Institut zur Qualitätsentwicklung im Bildungswesen
Forschungsdatenzentrum
WISSENSCHAFTLICHE EINRICHTUNG DER LÄNDER AN DER HUMBOLDT-UNIVERSITÄT ZU BERLIN E.V.

Aufbereitung der Datensätze am FDZ am IQB / Steps of Data Preparation

POSTANSCHRIFT Humboldt-Universität zu Berlin | Institut zur Qualitätsentwicklung im Bildungswesen (FDZ)
Unter den Linden 6 | 10099 Berlin
Tel: +49 [30] 2093 46552 | Fax: +49 [30] 2093 46598
fdz@iqb.hu-berlin.de | www.iqb.hu-berlin.de

KULTUSMINISTER KONFERENZ
HUMBOLDT-UNIVERSITÄT ZU BERLIN

Bibliographische Informationen / Bitte zitieren als:

Forschungsdatenzentrum am Institut zur Qualitätsentwicklung im Bildungswesen (FDZ am IQB) (2021). *Aufbereitung der Datensätze am FDZ am IQB / Steps of Data Preparation.* Berlin: IQB - Institut zur Qualitätsentwicklung im Bildungswesen. Verfügbar unter https://www.iqb.hu-berlin.de/fdz/Datenuebergabe/Vorlage_AbkStudi.pdf

Aufbereitung der Datensätze des Projekts "Titel" - Abkürzung

In dieser Übersicht werden die zentralen Schritte unserer Prüfung der Daten aus dem Projekt „Titel“ zusammengefasst. Dabei gehen wir auf Herausforderungen ein, die mit der Bereitstellung der Daten für Re- und Sekundäranalysen verbunden sind. Außerdem thematisieren wir Unklarheiten und Rückfragen zu den Daten und Begleitmaterialien an Sie als Datengebende. Wir bitten Sie, Ihre Rückmeldungen zu diesen Fragen in diesem Dokument vorzunehmen. Dazu haben wir in den einzelnen Tabellen eine Spalte mit dem Titel "Rückmeldung Datengebende" erstellt, in der Sie Ihre Antworten vermerken können. Darüber hinaus können Sie gern Kommentare oder farblich hervorgehobene Textfelder ergänzen.

Unten stehend finden Sie weiterführende Informationen zur Datenübergabe, zu unseren Schritten bei der Datenaufbereitung und zum weiteren Vorgehen bei der Archivierung und Bereitstellung der Daten aus dem Projekt "Titel". Das vorliegende Dokument umfasst neben dieser Übersicht drei weitere Tabellenreiter:

1. Dokumentation der Daten: In dieser Tabelle werden unsere Prüfungen und ggf. Rückfragen zur Verfügbarkeit von Begleitmaterialien zu den Datensätzen (z. B. Skalenhandbuch, Projektberichte) und zur Korrespondenz der Begleitmaterialien mit den übergebenen Datensätzen zusammengefasst.
2. Data Cleaning: In dieser Tabelle finden Sie unsere Rückmeldung und ggf. Fragen zur Verständlichkeit und Konsistenz der übergebenen Datensätze (z. B. Verständlichkeit der Variablen- und Wertelabels, einheitliche Kodierung fehlender Werte).
3. Datenschutzprüfung: In dieser Tabelle fassen wir datenschutzrechtlich sensible Variablen zusammen und machen Vorschläge, wie wir mit diesen Variablen bei der Bereitstellung für Sekundäranalysen umgehen würden (z. B. Zusammenfassen von seltenen Merkmalsausprägungen zu größeren Gruppen, Leeren von bestimmten Variablen).

Datenübergabe

Die Daten wurden am gemäß den AGB des Verbund Forschungsdaten Bildung über eine gesicherte Leitung im Portal meinfdb.forschungsdaten-bildung.de hochgeladen. Das FDZ am IQB hat die Daten am TT.MM.JJJJ abgerufen.

Damit auch anderen Forschenden auf Antrag Ihre Daten zur Verfügung gestellt werden können,

a) schließen wir standardmäßig einen Vertrag.
b) haben wir am TT.MM.JJJJ mit Ihnen den Datenbereitstellungsvertrag geschlossen.
c) Dieser Datenbereitstellungsvertrag befindet sich zurzeit bei Ihnen zur Unterzeichnung.

Sicherung und Nachnutzung der Daten

Zunächst wurde anhand der uns vorliegenden Informationen überprüft, ob alle übergebenen Daten auch zur Sicherung und Nachnutzung freigegeben sind. Diese Prüfung fiel folgendermaßen aus:

a) Die folgenden Datensätze sind für Re- und Sekundäranalysen verfügbar: "Datensatzname 1", "Datensatzname 2". Die folgenden Datensätze werden am FDZ des IQB archiviert und können nicht beantragt werden: "Datensatzname 1", "Datensatzname 2"

b) Für die Datensätze wurde eine Sperrfrist bis zum TT.MM.JJJJ (maximal 2 Jahre nach Datenübergabe) vereinbart. Die Datensätze stehen erst nach Ablauf dieser Sperrfrist für Re- und Sekundäranalysen zur Verfügung.

c) Die Datengebenden haben Fragestellungen formuliert, die zunächst für Sekundäranalysen gesperrt sind (bis maximal 2 Jahre nach Datenübergabe). Diese Sperrvermerke werden im Datenbereitstellungsvertrag bzw. in der Zusatzvereinbarung spezifiziert.

Schritte bei der Datenprüfung und vorgeschlagene Datenaufbereitung

Zunächst wurde überprüft, ob die übergebenen Datensätze so dokumentiert sind, dass sie von Forschenden nachgenutzt werden können. Dazu haben wir die Begleitmaterialien (z. B. Skalenhandbücher) hinsichtlich ihrer Nachvollziehbarkeit und Korrespondenz mit den übergebenen Datensätzen geprüft. Außerdem haben wir versucht, zentrale Ergebnisse, Stichprobengrößen und Kennwerte zu den Variablen aus den Begleitmaterialien (z. B. Projektberichte) zu replizieren. Unsere Ergebnisse und Rückfragen sind im Tabellenreiter **"1. Dokumentation der Daten"** hinterlegt. Anschließend haben wir die übergebenen Datensätze dahingehend geprüft, ob die Variablen und Werte in den Variablen ausreichend beschrieben und sinnvoll interpretiert werden können (siehe Tabellenreiter **"2. Data Cleaning"**). Abschließend haben wir die Datensätze hinsichtlich datenschutzrechtlich sensibler Variablen und Merkmalsausprägungen geprüft (siehe Tabellenreiter **"3. Datenschutzprüfung"**).

Aufgrund gesetzlicher Vorgaben zur Sicherung der Vertraulichkeit von Einzelangaben müssen in den Datensätzen alle Informationen entfernt werden, die eine **potenzielle Re-Identifikation** einzelner Personen oder Institutionen ermöglichen würden. Dies kann u. a. dann gegeben sein, wenn bei bestimmten Variablen Merkmalsausprägungen vorliegen, die von fünf oder weniger Personen stammen, oder wenn kleinteilige geographische oder detaillierte biographische Informationen vorliegen. Dazu gehören auch sogenannte String-Variablen, in denen die Antworten der Teilnehmenden über freie Angaben und Textfelder erfasst werden.

An unserem FDZ stellen wir unterschiedliche **Datensatzversionen** für Re- und Sekundäranalysen bereit, bei denen die Datenschutzvorgaben wie folgt berücksichtigt werden. Die von Ihnen übermittelten **Originaldaten** werden von uns archiviert und nicht verändert mit der Ausnahme, dass direkte *Identifier* (z. B. Namen, Adressen, Telefonnummern) gelöscht werden. Die Originaldatensätze werden nicht für Re- und Sekundäranalysen bereitgestellt. Als **Scientific Use Files (SUFs)** bezeichnen wir Datensatzversionen, die interessierte Forschenden nach Antragsstellung und Abschluss eines Datennutzungsvertrags per Download erhalten. Die SUF-Datensätze werden von uns auf Basis der Originaldaten erstellt und so rekodiert, dass keine datenschutzrechtlich sensiblen Variablen und Merkmalsausprägungen enthalten sind (siehe unsere Vorschläge im Tabellenreiter "3. Datenschutzprüfung"). Als **IZA-Datensätze** bezeichnen wir Datensatzversionen, in der bestimmte datenschutzrechtlich bedenkliche Variablen nicht rekodiert oder geleert werden. Diese Datensätze können von Datennutzenden nur im Fernrechnenmodus ausgewertet werden, welches vom Institut zur Zukunft der Arbeit (IZA) bereitgestellt wird. In diesem Modus senden die Datennutzenden ihre Auswertungssyntaxen an uns, wir führen sie aus und prüfen den daraus resultierenden Output hinsichtlich datenschutzrechtlicher Bestimmungen und geben den Output dann an die Datennutzenden zurück. In diesem Modus haben die Datennutzenden also keinen direkten Zugriff auf die Datensätze.

Beim Umgang mit datenschutzrechtlich sensiblen Variablen wird das Re-Identifikationspotenzial in Verbindung mit allen aus den Daten- und Projektberichten herleitbaren Informationen berücksichtigt und mit dem Analysepotenzial dieser Variablen in Verbindung gesetzt. Das FDZ am IQB ergreift in seinem Ermessen Maßnahmen, um einen Personenbezug der Datensätze so gut wie unmöglich zu machen und dabei gleichzeitig das analytische Potential der Datensätze beizubehalten. Alle Veränderungen in den SUF-Datensätzen werden durch das FDZ dokumentiert und im Datensatz kenntlich gemacht.

Weiteres Vorgehen

Bitte geben Sie uns bis zum TT.MM.JJJJ Rückmeldung, ob Sie bezüglich der Änderungen am Datensatz ein anderes als das vorgeschlagene Vorgehen präferieren bzw. noch Informationen nachreichen können. Ansonsten gehen wir wie beschrieben vor. Weiter bitten wir Sie, die in den Tabellen angeführten Fragen zu beantworten.

Nachdem Sie uns als Datengebende Ihr Einverständnis zur Durchführung der vorgeschlagenen Änderungen gegeben haben, wird das FDZ am IQB die entsprechenden Rekodierungen und Anpassungen der Datensätze vornehmen. Weiterhin wird in jeden Datensatz eine Versionsvariable eingefügt, um bei möglichen Änderungen nach Publikation der Daten die ausgelieferte Version zurückverfolgen zu können.
Nach Fertigstellung der Scientific Use Files erhält das Datenpaket einen Digital Object Identifier (DOI) und wird so eindeutig recherchier- und zitierbar. Die Daten und zugehörigen Dokumentationsmaterialien werden interessierten Forschenden über die Website des IQB auf Antrag kostenfrei zur Verfügung gestellt und können unter der Bedingung, dass sie ordnungsgemäß zitiert werden, für wissenschaftliche Analysen und Publikationen genutzt werden.

1. Dokumentation der Daten

Allgemeine Fragen und Anmerkungen

Frage/Anmerkung	Rückmeldung Datengebende
z. B.: Fehlen zentrale Dokumentationen, Ergebnisberichte, Skalenhandbücher, Methodenberichte, Publikationslisten etc.?	
z. B.: Wie lassen sich Teildatensätze miteinander verknüpfen? Welche ID-Variablen werden dafür benötigt?	
z. B.: Gibt es Angaben zu den internen Konsistenzen (z. B. Cronbachs Alpha) der eingesetzten Skalen, die in der Skalendokumentation ergänzt werden könnten?	
z. B.: Beinhalten die ID-Variablen Informationen über den Schulstandort bzw. regionale Kennziffern?	

Korrespondenz zwischen Skalenhandbuch und Datensätzen

Variablennamen im Skalenhandbuch, die nicht im Datensatz vorkommen	Variablennamen im Datensatz, die nicht im Skalenhandbuch vorkommen	Rückmeldung Datengebende

Fragen zur Dokumentation spezifischer Datensätze und Variablen (z. B. Replikation deskriptiver Statistiken)

Datenquelle	Variablenamen	Variablenlabel	Seite im Skalendhandbuch	Frage	Rückmeldung Datengebende

2. Data Cleaning

Allgemeine Fragen und Anmerkungen

Frage/Anmerkung	Rückmeldung Datengebende
Wenn in Variablen- und Wertelabel Sonderzeichen bzw. Umlaute (z. B. ß, ä, ö, ü) verwendet werden, ersetzen wir diese im Zuge der Datenaufbereitung durch reguläre Zeichen (z. B. ae, oe, ue).	

Fragen, Anmerkungen und Rekodiervorschläge zu spezifischen Variablen

Datensatzname	Variablename	Variablenlabel	Frage, Anmerkung, Rekodiervorschlag	Rückmeldung Datengebende

3. Datenschutzprüfung

Allgemeine Hinweise

Die übergebenen Datensätze wurden hinsichtlich datenschutzrechtlich sensibler Variablen und Merkmalsausprägungen geprüft. Datenschutzrelevante Variablen sind alle Variablen, die Rückschlüsse auf statistische Einheiten wie z. B. Personen oder Schulen zulassen. Als riskant gelten dabei nicht nur Informationen wie beispielsweise Name, Adresse oder Sozialversicherungsnummer, sondern auch Merkmale, die eine indirekte Identifikation ermöglichen. Zu diesen Merkmalen gehören

(1) besondere Kategorien personenbezogener Daten (z. B. politische Meinung, religiöse Überzeugungen, genetische und medizinische Daten, psychosomatische Beschwerden),

(2) herkunftsidentifizierende Informationen (z. B. Muttersprache, Geburtsland, Berufsbezeichnung der Eltern),

(3) regionale Kennziffern (z. B. Geburtsort, Wohnort, Schulstandort),

(4) Freitextangaben.

Eine erhöhte Re-Identifikationsgefahr ist u. a. dann gegeben, wenn bei bestimmten Variablen Merkmalsausprägungen vorliegen, die von fünf oder weniger Personen stammen. Wir schlagen vor, alle potenziell sensiblen Variablen zu rekodieren und haben zu diesen Variablen in der unten stehenden Tabelle Lösungvorschläge notiert, zu denen Sie uns Ihre Rückmeldung geben können.

Die Rekodierung bzw. Leerung datenschutzrechtlich sensibler Variablen wird in den Datensatzversionen der Scientific Use Files (SUFs), diese Datensätze erhalten unsere Antragsstellenden per Download) und ggf. in den IZA-Datensätzen (diese Datensätze stehen nur im Fernrechenmodus zur Verfügung, Analyseergebnisse werden in diesem Modus erst nach Kontrolle durch das FDZ bereitgestellt) vorgenommen. Die Originaldatensätze werden im Zuge der Aufbereitung nicht verändert und stehen nicht für Re- und Sekundäranalysen zur Verfügung. Wir werden die unten aufgeführten Lösungsvorschläge zum Umgang mit sensiblen Variablen nach Ihrer Zustimmung bzw. nach Ihren Vorschlägen selbst durchführen. Weitere Informationen finden Sie im Tabellenreiter "Übersicht".

Möglichkeiten des Umgangs mit sensiblen Variablen

Üblicherweise kann bei der Erstellung der Scientific Use Files (SUFs) mit datenschutzrechtlich bedenklichen Variablen wie folgt vorgegangen werden:

(1) Variablenunterdrückung: Variablen werden in den SUFs geleert (und beinhalten keine Werte), diese Variablen können nur über einen sicheren Datenzugang (Fernrechnen) zu den IZA-Datensätzen genutzt werden.

(2) Variablenrekodierung:

- *Lokale Suppression*: Personen, die seltene Merkmalsausprägungen aufweisen, erhalten auf der betroffenen Variable fehlende Werte.
- *Vergröberung*: Zusammenfassung seltener Kategorien zu größeren Merkmalsgruppen, dies kann u. a. wie folgt geschehen:
  - für betroffene Variablen werden Kategorien mit Fallzahlen von n ≤ 5 zu einer Kategorie „Sonstiges“ zusammengefasst
  - diskrete Merkmalsausprägungen, die selten vorkommen, werden mit ähnlichen Ausprägungen zusammengefasst und bilden gemeinsam eine neue Kategorie
  - Bildung von Grenzwerten: Ausprägungen auf einem Merkmal werden durch eine Ober- und Untergrenze abgeschnitten
- *Skalenbildung*: Einzelne Items werden zu Skalen zusammengefasst bzw. wenn Skalen vorliegen, werden die Einzelitems aus den Daten entfernt

(3) Eingeschränkter Datenzugang:

Die Datensätze werden nicht als SUF an die Nutzenden ausgeliefert; der Zugang wird nur über sichere Datenzugangswege (Fernrechnen) gewährt. Alternativ zur Sperrung der kompletten Datensätze könnte man auch nur solche Teildatensätze, die einen hohen Anteil datenschutzrechtlich bedenklicher Variablen aufweisen, ausschließlich über sichere Datenzugangswege verfügbar machen.

Spezifische Fragen, Anmerkungen und Rekodiervorschläge zu datenschutzrechtlich bedenklichen Variablen

Datensatz	Variablenname	Variablenlabel	Grund für Beabeitungsbedarf/ Gefährdungspotenzial	Lösungsvorschlag	Rückmeldung Datengebende	interne Rückmeldung zur Aufbereitung IQB-WiMi