Forschungsdatenmanagement in der Sekundäranalyse

Netscher, Sebastian; Trixa, Jessica

Veröffentlichungsversion / Published Version
Sammelwerksbeitrag / collection article

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:
GESIS - Leibniz-Institut für Sozialwissenschaften

Empfohlene Zitierung / Suggested Citation:
Netscher, S., & Trixa, J. (2019). Forschungsdatenmanagement in der Sekundäranalyse. In U. Jensen, S. Netscher, & K. Weller (Hrsg.), *Forschungsdatenmanagement sozialwissenschaftlicher Umfragedaten: Grundlagen und praktische Lösungen für den Umgang mit quantitativen Forschungsdaten* (S. 135-150). Opladen: Verlag Barbara Budrich. https://doi.org/10.3224/84742233.09

Auszug aus dem Buch:

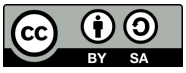
Uwe Jensen
Sebastian Netscher
Katrin Weller (Hrsg.)

Forschungsdatenmanagement sozialwissenschaftlicher Umfragedaten

Grundlagen und praktische Lösungen für den Umgang mit quantitativen Forschungsdaten

Verlag Barbara Budrich
Opladen • Berlin • Toronto 2019

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über
http://dnb.d-nb.de abrufbar.



Dieses Buch steht im Open-Access-Bereich der Verlagsseite zum kostenlosen Download bereit (https://doi.org/10.3224/84742233).
Eine kostenpflichtige Druckversion (Print on Demand) kann über den Verlag bezogen werden. Die Seitenzahlen in der Druck- und Onlineversion sind identisch.

ISBN 978-3-8474-2233-4 (Paperback)
eISBN 978-3-8474-1260-1 (eBook)
DOI 10.3224/84742233

Umschlaggestaltung: Bettina Lehfeldt, Kleinmachnow – www.lehfeldtgraphic.de
Lektorat: Nadine Jenke, Potsdam
Satz: Anja Borkam, Jena – kontakt@lektorat-borkam.de
Titelbildnachweis: Foto: Florian Losch
Druck: paper & tinta, Warschau
Printed in Europe

8. Forschungsdatenmanagement in der Sekundäranalyse

Sebastian Netscher und Jessica Trixa

Die Verfügbarkeit von Forschungsdaten ist für Forschende in den empirischen Sozialwissenschaften eine notwendige Voraussetzung ihres wissenschaftlichen Arbeitens. Bedingt durch moderne Arbeitsinstrumente stieg in den letzten Jahren nicht nur das Volumen an Forschungsdaten, sondern auch deren Komplexität stetig an (Ludwig/Enke 2013: 13). Auffällig ist dabei die wiederholte Erhebung von vergleichbaren Forschungsdaten in ähnlichen Kontexten durch unterschiedliche Forschungsprojekte. So neigen immer noch viele Forschende dazu, die für ihre Forschungsarbeit notwendigen Daten selbst zu erheben und aufzubereiten. Die Möglichkeit, bereits existierende Forschungsdaten systematisch zu evaluieren und im Rahmen einer Sekundäranalyse nachzunutzen, wird hingegen nur bedingt wahrgenommen.

Unter dem Begriff der Sekundäranalyse verstehen wir im Folgenden die „Methode, bereits vorhandenes Material (Primärerhebung) unabhängig von dem ursprünglichen Zweck und Bezugsrahmen der Datensammlung auszuwerten" (Friedrichs 1990: 353). Mit anderen Worten werden in der Sekundäranalyse „keine Daten erhoben, vielmehr wird auf bereits existierende Datenbestände zurückgegriffen" (Stier 1996: 234). Dieser Rückgriff ist jedoch an eine Reihe von Voraussetzungen geknüpft. Hierzu gehören u.a. die Zugänglichkeit und Verständlichkeit der Ausgangsdaten ebenso wie entsprechende Nachnutzungsrechte. Dementsprechend hängt die Möglichkeit, Forschungsdaten für die Sekundäranalyse nutzen zu können, auch davon ab, ob die Primärforschenden, d.h. die ursprünglichen Datenproduzierenden, entsprechende Maßnahmen in ihrem Forschungsdatenmanagement berücksichtigt haben.

Doch auch in der Sekundäranalyse ist ein geeignetes Forschungsdatenmanagement als Teil guter wissenschaftlicher Praxis von zentraler Bedeutung. Es fördert den reibungslosen Ablauf im Forschungsprojekt und ist eine grundlegende Voraussetzung zur erfolgreichen Umsetzung des Forschungsvorhabens. Es sichert Transparenz im Forschungsprojekt und ermöglicht die Replikation erzielter Forschungsergebnisse ebenso wie die Reproduktion generierter Forschungsdaten (Büttner/Hobohm/Müller 2011; Freese 2007; Fowler 1995). Darüber hinaus kommen Forschende mit Hilfe des Forschungsdatenmanagements in der Sekundäranalyse ggf. Auflagen Dritter nach, z.B. im Rahmen der Publikation ihrer Forschungsergebnisse (Pampel/Bertelmann 2011). So fordern etwa im deutschsprachigen Raum immer mehr Journals die Replizierbarkeit von Forschungsergebnissen ebenso wie die Reproduzierbarkeit von Forschungsdaten aktiv ein, wie z.B. die Politische Vierteljahresschrift oder die Zeitschrift für Soziologie. Schließlich bedeutet die Nachnutzung bereits existierender Forschungsdaten nicht per se, dass dabei keine ‚neuen' Forschungsdaten erzeugt werden. So kann das Zusammenspielen unterschiedlicher Individualdatensätze, etwa über Regionaleinheiten wie Länder oder über die Zeit hinweg, zur Erstellung neuer Daten in der Sekundäranalyse führen (vgl. z.B. Schnell/Hill/Esser 2013: 242f; Jensen 2012: 11). Die dem Zusammenspielen zugrunde liegenden Harmonisierungskonzepte (zur softwarebasierten Dokumentation der Harmonisierung von Variablen vgl. z.B. Winters/Netscher 2016: 5f.) sind dabei möglicherweise wiederum für Dritte zur weiteren Nutzung von Interesse.

Trotz dieser vielfältigen Gründe für ein Forschungsdatenmanagement in der Sekundäranalyse fehlen in den Sozialwissenschaften bislang allgemein anerkannte Standards. Um dieser Lücke zu begegnen, versucht das vorliegende Kapitel eine erste Beschreibung des Forschungsdatenmanagements in der Sekundäranalyse quantitativer sozialwissenschaftlicher Daten. Der Schwerpunkt dieses Kapitels liegt in der Reproduzierbarkeit der verwendeten

https://doi.org/10.3224/84742233.09

Daten. Die Replizierbarkeit der Forschungsergebnisse, etwa durch die Sicherung entsprechender Programmcodes zur Datenanalyse, werden in diesem Kapitel hingegen nicht weiter thematisiert. Interessierte Lesende finden grundlegende Hinweise und Verfahrensweisen zur Dokumentation der Datenanalyse beispielsweise bei Scott J. Long (2009) oder Thomas Ebel (2016).

Im folgenden Abschnitt (8.1) werden zunächst einige Voraussetzungen zur Nachnutzung bereits existierender Daten erörtert, wie deren Auffindbarkeit und Verständlichkeit oder bestehende Nachnutzungsrechte. Daran anschließend werden in Abschnitt 8.2 Gemeinsamkeiten und Unterschiede im Forschungsdatenmanagement der Primärerhebung und der Sekundäranalyse diskutiert. Aufbauend darauf unterbreiten wir in Abschnitt 8.3 einen Vorschlag zur einfachen Dokumentation der Erstellung von Forschungsdaten in der Sekundäranalyse im Rahmen einer Aufbereitungssyntax. Das Kapitel schließt mit einer kurzen Diskussion der Bedeutung und der Möglichkeiten des Forschungsdatenmanagements in der Sekundäranalyse (Abschnitt 8.4).

8.1 Die Sekundäranalyse von Forschungsdaten

Je nach Datenlage weist die Sekundäranalyse im Vergleich zur Primärerhebung einige Vorteile auf. Forschende sparen durch die Nachnutzung bereits erhobener Forschungsdaten sowohl finanzielle als auch zeitliche Ressourcen, etwa durch den Wegfall der kosten- und arbeitsintensiven Datenerhebung. Stattdessen sind die Forschungsdaten in der Sekundäranalyse direkt verfügbar und ermöglichen es, schnell Forschungsergebnisse zu erzielen und zu publizieren. Förderer können freiwerdende Ressourcen ihrerseits wiederum in neue Forschungsprojekte investieren (Schnell/Hill/Esser 2013; Friedrichs 1990). Zuletzt profitieren auch die Befragten von der Sekundäranalyse. Vor allem bei schwer erreichbaren Personengruppen oder zu erhebenden sensiblen Informationen stellt die Nachnutzung bereits existierender Forschungsdaten eine Alternative zur wiederholten Datenerhebung dar (vgl. zu praktischen und ethischen Aspekten z.B. Medjedovic 2014: 31ff.).

Die Durchführung einer Sekundäranalyse ist jedoch an verschiedene inhaltliche, analytische und rechtliche Voraussetzungen der Ausgangsdaten gebunden. Um diese näher zu beleuchten, erörtert der vorliegende Abschnitt zunächst die inhaltliche Nachnutzbarkeit von Forschungsdaten, d.h. den Prozess des Auffindens geeigneter Ausgangsdaten für die Sekundäranalyse. Darauf aufbauend werden deren analytische und rechtliche Nachnutzbarkeit thematisiert. Damit bereits erhobene Forschungsdaten in der Sekundäranalyse verwendet werden können, müssen diese sowohl verständlich sein als auch unter rechtlichen Aspekten des Datenschutzes ebenso wie des Urheberrechts zur Nachnutzung verfügbar sein. Schließlich müssen potentiell geeignete Ausgangsdaten für die Sekundäranalyse valide und je nach Forschungsansatz ggf. mit weiteren Daten kombinierbar sein.

8.1.1 Das Auffinden geeigneter Ausgangsdaten

Eine der größten Herausforderungen im Vorfeld der Sekundäranalyse stellt das Auffinden geeigneter Ausgangsdaten dar. Diese müssen inhaltlich nachnutzbar sein, d.h. eine adäquate Beantwortung der Forschungsfrage ermöglichen. Ausschlaggebende Merkmale für die inhaltliche Nachnutzbarkeit sind a) die für die Sekundäranalyse relevanten Informationen und

b) die dabei zugrunde liegende Untersuchungspopulation (Schnell/Hill/Esser 2013; Stier 1996; Friedrichs 1990).

Um inhaltlich nachnutzbare Daten für die Sekundäranalyse zu finden, stehen Forschenden unterschiedliche Möglichkeiten zur Verfügung. Erste Anhaltspunkte können individuelle Netzwerke, Konferenzen sowie publizierte Forschungsarbeiten liefern. Darüber hinaus bieten Forschungsdatenzentren, Repositorien und Datenarchive gute Anlaufstellen. Viele dieser Einrichtungen stellen Datenbestandskataloge bereit, die Forschende zur Recherche nach geeigneten Ausgangsdaten heranziehen können. Dabei ist die jeweilige Forschungsfrage für die Recherche nach geeigneten Ausgangsdaten entscheidend. Aus ihr werden entsprechende Suchbegriffe sowohl mit Bezug auf die notwendigen Informationen als auch mit Blick auf die anvisierte Untersuchungspopulation abgeleitet. Die Suchstrategie der Forschenden kann jedoch auch ein grundlegendes Kriterium dafür sein, dass die Recherche eine gute oder eine erfolglose Trefferquote erzielt. So werden in einigen Suchsystemen beispielsweise alle eingegebenen Suchwörter gemeinsam in die Suche einbezogen. Als Konsequenz werden häufig keine oder nur sehr wenige Treffer erzielt. Daher bietet es sich an, die wichtigsten Konzepte der eigenen Forschungsfrage mit Hilfe Boole'scher Operatoren zu verbinden, etwa durch die Nutzung des Boole'schen *UND* zur Verknüpfung der Schlüsselwörter zu einem Such-String oder des Boole'schen *ODER* zur erweiterten Suche. Ebenso kann der Einsatz von sogenannten Trunkierungen oder Maskierungen, wie beispielsweise das Sternchensymbol (*) oder das Fragezeichen (?), die Suche durch alternative Schreibweisen erleichtern (Kolle 2012: 57ff).

Schaukasten 8.1 konkretisiert anhand eines fiktiven Beispiels die Merkmale von Ausgangsdaten, die für die inhaltliche Nutzbarkeit unabdingbar sind. Ausgehend von den notwendigen Informationen müssen in den Ausgangsdaten alle relevanten Variablen enthalten sein, die zur Beantwortung der jeweiligen Forschungsfrage erforderlich sind. Beispielsweise setzt die Frage nach dem Erstarken rechter Parteien bei Nationalwahlen in Europa im Zuge der Wirtschaftskrise 2009 Angaben über das individuelle Wahlverhalten (abhängige Variable) sowie Informationen zu den Gründen der individuellen Wahlentscheidung (unabhängige Variable) voraus.

Wurden die relevanten Informationen überprüft, ist zu klären, ob die potentiellen Ausgangsdaten die für die Analyse notwendige Untersuchungspopulation repräsentieren. Dies betrifft erstens die Untersuchungsobjekte als solche, wie in unserem Beispiel etwa Wähler bei Nationalwahlen. Zweitens muss die Untersuchungspopulation der Ausgangsdaten den entsprechenden Untersuchungsraum abdecken, wie z.B. die Staaten in Europa. Drittens muss auch der Untersuchungszeitraum auf Basis des Erhebungszeitpunkts der Ausgangsdaten bzw. deren zeitlicher Horizont mit der zugrunde liegenden Forschungsfrage der Sekundäranalyse übereinstimmen. Dementsprechend müssen die Daten beispielsweise als Querschnitt, Längsschnitt oder Panel vorliegen oder in Form von Ereignisdaten aufbereitet sein. In Bezug auf unser Beispiel müssen geeignete Ausgangsdaten etwa einen Zeitraum vor, während und nach dem Eintreten der Wirtschaftskrise 2009 abdecken (vgl. Schaukasten 8.1).

Erweisen sich mehrere Ausgangsdatensätze als inhaltlich nachnutzbar, gilt es, die Vor- und Nachteile jedes einzelnen Datensatzes sorgfältig zu durchdenken und diese gegeneinander abzuwägen. Beide im Schaukasten vorgestellten Datensätze ermöglichen die inhaltliche Bearbeitung der aufgeworfenen Forschungsfrage. Sowohl die Comparative Study of Electoral Systems (CSES 2014) als auch die sogenannte Voter Study der European Parliament Election Study (EES) (Schmitt et al. 2015) liefern Angaben zum individuellen Wahlverhalten sowie zu den Gründen der individuellen Wahlentscheidung. Im Gegensatz zur CSES bezieht sich die EES aber auf die Europaparlamentswahl und fragt nur retrospektiv nach der Stimmabgabe bei der vorausgegangenen Nationalwahl.

Schaukasten 8.1: Das Auffinden geeigneter Ausgangsdaten für die Sekundäranalyse

Forschungsfrage: *Führte die Wirtschaftskrise 2009 in Europa zum Erstarken rechter Parteien bei Nationalwahlen?*

Inhaltlich geeignete Ausgangdaten müssen:

1. Informationen zur *individuellen Stimmabgabe bei der Nationalwahl* (abhängige Variable) sowie zu den *Motiven der Wahlentscheidung* (unabhängige Variable) beinhalten,
2. sich auf die *wahlberechtigte Bevölkerung in Europa vor, während und nach der Wirtschaftskrise 2009* beziehen.

Aufbauend auf diesen beiden Faktoren lassen sich unterschiedliche sozialwissenschaftliche Datensätze in Bezug auf ihre inhaltliche Nachnutzbarkeit beurteilen, wie z.B.:

a) die *European Parliament Election Study, Voter Study* (*EES*) zum Wahlverhalten von EU-Bürgern bei der Europaparlamentswahl:
   1. Untersuchungsaspekt:
      - abhängige Variable: *retrospektive Stimmabgabe bei der vorausgegangenen Nationalwahl*,
      - unabhängige Variable: *größte Herausforderungen im jeweiligen Land vor der Europaparlamentswahl,*
   2. Untersuchungspopulation: *wahlberechtigte Bevölkerung in den Staaten der Europäischen Union nach der Europaparlamentswahl 2009 bzw. 2014.*

b) Die *Comparative Study of Electoral Systems* (*CSES*) harmonisiert Nachwahlbefragungen zu den Nationalwahlen aus der ganzen Welt:
   1. Untersuchungsaspekt:
      - abhängige Variable: *Stimmabgabe bei der jeweiligen Nationalwahl*,
      - unabhängige Variable: *größte Herausforderungen im jeweiligen Land vor der Nationalwahl,*
   2. Untersuchungspopulation: *wahlberechtigte Bevölkerung bei den Nationalwahlen zwischen 2005 und 2011* (drittes Modul der CSES).

Quelle: Eigene Darstellung

Umgekehrt deckt die EES mit ihren 27 (bzw. 28) EU-Mitgliedsstaaten einen weitaus größeren geographischen Raum in Europa ab als die CSES. Für diese spricht wiederum der Erhebungszeitpunkt, da sie Daten vor, während und nach der Wirtschaftskrise 2009 beinhaltet. Insgesamt wären im vorliegenden Beispiel die Daten der CSES aufgrund ihres Informationsgehalts zum Wahlverhalten bei der Nationalwahl und dem Erhebungszeitraum jenen der EES vorzuziehen.

8.1.2 Die analytische und rechtskonforme Nutzbarkeit der Ausgangsdaten

Neben dem enthaltenen Analysepotential hängt die Nutzung der Ausgangsdaten auch von verschiedenen analytischen und rechtlichen Bedingungen ab. Forschende, die geeignete Ausgangsdaten für die Sekundäranalyse gefunden haben, müssen erstens klären, ob diese überhaupt analytisch sinnvoll genutzt werden können. Nur wenn die Ausgangsdaten eine entsprechende Qualität, etwa in Bezug auf die verwendeten Messkonzepte oder die Teilnahmebereitschaft der befragten Personen, aufweisen und ausreichend dokumentiert sind, können Dritte diese vollständig nachvollziehen und beurteilen. Die Beurteilbarkeit der Qualität beruht u.a. auf dem Vorliegen von Informationen über die methodische Generierung der Ausgangsdaten, deren Aufbereitung und Anonymisierung einschließlich der Dokumentation von Auffälligkeiten und Besonderheiten in den Daten. In Bezug auf das obige Beispiel liefert die CSES neben den Daten eine umfangreiche Dokumentation in Form eines Codebuchs, das die Variablen des Datensatzes vollständig, transparent und verständlich dokumentiert. Eine adäquate methodische und inhaltliche Dokumentation, etwa in Form eines Methodenberichts,

ermöglicht erst die sinnvolle Interpretation der Ausgangsdaten ebenso wie die Generalisierbarkeit der auf den Daten basierenden Forschungsergebnisse.

Zweitens müssen die Ausgangsdaten rechtskonform nachnutzbar sein. Dies betrifft vor allem die von den Urhebenden der Daten an die Nachnutzenden übertragenen Verwertungs- bzw. Nutzungsrechte. Häufig wird durch Lizenzen festgelegt, wer unter welchen Bedingungen und zu welchem Zweck auf die Daten zugreifen darf und in welcher Form diese nachgenutzt werden können. So erfordern ggf. nicht oder nur schwach anonymisierte Daten die Einschränkung des Zugangs auf bestimmte Personengruppen und bestimmte Nachnutzungszwecke. Fehlen entsprechende Nutzungsrechte beispielsweise durch restriktive Nutzungsbedingungen, kann dies unter Umständen dazu führen, dass Forschende die Ausgangsdaten zwar erhalten und analysieren können, eine Publikation der Forschungsergebnisse aber ausgeschlossen ist. Mit Hinblick auf unser CSES-Beispiel sind die aufgeführten Daten frei verfügbar und dürfen – eine adäquate Zitation vorausgesetzt – durch jede Person zu jeglichem Zweck nachgenutzt werden.

8.1.3 Die inhaltliche und methodische Validierung der Ausgangsdaten

Sind die inhaltlichen, analytischen und rechtlichen Bedingungen geklärt, müssen Forschende die Daten in Bezug auf die der Sekundäranalyse zugrunde liegende Forschungsfrage validieren. Dabei sollten mindestens drei Arbeitsschritte berücksichtigt werden. Erstens gilt es, die Dokumentation der Ausgangsdaten, d.h. die zugehörigen Methodenberichte, Codebücher usw., zu prüfen. Darin beschrieben sind u.a. die Feldphase der Datengenerierung, die Grundgesamtheit, das Stichprobenverfahren, die realisierten Interviews etc. Damit stellt die Dokumentation Transparenz in den Ausgangsdaten sicher und liefert somit wichtige Hinweise auf die Datenqualität.

Zweitens sollten die Randverteilungen zentraler soziodemographischer Variablen mit offiziellen bzw. amtlichen Statistiken verglichen werden, um sicherzustellen, dass die in den Ausgangsdaten enthaltene Untersuchungspopulation mit der in der Sekundäranalyse anvisierten Grundgesamtheit übereinstimmt. Zu bedenken sind dabei etwaige Abweichungen zwischen der offiziellen Statistik und der Untersuchungspopulation. So sind die Untersuchungsobjekte unseres Beispiels wahlberechtigte Personen in den Staaten Europas. Dementgegen beziehen sich die Daten der amtlichen Statistik, wie etwa von Eurostat (Eurostat: Database), zumeist auf die Wohnbevölkerung der einzelnen Länder. Zwischen den Randverteilungen der CSES und den Informationen von Eurostat können so Diskrepanzen im Hinblick auf die Untersuchungspopulationen entstehen, die bei der Validierung berücksichtigt werden müssen.

Drittens müssen Forschende alle in der Sekundäranalyse zu nutzenden Variablen prüfen. Dies betrifft beispielsweise Anmerkungen zu einzelnen Variablen in der Datendokumentation, etwa in Bezug auf Auffälligkeiten während der Datenerhebung. Ebenso sollten alle Variablen auf fehlende Werte bzw. Antwortverweigerungen sowie auf systematisches und inkonsistentes bzw. sich widersprechendes Antwortverhalten überprüft werden. Schließlich ist es notwendig, die Verteilungen aller relevanten Variablen zu kontrollieren. Verzerrte Verteilungen, häufig auftretende fehlende Werte, systematisches oder inkonsistentes Antwortverhalten usw. können die Analyse, die Ergebnisinterpretation sowie die Generalisierbarkeit der Forschungsergebnisse u.U. nachhaltig beeinflussen (vgl. zu Antworttendenzen in standardisierten Umfragen den Beitrag von Bogner/Landrock 2015; vgl. zur Datenkontrolle Jensen 2012: 32ff.).

8.1.4 Das Zusammenspielen unterschiedlicher Ausgangsdatensätze

Für die Sekundäranalyse werden häufig unterschiedliche Datensätze zusammengespielt, zumeist als *data* oder *record linkage* bezeichnet (Schnell/Hill/Esser 2013: 246). Dadurch können z.B. die zu analysierende Grundgesamtheit erweitert, statistische Inferenzen abgesichert oder die Ausgangsdaten um zusätzliche Informationen angereichert werden. Zusätzliche Informationen können hierbei beispielsweise administrative oder georeferenzierte Kontextdaten sein. In Bezug auf das obige Beispiel ließen sich die Individualdaten der CSES um den Datensatz der British Election Study (BES) aus dem Jahr 2010 erweitern (Whiteley/Sanders 2014). Analog können die Individualdaten um Aggregatdaten, beispielswiese zum Bruttosozialprodukt oder der Arbeitslosenquote auf Basis von Eurostatdaten (Eurostat: Database), ergänzt werden. Damit würde sich nicht nur die Anzahl an Untersuchungsobjekten auf der Individualebene erhöhen, sondern es käme auch zu einer Ausweitung des Untersuchungsraums um Großbritannien. Zudem wäre eine Anreicherung der Forschungsdaten um makroökonomische Kenngrößen möglich.

Das Zusammenspielen von unterschiedlichen Ausgangsdaten ist wiederum an diverse Voraussetzungen geknüpft. Hierzu zählen gleichermaßen die inhaltlichen, analytischen und rechtlichen Bedingungen der Nachnutzbarkeit aller Ausgangsdatensätze ebenso wie deren Validierung. Darüber hinaus treten neue Anforderungen auf, wie die Vergleichbarkeit der zu verwendenden Variablen, der zugrunde liegenden Messkonzepte sowie die Vergleichbarkeit der Stichproben bzw. Grundgesamtheiten. So muss beispielsweise vor dem Zusammenspielen der CSES-Daten mit den Kontextdaten von Eurostat sichergestellt werden, dass deren Untersuchungspopulationen vergleichbar sind. Während etwa die Individualdaten der CSES im Nachgang der jeweiligen Nationalwahl erhoben wurden, sind die Angaben von Eurostat quartalsweise erhältlich und müssen auf der Ebene der einzelnen Länder der CSES approximativ an die Individualdaten angeglichen werden.

Zur Sicherstellung der Vergleichbarkeit der verwendeten Variablen bzw. Messinstrumente sollten Forschende zunächst alle relevanten Fragen und Antwortvorgaben in den ursprünglichen Fragebögen der Studien prüfen (vgl. z.B. Winters/Netscher 2016: 9f.). Bezogen auf unser Beispiel sollten etwa die Messinstrumente bzw. Originalfragen der BES mit den Fragevorgaben der CSES abgeglichen werden, um sicherzustellen, dass diese vergleichbar sind und die darauf aufbauenden Variablen harmonisiert werden können. Dieser Prozess der Variablenharmonisierung ist vor allem dann unkompliziert, wenn die Fragen und Antwortkategorien in den Fragebögen eine hohe Übereinstimmung aufweisen. Dann kann jeder Kategorie einer Variablen eines Ausgangsdatensatzes, wie etwa der CSES, exakt eine Kategorie der entsprechenden Variablen der anderen Ausgangsdaten, z.B. der BES, zugewiesen werden. Von höherer Komplexität ist die Harmonisierung, wenn in den Ausgangsdaten zwar vergleichbare Messkonzepte genutzt wurden, die zugrunde liegenden Fragen und Antwortkategorien jedoch nicht übereinstimmen bzw. vergleichbar sind. Forschende müssen dann genau überlegen, inwiefern die unterschiedlichen Variablen und Codes in den Ausgangsdaten doch aneinander angeglichen werden können und in welchem Maß dies die Forschungsergebnisse, ihre Interpretation und Generalisierbarkeit beeinflusst (vgl. ebenda: 9f.).

8.2 Forschungsdatenmanagement in der Primär- und Sekundäranalyse

Ein Grundpfeiler der Forschung ist das Prinzip der Transparenz von Forschungsergebnissen und damit der Grundsatz der Reproduzierbarkeit von Forschungsdaten (Büttner/Hobohm/Müller 2011; Freese 2007; Fowler 1995). Um die Reproduzierbarkeit der Forschungsdaten in der Sekundäranalyse zu gewährleisten, empfiehlt sich ein systematisches Forschungsdatenmanagement, wie es in der Primärerhebung mittlerweile weit verbreitet ist. Darüber hinaus schließt die Nachnutzung bereits existierender Daten nicht aus, dass in der Sekundäranalyse nicht durchaus auch ‚neue' Daten generiert werden können, die für Dritte wiederum einen Nachnutzungswert besitzen. Dies lässt sich sehr einfach an unserem Beispiel erkennen. Bislang ist die BES (2010) nicht in die dritte Welle der CSES integriert. Beginnen Forschende nun im Rahmen ihres Projekts die beiden Datensätze zu harmonisieren und zu integrieren, so hat dieser integrierte Datensatz ggf. einen Mehrwert für andere Forschende, die ebenfalls die beiden Datensätze zusammenspielen möchten.

Entgegen dem Prinzip transparenter Forschung und der Möglichkeit, Forschungsergebnisse, wie etwa einen integrierten Datensatz oder das entsprechende Harmonisierungskonzept, nachzunutzen, fehlen in der sozialwissenschaftlichen Praxis bislang allgemein anerkannte Standards für ein Forschungsdatenmanagement in der Sekundäranalyse. Ansätze bieten einerseits das wissenschaftliche Paradigma der Reproduzierbarkeit (King 1995). Andererseits liefern Standards des Forschungsdatenmanagements in der Primärerhebung grundlegende Hinweise. Der folgende Abschnitt erörtert in diesem Zusammenhang zunächst Gemeinsamkeiten des Forschungsdatenmanagements in Primärerhebung und Sekundäranalyse. Darauf aufbauend wird aufgezeigt, wo sich das Forschungsdatenmanagement der Sekundäranalyse von jenem der Primärerhebung unterscheidet und welche zusätzlichen Anforderungen die Sekundäranalyse an das Forschungsdatenmanagement stellt.

8.2.1 Gemeinsamkeiten des Forschungsdatenmanagements

Gemeinsamkeiten der Primärerhebung und der Sekundäranalyse finden sich bei unterschiedlichen Aspekten des Forschungsdatenmanagements. Grundsätzlich unterstützt das Forschungsdatenmanagement auch in der Sekundäranalyse die erfolgreiche Umsetzung des Forschungsprojekts. So etwa im Rahmen der Organisation und Handhabung der Forschungsdateien (s. Kapitel 5). Regelungen zur Benennung und Versionierung von Arbeitsdateien, deren Sicherung gegen Verlust (Back-ups) oder die Definition von Zuständigkeiten im Rahmen des Forschungsdatenmanagements müssen in Forschungsprojekten der Sekundäranalyse genauso geklärt werden wie bei Primärerhebungen.

Gleichfalls müssen in Primärerhebung wie Sekundäranalyse die jeweils verwendeten Forschungsdaten adäquat dokumentiert werden. Dadurch wird auch in der Sekundäranalyse die Verständlichkeit, Nachvollziehbarkeit und Interpretierbarkeit der Forschungsdaten sichergestellt. Dies betrifft beispielsweise die Kodierung neuer Variablen, die dazu genutzten Methoden bzw. Standards sowie die eindeutige Benennung der enthaltenen Variablen und ihrer Codes.

Auch in Bezug auf die längerfristige Sicherung relevanter Forschungsdateien über das Projektende hinaus unterscheiden sich das Forschungsdatenmanagement der Primär- und Sekundäranalyse kaum voneinander. In beiden Fällen gilt es, alle zur Reproduktion bzw. Nachnutzung notwendigen Forschungsdaten zu dokumentieren, entsprechende begleitende Dokumentationen zu erstellen und diese Forschungsdateien am Projektende zu sichern bzw. zu

archivieren. Bei allen derartigen Aspekten kann das Forschungsdatenmanagement der Sekundäranalyse auf bereits existierende Vorgaben und Standards der Primärerhebung zurückgreifen und entsprechende Verfahrensweisen übernehmen oder anpassen.

8.2.2 Unterschiede im Forschungsdatenmanagement

Neben den Gemeinsamkeiten des Forschungsdatenmanagements existieren auch einige in der Primärerhebung besonders wichtige Aspekte, die hinsichtlich der Nachnutzung bereits existierender Forschungsdaten jedoch eher eine zweitrangige Rolle spielen. Dies zeigt sich beispielsweise in der Dokumentation der Datenerhebung. So muss das Forschungsdatenmanagement der Primärerhebung detailliert beschreiben, wie die Primärdaten erhoben und aufbereitet wurden. In der Sekundäranalyse geht es hingegen vor allem um die Nachvollziehbarkeit der verwendeten Forschungsdaten und deren Wiederauffindbarkeit durch Dritte. Dies impliziert beispielsweise, dass die Forschenden in der Sekundäranalyse die Wahl der Ausgangsdaten begründen und vor Beginn der Analyse deren Nachnutzbarkeit unter inhaltlichen, analytischen und rechtlichen Aspekten überprüfen.

Ähnlich finden sich auch im Hinblick auf datenschutzrechtliche Aspekte signifikante Unterschiede zwischen dem Forschungsdatenmanagement der Primärerhebung und der Sekundäranalyse. Im Rahmen der Primärerhebung werden neue Forschungsdaten durch das systematische Sammeln von Informationen generiert. Beinhalten diese Informationen sogenannte personenbezogene Angaben, dann bedarf es im Rahmen der datenschutzkonformen Nutzung solcher Daten zunächst der informierten Einwilligung der Befragten zu deren Erhebung, Speicherung und Verarbeitung, wie im ersten Abschnitt des vierten Kapitels näher ausgeführt. In der Sekundäranalyse wird hingegen auf bereits existierende Forschungsdaten zurückgegriffen, sodass die Sekundärforschenden keine weitere Einwilligung der ursprünglichen Befragten mehr einholen müssen. Selbstverständlich darf die informierte Einwilligung die Möglichkeit der Nachnutzung von Forschungsdaten im Rahmen der Sekundäranalyse nicht ausschließen. Diese ist aber Teil der Primärerhebung und für die Sekundärforschenden, soweit vorhanden, irrelevant.

Analog unterscheiden sich das Forschungsdatenmanagement der Primärerhebung und der Sekundäranalyse mit Blick auf Auflagen zur Anonymisierung. In der Regel arbeiten Sekundärforschende mit bereits anonymisierten Forschungsdaten und werden insofern nicht weiter mit Fragen des Datenschutzes und der Datenanonymisierung konfrontiert. Ausnahmen finden sich etwa im Hinblick auf die Nachnutzung nicht oder nur schwach anonymisierter Daten. Werden derartige Forschungsdaten in der Sekundäranalyse nachgenutzt, so sind die Sekundärforschenden selbstverständlich dazu verpflichtet, entsprechende Datenschutzmaßnahmen, wie beispielsweise das geschützte Speichern der Informationen, einzuhalten.

Zuletzt zeigen sich Differenzen mit Hinblick auf das Urheberrecht. Generell gilt, dass bei der Nachnutzung von Forschungsmaterialien, wie Forschungsdaten, Methoden oder Messkonzepten, das Urheberrecht und die damit verbundenen Nutzungsrechte zu beachten sind. Bestehende Unterschiede liegen aber gerade im Urheberrecht an den Forschungsdaten. Während dies in der Primärerhebung zumeist bei den Forschenden, deren Forschungseinrichtungen oder den Förderern des Forschungsprojekts liegt, greifen Sekundärforschende möglicherweise auf urheberrechtlich geschützte Werke anderer zurück. Dementsprechend bedarf es in der Sekundäranalyse der Zustimmung der Urheberrechtsinhaberinnen und Urheberrechtsinhaber zur Nachnutzung der Daten in Form entsprechender Verwertungsrechte, wie in Kapitel 7.2 im Kontext der Lizensierung näher erörtert.

8.2.3 Forschungsdaten in der Sekundäranalyse managen

Betrachtet man die Unterschiede, die das Forschungsdatenmanagement der Primärerhebung und das der Sekundäranalyse aufweisen, genauer, dann fällt auf, dass trotz bestehender Differenzen gewisse Parallelen existieren. Dementsprechend lassen sich – wie bereits erwähnt – aus dem Forschungsdatenmanagement der Primärerhebung durchaus Vorgehensweisen für das Forschungsdatenmanagement der Sekundäranalyse ableiten.

Dies zeigt sich z.B. im Prozess des Suchens und Auffindens geeigneter Forschungsdaten in der Sekundäranalyse. Analog zur Planung der Datenerhebung durch die Primärforschenden steht am Anfang der Sekundäranalyse der Prozess des Auffindens geeigneter Ausgangsdaten. Die Ergebnisse dieser Recherchen sollten im Rahmen des Forschungsdatenmanagements der Sekundäranalyse beschrieben werden. Sie dienen der Begründung der Datenauswahl im Forschungsprojekt und erzeugen Transparenz etwa im Rahmen der Veröffentlichung der Forschungsergebnisse. Hierzu zählen beispielsweise Angaben zu den verwendeten Suchbegriffen und Strategien oder zu den Gründen, einen bestimmten Ausgangsdatensatz nachzunutzen bzw. andere, alternative Daten nicht zu berücksichtigen.

Ähnliche Parallelen im Forschungsdatenmanagement der Primärerhebung und Sekundäranalyse zeigen sich in der Dokumentation der Forschungsdaten. In beiden Fällen dient diese der Transparenz und sichert die Verständlichkeit ebenso wie die Interpretierbarkeit der verwendeten Forschungsdaten. Die Dokumentation der Forschungsdaten in der Sekundäranalyse beginnt jedoch stets mit der Dokumentation der Ausgangsdaten als zitierfähige Nachweise der Datenquelle. In den letzten Jahren haben sich hierzu in den Sozialwissenschaften persistente Identifikatoren als Teil der Zitation etabliert, wie beispielsweise Digital Object Identifiers (DOI) oder Uniform Ressource Names (URN), wie im ersten Abschnitt des zehnten Kapitels weiter ausgeführt wird. Darüber hinaus müssen auch Sekundärforschende gewährleisten, dass ihre Forschungsdaten transparent und reproduzierbar sind. Entsprechend muss die Datenaufbereitung, wie etwa der Ausschluss von Teilen der Untersuchungspopulation der Ausgangsdaten, die Harmonisierung von Variablen oder die Datenvalidierung dokumentiert werden.

In Bezug auf die rechtlichen Aspekte des Forschungsdatenmanagements können auch in der Sekundäranalyse nicht oder nur schwach anonymisierte Forschungsdaten neu entstehen. Dies ist etwa der Fall, wenn durch das Zusammenspielen unterschiedlicher Ausgangsdatensätze bereits anonymisierte Forschungsdaten de-anonymisiert werden. So kann beispielsweise das Einspielen georeferenzierter Informationen, wie in Kapitel 12.3 beschrieben wird, dazu führen, dass natürliche Personen in den Individualdaten re-identifizierbar werden. In diesem Fall müssen Sekundärforschende selbstverständlich analog zur Primärerhebung die einschlägigen Regelungen des Datenschutzes beachten.

Schließlich erlangen auch Sekundärforschende Urheberrechte an den von ihnen erzeugten Forschungsdateien, wie etwa an Konzepten zur Harmonisierung von Variablen unterschiedlicher Ausgangsdatensätze. Dies ist vor allem dann von Relevanz, wenn derartige Konzepte für Dritte verfügbar gemacht werden sollen, egal ob zu Reproduktionszwecken oder zur weiteren Nutzung in neuen Forschungskontexten.

Analog zur Generierung nachnutzbarer Forschungsdaten in der Primärerhebung stellt sich letztlich auch in der Sekundäranalyse die Frage nach den Plänen zum längerfristigen Umgang mit zentralen Forschungsdateien. Sowohl die Sicherung zu Reproduktionszwecken als auch die Bereitstellung etwa von nachnutzbaren Harmonisierungskonzepten ist jedoch keineswegs trivial. Zum einen gilt es festzulegen, welche vom Projekt generierten Forschungsdateien längerfristig gesichert werden sollen. Sekundärforschende haben aufgrund des Urheberrechts und entsprechender Nutzungsbedingungen der Ausgangsdaten in der Regel nicht das Recht, diese Ausgangsdaten zu archivieren oder gar Dritten verfügbar zu machen. Dies ist im

Rahmen der Sekundäranalyse aber ohnehin häufig nicht notwendig, da durch eine adäquate Zitation der Ausgangsdaten sichergestellt wird, dass diese auch zukünftig eindeutig identifizierbar und auffindbar sind.

Zum anderen muss geklärt werden, wie lange die Forschungsdateien des eigenen Projektes aufbewahrt werden sollen und zu welchem Zweck dies geschieht. Dazu muss im Rahmen des Forschungsprojekts u.a. geplant werden, an welchem Ort, in welchen Formaten und mit welchen Sicherungsmaßnahmen die Forschungsdateien vor Verlust gesichert bzw. langfristig archiviert werden sollen. In Kapitel 3.2 werden diese Maßnahmen im Abschnitt zur Planung des Forschungsdatenmanagements über das Projektende hinaus genauer beschrieben. Es gilt auch festzulegen, wie der Zugang zu den Dateien für die originären Projetbeteiligten ebenso wie für Dritte sichergestellt werden kann. Zur Bereitstellung von Harmonisierungskonzepten zur Nachnutzung durch andere Forschende bieten sich hierfür beispielsweise Reproduktionsrepositorien, wie z.B. der *GESIS Replikationsserver*, an (Ebel 2016).

8.3 Die Aufbereitungssyntax als Dokumentation der Sekundäranalyse

Sollen relevante Forschungsdateien der Sekundäranalyse längerfristig aufbewahrt bzw. Dritten zur weiteren Nutzung verfügbar gemacht werden, stellt sich die Frage, wie die Generierung dieser Daten am besten beschrieben werden kann (Winters/Netscher 2016; Dickmann/Enke/Harms 2010). Generell können Sekundärforschende die Maßnahmen ihres Forschungsdatenmanagements in einem projektinternen Datenmanagementplan systematisch dokumentieren (vgl. Kapitel 2.4). Dieser kann gemeinsam mit der Aufbereitungssyntax, d.h. der Syntax zur Erstellung und Aufbereitung der Forschungsdaten, archiviert werden. Versehen mit einer eindeutigen Zitation des Datenmanagementplans und der Aufbereitungssyntax sowie mit entsprechenden Querverweisen in beiden Dateien, lassen sich die Forschungsdaten so reproduzieren. Die Vorteile an diesem Vorgehen liegen vor allem in der leichten Handhabung sowie in der besseren Verständlichkeit der Aufbereitungssyntax.

Nachteilig ist hingegen der Umstand, dass zur weiteren Verwendung der Aufbereitungssyntax stets deren Dokumentation in Form des Datenmanagementplans notwendig ist. Gerade mit Blick auf eine mögliche Nachnutzung durch Dritte sollten bereits in der Aufbereitungssyntax möglichst alle relevanten Informationen zur Reproduktion der Forschungsdaten gebündelt werden. Dazu kann diese um entsprechende Informationen zur Erstellung der Forschungsdaten für die Sekundäranalyse ergänzt werden, um so für Dritte transparent, verständlich sowie nachvollziehbar und damit längerfristig nachnutzbar zu bleiben.

Im Folgenden unterbreiten wir einen Vorschlag zur Struktur einer derartigen, erweiterten Aufbereitungssyntax und erörtern notwendige Informationen, die zusätzlich bereitgestellt werden sollten. Dabei orientieren wir uns an verschiedenen Leitlinienpapieren zur Reproduktion von Forschungsdaten, etwa von Datenrepositorien, wie beispielsweise Dataverse, oder von Fachzeitschriften der Politik- (Gherghina/Katsanidou 2013) und der Wirtschaftswissenschaften (Vlaeminck/Siegert 2012). Generell gilt, dass die bereitgestellten Informationen es Dritten erlauben müssen, die Generierung der Forschungsdaten im originären Forschungsprojekt nachzuvollziehen und zu verstehen. Hierzu zählen u.a. Angaben zum ursprünglichen Forschungsprojekt und den Sekundärforschenden, zu den Ausgangsdaten ebenso wie ggf. zur Harmonisierung und Integration unterschiedlicher Ausgangsdatensätze. Ziel der nachfolgenden Diskussion ist die Generierung der Forschungsdaten für die Sekundäranalyse. Aspekte der Datenanalyse werden hingegen, wie eingangs bereits erwähnt, nicht weiter erörtert.

8.3.1 Informationen zum Forschungsprojekt

Um die Aufbereitungssyntax für Dritte verständlich zu gestalten, sollte diese zunächst einige grundlegende Informationen zum Forschungsprojekt enthalten (Ebel 2016: 4ff). Wie in Schaukasten 8.2 exemplarisch dargestellt, liefern derartige Angaben einen allgemeinen Einblick und dienen dem Verständnis. Dritte, die mit der Aufbereitungssyntax weiterarbeiten wollen, erhalten so einen schnellen Überblick über das ursprüngliche Forschungsprojekt sowie den Sinn und Zweck der Syntaxdatei. Sie werden dadurch in die Lage versetzt, die Aufbereitungssyntax effektiv nutzen und die damit erzeugten Forschungsdaten reproduzieren zu können.

Zu den Projektinformationen gehören zunächst der Projekttitel und die Namen der Forschenden ebenso wie Angaben zu deren Instituten und ggf. zu den Förderern des Forschungsprojekts. Dadurch werden zum einen Forschungskooperationen oder Abhängigkeiten gegenüber Mittelgebern für Dritte ersichtlich. Zum anderen wird aber auch dem Arbeit- bzw. den Mittelgebern Rechnung gezollt. Darüber hinaus sollten Informationen zum ursprünglichen Forschungsanliegen, die zugrunde liegende Forschungshypothese sowie die damit verknüpften Publikationen und das Ziel der vorliegenden Aufbereitungssyntax bereitgestellt werden.

Schaukasten 8.2: Die Aufbereitungssntax zu Reproduktionszwecken: Informationen zum Forschungsprojekt

```
*******************************************************************************
**                                                                           **
**                Aufbereitungssyntax zu Reproduktionszwecken                **
**  ---------------------------------------------------------------------  **
**                                                                           **
**  1. Projektinformationen:                                                 **
**  ------------------------                                                 **
**  Projekttitel:      Rechte Parteien und Wirtschaftskrise in Europa        **
**  Forschende:        Sebastian Netscher & Jessica Trixa                    **
**  Kontakt:           forschende@gesis.org                                  **
**  Institut:          GESIS – Leibniz-Institut für Sozialwissenschaften     **
**  Förderer:          GESIS                                                 **
**  Projekt:           Zusammenhang zwischen dem Erstarken rechter Parteien  **
**                     bei den Nationalwahlen in Europa und der              **
**                     Wirtschaftskrise 2009.                                **
**  Hypothese:         Wähler stimmen für rechte Parteien, wenn sie die      **
**                     Wirtschaftskrise als persönliche Bedrohung empfinden. **
**  Publikation:       Sebastian Netscher und Jessica Trixa (2018): Das      **
**                     Erstarken rechter Parteien in Europa. Zeitschrift für **
**                     Wahlverhalten [in Publikation], doi.10.XXXX/XXXXX.    **
**  Syntax:            Die Syntax dient der Replikation der Forschungsdaten  **
**                     zur oben genannten Publikation.                       **
**  Autor der Syntax:  Sebastian Netscher & Jessica Trixa.                   **
**  Erstellungsdatum:  07.06.2017.                                           **
**  Dateiversion:      1.0.0.                                                **
**  Verfügbar unter:   [noch nicht publiziert, in Vorbereitung].             **
**  Zugang:            frei [geplant].                                       **
**  genutzte Software: Stata, Version 12.0.                                  **
**  Anmerkungen:       Zum Öffnen bzw. Speichern von Dateien müssen Befehle  **
**                     in den Zeilen 46, 93 & 281 angepasst werden.          **
```

Quelle: Eigene Darstellung

Des Weiteren sollte die Aufbereitungssyntax auch Angaben über die vorliegende Syntaxdatei selbst beinhalten (Dickmann/Enke/Harms 2010: 11). Dritte sollten Kenntnis darüber haben, wann die Syntaxdatei erstellt wurde, um welche Version es sich handelt, wer die Autoren der Aufbereitungssyntax sind und wie mit diesen für eventuelle Rückfragen in Kontakt getreten werden kann. In diesem Zusammenhang ist auch anzugeben, wo die jeweilige Syntaxdatei

archiviert wurde und wie sie zugänglich ist. Schließlich sollten Spezifika der Syntaxdatei kurz erörtert werden. Hierzu zählen beispielsweise Programmcodes der Syntaxdatei, die je nach Verwendung angepasst werden müssen, z.B. Pfadangaben zum Öffnen bzw. Speichern der Daten.

8.3.2 Die Ausgangsdaten der Sekundäranalyse

Neben den Angaben zum Forschungsprojekt und zur Syntaxdatei sollte die Aufbereitungssyntax Informationen zu den Ausgangsdaten beinhalten (Ebel 2016). Dazu zählt zunächst eine adäquate Zitation jedes einzelnen Ausgangsdatensatzes. Diese muss es Dritten ermöglichen, die Ausgangsdaten eindeutig identifizieren zu können, etwa mit Bezug auf die jeweilige Version jedes einzelnen Datensatzes, wie in Schaukasten 8.3 dargestellt. Dritten wird auf diese Art und Weise ermöglicht, die Ausgangsdaten wiederfinden und so die Forschungsdaten nachvollziehen zu können.

Schaukasten 8.3: Die Aufbereitungssysntax zu Reproduktionszwecken: Informationen zu den Ausgangsdaten

```
*******************************************************************************
**                                                                           **
**                Aufbereitungssyntax zu Reproduktionszwecken                **
**  ------------------------------------------------------------------------  **
…
**  2. Ausgangsdaten:                                                        **
**  -----------------                                                        **
**  - Individualdaten:                                                       **
**     ~ CSES: Comparative Study of Electoral Systems (2014):                **
**             CSES Module 3 Full Release. GESIS Datenarchiv, Köln.          **
**             ZA5181 Datenfile Version 4.0.0,                               **
**             doi:10.7804/cses.module3.2013-03-27.                          **
**             => Nachnutzungsrechte: Daten sind frei verfügbar.             **
**     ~ BES: Whiteley, P.F. und Sanders, D. (2014):                         **
**            British Election Study, 2010 (BES): Face-to-Face Survey        **
**            [computer file]. Colchester, Essex: UK Data Archive (Archiv).  **
**            Verfügbar unter: http://www.britishelectionstudy.com/data-     **
**                             objects/cross-sectional-data/.                **
**            Zuletzt besucht: 03.08.2017.                                   **
**            => Anmerkung: ~ keine Angaben zu den Wahlen in Nordirland.     **
**                          ~ Onlinesample, Repräsentativität kontrollieren.**
**            => Nachnutzungsrechte: Daten sind frei verfügbar.              **
**  - Kontextdaten: Eurostat (2015): Database.                               **
**                  Verfügbar unter: http://ec.europa.eu/eurostat/data/      **
**                                   database.                               **
**                  Zuletzt besucht: 03.08.2017.                             **
**                  => Anmerkung: Daten für ganz Großbritannien,             **
**                                inklusive Nordirland.                      **
**                  => Nachnutzungsrechte: Daten sind frei verfügbar.        **
…
```

Quelle: Eigene Darstellung

Darüber hinaus sollten Auffälligkeiten in den Ausgangsdaten im Rahmen der Aufbereitungssyntax erörtert werden. Dies betrifft beispielsweise Einschränkungen in deren Zugänglichkeit, Besonderheiten ihrer Dokumentation oder spezifische Nachnutzungsrechte. Werden dabei in der Sekundäranalyse mehrere Ausgangsdatensätze zusammengespielt, müssen die Vergleichbarkeit der Datensätze, etwa in Bezug auf die Untersuchungspopulationen, sowie deren Validität beschrieben werden. Zwar lassen sich im Hinblick auf die Validierung die einzelnen Arbeitsschritte anhand der Syntaxbefehle ablesen, dennoch sollten zur besseren Verständ-

lichkeit und Nachvollziehbarkeit kurze Erläuterungen das genaue Vorgehen und die Gründe dafür beschreiben. Dies beinhaltet neben Auffälligkeiten der Validierung auch eventuelle Einschränkungen in der Ergebnisinterpretation bzw. in deren Generalisierbarkeit. So weisen die Daten der BES in unserem Beispiel etwa keine Angaben zu Personen in Nordirland auf, was in der Interpretation und Generalisierung von Forschungsergebnissen entsprechend berücksichtigt werden muss.

8.3.3 Die Aufbereitung der Ausgangsdaten für die Sekundäranalyse

Schaukasten 8.4: Die Aufbereitungssyntax zu Reproduktionszwecken: Informationen zum Erstellen der Forschungsdaten

```
*************************************************************************
**                                                                     **
**               Aufbereitungssyntax zu Reproduktionszwecken           **
**  ---------------------------------------------------------------    **
…
**  3. Erzeugen der Forschungsdaten:                                   **
**  --------------------------------                                   **
**  -> Individualdatensätze: CSES & BES:                               **
**      a) Löschen nicht benötigter Informationen:                     **
**          - Untersuchungsobjekte: ~ Länder außerhalb Europa,         **
**                                  ~ alle Nichtwähler;                **
**          - in der Sekundäranalyse nicht genutzte Variablen.         **
**      b) Harmonisierung und Erstellen neuer Variablen:               **
**          - Harmonisierung von Variablen,                            **
**          - Erstellen einer ID-Variable (<id>)                       **
**            (Ländercodes für Kontextdaten).                          **
[Aufbereitung der Ausgangsdaten]
**  Harmonizing Respondent's Marital Status (<marital>):               **
**  - CSES: C2004 (D4):    This variable reports the respondent's current  **
**                         marital or civil union status. For instance, a  **
**                         person who is both divorced and living as       **
**                         married would be coded 1.                       **
**  - BES: aq64 (PostQ89): Can I just check, which of these applies to you  **
**                         at present? Please choose the first on the list  **
**                         that applies (CARD AY).                         **
**                                                                         **
**  marital          CSES                       BES                     **
**   1. married       1. married or living       1. married              **
**                       together as married     2. living with a partner **
**   2. widowed       2. widowed                 5. widowed              **
**   3. divorced /    3. divorced or separated   3. separated             **
**      separated                                4. divorced             **
**   4. single        4. single, never married   6. single (never married) **
**   5. others        5. [see variable notes]     -  ---                 **
**  -7. refused      7. refused                 -2. refused              **
**  -8. don't know   8. don't know              -1. don't know           **
**  -9. missing      9. missing                  -  ---                  **
…
**  -> Kontextdaten: Eurostat:                                         **
**      a) Löschen aller Länder, die nicht Teil der Individualdaten sind,  **
**      b) Erstellen einer ID-Variable (<id>)                          **
**         (Ländercodes für Kontextdaten).                             **
…
**  Zusammenspielen der aufbereiteten Ausgangsdaten:                   **
**  a) Zusammenspielen der Individualdatensätze (CSES & BES),          **
**  b) Zusammenspielen der kumulierten Individualdaten mit den         **
**     Kontextdaten von Eurostat anhand der erstellen ID-Variabel (<id>),  **
**  c) Labeln der Variablen und ihrer Ausprägungen im Forschungsdatensatz.  **
[Zusammenspielen der aufbereiteten Ausgangsdaten; Labeln der Variablen und
Codes]
[Validierung und Speicherung des Forschungsdatensatzes]
…
**  // ENDE DER DATEI //                                              **
```

Quelle: Eigene Darstellung

Beim Erstellen der Forschungsdaten lassen sich mehrere Schritte unterscheiden, wie in Schaukasten 8.4 illustriert. Mit Blick auf die Übersichtlichkeit der Daten sollten zunächst alle Informationen gelöscht werden, die für die angestrebte Sekundäranalyse unerheblich sind. Hierzu zählen etwa nicht benötigte Untersuchungsobjekte und Variablen der Ausgangsdatensätze. Die Sekundärforschenden reduzieren damit die Komplexität der Daten, ebenso wie ihren Dokumentationsaufwand und erleichtern die Handhabung des Forschungsdatensatzes. In Bezug auf unser Beispiel müssen zunächst etwa alle Nichtwähler aus den Daten entfernt werden, da sie keinen Einfluss auf das Wahlergebnis haben. Ebenso gilt es, alle Länder aus den CSES-Daten zu entfernen, die außerhalb Europas und damit außerhalb des interessierenden Untersuchungsraumes liegen.

Daran anschließend sollten alle neu zu generierenden Variablen erstellt werden, wie etwa eine eindeutige ID-Variable auf Länderebene, die das Zusammenspielen von Kontext- und Individualdaten ermöglicht. Werden mehrere Ausgangsdaten auf einer Ebene (z.B. Individualdaten) zusammengespielt, müssen darüber hinaus die einzelnen Messkonzepte verglichen und die entsprechenden Variablen harmonisiert werden. Der Schaukasten liefert hierzu ein Beispiel anhand des Familienstands der Befragten (*marital*). Um sicherzustellen, dass die zugrunde liegenden Messinstrumente und die entsprechenden Variablen der CSES und der BES vergleichbar sind, werden zunächst die ursprünglich gestellten Fragen gegenübergestellt. Anschließend werden die einzelnen Antwortvorgaben und ihre Codes angeglichen (vgl. Winters/Netscher 2016: 9f.).

Letztendlich müssen die einzelnen aufbereiteten Ausgangsdatensätze zusammengespielt werden. Dabei empfiehlt es sich, zunächst die Ausgangsdaten auf einer Ebene zusammenzuspielen, z.B. die Individualdatensätze der CSES und der BES, und so einen integrierten Datensatz mit allen Untersuchungsobjekten auf einer Ebene zu erstellen. Anschließend werden die Daten über die Ebenen hinweg zusammengeführt, wie beispielsweise der integrierte Individualdatensatz mit den Kontextdaten von Eurostat.

Bevor der so erstellte Forschungsdatensatz in der Sekundäranalyse genutzt werden kann, sollten Forschende alle Variablen sowie die darin enthaltenen Ausprägungen zur besseren Verständlichkeit *labeln*. Zu guter Letzt muss der Datensatz erneut mit Blick auf die Forschungsfrage der Sekundäranalyse validiert und alle dabei auftretenden Auffälligkeiten müssen in der Aufbereitungssyntax dokumentiert werden. Der so erzeugte und dokumentierte Forschungsdatensatz kann dann für die eigentliche Sekundäranalyse verwendet werden.

8.4 Diskussion

Wie eingangs erörtert, ist in der Primärerhebung sozialwissenschaftlicher Forschungsdaten das Forschungsdatenmanagement längst akzeptiert. Dementgegen fehlen in der Sekundäranalyse bislang allgemein anerkannte Standards. Dabei gelten hier dieselben Gründe, die in der Primärerhebung als Teil guter wissenschaftlicher Praxis anerkannt sind: Forschungsdatenmanagement unterstützt das Forschungsprojekt an sich. Es erleichtert den Forschenden den Umgang mit ihren Forschungsdaten, erzeugt Transparenz im Forschungsprojekt und unterstützt die Replizierbarkeit von Forschungsergebnissen ebenso wie von Forschungsdaten. Schließlich unterstützt das Forschungsdatenmanagement die Nachnutzung von z.B. Konzepten zur Harmonisierung unterschiedlicher Datensätze durch andere Forschende.

Die Nachnutzung bereits existierender Forschungsdaten im Rahmen der Sekundäranalyse ist an unterschiedliche Voraussetzungen gebunden. Zunächst müssen Ausgangsdaten gefunden werden, die inhaltlich nachnutzbar sind, d.h. alle für die Sekundäranalyse notwendigen

Informationen über die entsprechende Untersuchungspopulation enthalten. Analog müssen sowohl die rechtlichen als auch die analytischen Voraussetzungen zur Nachnutzung gegeben sein. Dies betrifft sowohl die Nutzungsrechte an den Ausgangsdaten als auch deren Transparenz, Verständlichkeit und Interpretierbarkeit. Ein gezieltes Forschungsdatenmanagement unterstützt die Sekundärforschenden dabei bei der Suche nach inhaltlich, rechtlich und analytisch nachnutzbaren Ausgangsdaten. Durch eine adäquate Dokumentation sichert das Forschungsdatenmanagement zudem eine dauerhafte Nachvollziehbarkeit des Suchprozesses, ebenso wie des Entscheidungsprozesses für bestimmte Ausgangsdaten.

Um die aufgefundenen Ausgangsdaten in der Sekundäranalyse nachzunutzen, müssen Sekundärforschende diese validieren und entsprechend ihres Forschungsanliegens aufbereiten. Dies umfasst die Kontrolle der Ausgangsdaten und deren Dokumentation. Werden für die Sekundäranalyse verschiedene Ausgangsdatensätze integriert, müssen Sekundärforschende diese ggf. harmonisieren und zusammenspielen. Dabei sollten sowohl die Ausgangsdaten, ihre Dokumentation und ihre Qualität, ebenso wie Konzepte der Datenharmonisierung und des Zusammenspielens unterschiedlicher Datensätze im Rahmen des Forschungsdatenmanagements detailliert beschrieben werden. Nur so kann sichergestellt werden, dass die Forschungsdaten und Konzepte als solches reproduziert bzw. durch Dritte weitergenutzt werden können.

Um die Reproduzierbarkeit bzw. Nachnutzbarkeit der verwendeten Forschungsdaten sicherzustellen, ist es notwendig, die entsprechenden Konzepte zur Aufbereitung, Harmonisierung und Integration unterschiedlicher Ausgangsdatensätze Dritten verfügbar zu machen. Ein systematisches Forschungsdatenmanagement unterstützt die Planung und Aufbereitung der Ausgangsdaten, deren Harmonisierung sowie das Zusammenspielen unterschiedlicher Ausgangsdatensätze. Die Dokumentation dieser Arbeitsschritte im Rahmen der Aufbereitungssyntax bietet den Sekundärforschenden dabei die Möglichkeit, diese Prozesse und Maßnahmen einfach zu beschreiben und Dritten zugänglich zu machen.

Literaturverzeichnis

Bogner, Kathrin/Landrock, Ute (2015): Antworttendenzen in standardisierten Umfragen. GESIS – Leibniz-Institut für Sozialwissenschaften. GESIS Survey Guidelines. https://doi.org/10.15465/gesis-sg_016 [Zugriff: 19.03. 2018].

Büttner, Stephan/Hobohm, Hans-Christoph/Müller, Lars (2011): Research Data Management. In: Büttner, Stephan/ Hobohm, Hans-Christoph/Müller, Lars (Hrsg.): Handbuch Forschungsdatenmanagement. Bad Honnef: Bock und Herchen, S. 13-24.

CSES - Comparative Study of Electoral Systems (2014): CSES Module 3 Full Release. GESIS Datenarchiv, Köln. ZA5181 Datenfile Version 4.0.0. https://doi.org/10.7804/cses.module3.2013-03-27 [Zugriff: 03.08.2017].

Dickmann, Frank/Enke, Harry/Harms, Patrick (2010): Technische Evaluation der Grid-Technologie für das Modellprojekt Kollaborative Datenauswertung und virtuelle Arbeitsumgebung. VirtAug. Forschungsverbund sozioökonomische Berichterstattung. soeb-Arbeitspapier 2010-1. http://www.soeb.de/fileadmin/redaktion/downloads/VirtAug/Expertise_VirtAug.pdf [Zugriff: 19.03.2018].

Ebel, Thomas (2016): Einreichungen von Syntaxen in datorium (Replikationsserver). GESIS Datenarchiv für Sozialwissenschaften. 29.02.2016. https://www.gesis.org/fileadmin/upload/Replikationsserver/Einreichung_Syntaxen_2016-02-29.pdf [Zugriff: 19.03.2018].

Fowler, Linda L. (1995): Replication as Regulation. Political Science and Politics 28, 3, S. 478-481.

Freese, Jeremy (2007): Replication Standards for Quantitative Social Science. Why Not Sociology. Sociological Methods & Research 36, S. 153-172.

Friedrichs, Jürgen (1990): Methoden der empirischen Sozialforschung 14. Opladen: Westdeutscher.

Gherghina, Sergiu/Katsanidou, Alexia (2013). Data Availability in Political Science Journals. European Political Science 12, S. 333-349.
Jensen, Uwe (2012): Leitlinien zum Management von Forschungsdaten. Sozialwissenschaftliche Umfragedaten. GESIS-Technical Reports 2012/07. https://www.ssoar.info/ssoar/handle/document/32065 [Zugriff: 03.08.2017].
Kolle, Christian (2012): Wissenschaftliche Literaturrecherche. In: Berninger, Ina/ Botzen, Katrin/Kolle, Christian/ Vogel, Dominikus/Watteler, Oliver (Hrsg.): Grundlagen sozialwissenschaftlichen Arbeitens. Opladen u.a.: Verlag Barbara Budrich, S. 33-61.
King, Gary (1995): Replication, Replication. Political Science and Politics 28, 3, S. 443-499.
Long, Scott J. (2009): The Workflow of Data Analysis Using Stata. College Station: Stata Press.
Ludwig, Jens/Enke, Harry (2013): Leitfaden zum Forschungsdaten-Management. Glückstadt: Werner Hülsbusch. https://univerlag.uni-goettingen.de/bitstream/handle/3/isbn-978-3-86488-032-2/leitfaden_DGRID.pdf [Zugriff: 19.03.2018]
Medjedovic, Irena (2014): Sekundäranalyse in der quantitativen Forschung. In: Medjedovic, Irena (Hrsg.): Qualitative Sekundäranalyse. Zum Potenzial einer neuen Forschungsstrategie in der empirischen Sozialforschung. Wiesbaden: Springer, S. 27-47.
Pampel, Heinz/Bertelmann, Roland (2011): Data Policies im Spannungsfeld zwischen Empfehlung und Verpflichtung. In: Büttner, Stephan/Hobohm, Hans-Christoph/Müller, Lars (Hrsg.): Handbuch Forschungsdatenmanagement. Bad Honnef: Bock und Herchen, S. 49-61.
Schmitt, Hermann/Hobolt, Sara B./Popa, Sebastian/Teperoglou, Eftichia (2015): European Parliament Election Study 2014, Voter Study. GESIS Datenarchiv, Köln. ZA5160 Data file Version 2.0.0. https://doi.org/10.4232/1.12300.
Schnell, Rainer/Hill, Paul B./Esser, Elke (2013): Methoden der empirischen Sozialforschung 10. München: Oldenbourg.
Stier, Winfried (1996): Empirische Forschungsmethode. Berlin: Springer.
Whiteley, Paul F./Sanders, David (2014): British Election Study, 2010: Face-to-Face Survey [computer file]. Colchester, Essex: UK Data Archive. http://www.britishelectionstudy.com/data-objects/cross-sectional-data/ [Zugriff: 03.08.2017].
Vlaeminck, Sven/Wagner, Gert G./Wagner, Joachim/Harhoff, Dietmar/Siegert, Olaf (2013): Replizierbare Forschung in den Wirtschaftswissenschaften erhöhen. Eine Herausforderung für wissenschaftliche Infrastrukturdienstleister. RatSWD Working Paper Series 224. https://www.ratswd.de/dl/RatSWD_WP_224.pdf [Zugriff: 09.03.2018].
Winters, Kristi/Netscher, Sebastian (2016): Proposed Standards for Variable Harmonization Documentation and Referencing. A Case Study Using QuickCharmStats 1.1. PLoS ONE 11(2): e0147795. https://doi.org/10.1371/journal.pone.0147795.

Linkverzeichnis

Dataverse Online: https://dataverse.org/best-practices/replication-dataset [Zugriff: 09.3.2018].
DOI –-Digital Object Identifier: https://doi.org [Zugriff: 30.05.2018].
Eurostat: Database: http://ec.europa.eu/eurostat/data/database [Zugriff: 03.08.2017].
GESIS – Leibniz-Institut für Sozialwissenschaften (2015): Replikationsserver. https://www.gesis.org/replikationsserver/home/ [Zugriff: 03.08.2017].
Politische Vierteljahresschrift: http://www.pvs.nomos.de/ [Zugriff: 30.05.2018].
URN – Uniform Resource Name: http://tools.ietf.org/html/rfc2141 [Zugriff: 20.06.2018].
Zeitschrift für Soziologie: https://www.degruyter.com/view/j/zfsoz.2016.45.issue-2/issue-files/zfsoz.2016.45.issue-2.xml [Zugriff: 30.05.2018].