Themenfeld 5:
Umgang mit Daten und Dokumentation des Forschungsprozesses

Die Ergebnisse von Forschungsarbeiten sind nur 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.) soll in welcher Weise dokumentiert werden, um später zur Klärung von Sachverhalten 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?

Ein notwendiges Kriterium für Wissenschaft besteht in der **Nachvollziehbarkeit** der erzielten Ergebnisse. 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 allerdings keine Kultur der Replikation von Ergebnissen herausgebildet, weshalb auch die Angemessenheit bisheriger Dokumentationen in der Regel nicht geprüft wurde. Ausgelöst durch verschiedene Betrugsfälle erwarten Zeitschriften und Forschungsförderungsinstitutionen wie die DFG heute genauere **Dokumentationen** und zum Teil auch das **Verfügbarmachen der verwendeten Daten**. Damit möchte man auch schnellere – jeweils darauf aufbauende – wissenschaftliche Fortschritte erzielen.

Dokumentation von Daten: 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 oder Kunden schließen kann. 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.

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 Anreize 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). 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.

Experimente: Gerade mit Experimenten kann man vieles zeigen, was man möchte. 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. 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).

Computerexperimente: 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 werden und mit Hilfe von Programmiercodes dokumentiert werden, die man als Web-Appendix herunterladen kann.

Schätzverfahren und Optimierungsalgorithmen: Heutzutage sind Schätzverfahren oder Optimierungs-Algorithmen sehr komplex geworden. Zur besseren Nachvollziehbarkeit sollte der Code zur Verfügung gestellt werden. In der Physik ist nämlich festgestellt worden, dass in dem 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).

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. 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 Holdout-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. Heutzutage im Zeitalter des Internet ist es ein leichtes, Daten 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). 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, ohne dass sie dies aber kontrolliert.

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 man auf die zugesagte Vertraulichkeit achten. 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.

Die Sicherung von Daten empfiehlt sich zunächst im eigenen Interesse. Arbeitet man mit Assistenten zusammen, sollte man sich die Rohdaten vor Beginn der Forschungsarbeit auf einem nicht veränderbaren Medium (z.B. DVD) übergeben lassen. Meist hilft ein Mehraugenprinzip dabei, dass keine Manipulationen entstehen. Außerdem sollte man das Skript auf demselben Medium einfordern, mit dem die Rohdaten für die Analyse transformiert worden sind. Hinzu kommen sollten später alle Analysen mit ihren Ergebnissen, die irgendwelchen Publikationen zugrunde liegen. Alle Medien sollten dann in einem Panzerschrank 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 daran denken, dass die Daten nicht in einer Cloud landen, da diese häufig auf US-Servern gespeichert sind und die US-Behörden verlangen können, dass diese Daten für sie einsehbar sind. Unklar ist bei Datenarchiven, wie man sicherstellen kann, dass im Falle von Vertraulichkeitsvereinbarungen, insbesondere wenn diese mit Vertragsstrafen bewehrt sind, es zu keinen Verletzungen kommen kann. Hier müssen eindeutige Schadensersatzregelungen durch die Universität erlassen werden.

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.

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

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

Darüber hinaus zur Orientierung empfehlenswert:

http://www.eur.nl/researchmatters/research_data/data_management/

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
E-Mail: info@vhbonline.org - URL: http://vhbonline.org