# Protokoll zur Dokumentation empirischer Arbeiten

zuletzt aktualisiert: 23. Juli 2015

**Universität zu Köln**

**Wirtschafts- und Sozialwissenschaftliche Fakultät**

**Institut für Soziologie und Sozialpsychologie**

**Soziologie II**

**Prof. Dr. Marita Jacob**

**Dr. Michael Kühhirt**
**Dr. Judith Offerhaus**

Sekretariat Petra Altendorf:
Telefon+49 221 470-5652
Telefax +49 221 470-5025
altendorf@wiso.uni-koeln.de
www.soziologie.uni-koeln.de/
Greinstr. 2 - 50939 Köln

Im Folgenden werden Richtlinien für die Dokumentation Ihrer empirischen Arbeit dargelegt, um die Reproduzierbarkeit Ihrer Datenanalyse zu sichern. Diese Dokumentation geben Sie auf CD/DVD gebrannt zusammen mit Ihrer schriftlichen Abschlussarbeit ab.

*Reproduzierbarkeit* und *Replizierbarkeit* sind grundlegende Anforderung an wissenschaftliche Arbeiten. Reproduzierbarkeit bezeichnet die Erzielung identischer Ergebnisse bei Wiederholung der exakt gleichen Analyse, also unter Verwendung identischer Daten, Methoden, Software und ggf. Programmierung. Unter Replizierbarkeit versteht man die Erzielung vergleichbarer Ergebnisse unter verschiedenen Graden von Änderungen der Originalanalyse, beispielsweise die Verwendung anderer Daten oder Analysemethoden und die Nutzung alternativer Operationalisierungen.

Bei der Erstellung einer Abschlussarbeit ist die Reproduzierbarkeit der Analyseschritte und Ergebnisse vorrangig. Empirische Analysen, die nicht reproduzierbar sind oder die bereits bei trivialen Änderungen nicht mehr repliziert werden können, sind unwissenschaftlich und werden am Lehrstuhl nicht angenommen. So ist beispielsweise eine unzureichende Beschreibung der Datenaufbereitung, fehlerhafter Software-Code oder gar das vollständige Fehlen jedweder Dokumentation der Analyse (z.B. Point&Click-Analysen) nicht zulässig.

Mit dem vorliegenden Protokoll erläutern wir Ihnen, wie Ihre empirischen Analysen und deren Dokumentation den folgenden drei zentralen Kriterien genügen:

1. Nach Anpassung des Arbeitsverzeichnisses läuft Ihr Analysecode *fehlerfrei* und *an einem Stück* auf *jedem* Rechner, auf dem die Originaldaten sowie die von Ihnen verwendete Software (inkl. Zusatzpakete) vorliegen bzw. installiert sind.
2. Der durch diesen Vorgang produzierte Output (allen voran Tabellen und Abbildungen) entspricht *genau* den Ergebnissen in Ihrer schriftlichen Arbeit.
3. Die Analyse ist an jeder Stelle *transparent dokumentiert*, so dass sie ohne persönlichen Kontakt gestartet, durchgeführt und verstanden werden kann. Hierbei zentral ist auch übersichtlich gegliederter und ausführlich kommentierter Softwarecode.

Um diese Anforderungen zu erfüllen, befolgen Sie am besten die nun folgenden Arbeitsschritte. Zusätzlich stellen wir Ihnen eine Muster-Dokumentation inkl. Dateivorlagen auf der Lehrstuhl-Homepage zur Verfügung (`template.zip`).

## I. Verzeichnisstruktur und -inhalt

Der erste Schritt für die Dokumentation Ihrer Arbeit besteht darin, ein Arbeitsverzeichnis an beliebiger Stelle auf Ihrem Rechner anzulegen. Dort richten Sie die in Abbildung 1 dargestellte Verzeichnisstruktur ein.[1] Die Abbildung gibt zudem Auskunft über den Inhalt der einzelnen (Unter-)Verzeichnisse.

Ihr Rechner
- Arbeitsverzeichnis der Abschlussarbeit (z.B. `/masterthesis`)
  - `code`
    *Hier werden sämtliche Code-Dateien abgelegt.*
  - `data`
    - `original`
      *Hier werden die Originaldaten, auf denen Ihre Analyse beruht,* unverändert *abgelegt.*
    - `prepared`
      *Hier werden von Ihnen erstellte Datensätze abgespeichert.*
  - `docu`
    - `codebooks`
      *Hier werden sämtliche Codebooks zu den Originaldaten, aber auch zu Ihren eigenen Analysedatensätzen abgelegt. Benennen Sie Codebooks stets so:* `cb_[datenname einsetzen]`
    - `questionaires`
      *Hier werden relevante Fragebögen für die Originaldaten abgelegt.*
  - `figures`
    *In dieses Verzeichnis legen Sie die Abbildungen für Ihre Arbeit.*
  - `manuscript`
    *Hier liegt eine PDF-Version Ihrer endgültigen Arbeit.*
  - `tables`
    *In dieses Verzeichnis gehören die Tabellen für Ihre Arbeit.*

**Abbildung 1:** *Struktur und Inhalt der Verzeichnisse für die Dokumentation Ihrer Abschlussarbeit*

In Ihrem Arbeitsverzeichnis können Sie im weiteren Verlauf auch weitere (Unter-)Verzeichnisse anlegen, z.B. für heruntergeladene Literatur oder die Verschriftlichung einzelner Kapitel. Als Dokumentation abge-

[1] Dazu können sie schlicht das Archiv `template.zip` an beliebiger Stelle abspeichern und anschließend entpacken. Zum Schluss ist lediglich der Name des Arbeitsverzeichnis nach Ihren Wünschen anzupassen. Verzichten Sie dabei auf allzu lange Namen sowie auf Leer- und Sonderzeichen.

ben werden Sie jedoch lediglich die oben aufgeführten Verzeichnisse und deren Inhalt.

**II. Originaldaten**

Nachdem Sie sich die Daten, die Sie analysieren möchten, besorgt haben, legen Sie diese in das Verzeichnis `/data/original`. Ändern Sie weder Inhalt noch Namen der Originaldaten! Dies geschieht erst über den von Ihnen geschriebenen Code. Als Originaldaten zählen auch offizielle Statistiken, beispielsweise der OECD, die Sie aus dem Internet bezogen haben.

Erkundigen Sie sich, ob Ihren Daten ein Digital Object Identifier (kurz: DOI) zugewiesen ist und dokumentieren Sie diesen in Ihrer `readme`-*Datei* (siehe unten). Der DOI dient dazu, Ihren Datensatz, inkl. Versionsnummer, zweifelsfrei zu bestimmen. Vor allem bei größeren Datensätzen wie dem Allbus oder dem Sozio-ökomischen Panel (SOEP) kommt es häufig vor, dass nach einer Erstveröffentlichung ein oder mehrere Updates des Datensatzes erscheinen, die Informationen ergänzen oder Fehler korrigieren. Wenn unbekannt ist, auf welcher Version der Daten Ihre Analysen aufbauen, können diese u.U. aufgrund von Abweichungen in den Rohdaten nicht reproduziert werden.

**III. Für die Dokumentation notwendige Dateien**

*a. Readme-Datei:* In Ihrem Arbeitsverzeichnis legen Sie eine Textdatei an, die

1. Ihre Dokumentationsverzeichnisse und -dateien beschreibt,
2. die verwendeten Daten (inkl. Versionsnummern und DOI) und deren Herkunft genau benennt
3. die genutzte Software (inkl. Versionsnummern) ausweist und
4. die für die Reproduktion Ihrer Analyse notwendigen Schritte darlegt.

Nutzen Sie für die Erstellung dieser Datei keine aufwendige Formatierung: eine einfache txt-Datei genügt. In der Musterdokumentation finden Sie ein Grundgerüst für eine solche Datei. Erstellen Sie daraus am Ende Ihrer Arbeit eine PDF-Datei mit dem Namen `readme.pdf`.

*b. Master Code-File:* Im Verzeichnis `/code` legen Sie eine Code-Datei mit dem Namen `00master.do` an, die der Ausgangspunkt für Ihre gesamte Datenaufbereitung und -analyse ist. Diese Datei

1. installiert von Ihnen genutzte Zusatzpakete der Analysesoftware
2. spezifiziert Makros für Ihre Projektverzeichnisse

3. zeigt Namen und Reihenfolge der von Ihnen genutzten Code-Dateien an und führt diese ggf. aus.

Auch für diese Datei finden Sie in der Musterdokumentation eine Vorlage, die Sie für Ihre Arbeit anpassen können.

*c. Codebook des Analysedatensatzes:* Unabhängig davon, ob Codebooks für die Originaldaten vorliegen, erstellen Sie für den Datensatz, mit dem Sie Ihre Analysen durchführen eine Datei cb_[Name Ihres Analysedatensatzes].txt, die die in diesem Datensatz enthaltenen Variablen und Ihre Ausprägungen grob beschreibt. Die Datei `02an_verify.do` im Ordner `/code` enthält unter `@4` Beispielcode, mit dem Sie in Stata über den `codebook`-Befehl automatisch eine solche Datei über eine `log-file` anlegen können.

## IV. Zur Erstellung reproduzierbarer Analysen

Bei der Erstellung des Codes für die Datenaufbereitung und -analyse ist es essentiell, dass Sie auf eine klare und transparente Struktur achten, Ihre einzelnen Arbeitsschritte ausführlich kommentieren und Code und Daten sorgfältig und fortlaufend auf Fehler überprüfen.
Beachten Sie dazu bitte die folgenden Hinweise:

- Unterteilen Sie Ihre Code-Dateien in solche, die der Datenaufbereitung dienen und solche, für die Datenanalyse.
- Geben Sie ersteren den Präfix `cr_` und letzteren den Präfix `an_`.
- Vermischen Sie *niemals* Datenaufbereitung und Datenanalyse in einer Code-Datei.
- Legen Sie für jeden Datensatz, den Sie erstellen eine eigene Code-Datei an und benennen Sie jeden Datensatz nach der ihn erstellenden Code-Datei (z.B. `01cr_andata` und `andata.dta`)
- Besteht Ihre Analyse aus sehr vielen Teilen, legen Sie für jede sinnvolle Analyseeinheit eine spezifische Code-Datei an.
- Nummerieren Sie die Code-Dateien in der Reihenfolge, in der diese ausgeführt werden.
- Unterteilen Sie jede Code-Datei in sinnvolle Abschnitte, die Sie durch `@` nummerieren und weitere Unterpunkte, die Sie mit Großbuchstaben kennzeichnen.
- Achten Sie darauf, dass einzelne Code-Zeilen nicht über 80 Zeichen enthalten und schreiben Sie einen Befehl über mehrere Zeilen mithilfe des Kommentarzeichens `///` oder dem Befehl `delimit`.

- Nutzen Sie die Kommentarfunktionen sowie den `note`-Befehl zur ausführlichen Beschreibung des Codes und der von Ihnen erstellten Variablen.
- Erstellen Sie mindestens eine Code-Datei, in der Sie den Analysedatensatz genau auf fehlerhafte Codierung, inhaltliche Inkonsistenzen u.ä. überprüfen und in der Sie das Codebook erstellen (siehe `/code/02an_verify.do`).
- Erstellen Sie mindestens eine Code-Datei für die ausführliche Beschreibung Ihres Analysesamples mithilfe univariater Statistiken zu wichtigen Hintergrund- und zentralen Analysevariablen (siehe `/code/03an_sample.do`).
- Automatisieren Sie Datenaufbereitung, -analyse und Ergebnisausgabe soweit wie möglich über Code-Schleifen u.ä und vermeiden Sie das „Abtippen“ von Zahlen aus dem Output.
- Exportieren Sie relevanten Output für Tabellen und Abbildungen direkt mit dem korrekten Namen in den vorgesehenen Ordnern `/tables` und `/figures` anstatt diese per Hand zu verschieben und umzubenennen (siehe `/code/04an_result.do`, für Beispielcode).
- Vermerken Sie in der Codedatei, wo genau in Ihrer Arbeit auf einen bestimmten Output zurückgegriffen wird (z.B. Tabelle 2, S. 27) und vermerken Sie umgekehrt in Ihrer Arbeit, in welcher Code-Datei und aus welchem Abschnitt der verwendete Output stammt (z.B. `04an_result.do@4B`).
- Benennen Sie Tabellen und Abbildungen stets so wie in Ihrer schriftlichen Arbeit (z.B. `abb1.jpg`). Falls eine Abbildung in der Arbeit aus mehreren Einzelteilen besteht, nutzen Sie für den Export Kleinbuchstaben, um die Einzeldateien zu differenzieren (z.B. `abb1a.jpg`).

Weiterführende Anleitungen für effizienten und transparenten Code finden sich bei beispielsweise bei Long (2008) und Kohler & Kreuter (2012).

## V. Test der Dokumentation

Vor Abgabe Ihrer Arbeit testen Sie bitte ob Ihre Dokumentation auf einem zweiten Rechner funktioniert. Haben Sie sämtliche Zusatzsoftware angegeben, auch in dem Master-Codefile? Laufen alle Code-Dateien fehlerfrei durch? Sind die Angaben in allen Dokumentationsdateien vollständig? Haben Sie die erforderlichen PDF-Dateien erstellt?

**Literatur**


Freese, J. (2007). Replication standards for quantitative social science - Why not sociology. *Sociological Methods & Research*, 36 (2), 153-172.

Haverford College (2014). *Project TIER: Teaching Integrity in Empirical Research*, Online unter: http://www.haverford.edu/TIER/ Letzter Abruf: 12.03.15

Ioannidis, J. P. A. (2005). Why most published research findings are false. *Plos Medicine*, 2 (8), 696-701.

Ioannidis, J. P. A. (2014). How to make more published research true. *Plos Medicine*, 11 (10).

Janz, N. (2015). Bringing the gold standard into the classroom: Replication in university teaching. *International Studies Perspectives*, Early View, DOI: 10.1111/insp.12104.

King, G. (1995). Replication, Replication. *PS: Political Science & Politics*, 28 (3), 444-452.

Kohler, U., & Kreuter, F. (2012). *Datenanalyse mit Stata: Allgemeine Konzepte der Datenanalyse und ihre praktische Anwendung*. München: Oldenbourg.

Long, S. J. (2008). *The Workflow of Data Analysis Using Stata*. Texas: Stata Press.

McCullough, B. D., McGeary, K. A., & Harrison, T. D. (2008). Do economics journal archives promote replicable research? *Canadian Journal of Economics-Revue Canadienne D Economique*, 41 (4), 1406-1420.

Miguel, E. et al. (2014). Promoting transparency in social science research. *Science*, 343 (6166), 30-31.

Nosek, B. A. et al. (2015). Promoting an open research culture. *Science*, 348(6242), 1422-1425.

Stodden, V. (2015). Reproducing statistical results. *Annual Review of Statistics and Its Application* 2, 1-19.

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist*, 61 (7), 726-728