(Stand 02.10.2018)

GfP 5 – Umgang mit Daten und Dokumentation des Forschungsprozesses

Die Ergebnisse von Forschungsarbeiten sind nur dann intersubjektiv überprüfbar, wenn auch ein Zugang zu den Daten gewährt wird bzw. im Nachgang auch deren Überprüfbarkeit sichergestellt wird. Wie kann dieser Zugang sichergestellt werden? Für welchen Zeitraum? Wie kann mit vertraulichen Daten umgegangen werden? Wie werden die Interessen der Autor/innen gewahrt, die die Daten unter Umständen mit viel Aufwand gesammelt haben und nun auch alleine davon profitieren wollen? Wie sollte eine revisionssichere Dokumentation von Forschungsarbeiten aussehen? Welche Sachverhalte - insbesondere bei empirischen Untersuchungen (z.B. Projektpläne, Publikationsprojektpläne, Wechsel von Autorenschaften, Anpassung des Rohdatensatzes durch Bereinigungen usw., inhaltliches vs. rechtliches Eigentum an Daten und Erkenntnissen, Art der Mitarbeit usw.) sollen in welcher Weise dokumentiert werden, um später zur Klärung möglicher Probleme beitragen zu können? Wie sind diese Unterlagen bzw. auch die verwendeten (Roh-)Daten zu archivieren – in welchem Format, mit welchen Sicherungsinstrumenten und für welchen Zeitraum? Wie kann sichergestellt werden, dass die Arbeitsschritte des Forschers aus den ursprünglichen Daten auch nachvollzogen werden können?

Eine notwendige Anforderung an wissenschaftliche Arbeiten besteht in der Nachvollziehbarkeit der erzielten Ergebnisse. Die Erfüllung dieser Anforderung ist bei jedem Forschungsprojekt von Anfang an zu verfolgen und langfristig zu gewährleisten. Die Herausgeber wissenschaftlicher Zeitschriften haben deshalb von je her darauf geachtet, dass der Forschungsprozess zur Erzielung der in einem Aufsatz berichteten Ergebnisse angemessen dokumentiert ist. Anders als in den Naturwissenschaften hat sich in den Wirtschafts- und Sozialwissenschaften allerdings keine Kultur der regelmäßigen Replikation von Ergebnissen herausgebildet, weshalb auch die Angemessenheit der Dokumentationen in der Regel nicht geprüft wurde. Ausgelöst durch verschiedene Betrugsfälle erwarten Herausgeber und Forschungsförderungsinstitutionen wie die DFG heute genauere **Dokumentationen** und zum Teil auch das **Verfügbarmachen der verwendeten Daten**. Neben der Redlichkeit wissenschaftlichen Vorgehens möchte man damit auch die Dynamik – jeweils darauf aufbauender – wissenschaftlicher Fortschritte fördern. Die Ausführungen in diesen GfP unterstreichen die Wichtigkeit der Thematik. Potentielle Konsequenzen eines Nichtbefolgens können dramatisch sein. Während es bis dato noch ausreichend gewesen sein mag, sich als Doktorand/in an Datenschutzbestimmungen zu halten, soll zukünftig eine Bestätigung verlangt werden, dass die Grundsätze ordnungsmäßiger Dokumentation und Datensicherung eingehalten wurden. Zudem bietet professionalisiertes Datenmanagement Nachwuchswissenschaftler/innen, Betreuer/innen und Ko-Autor/innen eine höhere Sicherheit vor Plagiats- und Reputationsrisiken. Und nicht zuletzt erfährt die Phase der Datenerhebung durch Datenmanagement eine Aufwertung, die sie zum intellektuellen Wert an sich macht. Daten(sätze) „zitierbar“ zu veröffentlichen erscheint wertvoll und würde eine effiziente Nachnutzung bestehender Datensätze gewährleisten.

Dokumentation von Daten:

Forschungsdaten sind alle Informationen, die ein legitimierter (und qualifizierter) Dritter benötigt, um die Ordnungsmäßigkeit der betreffenden Forschungsarbeit zu überprüfen. Mit ihnen verbunden sind auch die zu ihrem Verständnis erforderliche Dokumentation und Software, die Quellcodes, Transkripte, sowie andere relevante Forschungsdaten. Die Offenlegung dieser Informationen stellt sicher, dass einem legitimierten Dritten die Nachvollziehbarkeit von Forschungsergebnissen nicht unnötig erschwert wird. Um die Ergebnisse von empirischen Arbeiten besser beurteilen zu können, ist es vorteilhaft, die verwendeten Daten genau zu beschreiben. Bei öffentlich verfügbaren Sekundärdaten gibt man am besten die genauen Quellen (z.B. URL-Adresse) an, so dass andere die Daten ebenso analysieren können. Handelt es sich um selbst gesammelte Umfragedaten, gibt man an, was die Grundgesamtheit ist, wie das Sampling erfolgte, zu welchen Zeitpunkten welche Teile der Daten erhoben worden sind, wie die Antwortrate war sowie (wenn bekannt) die Verteilung von Befragten-Charakteristika in der Grundgesamtheit und im Sample zum Vergleich. Bei den heute üblichen Online-Panels gibt man die Menge der adressierten Personen an, die Menge der Abbrecher und die Menge der Befragten mit kompletten Datensätzen. Hat man Datensätze von Unternehmen zur Verfügung gestellt bekommen, z.B. alle Daten von Kunden und ihre Käufe, so ist die Struktur dieser Daten so gut wie möglich zu beschreiben, ohne dass Vertraulichkeitsvereinbarungen verletzt werden. Hier ist insbesondere darauf zu achten, dass man nicht auf z.B. die Identität des Unternehmens, Kunden oder Mitarbeiter schließen kann.[1] Bei Umfragen ist idealerweise der gesamte Fragebogen anzugeben, so dass der Leser sich selber ein Bild über die Operationalisierung von Konstrukten machen kann. Auf jeden Fall sollte berichtet werden, für welche Variablen insgesamt Daten erhoben worden sind, damit man abschätzen kann, ob für bestimmte Analysen ein „variable omission bias“ vorliegen kann. Das Ziel muss darin bestehen, so viel wie möglich über die Daten auszusagen, damit der Leser besser erkennen kann, ob die Ergebnisse mit eventuellen Besonderheiten der Daten zu tun haben können. Sekundärdaten aus kommerziellen Datenbanken sollten wie Unternehmensdaten behandelt werden.

Von den Rohdaten zu analysierbaren Daten:

Die gesammelten Rohdaten werden in der Regel aufbereitet, um zur Analyse verwendet werden zu können. Dieser Prozess ist genau zu dokumentieren, damit man die Ergebnisse reproduzieren kann. Z.B. sollte berichtet werden, wie man mit Missing Values und Ausreißern umgegangen ist. Sind Datensätze gelöscht worden, weil sie z.B. Fehler enthalten, so ist dies genauestens zu berichten. Gleiches gilt für Online-Umfragen, sofern offenbar zufällige oder willkürliche Antworten gelöscht worden sind, die Antwortende gegeben haben, um irgendwelche Belohnungen für das Ausfüllen von Fragebögen zu bekommen. Natürlich sind dann auch die Kriterien zu berichten, mit denen man auf zufällige bzw. willkürliche Antworten schließen kann. Ausreißer zu eliminieren ist in der Regel falsch, da dann die Normalverteilungsannahmen des Samples noch weniger gelten (Laurent 2013). Winsorizing (auch zum Zwecke der Überprüfung der Robustheit von Ergebnissen) kann durchaus legitim sein. Dies ist aber keine Frage des Datenmanagements, sondern eine Ermessensentscheidung des Forschers. Am besten lässt sich die Dokumentation mit einem Skript realisieren, das wie in einem Batch-Betrieb alle Schritte der Modifikationen des ursprünglichen Datensatzes aufzeichnet, so dass man jederzeit den analysierten Datensatz aus dem Rohdatensatz automatisch reproduzieren kann. Veränderungen am Original-Datensatz müssen dem interessierten und legitimierten Dritten nachvollziehbar geschildert werden.

Ergebnisse:

Ergebnisse mussten in den Zeiten knapper Print-Ressourcen häufig sehr knapp dargestellt werden, wobei dies auf viele nur noch online angebotene Zeitschriften nicht mehr zutreffen dürfte. Wünschenswert ist, dass dem Leser der Forschungsarbeit ein möglichst aussagekräftiges Bild über die Datenlage vermittelt wird. Beispiele und Konkretisierungen sollten nur illustrativ verwendet werden und sind nie abschließend zu verste-

[1] Zum Umgang mit Sperrklauseln, die insbesondere bei Forschungskooperationen mit Unternehmen relevant sein können, finden sich Hinweise in GfP 7.

hen. In jedem Fall sind die Daten zunächst an Hand der deskriptiven Maße für die Variablen wie Mittelwert, Standardabweichung, Minimum, Maximum oder Anteile (bei Dummy-Variablen) so zu beschreiben, dass der Leser sich selbst einen Reim darauf machen kann, ob die Daten typisch für das zu zeigende Phänomen sind bzw. warum bestimmte Ergebnisse eingetreten sind. Zum besseren Verständnis der Variablen sollte auch immer eine Korrelationstabelle angegeben werden. Für die Ergebnisse der (meist statistischen) Analysen ist es hilfreich, nicht nur die Information anzugeben, ob eine Variable signifikant ist, sondern auch den Wert des Koeffizienten, den Standardfehler, eine Test-Statistik wie den t-Wert und die Fehlerwahrscheinlichkeit (p-value) sowie eine Angabe der Ergebnisse eines Multikollinearitätstests. Um den Fokus nicht nur auf die statistische Signifikanz, sondern auch auf die inhaltliche Signifikanz zu legen, sollten auch Gütemaße für die gesamte Analyse angegeben werden. So wie man für lineare Regressionen in der Regel nicht die fast aussagelose Summe der Fehlerquadrate, sondern einen R^2-Wert angibt, der über Studien vergleichbar ist, so sollte man z.B. auch bei Maximum-Likelihood-Schätzungen nicht allein die Summe der logarithmierten Likelihoods angeben, sondern ein über die Studien vergleichbares Gütemaß wie die mittlere Likelihood oder einen Hold-out-Prognosefehler. Will man die Höhe von Koeffizienten vergleichbar machen, so sollte man zusätzlich entweder Elastizitäten, Mittelwertdifferenzen oder wenigstens standardisierte Koeffizienten angeben.

Datenarchiv:

In der Vergangenheit war es nicht üblich, Daten zur Verfügung zu stellen. Dies hatte etwas mit dem begrenzten Raum der Print-Zeitschriften zu tun. Heute finden sich bei Bedarf Wege, Daten im Zeitalter des Internets als Zusatz online zu stellen. Auf freiwilliger Basis konnte dies schon immer erfolgen, z.B. bei GESIS (www.gesis.org). Nun verlangen es aber immer mehr Zeitschriften, siehe z.B. die editorial policy von Marketing Science (Desai 2013). Elsevier bietet beispielsweise für bestimmte Zeitschriften ein sog. „Database Linking Tool“ an (Elsevier 2017). Die DFG verlangt, dass die Daten wenigstens 10 Jahre gespeichert sind (Deutsche Forschungsgemeinschaft, 1998, p. 55), um im Fall der Fälle darauf zurückgreifen zu können. Weiterhin sollten insbesondere projekt- oder fördermittelspezifische Vorgaben beachtet werden.

Viele Forscher fürchten, dass bei einer Veröffentlichung ihrer Daten ihr Wettbewerbsvorteil verloren geht. Übersehen wird dabei, dass Aufsätze mit Daten einen höheren Zitations-Impact haben (Albers 2012). Allerdings muss die zugesagte Vertraulichkeit sichergestellt werden, keine kleine Herausforderung. Mit Profiling können Dritte unter Umständen aus den Charakteristika von Befragten auf deren Identität zurückschließen. Nur wenn dies ausgeschlossen werden kann, sollte man Daten anonymisiert in einem Web Appendix zur Verfügung stellen. Zur Sicherung des Wettbewerbsvorteils kann man Daten auch unter der (schriftlich vereinbarten) Bedingung zur Verfügung stellen, dass diese nur für identische Replikationen genutzt werden können, nicht aber für weitere Analysen. Selbst bei Vertraulichkeitsvereinbarungen sollte man mit einem Unternehmen verhandeln, dass die Daten wenigstens nach einem zeitlich befristeten Embargo (z.B. nach 10 Jahren) freigegeben werden, wenn dort keine wettbewerbsschädliche Gefahr mehr vorhanden ist. Bei der Verwendung anonymisierter Sekundärdaten, z.B. vom ifo Institut, dem UK Office of National Statistics oder von der Deutschen Bundesbank, können oft nur aggregierte Ergebnisse ohne Rückschlüsse auf Einzelbeobachtungen veröffentlicht werden. Um Probleme für die Veröffentlichung zu vermeiden, müssen Doktorand/inn/en die Data Policy des Target-Journals beachten, zumal die hier veröffentlichten GfP keine bindende Wirkung auf Verlage und Datenprovider haben.

Die Sicherung von Daten empfiehlt sich zunächst im eigenen Interesse. Arbeitet man mit Koautoren zusammen, sollte man sich die Rohdaten vor Beginn der Forschungsarbeit auf einem nicht veränderbaren Medium (z.B. DVD) übergeben lassen. Das Mehraugenprinzip ist hierbei das Mittel der Wahl zur Verhinderung etwaiger Manipulationen. Außerdem sollte man das Skript auf demselben Medium einfordern, mit dem die Rohdaten für die Analyse transformiert worden sind. Hinzukommen sollten später alle Analysen mit ihren Ergebnissen, die irgendwelchen Publikationen zugrunde liegen. Entscheidend für die Nachvollziehbarkeit sind eine Versionierung von Datensätzen und die stetige Speicherung bei zusätzlicher Datenerhebung. Bei Ökobilanzen ist z.B. im Allgemeinen ein iteratives Vorgehen typisch, d.h. die Datensammlung lässt sich nicht auf den Beginn festlegen. Ebenso wissen die Auftraggeber oft zu Beginn nicht, welche Daten gebraucht werden.

Alle Medien sollten dann mit bester verfügbarer Technik mit beschränktem Zugang gesichert werden. Damit ist man für alle etwaigen Rückfragen gerüstet, solange Universitäten kein eigenes Datenarchiv anbieten. An der Erasmus-Universität wird verlangt, dass alle Daten unmittelbar nach ihrer Erhebung (in Rohform) zu hinterlegen sind, wobei der einzelne Forscher keinen Zugang für Änderungen besitzt, sondern höchstens Dekane oder dafür abgestellte Controller (http://www.eur.nl/researchmatters/research_data/data_management/). Man sollte allerdings bedenken, dass Daten, die in einer Cloud abgelegt werden, häufig auf US-Servern gespeichert sind und US-Behörden demzufolge u.U. verlangen können, dass diese Daten für sie einsehbar sind. Unklar ist, wie bei Datenarchiven sichergestellt werden kann, dass im Falle von Vertraulichkeitsvereinbarungen, insbesondere wenn diese mit Vertragsstrafen bewehrt sind, keine Rechte verletzt werden. Hier müssen eindeutige Schadensersatzregelungen durch die Universität erlassen werden. Vertraulichkeitsvereinbarungen sind so zu verfassen, dass sie dem berechtigten Interesse der akademischen Gemeinschaft an Nachvollziehbarkeit und breiter Streuung von Forschungsergebnissen [insbesondere der Publikationspflicht von Dissertationen] nicht zuwiderlaufen.

Doktorand/innen können durch verschiedene Handreichungen für das Thema sensibilisiert werden: Durch verbindliche Regeln im Rahmen der akademischen Betreuung durch Doktorvater bzw. -mutter (als Teil der Promotionsausbildung), durch Handreichung dieses Leitfadens, durch eine Checkliste für Betreuer und Doktoranden, durch Hinweise auf andere Guidelines und Webinare, als Bestandteil in allen VHB-Workshops, durch ProDok-Dozenten, durch eine entsprechende Bestätigung oder ein Musterbestätigungsschreiben – von den Doktorand/innen zu verlangen –, dass sie sich an die Grundsätze ordnungsmäßiger Dokumentation (insbesondere: Aufbewahrung Rohdatensatz, Erläuterung Modifikationen, endgültiger Datensatz) gehalten haben und diese Dokumente auch selbst archivieren.

Spezialfall Experimente:

Experimente haben in den letzten Jahren in den Wirtschafts- und Sozialwissenschaften immer mehr an Bedeutung gewonnen. Insofern sind Experimente genauso detailliert und gewissenhaft zu dokumentieren, wie es in den Naturwissenschaften mit den Laborbüchern üblich ist, wo nur handschriftliche Aufzeichnungen erlaubt sind, weil man diese später nicht verändern kann. Dies betrifft vor allem den exakten Zeitraum, in welchem ein Experiment mit wie vielen Probanden durchgeführt worden ist, um besser abschätzen zu können, ob Ergebnisse durch späteres Hinzufügen von Probanden signifikant geworden sind. Wichtig ist auch, dass man über alle Experimente berichtet, auch wenn einige nicht zu den erwarteten Ergebnissen geführt haben. Nicht-Befunde sind wichtig für das Verständnis, was funktioniert und was nicht. Solche Befunde können auf Wunsch dem Reviewer zur Verfügung gestellt werden, selbst wenn diese letztlich nicht in der Veröffentlichung erscheinen. Ein Experiment ist so ausführlich zu dokumentieren, dass es reproduzierbar ist. Dies beinhaltet auch die genauen Anweisungen an Probanden (im Wortlaut), welche als Anhang oder Web-Appendix zur Verfügung gestellt werden sollten. Außerdem empfiehlt es sich, über die Anzahl der Probanden, ihre Gewinnung und die gewährten Anreize zu berichten. Detaillierte Empfehlungen unterbreiten Simmons, Nelson und Simonsohn (2011,2012).

Spezialfall Algorithmen:

Gerade im Gebiet Operations Research wird gerne gezeigt, dass ein bestimmter Algorithmus besser abschneidet als bisher angewendete, was in der Regel mit computergenerierten Datensätzen geschieht. Insofern sollte die Art der Datengenerierung detailliert beschrieben und mit Hilfe von Programmiercodes dokumentiert werden, die man als Web-Appendix herunterladen kann. Schätzverfahren oder Optimierungs-Algorithmen sind mittlerweile sehr komplex geworden. Zur besseren Nachvollziehbarkeit sollte der Code zur Verfügung gestellt werden. In der Physik ist nämlich festgestellt worden, dass in den meisten veröffentlichten Algorithmen noch Fehler steckten. Dies gilt natürlich nicht für proprietäre Software, sofern diese jeder käuflich erwerben kann. Im Übrigen gibt es bekannte Fälle, in denen Forscher durch das öffentliche Anbieten ihrer Software zum Download einen hohen Impact in der Wissenschaft erzielt haben, so z.B. Ringle mit seinem Smart-PLS (www.smartpls.de).

Spezialfall Qualitative Datenerhebung:

Bei einer qualitativen Datenerhebung ist in einem Erhebungsprotokoll darzulegen, wie die analysierten Organisationen oder Individuen ausgewählt wurden und welche Daten wie erhoben wurden. Das Verfahren der Transkription ist zu beschreiben, ebenso wie Sicherstellung der Anonymität und der Freigabe der Daten. Die Datenauswertung ist zu beschreiben, im Falle einer inhaltsanalytischen Kodierung ist der Kodebaum darzulegen. Schließlich ist darzulegen, wie das Ziel einer objektivierten Auswertung erreicht wurde. Hierzu ist beispielsweise die Interraterreliabilität zu berechnen.[2]

Literaturhinweise:

Albers, Sönke (2009): Editorial: Well Documented Articles Achieve More Impact, BuR – Business Research, 2 (1), 8-9.

Desai, Preyas S. (2013): Editorial: Marketing Science Replication and Disclosure Policy, Marketing Science, 32 (1), 1–3.

Deutsche Forschungsgemeinschaft (1998): Proposals for Safeguarding Good Scientific Practice, Weinheim: Wiley-VCH.

Elsevier (Hrsg.) (2017): Database Linking Tool (January 14, 2017). Available at: https://www.elsevier.com/books-and-journals/enrichments/data-base-linking

Laurent, Gilles (2013): Respect the data!, International Journal of Research in Marketing, 30 (4), 323–334.

Lewis, J. (2009), Redefining Qualitative Methods: Believability in the Fifth Moment, International Journal of Qualitative Methods, 8, 1-14.

Miyata, H. and K. Ichiro (2009), Reconsidering Evaluation Criteria for Scientific Adequacy in Health Care Research: An Integrative Framework of Quantitative and Qualitative Criteria, International Journal of Qualitative Methods, 8, 64-75.

Simmons, Joseph P., Leif D. Nelson and Uri Simonsohn (2011): False-Positive Psychology. Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant, Psychological Science, 22 (11), 1359-1366.

Simmons, Joseph P. and Nelson, Leif D. and Simonsohn, Uri (2012): A 21 Word Solution (October 14, 2012). Available at SSRN: http://ssrn.com/abstract=2160588

Technische Universität Darmstadt (Hrsg.) (2015): Leitlinien zum Umgang mit digitalen Forschungsdaten an der TU Darmstadt (December 16, 2015). Available at: http://www.tu-darmstadt.de/media/dezernatvi/relaunch_2015/gute_wiss_praxis/Leitlinien_Forschungsdaten_2015

Darüber hinaus zur Orientierung empfehlenswert:
http://www.forschungsdaten.org/index.php/Data_Policies
http://www.eur.nl/researchmatters/research_data/data_management/

[2] Hilfreiche Hinweise zu Besonderheiten bei der Anfertigung qualitativer Forschungsarbeiten finden sich bei Lewis (2009), sowie Miyata, H. and K. Ichiro (2009).

Deutsche Forschungsgemeinschaft (2013). Vorschläge zur Sicherung guter wissenschaftlicher Praxis, 2. edition, Weinheim: Wiley-VCH. Available at: http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_1310.pdf

Deutsche Forschungsgemeinschaft (2015): Leitlinien zum Umgang mit Forschungsdaten. Verfügbar unter: http://www.dfg.de/foerderung/info_wissenschaft/2015/info_wissenschaft_15_66/

Philipps-Universität Marburg (Hrsg.) (2016a): Datenmanagementplan (November 9, 2016). Available at: http://www.uni-marburg.de/projekte/forschungsdaten/bilder/damanagementplan

Universität Bielefeld (Hrsg.) (2016): Grundsätze zu Forschungsdaten an der Universität Biele-feld (December 18, 2016). Available at: https://data.uni-bielefeld.de/policy

Universität Heidelberg (Hrsg.) (2016): Research Data Policy - Richtlinien für das Management von Forschungsdaten (September 15, 2016). Available at: http://www.uni-heidelberg.de/universitaet/profil/researchdata/

Verband der Hochschullehrer für Betriebswirtschaft e. V. Verbandsgeschäftsführerin: Tina Osteneck Geschäftsstelle: Reitstallstr. 7 – 37073 Göttingen – Deutschland Tel.: +49(0)551 – 797 78 566, Fax: +49(0)551 – 797 78 567 Email: info@vhbonline.org – https://vhbonline.org