt werden. **Qualitätssicherungsmechanismen direkt während der Dateneingabe** sind insbesondere:

- Die **Verwendung von Dateneingabemasken**, die so strukturiert sein können, dass das Risiko von Eingabefehlern gegenüber einer direkten Eingabe in eine Datentabelle verringert ist (wenn etwa das Eingabefeld aus Versehen verrutscht, oder man Daten in die falsche Spalte eingibt, weil die Variablennamen sehr ähnlich sind). Solche Masken lassen sich innerhalb einer Datenbankmanagementsoftware, aber auch in Tabellenkalkulationsprogrammen wie Excel oder Open Office erstellen. Eine andere Möglichkeit ist eine HTML-/browserbasierte Eingabemaske (die in der Regel die Daten an ein Datenbankmanagementsystem übergibt).
- Die **automatische Prüfung der eingegeben Daten durch den Computer**, z.B. auf fehlende oder nicht mögliche Eingaben, bei Eingabemasken auch das Ausgrauen von unmöglichen Antwortoptionen, das Überspringen nicht anwendbarer Variablen etc.; also dieselben Prüfmechanismen, die bei Verwendung von computerbasierten Erhebungsverfahren auch direkt bei der Erhebung möglich sind. Anders als bei diesen kann es aber unter Umständen sinnvoll sein, auch die Eingabe „unmöglicher“ Werte zuzulassen, etwa wenn solche Angaben tatsächlich von der Versuchsperson in einem Fragebogen gemacht worden sind. In diesem Fall ist es eventuell sinnvoller, erst einmal alle Daten einzugeben und anschließend auf seltsame Werte zu prüfen, die dann per Nachfrage beim Interviewer oder der Versuchsperson geklärt werden können.

Während der Datenerhebung können automatisierte Erhebungsinstrumente oder gut durchdachte Erfassungsstrategien die Datenqualität garantieren.

Bei der Dateneingabe ist auf ein sinnvolles, eineindeutiges Benennungsschema für die Variablen zu achten sowie ein gut vorstrukturiertes Transkriptions- oder Kodierungsmanual bei Audio-/Video-Daten oder bei offenem Antwortformat.
Außerdem kann während der Dateneingabe eine Qualitätssicherung durch die Verwendung von Dateneingabemasken oder eine automatische Prüfung durch den Computer erfolgen.

Nach der Datenerhebung

Trotz aller Schritte zur Fehlervermeidung bei Datenerhebung und -eingabe sind **Qualitätskontrollen nach der Dateneingabe** unerlässlich. Dazu gehören z.B. die Folgenden:

- **Mehrfacheingabe der Daten durch unabhängige Personen und anschließende Prüfung auf Konsistenz** (per Softwareskript, Tabellenkalkulationsprogramm, o.ä.). Dies ist eine aufwändige, aber sehr effektive Maßnahme insbesondere zur Reduzierung mehr oder weniger „zufällig" auftretender Fehler, wie z.B. Tippfehler.
- **Manuelle Durchsicht der Daten**: Um von Anfang an auftretende, systematische Fehler direkt zu erkennen, sollten die zuerst eingegebenen Fälle vollständig, anschließend dann zufallsstichprobenartig einzelne Fälle geprüft werden.
- **Geskriptete Prüfung der Daten**: Eine algorithmische Prüfung kann etwa auf unmögliche Werte (Werte außerhalb des möglichen Wertebereichs) oder Wertekombinationen (schwangere Männer, Kinder, die älter sind als ihre Eltern) erfolgen sowie auf „verdächtige" Werte (z.B. Arbeitslose mit hohem Einkommen); bei letzterem sind alle möglichen Heuristiken denkbar, die natürlich stark vom jeweiligen Untersuchungsgegenstand abhängen. Generell sollte man versuchen, das vorhandene Wissen zum Themenbereich zur Konstruktion solcher Prüfungsskripte zu nutzen. Die Prüfungsskripte sollten in jedem Fall gespeichert und dokumentiert werden.
- **Prüfung anhand von Deskriptivstatistiken**: Unmögliche Werte oder eine hohe Anzahl fehlender Werte können auch durch Häufigkeitstabellen oder -diagramme direkt aufgedeckt werden, unmögliche oder verdächtige Wertekombinationen durch Kreuztabellen. Als Indizien können auch Verteilungsformen oder Mittelwerte genutzt werden, die stark von der ungefähren Erwartung abweichen.
- **Prüfung der Identifikatorvariablen**: Wenn mehrere Datensätze vorhanden sind, die durch Identifikatorvariablen (z.B. Versuchspersonenkodes) miteinander in Beziehung gesetzt werden können, beispielsweise bei den Erhebungswellen einer Längsschnittstudie, sollte unbedingt geprüft werden, ob die Fallzahlen unter Berücksichtigung von Dropouts und die Identifikationskodes zwischen den Datensätzen übereinstimmen.

Fehlende Werte verdienen bei der Datenerhebung besondere Berücksichtigung. Zunächst ist es wichtig, einen eigenen Wert, der außerhalb des Bereichs gültiger Variablenwerte liegt, für fehlende Werte zu reservieren. Da sich die gültigen Wertebereiche von Variable zu Variable unterscheiden, können sich die Fehlerkodes unterscheiden, insbesondere zwischen numerischen und alphanumerischen Variablen (für numerische Variablen sind Werte üblich wie „9", „99", oder negative Werte wie „-1"; bei alphanumerischen Variablen können Sonderzeichen wie „#" verwendet werden). Dort wo es sinnvoll ist, sollten aber identische Kodes für fehlende Werte verwendet werden. Die Datenzelle sollte nicht einfach leer gelassen werden, da daraus nicht eindeutig hervorgeht, ob es sich um einen fehlenden Wert handelt, oder die Eingabe einfach vergessen wurde. Lassen sich verschiedene Gründe für „Missingness" unterscheiden, sollte diese Information kodiert werden, indem je eigene Werte vergeben werden, z.B.: „96: fehlend – Frage

nicht zutreffend“, „97: fehlend – Vp weiß keine Antwort“, „98: fehlend – Vp will keine Angabe machen“, „99: fehlend – unklarer Grund“.

Die Inspektion einer „Missingness-Matrix“, die jedem Datenpunkt einen Wert von 0 (Wert vorhanden) oder 1 (fehlender Wert) zuweist, kann Rückschlüsse auf den Mechanismus erlauben, der dazu geführt hat, dass ein Wert fehlt. Fehlen bei einem Online-Fragebogen etwa alle Werte ab einer bestimmten Fragebogenseite, hat die Versuchsperson die Bearbeitung vermutlich abgebrochen, was etwa durch fehlende Zeit oder Motivation begründet sein könnte.

Es existiert eine Reihe von Techniken zur *Imputation* fehlender Werte, also der Ersetzung fehlender Werte durch Schätzungen basierend auf den vorhandenen Informationen. Wird eine Imputationsmethode angewendet, sollte diese in der Studiendokumentation beschrieben sein, und der Datensatz vor Anwendung des Imputationsverfahrens sollte (z.B. in Form einer Master-Datei) hinterlegt sein.

Abgeleitete Variablen, die nach der Dateneingabe aus den Ursprungsdaten berechnet werden, sollten ebenso einer Prüfung unterzogen werden, wobei die oben beschriebenen Prüfverfahren verwendet werden können. Die Berechnung abgeleiteter Variablen sollte per Computerskript erfolgen und der Programmkode dokumentiert werden. Viele der hier beschriebenen Prüfmethoden lassen sich im Übrigen auch auf Metadaten anwenden, insbesondere wenn diese stark strukturiert sind. Die Verwendung von Metadatenerzeugungstools wie MyPsychData, Colectica und Nesstar Publisher hilft bei der Qualitätssicherung durch Eingabemasken und Werteprüfungen. Falls die Metadaten im XML-Format vorliegen, lässt sich die XML-Datei auch durch einen XML-Editor daraufhin prüfen, ob sie mit der Schemadatei bzw. der DTD konform ist (*XML-Validierung*). Bei unstrukturierten Metadaten und sonstigen Dokumenten ist die Qualitätskontrolle aufwändiger; hier könnte eine zufallsstichprobenartige Prüfung angewendet werden, beispielsweise im Rahmen eines Reviews der Daten und Dokumente vor Erreichung eines Meilensteins.

Alle der zuvor aufgeführten Maßnahmen zur Qualitätssicherung können und sollten im Rahmen von „**Pilotstudien**“ auf ihre Praktikabilität und Fehlerfreiheit getestet werden, etwa indem ein Online-Fragebogen von Kollegen ausgefüllt wird, einige „Dummy-Datensätze“ in die Datenbank eingegeben werden oder ein Kodierungsschema auf Testdaten angewendet wird. Denn auch bei der sorgfältigst überlegten Datenerfassungsprozedur wird die eine oder andere Schwierigkeit oder Widersprüchlichkeit unbedacht geblieben sein.

Nach der Dateneingabe sollten Qualitätskontrollen wie Mehrfacheingabe der Daten, Überprüfung anhand von Deskriptivstatistiken oder geskriptete Prüfungen durchgeführt werden. Besonderes Augenmerk sollte auf die Dokumentation fehlender und abgeleiteter (berechneter) Variablen gelegt werden.

3.2.2.5. Sicherheitskopien, Datenintegrität, Zugriffskontrolle und Aufbewahrung

In Forschungsdaten und die dazugehörige Dokumentation wird viel Arbeit investiert, manchmal über Jahrzehnte hinweg. Es ist es also wert, sich Gedanken über ihre Sicherung zu machen. Schäden, Datenverlust oder unerwünschte Veränderungen an Daten und Dokumenten können etwa entstehen durch Hardwareversagen, Software-Bugs, Computerviren, Stromausfälle oder durch Missgeschicke bei Erhebung und Verarbeitung.

Grundlegend ist zunächst die Festlegung von Backup-Prozeduren, also einer Systematik, von welchen Dateien bzw. Dokumenten wann wohin Sicherheitskopien angefertigt werden sollen. Bei wichtigen Dateien, die das Ergebnis umfangreicher Arbeiten sind, auf die aber nicht unbedingt häufig zurückgegriffen wird bzw. die nicht mehr verändert werden, wie zum Beispiel Master-Datensätze, sollte das Backup-System besonders robust sein, also gegen alle Eventualitäten abgesichert sein. Bei Arbeitsversionen von Dateien, die laufend verändert werden, sollte eine besonders regelmäßige Sicherungskopie erfolgen.

Es ist eine ganze Reihe verschiedener **Backup-Mechanismen** denkbar, die auf digitale Daten und Dokumente anwendbar sind:

- **Regelmäßige automatisierte Anfertigung von Sicherheitskopien** durch das Betriebssystem („batch job“). Dies ist eine in jedem Fall empfehlenswerte grundlegende Sicherungsmaßnahme. Eine umfassendere Variante davon ist die regelmäßige Anfertigung eines Systemabbildes, der auch die Wiederherstellung des Betriebssystemzustands als Ganzem erlaubt.
- **Verteilte Speicherung**: Die Anfertigung mehrerer Sicherungskopien an verschiedenen, auch geographisch getrennten Orten. Dies schützt vor größeren Unglücken wie einem Gebäudebrand. Bei Vorliegen mehrerer Sicherungskopien können diese außerdem wechselseitig auf Integrität geprüft werden, indem jeweils eine eigene Prüfsumme erstellt wird.
- **Verwendung verschiedener Speichermedien** (optische Speicher wie CD und DVD, externe Festplatten, Magnetbänder): Dies verschafft zusätzlichen Schutz gegen versehentliches Überschreiben und Probleme, die nur bestimmte Speichermedien betreffen.
- **Verwendung eines „Journaling File System“**: Ein besonders sicheres Dateisystem, das jegliche Dateiänderungen in einem reservierten Speicherbereich (dem „Journal“) aufzeichnet, so dass noch nicht gespeicherte Änderungen selbst bei Systemabstürzen und Stromausfällen wiederhergestellt werden können.

Institutionelle Rechenzentren bieten in der Regel bestimmte Sicherungsmechanismen und Speicherinfrastruktur an oder können bei der Implementierung solcher Mechanismen unterstützen. Gerade bei vertraulichen Daten muss aber auch eine Abwägung zwischen Datensicherung und

Datenschutz getroffen werden; man sollte die Daten dann jedenfalls nicht wahllos auf die verschiedensten Speicherorte verteilen.

Bei vertraulichen oder anderweitig schutzwürdigen Daten sollte also auch verstärkt auf **Maßnahmen zur Zugriffskontrolle** geachtet werden. Diese lassen sich bei digitalen Dokumenten grob in zwei Kategorien einteilen:

- **Physische Zugriffskontrollen**: Darunter fällt die Verwahrung von Datenträgern in Räumlichkeiten, zu denen nur Befugte Zutritt haben, die Aufzeichnung dessen, wer die Räumlichkeiten wann betreten und verlassen hat und wer wann welche Datenträger mitgenommen bzw. wieder zurückgebracht hat sowie der Verzicht auf unnötige Transporte der Datenträger.
- **Informationstechnische Zugriffskontrollen**: Keine Speicherung der Dateien in externen Netzen, Installation und regelmäßiges Update von Antiviren- und Firewall-Software, die Einrichtung von System- und Dateipasswörtern sowie eines Rechtemanagementsystems (je nach Rolle des Benutzers z.B. kein Zugriff, nur Lesezugriff, Lese- und Schreibzugriff, oder auch das Recht, Zugriffsrechte zu ändern), Verschlüsselung von Dateien (insbesondere vor dem Versenden über Computernetze).

Eine weitere effektive Maßnahme zum Schutz vertraulicher Daten besteht schließlich darin, alle Personen, die Zugang zu den Daten haben, zuvor Vertraulichkeitserklärungen unterschreiben zu lassen.

Als letzter Aspekt des Datenmanagements während des laufenden Forschungsvorhabens gilt es, Maßnahmen zur mittel- bis längerfristigen Lagerung der Datenträger zu treffen:

- Die Auswahl eines geeigneten Lagerungsortes. Die Räumlichkeiten sollten idealerweise kühl, trocken, einsturz-, überflutungs- und brandsicher sein.
- Alle zwei bis fünf Jahre sollte eine Migration auf neue Daten erfolgen, da sowohl bei optischen als auch magnetischen Datenträgern Degradation stattfinden kann.
- Die Datenintegrität sollte, etwa anhand von Prüfsummen, in regelmäßigen Abständen getestet werden.
- Die zur Speicherung gewählten Dateiformate sollten nicht proprietär und auf absehbare Zeit weiter in Verwendung sein (s. Kapitel 3.2.3).
- Die Datenträger sollten gemäß einer langfristig nachvollziehbaren Systematik geordnet und beschriftet sein.

Alle in diesem Abschnitt beschriebenen Maßnahmen zur Anfertigung von Sicherungskopien, Zugriffskontrollen und zur langfristigen Aufbewahrung sollten selbstverständlich in analoger Weise auch bei wichtigen nicht-digitalen Dokumenten Anwendung finden. Zur zusätzlichen Sicherung bietet sich auch eine Digitalisierung der Dokumente an.

Die Sicherung von Daten und Dokumenten sollte robust sein und regelmäßig erfolgen. Die genaue Festlegung von Back-up-Prozeduren bietet sich an, beispielsweise die automatisierte Erstellung von Sicherheitskopien, verteilte räumliche Speicherung, Verwendung verschiedener Speichermedien oder ein „Journal File System".

Dabei sollte gleichzeitig auf Maßnahmen zum Datenschutz wie Zugriffskontrollen oder Verfassung von Vertraulichkeitserklärungen geachtet werden.

Zur mittel- und langfristigen Sicherung können optimalerweise bereits ein geeigneter Lagerort gewählt, Datenmigrationen und Prüfungen der Datenintegrität durchgeführt und nicht-proprietäre Datenformate verwendet werden.

3.2.3 Nach der Datenerhebung und -analyse

Die wichtigsten nach Abschluss der „Kernarbeiten" des Forschungsprojekts anstehenden Tätigkeiten sind die Langzeitarchivierung der Forschungsdaten und Data Sharing. Langzeitarchivierung von Forschungsdaten ist ein umfangreiches Tätigkeitsfeld der Informationswissenschaft. Hier sollen primär hilfreiche Hinweise für Forschende gegeben werden. Für umfassendere Informationen zum Thema siehe die Homepages des *Kompetenznetzwerk digitale Langzeitarchivierung (nestor)* [25] und des DFG-Projekts *KoLaWiss* [26], sowie die von Altenhöner und Oellers (2012), Neuroth (2012), und Neuroth et al. (2010) herausgegebenen Handbücher.

3.2.3.1. Vorbereitende Maßnahmen für die Langzeitarchivierung

In Kapitel 3.2.1 wurde bereits erwähnt, dass die Richtlinien der DFG zur Sicherung guter wissenschaftlicher Praxis (DFG, 1998) eine Aufbewahrung der Primärdaten für mindestens zehn Jahre fordern. Wie Klump (2011, S. 117) aber anmerkt, ist der „kritischste Moment" im Lebenszyklus von Forschungsdaten, „wenn das Projekt endet, denn hier endet meistens auch die Finanzierung weiterer Maßnahmen zur Datenerhaltung und das Interesse der Forscher ist bereits auf das nächste Projekt gerichtet". Diese Problematik kann in dem Maß abgemildert werden, in dem bereits während des Projekts solides Datenmanagement und Dokumentation stattgefunden haben. Dennoch sollten geplante Maßnahmen zur *Langzeitarchivierung* (LZA) und zum Data Sharing möglichst zügig umgesetzt werden, damit noch vorhandenes Hintergrundwissen gut genutzt werden kann und alle Projektbeteiligten noch erreichbar sind.

Mit Abschluss des Forschungsprojekts kann eine Revision der Aufbewahrungsmaßnahmen für Daten und Dokumente erfolgen, bei der gegebenenfalls weitere der im letzten Abschnitt des vorigen Kapitels beschriebenen Maßnahmen beschlossen werden können. Eine weitere wichtige Voraussetzung für längerfristige Aufbewahrung, und damit auch Data Sharing, ist die Konvertierung der Dateien in geeignete Formate. „Geeignet" heißt hier im Wesentlichen, dass die Fähig-

[25] http://www.langzeitarchivierung.de/ [09.05.02013]
[26] http://kolawiss.uni-goettingen.de/ [09.05.2013]

keit, den Dateiinhalt interpretieren zu können, langfristig gesichert ist. **Merkmale für Langzeitarchivierung geeigneter Dateiformate** sind (vgl. Vlaeminck, 2008, Abschnitt 3.2.1, sowie van den Eynden et al., 2011, S.12):

- **Offenheit**: Die Spezifikation des Dateiformats (Beschreibung des Aufbaus der Datei) ist frei zugänglich bzw. nicht-proprietär und *open source*, und es sind keine unter Lizenzbeschränkungen liegenden Algorithmen (z.B. Kompressionsalgorithmen) oder Objekte (z.B. Schriftarten oder Bilder) eingebunden;
- **Hoher Verbreitungsgrad**: Das Dateiformat ist weit verbreitet und bleibt daher voraussichtlich auch längere Zeit weithin in Verwendung;
- **Geringe Komplexität**: Es sind keine komplexen Operationen zur Interpretation des Formats nötig. Dies bedeutet etwa, dass keine Dekompression nötig ist und dass der Inhalt unmittelbar von Menschen lesbar ist (kein Maschinenkode);
- **Keine integrierten Schutzmechanismen** (Passwortschutz, Kopierschutz, Verschlüsselung), da diese Erhaltungsmaßnahmen (z.B. Sicherheitskopien) langfristig gesehen erschweren;
- **Möglichst umfassende Selbstdokumentation**: Die zur Interpretation des Dateiinhalts nötigen Metadaten sind in die Datei integriert;
- **Robustheit**: Das Format ist wenig anfällig gegenüber einzelnen Bitfehlern (die z.B. aufgrund physischer Degradation oder bei der Migration auf neue Speichermedien auftreten können);
- **Keine Abhängigkeiten** von bestimmter Hardware, Betriebssystemen, Lesesoftware (Reader), oder sonstigen externen Ressourcen;
- **Verlustfreiheit**: Es gehen bei der Erstellung der Datei keine relevanten Informationen verloren (z.B. keine „lossy compression" bei Audio- oder Videodateien).

Die wenigsten Dateiformate werden alle diese Kriterien erfüllen, so dass eher eine Beurteilung daraufhin erfolgen sollte, zu welchem Grad ein Format den einzelnen Kriterien entspricht. Das UK Data Archive (van den Eynden et al., 2011) empfiehlt für quantitative Datensätze die Verwendung einer (tab-, komma-)delimitierten Plaintextdatei zusammen mit einem *command file* eines Statistikprogramms, das auch Metadaten enthalten kann sowie eine strukturierte Textdatei oder XML-Datei mit Metadaten. Für die weitere Dokumentation werden die Dateiformate RTF, ODT, und PDF bzw. PDF/A empfohlen (bei letzterem handelt es sich um eine Norm der ISO zu für Langzeitarchivierung geeigneten PDF-Dokumenten). **Empfehlungen zu geeigneten Formaten für die verschiedensten Ressourcen** (Primärdatensätze, qualitative/Textdaten, Bild-, Audio-, Videodateien, Dokumentation, …) finden sich in den folgenden Listen:

- van den Eynden et al. (2011, S. 12): Kompakte, übersichtliche Tabelle des UK Data Archive;

- Liste bevorzugter Formate der Library of Congress [27]: Umfangreich, mit zahlreichen inhaltlichen Erläuterungen;
- Vlaeminck (2008, Abschnitt 6.1): Umfangreiche Zusammenstellung der Empfehlungen verschiedener Gedächtnisorganisationen.

Bei Plaintext-Dateien ist auf ein subtiles Detail zu achten: Die *Textenkodierung*. Diese spezifiziert, wie die Bits einer Textdatei in Symbole übersetzt werden. Um eine Plaintext-Datei (also auch als Plaintext abgespeicherte Datensätze!) richtig darstellen zu können, muss also die richtige Textenkodierung in der Software, die die Datei einliest, spezifiziert sein. Leider gibt es eine ganze Reihe von Enkodierungsschemata (*code pages*), die zwar größtenteils, aber nicht völlig identisch sind. Dies ist zum Beispiel der Grund, warum anstelle von Umlauten manchmal kryptische Zeichen wie „Â½“ angezeigt werden. Ein Minimalstandard, der bei allen code pages identisch ist, ist allerdings das 128 Zeichen umfassende *ASCII Character Set* (Tabelle 2). Um Probleme durch die Verwendung unterschiedlicher Textenkodierungen zu vermeiden, sollten die für die Langzeitarchivierung vorgesehenen Plaintextdateien, soweit es möglich ist, lediglich Symbole aus diesem Zeichensatz enthalten. Eine sich zunehmend verbreitende Enkodierung, die neben der Darstellung von Umlauten auch die Darstellung nicht-lateinischer Schriftzeichen ermöglicht, ist *UTF-8*.

Tabelle 2. ASCII-Zeichensatz [28]

	...0	**...1**	**...2**	**...3**	**...4**	**...5**	**...6**	**...7**	**...8**	**...9**	**...A**	**...B**	**...C**	**...D**	**...E**	**...F**
0...	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
1...	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2...	SP	!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
3...	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4...	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5...	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6...	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7...	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	DEL

Anmerkung: Die hexadezimale Position des Zeichens ergibt sich aus der Zusammensetzung der Ziffern in der ersten Spalte und Zeile. Die Zeichen 00 bis 1F sind Steuerzeichen (z.B. Zeilenumbruch), das Zeichen 20 das Leerzeichen und 7F das Löschzeichen.

[27] http://www.digitalpreservation.gov/formats/ [09.05.2013]
[28] Übernommen von http://de.wikipedia.org/wiki/ASCII [09.05.2013]

Da der „kritischste Moment“ im Lebenszyklus von Forschungsdaten das Ende des Forschungsprojekts darstellt, sollten Maßnahmen zur Langzeitarchivierung und zum Data Sharing möglichst umgehend umgesetzt werden.

Für die Langzeitarchivierung sollten geeignete Dateiformate (offen, weit verbreitet, wenig komplex, robust, unverschlüsselt, weitgehende Selbstdokumentation, verlustfrei erstellbar) gewählt werden. Listen mit entsprechenden Empfehlungen für verschiedene Ressourcen finden sich online.

Für quantitative Forschungsdaten bieten sich Plaintext-Dateien an, wobei auf die Textenkodierung (empfehlenswerter Minimalstandard: ASCII Character Set oder UTF-8) zu achten ist.

3.2.3.2. Auswahl der zu archivierenden Daten und Dokumente und des Archivs

Nicht alle im Rahmen der Forschungsarbeit erzeugten Daten und Dokumente sollten langzeitarchiviert werden. Angesichts der immer weiter wachsenden Speicher- und Verarbeitungskapazität moderner Computertechnik stellt sich zwar die Frage, ob man der Sicherheit halber nicht einfach alles archiviert. Gegen diese Sichtweise spricht aber eine Reihe von Argumenten (Whyte & Wilson, 2010):

- Angesichts einer immer stärkeren Digitalisierung der Forschungsarbeit und einem anhaltenden Wachstum des „Outputs“ der Forschung ist es nicht unwahrscheinlich, dass die Menge der durch die Forschung erzeugten digitalen Inhalte schneller wächst als die Speicher- und Verarbeitungskapazitäten.
- Langzeitarchivierung erfordert die Anfertigung von Sicherheitskopien. Die reale Menge der zu speichernden Inhalte ist also um einen Faktor von mindestens zwei größer als die für die Archivierung eingereichten Inhalte.
- Die Suche nach bzw. Entdeckung von relevanten Inhalten wird schwieriger.
- Langzeitarchivierung erfordert eine aktive Pflege des Archivguts durch geschultes Personal, und ist damit kostspielig.

Es stellt sich also die Frage, was wie lange wo archiviert werden soll (die Frage nach der Nachnutzung wird im Abschnitt „Data Sharing“ behandelt).

Zunächst zur Frage, was archiviert werden soll (was *archivierungswürdig* ist). Ausgangspunkt sollte hier eine Auswahl relevanter Forschungsdaten sein. Darauf basierend kann dann entschieden werden, welche Metadaten bzw. Dokumentation mit archiviert werden muss. Neben eventuell bestehenden rechtlichen Verpflichtungen und Vorgaben durch Forschungsförderer sollte die Bedeutsamkeit für potentielle Nachnutzer das oberste Leitkriterium sein (Weichselgartner et al., 2011a). Dabei handelt es sich aber um ein ziemlich abstraktes Kriterium; außerdem ist in der Regel nicht absehbar, ob nicht in Zukunft neue Forschungsperspektiven aufkommen, die einen zuvor uninteressanten Datensatz plötzlich interessant erscheinen lassen. Eine

konkreter formulierte Liste von **Kriterien zur Auswahl zu archivierender Daten** könnte wie folgt aussehen (basierend auf Whyte & Wilson, 2010, und Weichselgartner et al., 2011a):

- **Absehbarer Wert der Daten**: Welcher Wert der Daten ist bereits jetzt absehbar? Dieser Nutzen kann wissenschaftlicher Art sein (zum Beispiel basierend auf der Aufmerksamkeit, die Publikationen basierend auf den Daten erhalten, aber auch auf Nutzungspotentialen der Daten für die Lehre) oder kultureller und historischer Art (etwa als Dokumentation historisch bedeutsamer Ereignisse wie dem Mauerfall oder von Naturkatastrophen).
- **Einzigartigkeit/Nicht-Replizierbarkeit der Erhebung**: Etwa, weil die Daten ein nicht wiederholbares Ereignis „eingefangen" haben oder weil die Datenerhebung äußerst kostspielig war. Bei der Bewertung hilft auch eine Berücksichtigung des Datentyps: So unterscheidet das Research Information Network (2008) (unter anderem) zwischen Daten aus Beobachtungsstudien, Experimenten und Modellrechnungen. Erstere sind in der Regel kaum exakt replizierbar. Experimente sollten zumindest prinzipiell reproduzierbar sein. Bei Modellrechnungen dagegen sind die Daten im Prinzip nach Belieben wiederherstellbar, so dass das eigentlich Wertvolle, zu Archivierende in der Methodenbeschreibung liegt.
- **Voraussichtlich entstehende Kosten der Langzeitarchivierung und Nachnutzung**: Diese sind beispielsweise höher bei umfangreichen Datensätzen, bei Daten, die besondere Schutzmaßnahmen erfordern oder wenn die Notwendigkeit zum Erwerb von Lizenzen für eine Langzeitarchivierung und Weitergabe besteht. Die Kosten können gegen den voraussichtlichen Nutzen abgewogen werden.
- **Vollständigkeit der Dokumentation und Datenqualität**: Wie hoffentlich deutlich geworden sein sollte, sinkt die Interpretierbarkeit und damit der Nutzwert der Daten mit der Qualität ihrer Dokumentation.
- **Schranken für die Nachnutzung**: Solche Schranken können entstehen durch die Notwendigkeit, Lizenz- bzw. urheberrechtliche Bestimmungen einzuhalten oder Anonymisierungsschritte vorzunehmen, welche das analytische Potential der Daten beeinträchtigen, aber auch dadurch, dass die Daten nicht in für die Langzeitarchivierung geeignete Formate konvertierbar sind.
- **Übereinstimmung mit den Aufnahmekriterien relevanter Archive**: Institutionelle Archive haben in der Regel eigene Policies darüber, welche Daten sie aufnehmen. Diese können z.B. Anforderungen zur Daten- und Metadatenqualität und -formatierung, die Freiheit der Daten und Dokumente von urheberrechtlichen Ansprüchen sowie die Einschränkung auf Datensätze bestimmter Disziplinen (beispielsweise Psychologie im Falle von PsychData), Erhebungsorte, historischer Perioden, und methodischer Ansätze

(PsychData spezialisiert sich etwa auf quantitative Forschungsdaten, während das Bremer *Archiv für Lebenslaufforschung* [29] qualitative Interviewdaten aufnimmt) sein.

Eine besondere Schwierigkeit ergibt sich bei der **Beurteilung der Archivierungswürdigkeit von Datensätzen, die unter Verwendung von Sekundärdaten erzeugt wurden**, also Daten, die in mehr oder weniger vergleichbarer Form schon an anderer Stelle vorhanden sind. Unter der Annahme, man verfügt über die Lizenzen bzw. Berechtigungen, die verwendeten Sekundärdaten weiterzugeben, kann man sich an folgenden Richtlinien orientieren (ICPSR, 2012, S. 42):

- **Basiert der eigene Datensatz allein auf öffentlich zugänglichen Sekundärdaten**, ist es sinnvoll, lediglich eine exakte Methodenbeschreibung zu archivieren, inklusive Informationen zum Auffinden der Sekundärdaten (Angabe eines Persistent Identifier).
- Sind die verwendeten **Sekundärdaten bislang nicht für die Öffentlichkeit zugänglich**, spricht dies für die Archivierung, auch wenn der Datensatz keinerlei selbst erhobene Primärdaten enthält; besteht der Datensatz teilweise aus Pri-mär- teilweise aus öffentlich zugänglichen Sekundärdaten, hängt die Archivierungswürdigkeit des Sekundärdatenteils von weiteren Erwägungen ab:
    - **Verbindung von Primärdaten mit Zensusdaten**: Da letztere in der Regel sehr umfangreich sind und ihre Handhabung anspruchsvoll und zeitaufwändig, sollten die Sekundärdaten mit archiviert werden, auch wenn eine erneute Linkage durch Nachnutzer machbar (aber eben aufwändig) wäre.
    - Eine **Verbindung mit Sekundärdaten, die ohne besondere Kenntnisse nicht möglich ist**: Beispielsweise wenn die Linkage auf Grundlage einer komplexen Kombination von Variablen erfolgt oder wenn (bei Geodaten) bestimmte Ortskenntnisse für die Linkage nötig sind; auch hier erscheint eine Mitarchivierung der Sekundärdaten angezeigt.
    - Eine **Verbindung mit Sekundärdaten, die ohne besondere Kenntnisse möglich ist**: Im einfachsten Fall anhand einer Identifikatorvariable, die in beiden Datensätzen vorhanden ist. In diesem Fall sollten die Sekundärdaten nicht mitarchiviert, aber klare Instruktionen zur Durchführung der Linkage sowie Informationen zum Auffinden der Sekundärdaten gegeben werden. Wurden basierend auf den Sekundärdaten abgeleitete Variablen berechnet, sollten diese aber mitarchiviert werden, vor allem wenn die Variablenbildung mit einem höheren Aufwand verbunden war.

Als nächstes steht die Wahl eines Repositoriums für die ausgewählten Daten und Dokumente an. Ein Forschungsdaten-Repositorium lässt sich definieren als „eine Organisation (…), die die Verantwortung für den Langzeiterhalt (…) digitaler Objekte sowie für ihre Interpretierbarkeit zum

[29] http://www.lebenslaufarchiv.uni-bremen.de/ [09.05.2013]

Zwecke der Nutzung durch eine bestimmte Zielgruppe (…) übernommen hat“ (Aschenbrenner & Neuroth, 2011). Je nach verantwortlicher Organisation lassen sich verschiedene Arten von Archivierungslösungen unterscheiden: Archivierung durch die eigene Arbeitsgruppe bzw. Fakultät oder durch den Forschungsverbund, Übergabe an ein von Universität, Rechenzentrum oder Universitätsbibliothek als Dienstleistung betriebenes Repositorium (vergleichbar den inzwischen weit verbreiteten Hochschulschriftenservern), Einreichung des Datensatzes zusammen mit einer darauf basierenden Publikation bei einem Verlag, welcher die Archivierung vornimmt oder die Übergabe an ein organisatorisch eigenständiges Archiv. Letztere können disziplinspezifisch (z.B. PsychData) oder disziplinübergreifend ausgerichtet sein.

Die Selbstarchivierung etwa im Rahmen der eigenen Arbeitsgruppe bietet große Flexibilität, ist allerdings gegebenenfalls mit dem Aufwand verbunden, noch nicht vorhandene Infrastrukturen zu schaffen. Dies umfasst neben einer Speicher- und Serverinfrastruktur auch die Installation und Einrichtung einer geeigneten Repositoriensoftware. Aschenbrenner und Neuroth (2011) geben eine Einführung in mögliche Softwarelösungen. Die langfristig gesehen größere Herausforderung wäre aber sicherzustellen, dass das Repositorium auch auf lange Sicht in Betrieb gehalten wird, was insbesondere bei mehr oder weniger temporären Konstellationen wie Forschungsverbünden schwierig ist. Auch die Archivierung durch Verlage ist zumindest in der Psychologie oft schwierig, etwa weil Gutachtern Richtlinien zur Beurteilung der Datensätze fehlen (vgl. Weichselgartner, 2011a, S. 5).

Für die meisten Arbeitskontexte in der Psychologie, wo (anders als etwa in Bereichen wie der Hochenergiephysik) die für die Einrichtung eigener Repositorien nötigen Ressourcen nicht vorhanden sind, bietet sich daher die Übergabe an ein spezialisiertes Datenarchiv an. **Vorteile einer Archivierung in einem spezialisierten Archiv** sind:

- Die **Beratung und Unterstützung bei der Datenvorbereitung und -übergabe** durch geschultes Personal;
- Eine in der Regel **höhere institutionelle Stabilität** (die Daten sind langfristig gesichert);
- **Professionelle Infrastrukturen und Arbeitsabläufe** zur Speicherung, Instandhaltung, Pflege, und Verfügbarmachung der Daten;
- **Klare Definition von Rollen und Verantwortlichkeiten;**
- **Sicherstellung der Einhaltung rechtlicher Vorgaben**, sowie die Übernahme von Garantien über die Bedingungen der Nachnutzung der übergebenen Forschungsdaten;
- **Dienstleistungen zur Erhöhung der Visibilität/Auffindbarkeit des archivierten Datensatzes** (Vergabe eines Persistent Identifier, Durchsuchbarkeit des Datenbestandes über das Webangebot des Archivs anhand von Metadaten, Verknüpfung des Datenbestands mit Meta-Suchportalen und Web-Verzeichnissen).

Eine detaillierte Darstellung der durch PsychData angebotenen Dienstleistungen wird in Kapitel 4 gegeben.

Mittlerweile wurden auch **Kriterienkataloge formuliert, um die Vertrauenswürdigkeit digitaler Archive zu beurteilen**. Diese basieren in der Regel auf dem *Open Archival Information System (OAIS)*-Referenzmodell, einem ISO-Standard über die Organisationsstruktur von Archiven. Beispiele für Kriterienkataloge sind *TRAC (Trustworthy Repositories Audit & Certification: Criteria and Checklist)*, *DRAMBORA (Digital Repository Audit Method Based on Risk Assessment)* und *DAS (Data Seal of Approval)*. Eine deutschsprachige Checkliste wurde von nestor (2008) entwickelt. Die Kriterien dieser Liste, auf der basierend auch die DIN-Norm 31644 für vertrauenswürdige digitale Langzeitarchive entwickelt wurde, seien hier auszugsweise wiedergegeben:

- Das **Archiv hat definierte Ziele** (z.B. definierte Zielgruppe, Kriterien für die Auswahl der zu archivierenden Daten);
- Das **Archiv ermöglicht eine angemessene Nutzung des Archivguts** (Zugang, Interpretierbarkeit);
- **Berücksichtigung gesetzlicher und vertraglicher Regelungen** durch das Archiv;
- **Angemessene Organisationsform des Archivs** (z.B. Finanzierung sichergestellt, langfristige Planungen sowie Regelungen für eine Fortführung der Aktivitäten für den Fall einer Abwicklung des Archivs sind vorhanden);
- Ein **angemessenes Qualitätsmanagement ist implementiert** (Definition von Prozessen und Verantwortlichkeiten, Dokumentationsprozeduren sind vorhanden);
- **Sicherstellung der Integrität und Authentizität des Archivguts** in allen Verarbeitungsstufen (Übernahme, Lagerung, Nachnutzung);
- **Übernahme, Aufbewahrung und Nachnutzung des Archivguts erfolgt nach definierten Vorgaben** (z.B. was ist in welcher Form einzureichen: Spezifikation des *Submission Information Package*, SIP);
- Das **Archiv betreibt ein aktives Datenmanagement zur Unterstützung seiner Aufgaben** (z.B. Vergabe von Identifikatoren für das Archivgut, Erhebung von Metadaten zur technischen Beschreibung des Archivguts und zu Nutzungsrechten und –bedingungen);
- Das **Archiv verfügt über eine angemessene IT-Infrastruktur.**

Zu guter Letzt stellt sich die Frage, was mit nicht länger aufzubewahrenden Daten und Dokumenten geschehen soll. Während nicht-sensible Daten auf herkömmliche Weise gelöscht werden können, sollte bei sensiblen Daten eine einschneidendere Prozedur zur Löschung bzw. Zerstörung angewendet werden. Denn die „normale“ Löschung einer Datei zerstört nicht den Inhalt der Datei selbst, sondern nur die Referenz auf den Dateiinhalt im Dateisystem, so dass eine Wiederherstellung möglich bleibt. Eine tatsächliche Auslöschung der Daten auf dem Datenträger kann durch Überschreiben der entsprechenden Festplattensektoren mit Hilfe spezialisierter Software

erreicht werden. Die allersicherste, wenngleich natürlich radikale Option ist die physische Zerstörung der Datenträger. Hierfür kann man (sowohl bezüglich digitaler als auch nichtdigitaler Speichermedien) zum Beispiel auf Entsorgungsanlagen innerhalb der eigenen Institution zurückgreifen.

Für die Auswahl der archivierungswürdigen Daten und Dateien können das Nachnutzungspotential und die Einzigartigkeit der Daten, absehbare Kosten der Archivierung, Vollständigkeit der Daten und Dokumentationen, Einschränkungen der Nachnutzbarkeit und Übereinstimmung mit Aufnahmekriterien relevanter Archive als Kriterien herangezogen werden.

Verschiedene Vorteile sprechen für die Archivierung von Daten in einem spezialisierten Datenarchiv:

- **Beratung und Unterstützung,**
- **Institutionelle Stabilität,**
- **Professionelle Infrastrukturen & Arbeitsabläufe,**
- **Klar definierte Rollen & Verantwortlichkeiten,**
- **Einhaltung rechtlicher Vorgaben,**
- **Bessere Visibilität und Recherchierbarkeit.**

Für die Archivauswahl existieren Kriterienkataloge, an denen man sich orientieren kann.

Bei der Löschung von Daten ist besonders bei sensiblen Daten auf eine komplette „Zerstörung“ zu achten.

3.2.3.3. Berücksichtigung rechtlicher Bestimmungen zu Datenschutz und Urheberrecht

Die rechtlichen Bestimmungen, die den Umgang mit Forschungsdaten berühren, gewinnen eine verstärkte Bedeutung im Kontext der Langzeitarchivierung von Daten und Data Sharing, da die Daten dadurch nicht mehr allein der erhebenden Forschungsgruppe zugänglich sind. Da die Archivierung und Nachnutzung digitaler Forschungsdaten ein relativ junger Trend sind, bestehen diesbezüglich (wie auch bezüglich der Handhabung digitaler Inhalte allgemein) noch rechtliche Grauzonen und Spannungsfelder und es ist in naher Zukunft mit Anpassungen in der Gesetzgebung zu rechnen. Der Grundkonflikt liegt im Wesentlichen zwischen den ethischen und rechtlichen Begründungen für die Förderung wissenschaftlicher Forschung einerseits und datenschutzrechtlichen und urheberrechtlichen Erwägungen andererseits. Hier sei nur ein Beispiel genannt: Einerseits wird die Forderung nach Aufbewahrung von Primärdaten zur Sicherung guter wissenschaftlicher Praxis erhoben, andererseits sollten personenbezogene Daten gelöscht werden, sobald der primäre Forschungszweck erreicht ist.

Weiter verkompliziert ist die Lage bei multinationalen Forschungsprojekten sowie unter Umständen auch bei einer Nachnutzung der Daten in internationalem Kontext. Zwischen Datenschutzbestimmungen und Urheberrecht verschiedener Jurisdiktionen bestehen teils erhebliche Unterschiede. Beispielsweise ist die Schwelle für die Zuerkennung von *Copyright*-Schutz in

Großbritannien niedriger als für die von Urheberschutz in Deutschland (vgl. de Cock Buning, van Dinther, Jeppersen de Boer, & Ringnalda 2011a). Im Folgenden soll auf einige grundlegende Begrifflichkeiten und Probleme bezüglich Datenschutz- und Urheberrecht bei Forschungsdaten in Deutschland eingegangen werden. Für die Klärung komplizierter, insbesondere mehrere nationale Jurisdiktionen berührender Sachverhalte sollte eine Rechtsberatung erwogen werden. Die folgende Literatur, auf der auch der folgende Abschnitt aufbaut, kann für eine **vertiefende Einführung** herangezogen werden:

- Spindler und Hillegeist (2009): Eine ausführliche Untersuchung verschiedener rechtlicher Fragestellungen, die sich (in Deutschland) bezüglich der Langzeitarchivierung von Forschungsdaten ergeben, mit besonderem Fokus auf Daten aus der medizinischen Forschung;
- Spindler und Hillegeist (2011): Eine kompakte Diskussion des deutschen Datenschutz- und Urheberrechts im Hinblick auf Forschungsdaten;
- de Cock Buning et al. (2011a): Vergleichende Untersuchung der Bedeutung urheberrechtlicher Regelungen in den Niederlanden, Dänemark, Deutschland, und Großbritannien für den Umgang mit Forschungsdaten; zu den Ergebnissen für die einzelnen Länder wurden jeweils auch eigene Berichte erstellt;
- de Cock Buning, van Dinther, Jeppersen de Boer, & Ringnalda (2011b): Strukturierte Darstellung der Urheberrechtssituation bezüglich Forschungsdaten in Deutschland;
- Metschke und Wellbrock (2002): Eine ausführliche Abhandlung zu Fragen des Datenschutzes in Wissenschaft und Forschung in Deutschland.

Datenschutz

Sofern es sich bei den zu archivierenden Forschungsdaten um personenbezogene Daten handelt, sind Bestimmungen aus dem Bundesdatenschutzgesetz (BDSG), den Landesdatenschutzgesetzen der betroffenen Länder sowie dem Sozialgesetzbuch X (Schutz von Sozialdaten) relevant. Zunächst seien einige **für die Datenschutzproblematik wichtige Begrifflichkeiten** bestimmt:

- **Personenbezug**: Personenbezogene (auch: persönliche) Daten sind laut BDSG Angaben, die eindeutig einer bestimmten natürlichen Person zugeordnet sind oder ihr ohne unverhältnismäßigen Aufwand zugeordnet werden können („personenbeziehbare“ Daten). Das aus dem Grundgesetz abgeleitete Recht auf informationelle Selbstbestimmung besagt, dass das Recht über Preisgabe und Verwendung persönlicher Daten bei der Person selbst liegt. Personenbezogene Daten sollten zerstört (bzw. anonymisiert) werden, sobald dies der Forschungszweck erlaubt, spätestens aber, sobald der Forschungszweck erreicht ist, es sei denn, eine anderweitig lautende, hinreichend spezifisch formulierte, ausdrückliche Einwilligungserklärung liegt vor.

- **Direkter Identifikator**: Jedes Merkmal, das explizit der Identifikation einer bestimmten Person dient, z.B. eine Postanschrift, eine Telefonnummer oder ein Autokennzeichen. Dies ist insbesondere bei Langzeitstudien relevant, bei denen solche Merkmale gebraucht werden um die Teilnehmer erneut zu kontaktieren.
- **Indirekter Identifikator**: Jede Kombination von Merkmalen, die zwar jeweils für sich genommen nicht der Identifikation einer Personen dienen, die dies aber in ihrer Gesamtheit ermöglichen. Bestimmte Variablen eignen sich in besonderem Maße als indirekte Identifikatoren, beispielsweise Angaben zu Geburtsjahr, besuchter Schule, Arbeitsstellen, zum genauen Ort und Zeitpunkt bestimmter Ereignisse (wie sie etwa in Interviews gemacht werden) oder Extremwerte von Variablen (z.B. von Einkommen oder Körpergröße). Prinzipiell können aber auch sehr generische Merkmale wie das Geschlecht zu einer indirekten Identifikation beitragen. Die Problematik der indirekten Identifikation ist insbesondere auch im Zusammenhang mit der Verknüpfung verschiedener Datenquellen zu berücksichtigen.
- **Anonymisierung**: Personenbezogene Daten gelten als anonymisiert, sobald sie so abgeändert wurden, dass eine direkte oder indirekte Identifizierung der einzelnen Merkmalsträger unmöglich ist. Eine im Wesentlichen gleichbedeutende „faktische Anonymisierung“ liegt vor, wenn eine Reidentifizierung der Person nur mit völlig unverhältnismäßigem Aufwand möglich wäre. Anonymisierte Daten unterliegen keinen datenschutzrechtlichen Bestimmungen.
- **Pseudonymisierung**: Daten gelten als pseudonymisiert, wenn die identifizierenden Merkmale anhand einer Zuordnungsvorschrift so verändert wurden, dass die Reidentifizierung nur unter Kenntnis dieser Zuordnungsvorschrift möglich ist. Dies könnte etwa eine kryptographische Verschlüsselung der identifizierenden Merkmale oder die Angabe eines durch die Versuchsperson selbst erzeugten Kodes sein.

Die einfachste Lösung der Datenschutzproblematik besteht zweifelsfrei darin, lediglich anonymisierte Forschungsdaten zu archivieren und zur Nachnutzung freizugeben. Mögliche **Anonymisierungstechniken bei quantitativen Primärdatensätzen** sind etwa die Folgenden:

- **Vollständige Entfernung relevanter Variablen aus dem Datensatz**: Insbesondere natürlich direkte Identifikatoren, aber auch Variablen die ein starkes Risiko indirekter Identifikation transportieren. Dies ist die sicherste Methode, kann aber natürlich das analytische Potential der Daten erheblich beeinträchtigen.
- **Zusammenfassung von oder zu Kategorien**: Eine Reduktion der Datenpräzision durch Zusammenfassung kategorialer Variablen zu breiteren Kategorien, oder von quantitativen Variablen zu kategorialen Variablen. Hier ist es sinnvoll zu prüfen, welche Wertebereiche besonders kritisch sind, und speziell diese Bereiche weiter zusammenzufassen.

Damit kann der zwangsläufig einhergehende Informationsverlust verringert werden. Oft ist beispielsweise die Zusammenfassung der Extrembereiche zu einer „größer/kleiner als"-Kategorie ausreichend („top coding").

- **Zusammenfassung verschiedener Variablen zu einem abgeleiteten Wert**: Beispielsweise die Ersetzung der explizite Angaben von Wohnort und Arbeitsplatz durch eine Angabe zur Distanz zwischen beiden Orten.
- **Datenverfremdende Techniken** wie „Swapping" (Vertauschung kritischer Variablenwerte zwischen möglichst ähnlichen Probanden) oder die Hinzufügung eines geringfügigen zufälligen Fehlerterms zu kontinuierlichen Variablen. Dies dient insbesondere zur Erschwerung der indirekten Identifikation, sollte aber mit Vorsicht angewandt werden.
- **Reduktion des Datensatzes auf eine zufällig gezogene Teilstichprobe**: Eine vor allem bei großen Datensätzen sinnvolle Technik, da einerseits die Größe der Teilstichprobe weiterhin für statistische Analysen ausreichend ist, andererseits ein geringeres Risiko für indirekte Identifikation besteht.

Qualitative Daten bedürfen eigener Anonymisierungstechniken, die unter Umständen deutlich mehr Aufwand erfordern als quantitative Daten. Bei Audio- und Videodaten müssen etwa Gesichter und Stimmen verfremdet werden. Bei Interview- oder sonstigen textuellen Daten sollten Namen durch Pseudonyme oder generische Bezeichnungen wie „Mutter" und „Sohn" ersetzt werden und weitere Angaben geschwärzt oder verfremdet werden, welche eine Identifikation ermöglichen (beispielsweise Angaben zu bestimmten Ereignissen oder Orten in der Biographie einer Person).

Falls eine Anonymisierung nicht ohne eine inakzeptable Verminderung des Analysepotentials der Daten möglich ist, bestehen einige **Möglichkeiten, die personenbezogenen Daten der Versuchsteilnehmer auch bei Langzeitarchivierung und Weitergabe an Nachnutzer zu schützen**. Diese Maßnahmen können aber letztlich nur das Vertrauen der Probanden in die Sicherheit ihrer Daten und damit ihre Bereitschaft zur Einwilligung in die Weitergabe erhöhen, jedoch nie deren informierte Einwilligung ersetzen:

- Die **Pseudonymisierung der Daten und Aufbewahrung der Zuordnungsvorschrift in einer vom Archiv getrennten Institution**, z.B. am Institut des Hauptverantwortlichen für die Primärdatenerhebung oder bei einem spezialisierten *Datentreuhänder* (z.B. einem Notar). Unter besonderen Schutzvorkehrungen würde Forschern mit berechtigtem Interesse dann eine Verbindung der Datensätze gestattet. Die Datentreuhänderlösung ist im Übrigen auch für die Durchführung von Langzeitstudien eine sinnvolle Option, um datenschutzrechtliche Bedenken auszuräumen.

- Die **Abgabe von Verschwiegenheitserklärungen oder anderweitigen Erklärungen zur Aufrechterhaltung des Datenschutzes** durch Datennutzer als Voraussetzung für den Zugang zu den Daten.
- **Datenzugriff über physische oder virtuelle „Datenenklaven“**: Der Zugang ist nur vor Ort in gesicherten Räumlichkeiten oder über eine gesicherte Online-Analyseumgebung, die lediglich das Herunterladen aggregierter Daten erlaubt, möglich.

Die Generierung sowohl von vollständig anonymisierten, für die Öffentlichkeit zugänglichen „public use files“ und nichtanonymisierten bzw. lediglich pseudonymiserten „restricted use files“ stellt schließlich einen Kompromiss zwischen Zugänglichkeit und Schutz persönlicher Informationen dar.

Mögliche Anonymisierungstechniken bestehen in der vollständigen Löschung von Variablen, der Zusammenfassung kritischer Variablen zu Kategorien oder zu abgeleiteten Werten, datenverfremdenden Techniken oder der Reduktion auf eine zufällig gezogene Teilstichprobe.

Maßnahmen zum Schutz anonymisierter Daten können die Pseudonymisierung und Aufbewahrung der Zuordnungsvorschrift in einer spezialisierten Institution, die Abgabe von Verschwiegenheitserklärungen oder der Datenzugriff über sogenannte „Datenenklaven“ (gesicherter Zugang) sein.

Urheberrecht

Der zweite Rechtsbereich, der bei Langzeitarchivierung von Forschungsdaten und Data Sharing relevant ist, ist das **Urheberrecht**. Dieses ist in Deutschland im Urheberrechtsgesetz (UrhG) geregelt, das auch die sogenannten „verwandten Schutzrechte“ umfasst, welche bestimmte Leistungen abdecken, die nicht durch Urheberschutz im engeren Sinne erfasst sind.

Zunächst stellt sich die Frage, **ob die erhobenen Daten selbst unter urheberrechtlichem Schutz stehen** könnten. Ein wichtiger Rechtsgrundsatz ist hierbei, dass bloße Fakten nicht unter urheberrechtlichen Schutz fallen können; lediglich aufgrund der Form ihrer Darbietung können Schutzrechte entstehen, und zwar auf zweierlei Wegen:

- Ein Datensatz kann ein **urheberrechtlich geschütztes Werk im engeren Sinn** sein. Dazu muss es sich bei dem Datensatz um eine persönliche intellektuelle Schöpfung mit einem gewissen Mindestniveau an Originalität („Schöpfungshöhe“), welches individuelle Entscheidungen des Schöpfers über die Form des Werkes zum Ausdruck bringt, handeln. Da Datensätze in der Regel aber gerade anhand bestimmter, durch die wissenschaftliche Gemeinschaft geprägte Konventionen gestaltet werden, ist diese Individualität im Allgemeinen nicht gegeben. Daher kann, gerade in der quantitativ orientierten Forschung, davon ausgegangen werden, dass Datensätze keine urheberrechtlich geschützten Werke

sind (was ja auch durchaus im Sinne einer offenen Forschung ist). Würde aber aufgrund besonderer Umstände doch der Urheberschutz greifen, lägen das Urheberrecht und damit die Nutzungsrechte beim Schöpfer, das heißt, den für die Datenerhebung verantwortlichen Forschern. Urheberrechte können nach deutschem Recht ausschließlich bei natürlichen Personen liegen. Nutzungsrechte können allerdings (etwa im Rahmen von Arbeitsverträgen) auch an juristische Personen eingeräumt werden.

- Eine andere Möglichkeit wäre, dass ein Datensatz unter das **verwandte Schutzrecht für Datenbanken** fällt. Dieses greift, sobald folgende Kriterien erfüllt sind: Die Datensammlung müssen auf eine systematische Weise geordnet sein, es muss ein Zugriff auf individuelle Datenpunkte möglich sein, und die Schaffung der Datenbank muss eine „wesentliche Investition" (im Sinne von finanziellen Mitteln, Zeit, oder Arbeitskraft erfordert haben. Diese Investition bezieht sich aber auf die Beschaffung *bereits vorhandener Daten* oder die Gestaltung der Datenbankstruktur, jedoch gerade *nicht* auf den Aufwand zur Erhebung noch nicht bestehender Daten. Da aber gerade dies den wesentlichen Aufwand bei einer Erhebung von Forschungsdaten ausmacht, werden Forschungsdatensätze im Allgemeinen auch nicht unter Datenbankschutz fallen. Sollte dies doch der Fall sein, lägen die Schutz- und damit Nutzungsrechte bei demjenigen, der die „wesentliche Investition" getätigt hat. Dies ist bei Datenbanken, die von Angestellten im Rahmen von Beschäftigungsverhältnissen (also auch den meisten Forschern) erstellt werden, der Arbeitgeber; verwandte Schutzrechte können sowohl bei natürlichen als auch bei juristischen Personen liegen.

In den meisten Fällen werden Datensätze also (anders als etwa wissenschaftliche Literatur!) unter keinerlei urheberrechtlichen Schutzregelungen fallen. Dies mag auch einer der Gründe für die derzeit noch relativ verbreitete Zurückhaltung bei der Veröffentlichung von Datensätzen sein. Daher ist es wichtig, über Datenübergabeverträge zwischen Primärdatenerzeuger und Archiv sowie über Nutzungsverträge zwischen Archiv und Nachnutzer die Konditionen der Weitergabe von Forschungsdaten zu regeln.

Im seltenen Fall eines urheber- oder datenbankrechtlich geschützten Datensatzes ist es erforderlich, dass der oder die Rechteinhaber/-in die maßgeblichen Nutzungsrechte (Weitergabe, Anfertigung von Kopien) in Form einer entsprechenden *Lizenz* an das Archiv abtritt, was als Teil des Datenübergabevertrags geregelt werden kann. Eine in letzter Zeit an Popularität gewinnende Lizenzform sind die sogenannten *Creative Commons*-Lizenzen. Dabei handelt es sich um von der gleichnamigen, gemeinnützigen Organisation entwickelte Standard-Lizenzverträge, die es Urheberrechteinhabern auf einfach Weise erlauben sollen, der Allgemeinheit Nutzungsrechte einzuräumen. Je nach Variante werden dabei bestimmte Einschränkungen erhoben, etwa eine Zitierungspflicht oder eine Verpflichtung auf nichtkommerzielle Nutzung.

Ob Metadaten und Dokumentation zu den Forschungsdaten unter Urheberschutz fallen, müsste gegebenenfalls im Einzelfall geprüft werden, sollte allerdings auch eher selten der Fall sein, da die Dokumentation ebenfalls durch ein relativ standardisiertes Vorgehen geprägt ist. Eine wichtige Ausnahme stellt **Programmkode** dar, der im Lauf der Erhebung verfasst wurde. Dieser kann eine beträchtliche Komplexität erreichen, beispielsweise ein individuell entwickeltes Programm zur Darbietung von Stimuli und Registrierung von Reaktionen in einem psychologischen Experiment. In diesem Fall besteht mit hoher Wahrscheinlichkeit Urheberschutz, der in vertraglichen Einigungen über Datenarchivierung und -nachnutzung berücksichtigt werden müsste.

Ein weiterer, bereits erwähnter Sonderfall liegt bei der Verwendung psychologischer Test- und Fragebogeninstrumente vor, deren Nutzungsrechte etwa bei Testverlagen liegen. Die Items und Anwendungsprozedur solcher Instrumente werden in der Regel als Teil der Dokumentation erfasst. Hier sollte eine Konsultation mit dem Rechteinhaber über die Möglichkeit der Weitergabe erfolgen. Schließlich kann ein Datensatz auch als Grundlage für die Konstruktion solcher Testverfahren dienen, deren Nutzungsrechte dann später an Verlage abgetreten werden. Falls Langzeitarchivierung und Data Sharing des Datensatzes vorgesehen sind (wofür gerade bei den großen Normdatensätzen viel spricht), sollte man auf keinen Fall vergessen, sich entsprechende Nutzungsrechte im Vertrag mit dem Verlag vorzubehalten.

In der Regel sind Datensätze keine urheberrechtlichen geschützten Werke im engeren Sinn und fallen auch nicht unter das Schutzrecht für Datenbanken. Ausnahmen können z.B. für aufwändig entwickelten Programmkode (zur Datenerfassung und -verarbeitung) gelten.

Falls doch Urheberschutz besteht, kann mit einer Creative Commons Lizenz festgelegt werden, welche Nutzungsrechte Anderen eingeräumt werden.

3.2.3.4. Zugänglichmachung der Daten

Mit der Übergabe von Daten und Dokumentation an ein Archiv sollte wie oben erörtert ein Datenübergabevertrag abgeschlossen werden, der unter anderem auch regelt, unter welchen Bedingungen das Archiv die Daten Dritten zugänglich machen kann. Solche Nutzungsbedingungen dienen zum einen der Absicherung gesetzlicher Vorgaben, wie sie sich zum Beispiel aus dem Datenschutzrecht ergeben. Darüber hinaus sollten sie aber idealerweise einerseits das legitime Interesse des Datengebenden an einer hinreichenden Würdigung der eigenen Arbeit wahren (insbesondere angesichts der Tatsache, dass Datensätze per se in der Regel unter keinen besonderen Leistungsschutz fallen, siehe oben) und andererseits für Nachnutzer keine inakzeptabel hohen Zugangsbarrieren errichten. Wie genau dieser Kompromiss aussieht, hängt auch von den besonderen Umständen innerhalb der einzelnen Forschungsdisziplinen ab. Forschungsdatenarchive verfügen in der Regel über eigene Leitlinien zur Gestaltung solcher Verträge (die natürlich auch Verhandlungsspielraum lassen können).

Ein gängiges Modell sieht etwa die Einhaltung einer gewissen Sperrfrist, z.B. von einem Jahr, ein, in der das Archiv die Daten aufbewahrt, ohne sie für Dritte zugänglich zu machen. Erst nach Ablauf der Frist sind die Daten zur Nachnutzung freigegeben. Dadurch bleibt den Datengebern ausreichend Zeit, die erhobenen Daten gemäß der Zielstellung ihres Forschungsvorhabens auszuwerten und Befunde zu publizieren. Mögliche Modelle für den Zugangsweg zu den Daten sind z.B. für jedermann über die Website des Archivs herunterladbare Datensätze, Download nach Registrierung, Download erst nach Abschluss eines Datennehmervertrags, vor dessen Zustandekommen gegebenenfalls das Forschungsinteresse oder eine sonstige Begründung für die Nachnutzung dargelegt werden muss oder Zugang nach Abschluss eines Vertrags über eine geschützte Datenenklave.

Als Mindestforderung wird in nahezu allen Fällen von Nachnutzern verlangt, die nachgenutzten Daten in Publikationen, die darauf aufbauen, zu zitieren. Damit soll langfristig eine Gleichstellung von Daten mit Publikationen als eigenständig zu würdigendes Ergebnis wissenschaftlicher Arbeit erreicht werden (vgl. Kapitel 2.2). Für wissenschaftliche Literatur existiert eine gut etablierte Infrastruktur zum Auffinden und Referenzieren genutzter Publikationen, die ein einfaches Zitieren erlaubt. Für Datensätze befindet sich solch eine Infrastruktur zur Zeit erst im Aufbau.

Ein zentraler Grundbaustein dieser Infrastruktur ist die Vergabe von *Persistent Identifiers*. Dabei handelt es sich um abstrakte Namen (Identifier) für (in aller Regel digitale) Objekte, die zum Auffinden der Objekte dienen, und zwar unabhängig vom Ort, an dem sie sich befinden. Darin liegt der Hauptunterschied zu URLs: Diese geben einen bestimmten Ort im WWW an, an dem sich das Objekt befinden kann. Im schnelllebigen Internet sind aber Änderungen von Webadressen, z.B. bei der Restrukturierung des Webauftritts eines Archivs, an der Tagesordnung. Ein Datensatz wäre dann weiterhin auf der Website des Archivs auffindbar, nur eben unter einer anderen URL. Ein Persistent Identifier soll langfristige (persistente) Auffindbarkeit unter diesen Bedingungen garantieren.

Die Etablierung von Persistent Identifiers bedarf immer eines dahinterstehenden Systems, das die Verknüpfung zwischen Objekt und Identifier aufrechterhält. Ein derzeit populäres System für die Identifikation von wissenschaftlichen Publikationen, das *Digital Object Identifier* (DOI)-System, findet auch zunehmend als Identifier-System für Forschungsdaten Verwendung. Das DOI-System wird von der *International DOI Foundation* (IDF) weiterentwickelt und verwaltet. Die IDF koordiniert auch die Arbeit der DOI-Registrierungsagenturen. Diese unterhalten Verzeichnisse, in denen die DOIs zusammen mit einer URL und weiteren Metadaten zum Objekt, z.B. Titel und Publikationsjahr, hinterlegt sind. Über eine Anmeldung bei der Registrierungsagentur kann ein „Registrant“ (ein Verlag, ein Datenarchiv, eine Forschergruppe…) Einträge in diesem Verzeichnis anlegen. Der Registrant übernimmt auch die aktive Pflege der Einträge, so dass die Verbindung zwischen Persistent Identifier und Objekt aktuell bleibt.

Über das Resolver-Interface der IDF [30] wird man durch die Eingabe eines DOI direkt zu dem jeweils aktuellen Ort im WWW weitergeleitet. Dies ist bei Publikationen und Datensätzen in der Regel eine „Landing Page", die Informationen zu dem Objekt enthält sowie die Möglichkeit, das Objekt herunterzuladen, sofern man über die Berechtigung verfügt. Beispielsweise gelangt man durch Eingabe des DOI „10.5160/psychdata.brpr88pe99" zu der Beschreibung der Eichstichprobe für den Trierer Persönlichkeitsfragebogen auf der Website von PsychData (Becker, 2004). Die Angabe von DOIs kann damit die Rolle der bibliographischen Angaben in einer „klassischen" Literaturzitation übernehmen, nämlich den Verweis auf eine eindeutig identifizierbare Quelle.

Die DOI-Registrierungsagenturen sind jeweils auf bestimmte Arten von Objekten spezialisiert. Für Datensätze zuständig ist die Registrierungsagentur DataCite [31], ein internationaler Verbund von Forschungsinfrastruktureinrichtungen. Viele Forschungsdatenarchive bieten die Registrierung von DOIs für überlassene Datensätze als Dienstleistung an. PsychData nimmt die Registrierung etwa über die von GESIS und der *Zentralbibliothek für Wirtschaftswissenschaften* (ZBW) betriebene Agentur *da|ra* vor, die Mitglied von DataCite ist.

Neben dem DOI-System existiert auch noch eine Reihe anderer Systeme für Persistent Identifiers, etwa das *Uniform Resource Name* (URN)- oder das *Persistent Uniform Resource Locator* (PURL)-System.

Das zweite Grundelement der Infrastruktur zur Zitierbarmachung von Datensätzen ist der Aufbau von umfangreichen, gezielt absuchbaren Katalogen. In Kapitel 3.2.1 wurden bereits einige Verzeichnisse und Kataloge für Datenarchive aufgezählt. Für das Absuchen umfangreicher Kataloge ist ein gewisses Mindestmaß an Vereinheitlichung der Metadaten nötig, da letztlich nur diejenigen Informationselemente über die Datenarchive hinweg absuchbar sind, die über alle Archive hinweg konsistent strukturiert sind. Der bereits erwähnte Minimalstandard Dublin Core ist zum Beispiel die Grundlage für das *Protocol for Metadata Harvesting* der *Open Archives Initiative* (*OAI-PMH*) (Rusch-Feja, 2001), das ein automatisiertes „Ernten" der Metadaten in den Repositorien, die eine Computerschnittstelle dafür anbieten, erlaubt. Auf Grundlage der so von zahlreichen Repositorien eingesammelten Metadaten können umfangreiche Kataloge erstellt werden. DataCite stellt beispielsweise eine solche OAI-PMH Schnittstelle zur Verfügung

Andererseits muss ein Kompromiss zwischen dem Umfang des Kataloges und der Spezifität der Metadaten gefunden werden. DataCite hat das Minimal-Metadatenschema der IDF um für Forschungsdaten spezifische Metadatenelemente erweitert. Diesem Schema hat wiederum GESIS weitere Elemente hinzugefügt, z.B. Angaben, die speziell für die Sozialwissenschaft relevant sind und die am DDI-Schema orientiert sind sowie Angaben zur Datenverfügbarkeit (direkter Down-

[30] http://dx.doi.org/ [12.05.2013]
[31] http://www.datacite.org/ [12.05.2013]

load, Datenbestellung, Nutzung ausschließlich vor Ort). PsychData hat ein psychologiespezifisches, ebenfalls an DDI orientiertes Schema entwickelt.

Die Entwicklung von Infrastrukturen zum Auffinden von Forschungsdaten ist derzeit noch im Frühstadium ihrer Entwicklung. Es ist aber in absehbarer Zeit mit der Herausbildung etablierter Such- und Nachweissysteme zu rechnen, die den jeweils spezifischen Bedürfnissen der Fachgemeinschaften entsprechen und die Grundlage für eine Honorierung veröffentlichter Forschungsdatensätze durch Zitationen sind.

Datenübergabeverträge (an Archive) regeln Nutzungsbedingungen und schaffen so Rechtssicherheit für beide Parteien. Solche Datenübergabeverträge können Sperrfristen und zugelassene Zugangswege zu Daten festlegen und enthalten meist eine Verpflichtung zur Zitation der Forschungsdaten.

Um eine Infrastruktur zur Referenzierung von Forschungsdaten ähnlich der von Publikationen zu entwickeln, bedarf es einerseits der Etablierung von Persistent Identifiers, andererseits des Aufbaus von gezielt absuchbaren Katalogen zur Datenrecherche.

4. Datenmanagement und -archivierung mit PsychData

In diesem Kapitel soll eine Einführung in und Anleitung zur Nutzung der Dienste des psychologischen Forschungsdatenarchivs PsychData gegeben werden. Kapitel 4.1 gibt einen Überblick zur Entstehung und Leistungsspektrum von PsychData sowie seiner Einbindung in die anderen Informationsdienstleistungen des ZPID. In Kapitel 4.2 werden Anleitungen zur Nutzung der zentralen PsychData-Dienste gegeben: Das Online-Dokumentationstool MyPsychData, die Übernahme und Archivierung von Primärdatensätzen sowie die Nachnutzung dieser Datensätze.

Weitere Informationen sowie Neuigkeiten zum Thema Forschungsdaten in der Psychologie finden Sie auf der Website von PsychData[32], die auch der zentrale Zugangsort für die im Folgenden beschriebenen Angebote ist.

4.1 PsychData: Hintergrund und Dienstleistungsangebot

Das vom Leibniz-Zentrum für Psychologische Information und Dokumentation (ZPID) [33] betriebene Forschungsdatenarchiv PsychData ist das derzeit einzige dezidiert für die Psychologie eingerichtete Datenarchiv für quantitative Daten. Angeregt wurde seine Entwicklung durch Forschende aus der Psychologie, die aufwändig erhobene Längsschnittstudien der Nachwelt erhalten wollten, aber ihre Daten in existierende Archive nicht fachgerecht einspielen konnten, da diese an die Verhältnisse in der Psychologie nicht angepasst waren. Die Entwicklung begann 2002 mit DFG-Unterstützung, 2003 wurden die ersten Forschungsdaten in das Archiv übernommen. Weitere Tätigkeitsfelder neben der Übernahme, Pflege, und Verfügbarmachung von Forschungsdaten sind die Integration der Dienste mit anderen ZPID-Produkten, Beratung und Hilfestellung für Forschende sowie die Vertretung des Fachs Psychologie in den maßgeblichen forschungspolitischen Institutionen. PsychData gehört zu der vom Rat für Sozial- und Wirtschaftsdaten (RatSWD) [34] empfohlenen und zertifizierten Dateninfrastruktur in den Sozial-, Verhaltens- und Wirtschaftswissenschaften. Die Ansiedlung am seit 1972 bestehenden ZPID, das als Leibniz-Institut von Bund und Ländern finanziert wird, sorgt für eine stabile institutionelle Verankerung und stellt den Fortbestand des Archivs sicher (siehe Weichselgartner, Günther & Dehnhard, 2011b, und Weichselgartner, 2011b, für weitere Hintergrundinformationen).

Gemäß seinem Archivauftrag ist PsychData offen für quantitative Forschungsdaten aus der psychologischen Forschung, die als Grundlage von mindestens einer qualitätsgeprüften Veröffentlichung verwendet wurden. Mit „quantitativ“ sind hier Messdaten gemeint, d.h. numerische Abbildungen empirischer Relationen; z.B. psychophysiologische, Reaktionszeit- oder Fragebogendaten (das an der Universität Bremen angesiedelte Archiv für Lebenslaufforschung kann zur

[32] http://psychdata.zpid.de/ [13.05.2013]
[33] http://www.zpid.de/ [13.05.2013]
[34] http://www.ratswd.de/ [13.05.2013]

Archivierung qualitativer Daten genutzt werden, s. Fußnote 28). Als Forschungsdaten werden Daten, die in digitaler, maschinenlesbarer Form vorliegen, archiviert. Zugrundeliegende Rohdaten wie z.B. Videoaufzeichnungen werden derzeit nicht von PsychData übernommen, können und sollten aber natürlich in der archivierten Studiendokumentation (siehe unten) referenziert werden. Die Studiendokumentation dient primär der langfristigen Interpretier- und Nutzbarkeit der Daten. In diesem Rahmen kann auch in begrenztem Umfang qualitatives und Rohdatenmaterial archiviert werden (z.B. zu Illustrationszwecken).

PsychData ist auch für die Archivierung von Forschungsdaten seit längerem abgeschlossener Studien offen und arbeitet aktiv an der retrospektiven Dokumentation und Digitalisierung erhaltenswerter Studien aus der Psychologie. So wurden beispielsweise bereits Normstichproben und Längsschnitterhebungen aus den 80er und 90er Jahren erfasst. Da dies jedoch sehr arbeitsintensiv ist, insbesondere bei umfangreichen Längsschnittstudien, sind die Kapazitäten von PsychData zur Unterstützung bei der retrospektiven Dokumentation oft ausgelastet. Wenn bereits während des laufenden Forschungsvorhabens eine aktive Dokumentation erfolgt, ist die anschließende Archivierung dagegen in der Regel eine einfache und schnelle Angelegenheit. Um die forschungsbegleitende Dokumentation zu erleichtern (unabhängig davon, ob die Daten später an PsychData zur Nachnutzung übergeben werden sollen oder nicht), steht das webbasierte Dokumentationstool MyPsychData zur Verfügung, das in Kapitel 4.2.1 näher beschrieben wird.

Ein **Archivierungspaket in PsychData (im Folgenden auch „Studie“ genannt) umfasst vier Arten von Materialien**: Den Forschungsdatensatz, Metadaten auf Variablenebene („Kodebuch“), Metadaten auf Studienebene sowie (optional) sonstige Dokumentationsmaterialien. Das Schema, das den einzureichenden Metadaten zugrundeliegt, wurde von PsychData in Orientierung an den Standards Dublin Core und DDI 2.0 entwickelt und um spezifisch für die Psychologie relevante Elemente ergänzt.

- **Forschungsdatensätze**: Hierbei muss es sich um rechteckige Datenmatrizen (s. Kapitel 3.2.2, Abschnitt „Dateiorganisation…“) handeln; komplexere, z.B. hierarchisch geschachtelte Datenstrukturen sind nicht zulässig. Diese Beschränkung der logischen Struktur der Daten erleichtert die Implementation von Qualitätskontrollen wie dem automatisierten Abgleich der Forschungsdaten mit den Metadaten auf Variablenebene sowie von Suchprozeduren auf Variablenebene. Die Archivierung von Längsschnittstudien ist aber dennoch problemlos möglich, da einer Studie auch mehrere Datensätze zugeordnet sein können. Alternativ ist auch die Überführung der Daten in das „long“-Format oder durch die Übergabe mehrerer Archivierungspakete möglich (eines pro Erhebungswelle, wobei Querverweise zwischen den Wellen als Teil der Studiendokumentation oder als „sonstige Materialien“ archiviert werden). Je nach Bedarf können auch mehrere Versionen eines

Datensatzes als Teil eines Pakets archiviert werden, z.B. ein Datensatz mit Ausgangsdaten und einer mit aggregierten Daten.

Forschungsdatenmatrizen werden bei PsychData als delimitierte Plaintext-Datei (s. Kapitel 3 für eine Erläuterung dieser Begriffe) mit den Variablennamen als Kopfzeile archiviert, können aber auch in einem anderen Format, etwa als SPSS-Datei, an PsychData eingereicht werden, wo sie dann konvertiert werden.

- **Kodebuch**: Datei mit Metadaten auf Variablenebene, die nach einer spezifischen Syntax strukturiert sein muss (s. Kapitel 4.2). Das Kodebuch wird als Plaintext-Datei archiviert und kann gemäß den unten beschriebenen Syntaxregeln relativ einfach in einem Texteditor selbst erstellt werden. Alternativ lässt sich das Kodebuch auch über MyPsychData erstellen, wo auch eine automatische Fehler- und Konsistenzprüfung erfolgt. Liegen die Variablen-Metadaten bereits in anderer Form vor, z.B. innerhalb einer SPSS-Datei oder in einem anderen Kodebuchformat, können sie in den meisten Fällen auch von den Mitarbeitern bei PsychData konvertiert werden.
- **Studienmetadaten**: Die Metadaten auf Studienebene umfassen etwa Elemente zu Hintergrundliteratur, an der Forschung beteiligte Personen, zu Zuwendungsgebern, zum theoretischen Hintergrund und zur Erhebungsmethode. Sie werden entweder mit dem über die PsychData-Website herunterladbaren Metadatenformular oder über MyPsychData übermittelt, so dass sie bereits hinreichend vorstrukturiert sind für die Archivierung.
- **Sonstige Materialien**: Gibt es weitere, für die Interpretation und Nachnutzung der Daten wichtige Dokumente, z.B. Programmkode, Skizzen oder Konkordanztabellen für die Items von Längsschnittstudien, können diese in jeweils geeigneten Formaten mitarchiviert werden. Zur Wahl des Formats und zur Konvertierung sollte eine Absprache mit PsychData erfolgen.

Bei PsychData archivierte Datensätze werden als „scientific use files“ bereitgestellt. Dementsprechend wird sowohl die Übergabe als auch die Nachnutzung von Daten-sätzen in PsychData über den Abschluss von Nutzungsverträgen geregelt, die sicherstellen, dass die Daten zu wissenschaftlichen Zwecken verwendet und keine Deanonymisierungsversuche unternommen werden. Zusätzlich ist festgelegt, dass in Publikationen eine Zitation der Datenquelle erfolgen muss. Datengeber können gegebenenfalls über eine Zusatzvereinbarung bei Vertragsabschluss eine Sperrfrist für die Freigabe zur Nachnutzung angeben. Der Zugang zu den Datensätzen und Kodebüchern für Nachnutzer erfolgt über den Versand von CD-ROMs; die Studienmetadaten sind jedoch frei über die Website von PsychData einsehbar. Die Daten werden ausschließlich in (faktisch) anonymisierter Form zur Nachnutzung bereitgestellt. Idealerweise sollten die Daten bereits anonymisiert übergeben werden.

Die von PsychData verwendete technische Infrastruktur zur Archivierung der Daten und Metadaten basiert auf quelloffener Software (unter anderem Unix, MySQL, Apache, PHP). Durch die Nutzung von mehreren, räumlich getrennten Servern, definierten Backup-Prozeduren unter Verwendung sowohl magnetischer als auch optischer Medien, Prüfsummen, und einer abgestuften Zugriffskontrolle wird die physische Integrität und Authentizität des Archivguts sichergestellt.

Für alle archivierten Forschungsdatensätze wird von PsychData über den Registrierungsservice da|ra ein DOI angelegt, der eine einfache Zitation des Datensatzes ermöglicht und der den Datensatz über Metadatensuchmaschinen der DOI-Registrierungsagenturen auffindbar macht (vgl. Kapitel 3.2.3). Außer über allgemeine Suchmaschinen wie Google lässt sich PsychData (zusammen mit anderen psychologierelevanten Forschungsdatenarchiven) auch gezielt über die Psychologie-Suchmaschine PsychSpider des ZPID absuchen (s. Kapitel 3.2.1, Abschnitt „Suche nach…"). Die Auffindbarkeit wird außerdem dadurch erhöht, dass die archivierten Datensätze mit anderen Produkten des ZPID integriert sind. Es bestehen wechselseitige Verweise

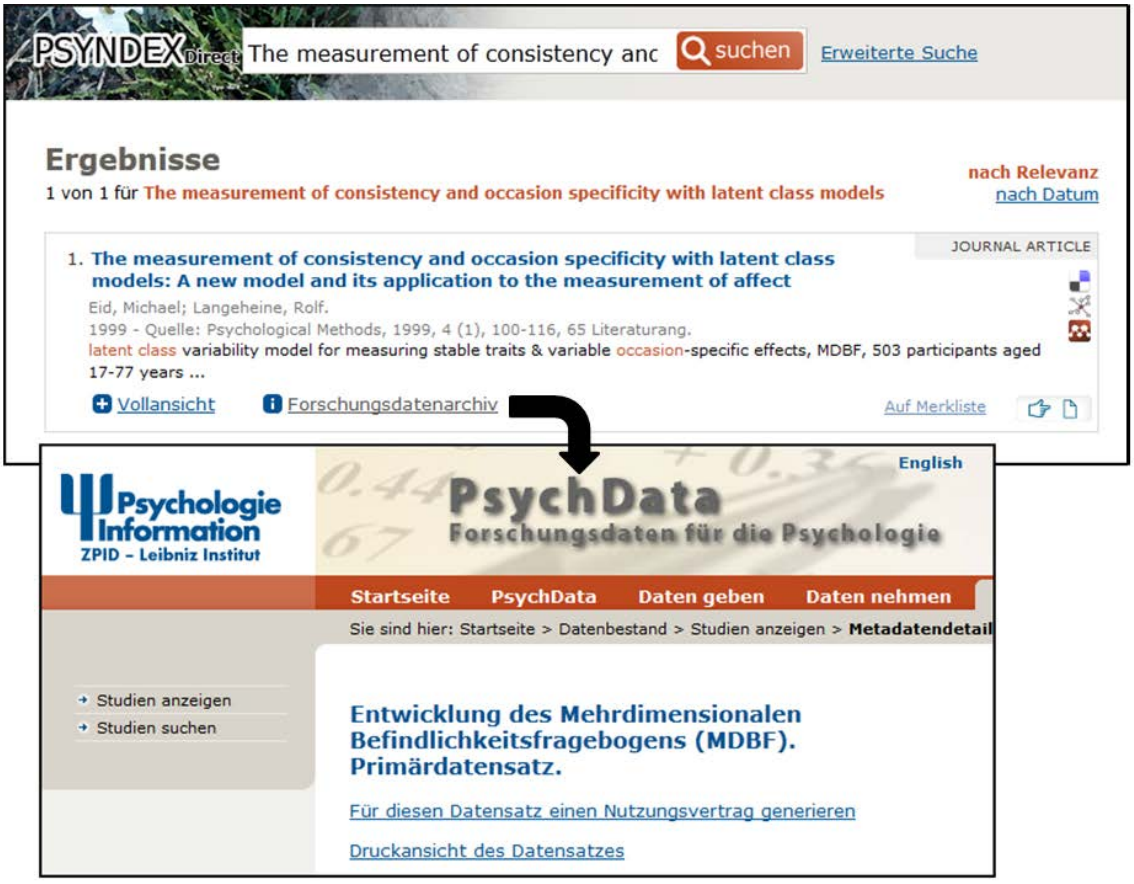


Abbildung 4. Verknüpfung zwischen PSYNDEX und PsychData

zwischen PsychData-Datensätzen und den sich darauf beziehenden Publikationen, die in der Literaturdatenbank PSYNDEX [35] verzeichnet sind (siehe Abbildung 4). Außerdem sind, sofern

[35] http://psyndexdirect.zpid.de/ [13.05.2013]

ein Datengeber über ein öffentliches Profil in der Autorendatenbank Psych-Authors [36] verfügt, dessen Datensätze auch im Autorenprofil verzeichnet.

PsychData ist ein speziell für die Psychologie eingerichtetes Datenarchiv für quantitative digitale Forschungsdaten. Die Archivierung von Daten kann retrospektiv oder forschungsbegleitend erfolgen und umfasst als Bestandteile die Forschungsdaten (als ASCII-kodierte rechteckige Datenmatrizen), Kodebuch, Studienmetadaten und sonstige Dokumente. Für alle bei PsychData archivierten Forschungsdatensätze wird ein DOI registriert.

Rechtliche Aspekte, besonders bzgl. Datenschutz, werden in Datengeber- und Datennehmerverträgen geregelt.

4.2 Manual zur Nutzung der PsychData-Dienstleistungen

4.2.1 MyPsychData

MyPsychData [37] (früher: PsychDataExtern) ist ein webbasiertes Tool, das Sie bei der forschungsbegleitenden Dokumentation ihrer Daten unterstützen soll. Durch ein strukturiertes Webinterface können parallel zu einem laufenden Forschungsprojekt die Daten und die relevanten Metadaten direkt erfasst werden, wenn sie anfallen. Die eingegebenen Daten und Metadaten werden automatisch auf Konsistenz, z.B. bezüglich des Wertebereichs, geprüft und sind so formatiert, dass sie nach Abschluss des Forschungsprojekts und eines Datengebervertrags direkt von PsychData zur Archivierung übernommen werden können. Auch das Ausfüllen eines gesonderten Studienmetadatenformulars, das bei der „herkömmlichen“ Übergabeprozedur (s. Kapitel 4.2.2) nötig ist, entfällt.

Weiterhin bietet MyPsychData die Möglichkeit der kooperativen Bearbeitung von Daten und Metadaten, indem Sie anderen MyPsychData-Nutzern Zugriffsrechte einräumen können. Damit können Daten und Metadaten kooperativ an einem zentralen Aufbewahrungsort verwaltet werden, so dass keine Probleme mit nicht synchronisierten Dateiversionen entstehen. Voraussetzung für eine Freigabe an andere Nutzer ist jedoch, dass die Daten ausreichend anonymisiert sind. Die über MyPsychData gehosteten Daten und Metadaten befinden sich auf den Servern des ZPID in einer sicheren und vertrauenswürdigen Speicherumgebung.

Legen Sie zur Nutzung von MyPsychData zunächst über den Link auf der Startseite ein neues Benutzerprofil an. Sobald Ihr Konto freigeschaltet wurde, gelangen Sie über den Login zur Hauptbenutzeroberfläche (s. Abbildung 5). Über die Hauptnavigationsleiste links können Sie zwischen den Funktionen der Oberfläche wechseln.

[36] http://www.psychauthors.de/ [13.05.2013]
[37] http://mypsychdata.zpid.de/ [14.05.2013]

Abbildung 5. MyPsychData-Webinterface

Über die Schaltfläche „Studie auswählen“ können Sie eine neue „Studie“ (im Sinne des oben definierten Archivierungspakets) anlegen. Dort werden auch alle bereits erstellten Studien zusammen mit den jeweils dazugehörigen Forschungdatensätzen angezeigt. Um die Studienmetadaten zu bearbeiten, Kodebücher zu erstellen und Daten hochzuladen, „aktivieren“ Sie durch einfaches Anklicken eine Studie. Dadurch werden weitere Menüpunkte in der Hauptnavigationsleiste zugänglich. Über die Schaltfläche „Studiendokumentation“ können Sie die Studienmetadaten eingeben, die in einzelnen Formularfeldern abgefragt werden. Erläuterungen, die neben allen Metadatenelementen angegeben sind, helfen beim Ausfüllen. Eine besonders wichtige Angabe ist hier der Anonymisierungsstatus der Daten (unter dem Reiter „Angaben zu den Daten“), da vor einer eventuellen späteren Freigabe der Daten zur Nachnutzung und auch bei einer Freigabe für andere Nutzer innerhalb von MyPsychData eine Anonymisierung erfolgen muss.

Über die Schaltfläche „Kodebücher“ der Hauptnavigationsleiste können Sie ein oder mehrere Kodebücher für die derzeit aktivierte Studie anlegen. Erst wenn ein Kodebuch angelegt ist, kann der dazugehörige Datensatz in MyPsychData geladen werden (da erst dann eine Fehlerprüfung auf Grundlage des Kodebuchs möglich ist). Jedem Kodebuch ist genau ein Datensatz zugeordnet. Die logischen Relationen der Objekte in MyPsychData sind in Abbildung 6 veranschaulicht.

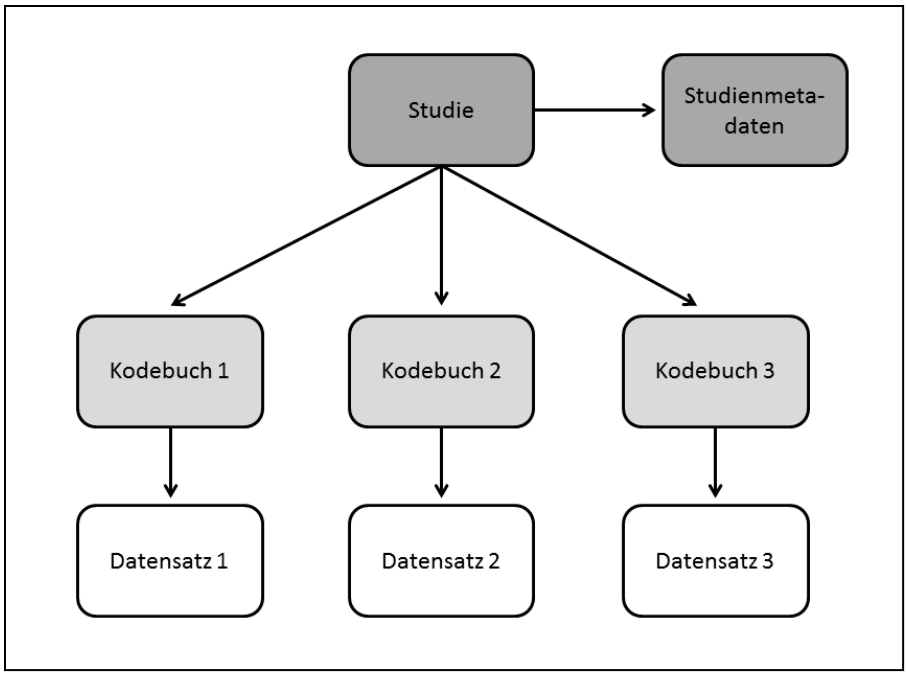


Abbildung 6. Logische Relationen zwischen den Objekten in MyPsychData

Sobald Sie ein neues Kodebuch angelegt haben, können Sie es mit Inhalt füllen, indem Sie über die Hauptnavigationsleiste eine Kodebuch-Datei hochladen. Es muss sich dabei um eine txt-Datei handeln, die gemäß der in Kapitel 4.2.2 beschriebenen Syntax aufgebaut ist; die Syntaxbeschreibung ist auch über den Link „Kodebuchregeln“ unter dem Upload-Formular für Kodebücher in MyPsychData verfügbar. Alternativ können Sie die Angaben auch direkt in MyPsychData machen. In diesem Fall müssen Sie mit der Syntax nicht vertraut sein, da die Eingabe über vorstrukturierte Web-Formularfelder mit Erläuterungen erfolgt. Da die Angaben aber jeweils einzeln für jede Variable gemacht werden müssen, kann die Offline-Erstellung über einen Texteditor vor allem bei einer großen Anzahl ähnlicher Variablen wesentlich schneller sein.

Über die Schaltfläche „anzeigen“ können Sie sich den Inhalt des Kodebuchs anzeigen lassen und dieses auch als syntaxkonforme txt-Datei exportieren. Statt ein Kodebuch von Grund auf neu zu erstellen, kann auch ein bestehendes Kodebuch kopiert werden. Dies ist vor allem nützlich, wenn dieselben Variablen erneut erhoben werden, beispielsweise bei wiederholten Befragungen. Bestehende Kodebücher können über die Hauptnavigationsleiste auch wieder aus MyPsychData gelöscht werden. Bitte beachten Sie aber, dass der mit diesem Kodebuch verbundene Forschungsdatensatz dabei mitgelöscht wird.

Nachdem das Kodebuch erstellt ist, können Sie über „Forschungsdaten“ eine mit diesem Kodebuch konforme Datenmatrix in MyPsychData hochladen. Es muss sich bei der hochgeladenen Datei um eine tabulator-delimitierte txt-Datei handeln, deren Kopfzeile die im Kodebuch angegebenen Variablennamen in derselben Reihenfolge enthält. Für Unterstützung bei der Konvertierung aus anderen Dateiformaten wenden Sie sich bitte an das PsychData-Team. Beim Hochla-

den erfolgt eine Prüfung auf Konsistenz mit den Kodebuch-Angaben (Variablennamen, Wertebereich, fehlende Werte, …). Bei Inkonsistenzen werden die entsprechenden Fehler angezeigt (die Daten werden dann nicht hochgeladen, es muss also nichts innerhalb von MyPsychData gelöscht werden). Der Forschungsdatensatz kann erst dann hochgeladen werden, wenn alle gemeldeten Fehler bereinigt sind. Somit wird ein Mindest-Qualitätsstandard an die Forschungsdaten und ihre zugehörigen Variablenbeschreibungen geleistet. War das Hochladen erfolgreich, können Sie sich die Datenmatrix sowie die Häufigkeitsverteilung der einzelnen Variablen anzeigen lassen und die Datenmatrix exportieren. Ein hochgeladener Datensatz kann wieder gelöscht werden. Das zugrundeliegende Kodebuch ist davon nicht betroffen.

Erstellte Studien und die damit zusammenhängenden Metadaten, Kodebücher und Daten können zur Betrachtung und Bearbeitung für andere bei MyPsychData registrierte Nutzer freigegeben werden. Wählen Sie dazu unter „Data Sharing“ in der Hauptnavigationsleiste die entsprechende Studie aus. Es wird eine Liste aller registrierten Benutzer angezeigt, für die Sie einzeln bestimmte Nutzerrechte festlegen können: Lesen, Lesen und Schreiben oder Vollzugriff (Lesen, Schreiben und Recht zur Rechtevergabe). Bitte anonymisieren Sie die Datensätze, bevor Sie sie zum Zugriff für andere freigeben.

„Data Sharing“ bezieht sich hier nur auf das MyPsychData-interne Data Sharing; für eine vollumfängliche Übergabe zur Archivierung an PsychData ist der Abschluss eines Datengebervertrags nötig. Hierfür kontaktieren Sie bitte das PsychData-Team unter dem Menüpunkt „Kontakt“.

MyPsychData ist ein webbasiertes Tool zur (forschungsbegleitenden) Datendokumentation und Selbstarchivierung. Nach der kostenfreien Registrierung muss zunächst eine Studie angelegt werden, dann können Kodebücher erstellt und Forschungsdaten hochgeladen werden. Automatische Qualitätskontrollen finden statt (z.B. Konsistenzprüfung von Kodebuch und Forschungsdaten).

Innerhalb von MyPsychData können Zugriffrechte an andere Benutzer vergeben werden, so dass eine kooperative Bearbeitung von Daten möglich ist.

4.2.2 Daten geben

4.2.2.1. Ablauf der Datenübergabe

Wenn Sie Ihre Daten bei PsychData archivieren lassen wollen, wird im Folgenden der Ablauf erklärt (s. Abbildung 7). Sind bereits Daten und/oder Metadaten in MyPsychData eingestellt, können entsprechend einige oder alle der Schritte 3 bis 5 entfallen.

1. Kontaktaufnahme mit PsychData,
2. Abschluss des Datengebervertrags,
3. Erstellung der Studienmetadaten,
4. Erstellung des Kodebuchs (bzw. der Kodebücher),

5. Umwandlung der Forschungsdaten in eine archivierbare Form,
6. Zusammenstellung weiterer relevanter Dateien,
7. Kontrolle der Metadaten und Forschungsdaten durch PsychData,
8. Abschließende Bestätigung des Datengebers,
9. Speicherung der Forschungsdaten im PsychData-Archiv,
10. Veröffentlichung der Metadaten auf der Website von PsychData und Vergabe eine DOI (zur Zitation).

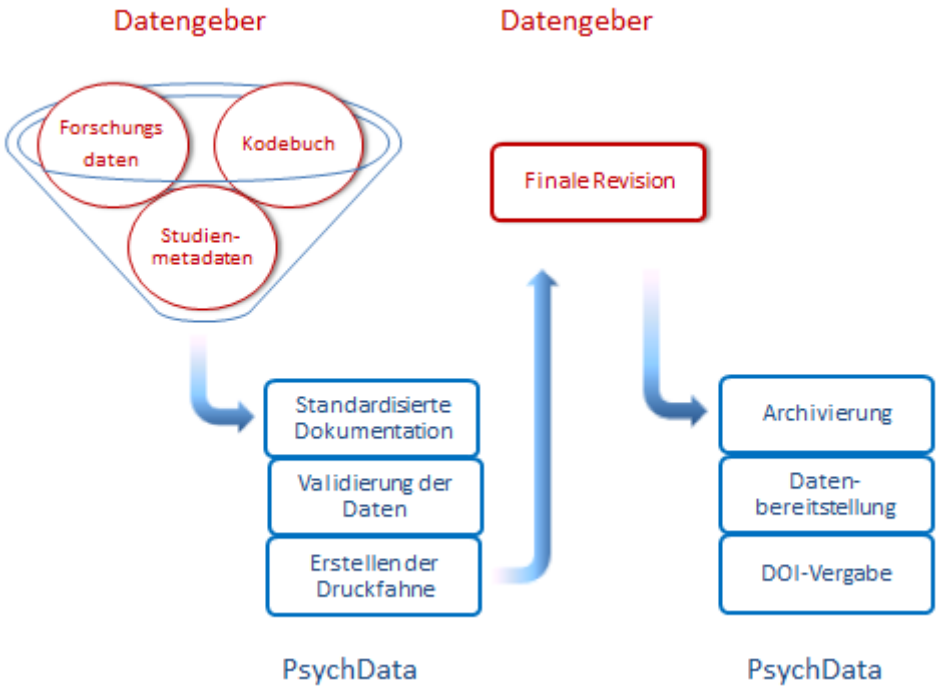


Abbildung 7. Arbeitsteilung zwischen Datengeber und PsychData

Sie nehmen per Telefon, E-Mail oder über das Web-Kontaktformular unter „Daten geben > Vertrag“ Kontakt mit PsychData auf, um die grundsätzliche Eignung Ihres Datensatzes für PsychData abzuklären. Danach wird durch Abschluss eines Datengebervertrags Rechtssicherheit hergestellt (vgl. die Ausführungen in Kapitel 4.1 zu den Datensätzen, für die PsychData offen ist und zu den vertraglichen Regelungen). Ein Blankovertrag für die Datenübernahme ist auf der Website einsehbar. Zusatzvereinbarungen können individuell ausgehandelt werden. Sobald alles geklärt ist, erfolgt ein postalischer Austausch der unterschriebenen Verträge.

Anschließend müssen die Metadaten, die Ihre Studie beschreiben, zusammengestellt werden. Dazu können Sie das über die Website herunterladbare PDF-Metadatenformular nutzen, das Erläuterungen zu den einzelnen Metadatenelementen enthält. Das ausgefüllte Formular lässt sich (auch mit dem herkömmlichen Adobe Reader) speichern und per E-Mail an PsychData senden. Alternativ können die Metadaten in MyPsychData erfasst werden.

Die dann notwendige Erstellung eines mit der PsychData-Kodebuchsyntax konformen Kodebuchs wird im folgenden Unterabschnitt erläutert.

Als nächstes ist in Zusammenarbeit zwischen Ihnen und PsychData zu gewährleisten, dass die Daten entsprechend den Archivstandards aufbereitet sind. Die Daten sollten als tabulatorgetrennte Textdatei vorliegen, wobei jeder Fall eine Zeile ist. Achten Sie beim Export darauf, dass eventuell in den Daten vorhandene Tabulatorzeichen vorher durch andere Zeichen ersetzt wurden, z.B. in freien Antwortfeldern. Andere Formate können nach Absprache auch vom PsychData-Team umgewandelt werden.

Sofern Sie Ihre Studienmetadaten, Forschungsdaten und/oder Ihr Kodebuch bereits in MyPsychData eingegeben und verwaltet haben, teilen Sie dies den PsychData-Mitarbeitern mit. Die Datenüberführung erfolgt dann automatisch.

Dann werden Kodebuch, Forschungsdatenmatrix und Studienmetadaten von PsychData-Mitarbeitern auf Konsistenz geprüft. Angaben in den Metadaten werden, wenn notwendig, in kontrolliertes Vokabular überführt (z.B. Schlagwörter). Falls erforderlich werden auch noch Anonymisierungsschritte durchgeführt. Zu allen Änderungen wird Rücksprache mit dem Datengeber gehalten.

Nachdem die Daten kontrolliert wurden, wird die Studiendokumentation den Datengebern ein letztes Mal zur abschließenden Prüfung vorgelegt. Erst wenn diese Freigabe erfolgt ist, werden die Daten in den permanenten Bereich des Datenarchivs übernommen und die Metadaten auf der PsychData-Webseite zusammen mit dem für den Datensatz registrierten DOI veröffentlicht.

Zur Datenübergabe an PsychData wird zunächst ein Datengebervertrag abgeschlossen und anschließend alle für die Archivierung notwendigen Materialien (Forschungsdaten, Kodebuch, Studienmetadaten) an PsychData übergeben.

Nach der Überarbeitung der Materialien erhält der Datengeber die Studiendokumentation zur abschließenden Überprüfung. Erst dann werden die Daten zur Nachnutzung bereitgestellt und erhalten einen DOI zur Zitation.

4.2.2.2. Erstellung eines syntaxkonformen Kodebuchs

Ein PsychData-Kodebuch ist eine einfache Plaintext-Datei, die entweder selbst im Texteditor oder über MyPsychData erstellt werden kann. Als Anleitung für Ersteres wird im Folgenden die Struktur des Kodebuchs erläutert.

Das **Kodebuch hat die folgende verbindliche Grundstruktur**:

- Jede Variable im Forschungsdatensatz ist durch einen Block von Metadatenelementen repräsentiert (siehe unten); die Elemente innerhalb eines Blocks sind durch Zeilenumbrüche getrennt;

- Die Blöcke stehen in exakt der Reihenfolge, in der die Variablen im Forschungsdatensatz angeordnet sind;
- Die einzelnen Blöcke sind voneinander durch eine Leerzeile getrennt.

Der **Aufbau des Metadatenblocks zu einer Variablen** ist wie folgt. <In spitzen Klammern Stehendes> ist ein Platzhalter zur Benennung des Metadatenelements. Die Angabe in runden Klammern hinter dem Element gibt an, wie viele Wiederholungen des Elements (in einer jeweils eigenen Zeile) vorkommen dürfen. Anführungsstriche und geschweifte Klammern sind Teil der Syntax und müssen so verwendet werden, wie sie hier stehen:

```
<Variablenname> (1)
<Variablenbeschreibung> (1)
"<Variablenitem>" (1)
{<Wertemenge>} (1)
{<Fehlermenge>} (1)
<Wert> "<Wertelabel>" (1 oder mehr)
<Fehlender Wert> "<Label fehlender Wert>" (1 oder mehr)
```

Anhand der folgenden Beispielvariablen wird nun die Syntax innerhalb der einzelnen Metadatenelemente weiter erläutert, sowie gegebenenfalls, welche inhaltlichen Angaben das jeweilige Element enthalten sollte:

```
FAM_ZU
Zufriedenheit mit der Familie (Filterfrage)
"Wie zufrieden sind Sie im Augenblick mit Ihrer Familie?"
{1;2;3;4}
{97;98;99}
1 "nicht zufrieden"
2 ""
3 ""
4 "sehr zufrieden"
97 "fehlender Wert (Vp verweigert)"
98 "fehlender Wert (Vp weiß es nicht)"
99 "fehlender Wert (Frage nicht zutreffend)"
```

REAKZEIT
Reaktionszeit
"Sobald Sie einen rosa Elefanten sehen, drücken Sie auf den Knopf."
{0-1000}
{9999}
0-1000 "Reaktionszeit in msec"
9999 "Fehlender Wert"

GEDICHT
Liebstes Gedicht der Versuchsperson
"Bitte geben Sie den Titel Ihres liebsten Gedichtes an."
{Zeichenkette}
{9991;9992}
Zeichenkette "Titel des Lieblingsgedichts"
9991 "Fehlender Wert: Keine Angabe"
9992 "Fehlender Wert: Unleserlich"

Der **Variablenname** darf nur Großbuchstaben, Zahlen und den Unterstrich beinhalten:

FAM_ZU

Das **Variablenlabel** kann, abgesehen davon, dass nur die durch die Textenkodierung zugelassenen Zeichen sowie kein Zeilenumbruch verwendet werden dürfen, eine beliebige Zeichenkette sein. Natürlich sollte sie, dem Sinn und Zweck eines Variablenlabels entsprechend, möglichst kompakt und zugleich aussagekräftig sein. Hier sollte auch das Erhebungsinstrument festgehalten werden, falls es sich um ein standarisiertes Messinstrument handelt sowie nützliche Zusatzinformationen wie Angaben zu Filtervariablen, Matchvariablen, der Messzeitpunkt bei mehrmaligen Erhebungen, etc.

FAM_ZU
Zufriedenheit mit der Familie (Filterfrage)

Das **Variablenitem** ist eine Zeichenkette wie das Variablenlabel, wird aber durch hochgestellte Anführungszeichen eingeschlossen. Es sollte den exakten Inhalt des Fragebogenitems oder der Instruktionsanweisung enthalten. Zu der exakten Beschreibung gehört auch die Messgröße (z.B. cm, Alter in Jahren, msec) und im Fall von abgeleiteten Variablen die genaue Berechnungsformel. Existiert keine solche Beschreibung, werden nur Anführungszeichen ohne Inhalt angegeben.

FAM_ZU
Zufriedenheit mit der Familie (Filterfrage)
"Wie zufrieden sind Sie im Augenblick mit Ihrer Familie?"

Die **Wertemenge** enthält alle zulässigen, gültigen (=nicht fehlenden) Werte der Variablen. Die Wertemenge ist durch geschweifte Klammern eingeschlossen. Jeder Wert der Wertemenge muss auch bei den Wertelabels aufgeführt werden (s.u.). Es gibt **drei mögliche Varianten**, eine Wertemenge anzugeben:

1. **Variable mit Wertekategorien**: Eine Auflistung der Variablenwerte für alle Kategorien, getrennt durch Semikola. Jede Variable, für die eine Wert-Label-Zuordnung vorliegt, sollte als solch eine Menge einzelner Elemente angegeben werden. Werte können sowohl numerisch als auch nichtnumerisch sein (ersteres ist im Allgemeinen aber vorzuziehen).

   FAM_ZU
   Zufriedenheit mit der Familie
   "Wie zufrieden sind Sie im Augenblick mit Ihrer Familie?"
   {1;2;3;4}

2. **Intervall**: Durch Bindestrich getrennte Angabe von Minimal- und Maximalwert. In der Regel zur Verwendung bei mit Messgrößen behafteten, stetigen Zahlen, kann aber auch für Integer-Variablen (z.B. Anzahl der Kinder) verwendet werden.

   REAKZEIT
   Reaktionszeit
   "Sobald Sie einen rosa Elefanten sehen, drücken Sie auf den Knopf."
   {0-1000}

3. **Zeichenkette**: Das Wort „Zeichenkette“. Der gültige Wertebereich umfasst dann beliebige alphanumerischen Zeichenketten (d.h. sowohl Buchstaben als auch Zahlen). Zur Verwen-

dung bei offenen Fragen mit freiem Inhalt, aber auch zur Angabe spezieller Formate wie Datumsangaben (die dann im Wertelabel-Element genauer spezifiziert werden sollten).

GEDICHT
Liebstes Gedicht der Versuchsperson
"Bitte geben Sie den Titel Ihres liebsten Gedichtes an."
{Zeichenkette}

Die **Menge fehlender Werte** wird genauso angegeben wie die Wertemenge für Variablen mit Wertekategorien, darf aber keine Werte aus der Wertemenge enthalten. Bei Zeichenkettenvariablen ist besonders auf die Wahl eines Werts zu achten, der unter den gültigen Werten nicht vorkommen kann, da hier in MyPsychData keine automatische Kontrolle stattfinden kann. Jeder Wert der Fehlermenge muss auch bei der Fehlerwert-Fehlerlabel Zuordnung vorkommen (s.u.).

Im PsychData-Kodebuch müssen zu jeder Variablen definierte fehlende Werte angegeben werden. Auch bei Variablen, die keine fehlenden Werte enthalten, ist es momentan noch notwendig, diese im Kodebuch (dann willkürlich) zu definieren. Grund hierfür ist der aktuelle Entwicklungsstand der automatischen Kodebuch-Kontrollen, die ablaufen, wenn das Kodebuch hochgeladen wird. Diese erfordern die Angabe von fehlenden Werten bei allen Variablen, um korrekt zu funktionieren. Systemdefinierte fehlende Werte (Sysmis), d.h. leere Felder im Datensatz müssen in einen definierten Fehlerwert umgewandelt werden, da man nie sicher sein kann, ob es sich um Eingabefehler handelt oder um tatsächlich fehlende Werte. Bei Formatumwandlungen können zusätzlich Fehler entstehen, wenn Werte „verrutschen“. Außerdem bietet die Wertedefinition den Vorteil, zwischen verschiedenen Arten von fehlenden Werten differenzieren und so auch eine Fehleranalyse durchführen zu können.

FAM_ZU
Zufriedenheit mit der Familie (Filterfrage)
"Wie zufrieden sind Sie im Augenblick mit Ihrer Familie?"
{1;2;3;4}
{97;98;99}

Für jeden in der Wertemenge enthaltenen Wert muss ein **Wertelabel** definiert sein. Es wird immer zuerst der Wert, dahinter nach einem Leerzeichen das Label in hochgestellten Anführungszeichen angegeben. Existiert kein Label für einen Wert, werden einfach die Anführungszeichen ohne Inhalt angegeben. Es muss aber mindestens für einen Wert ein Label definiert sein (beachte die folgenden

Ausführungen zu Variablen mit Zeichenketten und Intervall-Wertebereich). Gemäß der drei Varianten bei der Wertemenge sind die zulässigen Optionen hier:

1. **Wertelabel bei Variable mit Wertekategorien**: Pro Wertkategorie werden Wert und Label zusammen in einer eigenen Zeile angegeben.

```
FAM_ZU
Zufriedenheit mit der Familie (Filterfrage)
"Wie zufrieden sind Sie im Augenblick mit Ihrer Familie?"
{1;2;3;4}
{97;98;99}
1 "nicht zufrieden"
2 ""
3 ""
4 "sehr zufrieden"
```

2. **Wertelabel bei Intervall-Wertebereich**: Es wird das Intervall und anschließend eine Spezifikation der Messeinheit bzw. -größe in Anführungszeichen angegeben.

```
REAKZEIT
Reaktionszeit
"Sobald Sie einen rosa Elefanten sehen, drücken Sie auf den Knopf."
{0-1000}
{9999}
0-1000 "Reaktionszeit in msec"
```

3. **Wertelabel bei Zeichenketten-Variablen**: Es wird das Wort „Zeichenkette“ und anschließend eine Beschreibung des Variableninhalts angegeben.

```
GEDICHT
Liebstes Gedicht der Versuchsperson
"Bitte geben Sie den Titel Ihres liebsten Gedichtes an."
{Zeichenkette}
{9991;9992}
Zeichenkette "Titel des Lieblingsgedichts"
```

Bei den **Labels für fehlende Werte** wird schließlich genauso verfahren wie bei den Labels für Variablen mit Wertekategorien, d.h. pro fehlendem Wert eine Zeile mit Wert und Label in Anführungszeichen. Idealerweise sollte für jeden echten Fehlenden-Wert-Typ eine aussagekräftiges Label, z.B. in der Form „Fehlender Wert: Ursache“ angegeben werden.

FAM_ZU
Zufriedenheit mit der Familie (Filterfrage)
"Wie zufrieden sind Sie im Augenblick mit Ihrer Familie?"
{1;2;3;4}
{97;98;99}
1 "nicht zufrieden"
2 ""
3 ""
4 "sehr zufrieden"
97 "Fehlender Wert (Vp verweigert)"
98 "Fehlender Wert (Vp weiß es nicht)"
99 "Fehlender Wert (Frage nicht zutreffend)"

Zur Veranschaulichung abschließend einige **weitere Beispiele für syntaxkonforme Variablen**:

Tabelle 3. PsychData-Kodebuchsyntaxbeispiele für verschiedene Variablentypen

Variablentyp	**Kodebuchsyntax**
Nominale Variable, numerisch kodiert	SEX Geschlecht der Versuchsperson "Bitte geben Sie Ihr Geschlecht an." {0;1} {9} 0 "weiblich" 1 "männlich" 9 "fehlender Wert (keine Angabe)"
Nominale Variable, durch Buchstaben kodiert	SEX Geschlecht der Versuchsperson "Bitte geben Sie Ihr Geschlecht an."

	{m;w} {f;k} w "weiblich" m "männlich" f "fehlender Wert: Angabe fehlt" k "fehlender Wert: Vp will keine Angabe machen"
Freie Antwort	SEX_FR Geschlecht der Versuchsperson "Bitte geben Sie Ihr Geschlecht an." {Zeichenkette} {#} Zeichenkette "Freie Beschreibung des Geschlechts" # "fehlender Wert: keine Angabe"
Ganzzahlige Variable	KINDER Anzahl der Kinder von Versuchsperson "Bitte geben Sie die Zahl Ihrer Kinder an." {1-100} {999} 1-100 "Anzahl der Kinder" 999 "fehlender Wert: keine Angabe"
Stetige Variable	GELD Kleingeld der Versuchsperson "Wieviel Kleingeld haben Sie in der Hosentasche?" {0,00-100,00} {-1;-2} 0,00-100,00 "Kleingeld in Euro" -1 "fehlender Wert: keine Angabe" -2 "fehlender Wert: Vp hat keine Hosen an"
Spezielle Formatkonvention	ALT_EPSY Alter im Format der Entwicklungspsychologie "Berechnung s. Datei 'abg_var.txt'" {Zeichenkette}

	{999} Zeichenkette "Alter im Format Jahre ; Monate" 999 "fehlender Wert: keine Angabe"
Variable ohne Itembeschreibung	LAUFNR Laufende Durchnummerierung der Vpn "" {1-100000} {-1} 1-100000 "Laufnummer der Vp" -1 "Fehlender Wert"

Das PsychData-Kodebuchschema besteht aus den grundlegenden Elementen Variablenname, aussagekräftiges Variablenlabel, Fragetext oder Instruktionsanweisung (wenn vorhanden), gültige Wertemenge, definierte fehlende Werte, Labels für gültige und fehlende Werte.

4.2.3 Daten nehmen

Wenn Sie Interesse an der Nachnutzung psychologischer Forschungsdaten haben, können Sie sich anhand der auf der PsychData integrierten Suchfunktionalitäten zunächst ein Bild von den vorhandenen Studien machen.

Zur Suche nach interessanten Forschungsdaten innerhalb von PsychData stehen unter dem Menüpunkt „Datenbestand“ auf der PsychData-Website zwei Wege offen:

- Der Datenbestand kann über die Psychologie-Suchmaschine PsychSpider (s. Kapitel 3.2.1) per Suchbegriff abgesucht werden. Es kann auch über die gesamte Kollektion „Forschungsdaten“ (PsychData und andere psychologierelevante Forschungsdatenarchive) gesucht werden;
- Unter „Datenbestand“ sind die archivierten Studien nach Teildisziplinen der Psychologie geordnet aufgelistet.

In der Detailansicht zu einer Studie sind die Studienmetadaten tabellarisch dargestellt. Diese Angaben sind in der Regel umfangreich genug, um eine Beurteilung der Relevanz für die eigene Forschung zu ermöglichen.

Sollten Sie einen für Sie relevanten Datensatz ausfindig gemacht haben, können Sie die Studie durch Abschluss eines Datennehmervertrags bestellen. Durch das Ausfüllen des Kontaktformulars unter „Daten nehmen > Vertrag“ kann (unter Angabe der Studie, die man nachnutzen

möchte) ein Datennehmervertrag erzeugt werden. Auf dieser Webseite ist auch ein Blankovertrag herunterladbar. Die **Kernvorgaben des Datennehmervertrags** sind:

- Das überlassene Material darf ausschließlich für wissenschaftliche Forschung und Lehre genutzt werden;
- Das überlassene Material darf nicht an Dritte weitergegeben werden; bei einer Verwendung der Daten in einer Forschungsprojektgruppe oder Lehrveranstaltung ist sicherzustellen, dass von den beteiligten Personen die Nutzungsbedingungen eingehalten werden;
- Bei jeder Veröffentlichung, die ganz oder teilweise auf dem überlassenen Datenmaterial und den zugehörigen Materialien beruht, müssen sowohl die Datengeber als auch das ZPID kenntlich gemacht werden (Zitationspflicht);
- Das ZPID ist über Publikationen zu informieren, die auf dem überlassenen Datenmaterial und den zugehörigen Materialien beruhen;
- Es dürfen keine Versuche der Reidentifikation und Kontaktierung der Befragten unternommen werden.

Schicken Sie den ausgefüllten und unterschriebenen Vertrag bitte in zweifacher Ausfertigung postalisch an PsychData. Die Forschungsdaten und zugehörigen Kodebücher werden Ihnen dann auf einer revisionssicheren CD-ROM zugeschickt. Sollte es bei der Nachnutzung und Interpretation der Daten Schwierigkeiten geben, können Sie sich jederzeit an das PsychData-Team wenden.

Die über PsychData bereitgestellten Forschungsdaten können mit der Suchmaschine PsychSpider abgesucht oder auf der Webseite thematisch geordnet durchgeschaut werden.

Die Daten dürfen nur für die wissenschaftliche Forschung und Lehre verwendet werden. Um Daten nachzunutzen, muss ein Datennehmervertrag unterzeichnet werden.

5. Literatur

Altenhöner, R. & Oellers, C. (Hrsg.). (2012). Langzeitarchivierung von Forschungsdaten. Standards und disziplinspezifische Lösungen. Berlin: Scivero Verlag.

APA. (2010). Ethical Principles of Psychologists and Code of Conduct. Washington, DC: American Psychological Association. Zugriff am 01.05.2013. Verfügbar unter http://www.apa.org/ethics/code/principles.pdf

Aschenbrenner, A. & Neuroth, H. (2011). Forschungsdaten-Repositorien. In S. Büttner, H.-C. Hobohm & L. Müller (Hrsg.), Handbuch Forschungsdatenmanagement (S. 101–114). Bad Honnef: Bock + Herchen.

Bartholomäus, U. & Schnabel, U. (1997). Betrüger im Labor: Die deutsche Forschung hat ihren Fall. Wie schützt man sich vor Fälschern? Die Zeit, 13.06.1997. Zugriff am 20.06.2013. Verfügbar unter http://www.zeit.de/1997/25/falsch.txt.19970613.xml

Becher, T. & Trowler, P. R. (2001). Academic tribes and territories. Intellectual enquiry and the culture of disciplines (2. Aufl.). Buckingham: Open University Press.

Becker, P. (2004). Trierer Persönlichkeitsfragebogen (TPF). Primärdaten der Eichstichprobe (Version 1)[Files auf CD-ROM]. Trier: Psychologisches Datenarchiv PsychData des Leibniz-Zentrums für Psychologische Information und Dokumentation ZPID. Zugriff am 15.05.2013. Verfügbar unter http://dx.doi.org/10.5160/psychdata.brpr88pe99

Bortz, J. & Döring, N. (2002). Forschungsmethoden und Evaluation. Für Human- und Sozialwissenschaftler (3. Aufl.). Berlin: Springer.

Breckler, S. J. (2009). Dealing with data. Monitor on Psychology, 40 (2), 41.

Bühner, M. (2011). Einführung in die Test- und Fragebogenkonstruktion (3. Aufl.). München: Pearson Studium.

Büttner, S., Hobohm, H.-C. & Müller, L. (Hrsg.). (2011). Handbuch Forschungsdatenmanagement. Bad Honnef: Bock + Herchen. Zugriff am 15.05.2013. Verfügbar unter http://nbn-resolving.de/urn/resolver.pl?urn:nbn:de:kobv:525-opus-2412

Curtin University Library. (2013). Research data management. Zugriff am 15.05.2013. Verfügbar unter http://libguides.library.curtin.edu.au/research-data-management

de Cock Buning, M., van Dinther, B., Jeppersen de Boer, C. G. & Ringnalda, A. (2011b). The legal status of research data in Germany. Annex 3 to the Knowledge Exchange report „The legal status of research data in the Knowledge Exchange partner countries". Utrecht: Centre for In-

tellectual Property Law. Zugriff am 15.05.2013. Verfügbar unter http://www.knowledge-exchange.info/default.aspx?id=461

de Cock Buning, M., van Dinther, B., Jeppersen de Boer, C. G. & Ringnalda, A. (2011a). The legal status of research data in the Knowledge Exchange partner countries. Utrecht: Centre for Intellectual Property Law. Zugriff am 15.05.2013. Verfügbar unter http://www.knowledge-exchange.info/default.aspx?id=461

DCC. (n.d.). DCC Curation Lifecycle Model. Zugriff am 15.05.2013. Verfügbar unter http://www.dcc.ac.uk/resources/curation-lifecycle-model

DFG. (1998). Vorschläge zur Sicherung guter wissenschaftlicher Praxis. Empfehlungen der Kommission „Selbstkontrolle in der Wissenschaft“. Weinheim: Wiley-VCH.

DFG. (2009). Empfehlungen zur gesicherten Aufbewahrung und Bereitstellung digitaler Forschungsprimärdaten. Bonn: Deutsche Forschungsgemeinschaft. Zugriff am 03.05.2013. Verfügbar unter http://www.dfg.de/download/pdf/foerderung/programme/lis/ua_inf_empfehlungen_200901.pdf

DFG. (2012). DFG-Vordruck 54.01 – 10/12: Leitfaden für die Antragstellung - Projektanträge. Bonn: DFG. Zugriff am 15.05.2013. Verfügbar unter http://www.dfg.de/formulare/54_01/54_01_de.pdf

DGPs. (2004). Revision der auf die Forschung bezogenen ethischen Richtlinien. Münster: Deutsche Gesellschaft für Psychologie. Zugriff am 15.05.2013. Verfügbar unter http://www.dgps.de/dgps/aufgaben/ethikrl2004.pdf

Dietz, P., Striegel, H., Franke, A. G., Lieb, K., Simon, P. & Ulrich, R. (2013). Randomized Response Estimates for the 12-Month Prevalence of Cognitive-Enhancing Drug Use in University Students. Pharmacotherapy, 33, 44–50.

Donnelly, M. & Jones, S. (2011). Checklist for a Data Management Plan. Version 3.0. Edinburgh: Digital Curation Centre. Zugriff am 15.05.2013. Verfügbar unter http://www.dcc.ac.uk/webfm_send/431

Duden (1976). Das große Wörterbuch der deutschen Sprache. Mannheim, Wien, Zürich: Bibliografisches Institut.

EDINA. (n.d.). Research Data MANTRA [online course]. Edinburgh: EDINA and Data Library, University of Edinburgh. Zugriff am 15.05.2013. Verfügbar unter http://datalib.edina.ac.uk/mantra

Enserink, M. (2012, 28. November). Final Report: Stapel Affair Points to Bigger Problems in Social Psychology. ScienceInsider. Zugriff am 15.05.2013. Verfügbar unter http://news.sciencemag.org/scienceinsider/2012/11/final-report-stapel-affair-point.html

ESRC. (2013). ESRC Research Data Policy. Swindon: Economic and Social Research Council. Zugriff am 15.05.2013. Verfügbar unter http://www.esrc.ac.uk/_images/Research_Data_Policy_2010_tcm8-4595.pdf

Freedland, K. E. & Carney, R. M. (1992). Data management and accountability in behavioral and biomedical research. American Psychologist, 47, 640–645. Zugriff am 15.05.2013. Verfügbar unter http://dx.doi.org/10.1037/0003-066X.47.5.640

Gallagher Tuleya, L. (2007). Thesaurus of psychological index terms (11. Aufl.). Washington, DC: American Psychological Association.

GESIS. (n.d.). Forschungsdatenmanagement für Wissenschaftler/-innen. Köln: GESIS. Zugriff am 15.05.2013. Verfügbar unter http://www.gesis.org/archive-and-data-management-training-and-information-centre/forschungsdatenmanagement/

Goldacre, B. (2012). Bad pharma. How drug companies mislead doctors and harm patients. London: Fourth Estate.

Hey, A. J. G. & Trefethen A. E. (2003). The Data Deluge: An e-Science Perspective. In F. Berman, G. Fox & A. J. G. Hey (Hrsg.), Grid computing. Making the global infrastructure a reality (S. 809–824). New York: Wiley.

Hey, T., Tansley, S. & Tolle, K. (2009). The fourth paradigm. Data-intensive scientific discovery. Redmond, WA: Microsoft Research.

ICPSR. (2012). Guide to social science data preparation and archiving. Best practice through the data life cycle (5. Aufl.). Ann Arbor, MI: Inter-University Consortium for Political and Social Research. Zugriff am 15.05.2013. Verfügbar unter http://www.icpsr.umich.edu/icpsrweb/content/deposit/guide/

International Conference on Harmonisation. (1996). ICH E6(R1): Guideline for Good Clinical Practice. Genf: International Conference on Harmonisation. Zugriff am 15.05.2013. Verfügbar unter http://www.ich.org/products/guidelines/efficacy/efficacy-single/article/good-clinical-practice.html

Jensen, U., Katsanidou, A. & Zenk-Möltgen, W. (2011). Metadaten und Standards. In S. Büttner, H.-C. Hobohm & L. Müller (Hrsg.), Handbuch Forschungsdatenmanagement (S. 83–100). Bad Honnef: Bock + Herchen.

JISC. (n.d.). Research lifecycle diagram. Zugriff am 15.05.2013. Verfügbar unter http://www.jisc.ac.uk/whatwedo/campaigns/res3/jischelp.aspx

Jones, S. (2011). How to Develop a Data Management and Sharing Plan. Edinburgh: Digital Curation Centre. Zugriff am 02.05.2013. Verfügbar unter http://www.dcc.ac.uk/resources/how-guides

Key Perspectives. (2010). Data dimensions: disciplinary differences in research data sharing, reuse and long term viability. SCARP Synthesis Study. Edinburgh: Digital Curation Centre. Zugriff am 15.05.2013. Verfügbar unter http://www.dcc.ac.uk/scarp

King, G. (1995). Replication, Replication. PS: Political Science and Politics, 28, 444–452. Zugriff am 15.05.2013. Verfügbar unter http://gking.harvard.edu/files/abs/replication-abs.shtml

Klump, J. (2011). Langzeiterhaltung digitaler Forschungsdaten. In S. Büttner, H.-C. Hobohm & L. Müller (Hrsg.), Handbuch Forschungsdatenmanagement (S. 115–122). Bad Honnef: Bock + Herchen.

Kommission Zukunft der Informationsinfrastruktur. (2011). Gesamtkonzept für die Informationsinfrastruktur in Deutschland. Empfehlungen im Auftrag der Gemeinsamen Wissenschaftskonferenz des Bundes und der Länder. Berlin: Wissenschaftsgemeinschaft Gottfried Wilhelm Leibniz. Zugriff am 15.05.2013. Verfügbar unter http://www.leibniz-gemeinschaft.de/fileadmin/user_upload/downloads/Infrastruktur/KII_Gesamtkonzept.pdf

Ludwig, J. & Enke, H. (Hrsg.). (2013). Leitfaden zum Forschungsdaten-Management. Ergebnisse aus dem WissGrid-Projekt (1. Aufl.). Glückstadt: vwh. Zugriff am 15.05.2013. Verfügbar unter http://www.wissgrid.de/publikationen/Leitfaden_Data-Management-WissGrid.pdf

Max-Planck-Gesellschaft. (2009). Regeln zur Sicherung guter wissenschaftlicher Praxis. Berlin: Max-Planck-Gesellschaft. Zugriff am 15.05.2013. Verfügbar unter http://www.mpg.de/229457

McFadden, E. (2007). Management of data in clinical trials (2. Aufl.). Hoboken, NJ: Wiley-Interscience.

Merton, R. K. (1973). The normative structure of science. In R. K. Merton (Hrsg.), The sociology of science. Theoretical and empirical investigations (S. 267–278). Chicago: University of Chicago Press.

Metschke, R. & Wellbrock, R. (2002). Datenschutz in Wissenschaft und Forschung (3. Aufl.). Berlin: Berliner Beauftragter für Datenschutz und Informationsfreiheit. Zugriff am 15.05.2013. Verfügbar unter http://www.datenschutz-berlin.de/attachments/47/Materialien28.pdf?1166527077

MIT Libraries. (n.d.). Data Management and Publishing. Zugriff am 15.05.2013. Verfügbar unter http://libraries.mit.edu/guides/subjects/data-management/

Moosbrugger, H. & Kelava, A. (2012). Testtheorie und Fragebogenkonstruktion (2. Aufl.). Berlin: Springer.

nestor. (2008). nestor-Kriterien – Kriterienkatalog vertrauenswürdige digitale Langzeitarchive (Version II). Frankfurt am Main: nestor. Zugriff am 15.05.2013. Verfügbar unter http://nbn-resolving.de/urn:nbn:de:0008-2008021802

Neuroth, H. (Hrsg.). (2012). Langzeitarchivierung von Forschungsdaten. Eine Bestandsaufnahme. Glückstadt: vwh. Zugriff am 15.05.2013. Verfügbar unter http://nestor.sub.uni-goettingen.de/bestandsaufnahme/

Neuroth, H., Oßwald, A., Scheffel, R., Strathmann, S. & Huth, K. (Hrsg.). (2010). nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung. Version 2.3. Frankfurt am Main: nestor. Zugriff am 15.05.2013. Verfügbar unter http://nestor.sub.uni-goettingen.de/handbuch/

NIH. (2003). NIH Data Sharing Policy and Implementation Guidance. Bethesda, MD: National Institutes of Health. Zugriff am 15.05.2013. Verfügbar unter http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm

NSF. (n.d.). Proposal and award policies and procedures guide. Arlington, VA: National Science Foundation. Zugriff am 15.05.2013. Verfügbar unter http://www.nsf.gov/pubs/policydocs/pappguide/nsf13001/nsf13_1.pdf

OECD. (2007). OECD principles and guidelines for access to research data from public funding. Paris: Organisation für wirtschaftliche Zusammenarbeit und Entwicklung. Zugriff am 15.05.2013. Verfügbar unter http://www.oecd.org/science/sci-tech/38500813.pdf

Oxford University Administration and Services. (2012). Research Data Management. Zugriff am 15.05.2013. Verfügbar unter http://www.admin.ox.ac.uk/rdm/

Pampel, H. & Bertelmann, R. (2011). „Data Policies“ im Spannungsfeld zwischen Empfehlung und Verpflichtung. In S. Büttner, H.-C. Hobohm & L. Müller (Hrsg.), Handbuch Forschungsdatenmanagement (S. 49–61). Bad Honnef: Bock + Herchen.

Pennock, M. (2007). Digital Curation: A Life-Cycle Approach to Managing and Preserving Usable Digital Information. Bath: UKOLN. Zugriff am 15.05.2013. Verfügbar unter http://www.ukoln.ac.uk/ukoln/staff/m.pennock/publications/docs/lib-arch_curation.pdf

Plant, R. R. (2011). Guidance Notes for Completing a “Checklist for a Data Management Plan v3.0” For Researchers in the Psychological Sciences. York: DMTpsych Project, Department of Psychology, The University of York. Zugriff am 15.05.2013. Verfügbar unter http://www.sheffield.ac.uk/psychology/research/groups/dmsppsych/onestop

Research Information Network. (2008). Stewardship of digital research data: a framework of principles and guidelines. London: Research Information Network. Zugriff am 15.05.2013. Verfügbar unter http://www.rin.ac.uk/system/files/attachments/Stewardship-data-guidelines.pdf

Rising, K., Bacchetti, P. & Bero, L. (2008). Reporting bias in drug trials submitted to the Food and Drug Administration: review of publication and presentation. PLoS Med., 5 (11), e217. Zugriff am 15.05.2013. Verfügbar unter http://dx.doi.org/10.1371/journal.pmed.0050217

Rusch-Feja, D. (2001). Die Open Archives Initiative (OAI). Neue Zugangsform zu wissenschaftlichen Arbeiten? BIBLIOTHEK Forschung und Praxis, 25, 291-300. Zugriff am 15.05.2013. Verfügbar unter http://www.b2i.de/themenportale/bibliothekforschungundpraxis/bestandsuebersicht-alle-artikel/2001/

Sedlmeier, P. & Renkewitz, F. (2008). Forschungsmethoden und Statistik in der Psychologie. München: Pearson Studium.

Spindler, G. & Hillegeist, T. (2009). KoLaWiss-Projekt Arbeitspaket 4: Recht. Göttingen: SUB Göttingen. Zugriff am 15.05.2013. Verfügbar unter http://kolawiss.uni-goettingen.de/projektergebnisse/AP4_Report.pdf

Spindler, G. & Hillegeist, T. (2011). Rechtliche Probleme der elektronischen Langzeitarchivierung von Forschungsdaten. In S. Büttner, H.-C. Hobohm & L. Müller (Hrsg.), Handbuch Forschungsdatenmanagement (S. 63–69). Bad Honnef: Bock + Herchen.

Strasser, C., Cook, R., Michener, W. & Budden, A. (n.d.). Primer on Data Management. Zugriff am 15.05.2013. Verfügbar unter https://www.dataone.org/sites/all/documents/DataONE_BP_Primer_020212.pdf

Treloar, A. & Harboe-Ree, C. (2008). Data management and the curation continuum: how the Monash experience is informing repository relationships. Zugriff am 15.05.2013. Verfügbar unter http://www.valaconf.org.au/vala2008/papers2008/111_Treloar_Final.pdf

UK Data Archive. (2012). Costing tool: data management planning. Essex: UK Data Archive. Zugriff am 15.05.2013. Verfügbar unter http://data-archive.ac.uk/create-manage/planning-for-sharing/costing

University of Edinburgh Information Services. (2013). Research data management guidance. Zugriff am 15.05.2013. Verfügbar unter http://www.ed.ac.uk/schools-departments/information-services/services/research-support/data-library/research-data-mgmt

van den Eynden, V., Corti, L., Woollard, M., Bishop, L. & Horton, L. (2011). Managing and sharing data. Best practice for researchers (3. Aufl.). Colchester: UK Data Archive. Zugriff am 15.05.2013. Verfügbar unter http://www.data-archive.ac.uk/media/2894/managingsharing.pdf

Vickers, A. J. (2006). Whose data set is it anyway? Sharing raw data from randomized trials. Trials, 7, 15. Zugriff am 15.05.2013. Verfügbar unter http://dx.doi.org/10.1186/1745-6215-7-15

Vlaeminck, S. (2008). AP 2 - Erfassen und Kategorisieren relevanter Datenbestände. Göttingen: SUB Göttingen. Zugriff am 15.05.2013. Verfügbar unter http://kolawiss.uni-goettingen.de/projektergebnisse/AP2_Report.pdf

Weichselgartner, E. (2011a). Forschungsdaten in der Psychologie: Disziplinspezifische und disziplinübergreifende Bedürfnisse. Zusammenfassung des Forums (2) der 5. Konferenz für Sozial- und Wirtschaftsdaten (RatSWD Working Paper Series Nr. 187). Berlin: RatSWD. Zugriff am 15.05.2013. Verfügbar unter http://www.ratswd.de/publikationen/working-papers

Weichselgartner, E. (2011b). Disziplinenspezifische Aspekte des Archivierens von Forschungsdaten am Beispiel der Psychologie (RatSWD Working Paper Series Nr. 179). Berlin: RatSWD. Zugriff am 15.05.2013. Verfügbar unter http://www.ratswd.de/publikationen/working-papers

Weichselgartner, E., Günther, A. & Dehnhard, I. (2011a). Archivierung von Forschungsdaten. In S. Büttner, H.-C. Hobohm & L. Müller (Hrsg.), Handbuch Forschungsdatenmanagement (S. 191–202). Bad Honnef: Bock + Herchen.

Weichselgartner, E., Günther, A. & Dehnhard, I. (2011b). Stärkung der Forschungskooperation und des Datenmanagements in der Psychologie mit PsychData (RatSWD Working Paper Series Nr. 214). Berlin: RatSWD. Zugriff am 15.05.2013. Verfügbar unter http://www.ratswd.de/publikationen/working-papers

Whyte, A. & Wilson, A. (2010). How to Appraise & Select Research Data for Curation. Edinburgh, Melbourne: Digital Curation Centre & Australian National Data Service. Zugriff am 15.05.2013. Verfügbar unter http://www.dcc.ac.uk/resources/how-guides/appraise-select-data

Wicherts, J. M., Borsboom, D., Kats, J. & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. American Psychologist, 61, 726–728. Zugriff am 15.05.2013. Verfügbar unter http://dx.doi.org/10.1037/0003-066X.61.7.726

World Medical Association. (2008). Declaration of Helsinki. Ethical Principles for Medical Research Involving Human Subjects. Ferney-Voltaire: World Medical Association. Zugriff am

15.05.2013. Verfügbar unter http://www.wma.net/en/30publications/10policies/b3/index.html

Yott, P. (2005). Introduction to XML. Cataloging & Classification Quarterly, 40, 213–235. Zugriff am 15.05.2013. Verfügbar unter http://dx.doi.org/10.1300/J104v40n03_10

ZPID. (2011). PSYNDEX Terms: Deskriptoren/Subject Terms zur Datenbank PSYNDEX (9. Aufl.). Trier: ZPID.