Datenschutzprobleme: Normen und Praxis beim internationalen Datenservice

Karhausen, Mark; Mochmann, Ekkehard

Veröffentlichungsversion / Published Version
Sammelwerksbeitrag / collection article

Datenschutzprobleme: Normen und Praxis beim internationalen Datenservice

Mark Karhausen
Ekkehard Mochmann

1. Vorbemerkung

Datenschutz und freier Datenzugang sind Anspruchsgrundsätze, die miteinander kollidieren können. In Institutionen, in denen empirische Daten archiviert und ihre Aufbereitung für weitere Analysen (Sekundäranalysen) betrieben wird, ist dieses Problem nicht unbekannt. Durch entsprechende Benutzungsordnungen und Akquisitionspolitik wird versucht, beide Anspruchsgrundsätze in den Griff zu bekommen. Ein besonderes Problem hat sich dadurch ergeben, daß zwischen Institutionen im internationalen Bereich ein Datentransfer vereinbart wurde. Datenaustausch über Ländergrenzen ist dann problematisch, wenn Empfänger und Sender aus verschiedenen Ländern unterschiedliche Datenschutzregelungen berücksichtigen müssen. Die Verabschiedung von Normen (code of ethics) und Verhaltensweisen (code of conduct) ist als ein erster Schritt in der Berücksichtigung der schutzwürdigen Belange von Betroffenen zu werten. Die Praxis zeigt jedoch, daß es eine Fülle von Problemen in der praktischen Umsetzung solcher Verabredungen zu lösen gilt :

a) In unserer Auseinandersetzung mit den vorgeschlagenen Maßnahmen und deren Realisierung gehen wir von den Erfahrungen einer Datenbank für Umfragen aus. Unsere Stellungnahme ist deswegen ebenso subjektiv wie die der Autoren, auf die wir uns beziehen, und nur als Diskussionsgrundlage gedacht.

b) Viele der Vorschläge sind normativ. Als solche sind sie grundsätzlich zu begrüßen (einzelne sind nach deutschem Recht nicht durchführbar, siehe unten).

c) Bei der praktischen Umsetzung ergeben sich besonders dann Probleme, wenn im Archiv erst entschieden werden muß, ob eine Personenbestimmung möglich ist. In der Archivtätigkeit können kaum alle Variablenkombinationen und - reduktionen durchgespielt werden, um diesen Nachweis zu führen (zu zeitaufwendig und zu teuer).

d) Das Pferd wird (trotz guten Willens) mit den vorgeschlagenen Maßnahmen von hinten aufgezäumt. Es wäre zu überlegen, ob als vorbeugende Maßnahme :

- der "principal investigator" vom Befragten eine generelle Nutzungserlaubnis explizit einholt - auch unter der Möglichkeit der Personenbestimmung -
  oder
- wo dies nicht geht (unrealistisch), sich das Archiv vom Benutzer vertraglich zusichern läßt, daß eine Personenbestimmung nicht beabsichtigt ist bzw. wenn sie - zufällig - auftritt, dies dem Archiv sofort anzuzeigen und eine Weiterverwendung auszuschließen.

2. Normative Regelungen

A. Im IASSIST (International Association for Social Science Information Service and Technology) Newsletter Vol 1, No 4 (1977) sind Normen des ICPSR (International Consortium for Political Social Research) aufgeführt, die im folgenden zusammengefaßt sind :

Preservation of Confidentiality

The issue of confidentiality arises primarily, but not exclusively, in the case of data collections that include information that is or is seen as potentially damaging or threatening to respondents, of when a promise of confidentiality has been given respondents in the process of data collection, and when information is included that allows or potentially allows identification of individual respondents. The confidentiality issue may appear most pressing in the case of data pertinent to elite populations. Such data are now increasingly available, and elite populations

are most visible, are or are seen as at greater risk, and are most easily identifiable by means of a few key variables rhan are respondents in a mass survey. Also contributing to the gravity of the issue is the sensitivity of the public as a whole, and public figures more particularly, to the "potential" uses of data obtained through personal interviews. The increased research focus on elite populations and the resultant increased availability of data combined with the need to facilitate extended research use of data collections makes it necessary for the Consortium to develop policies and procedures for legitimate data dissemination without jeopardizing the rights of respondents to confidentiality.

It should be noted, of course, that the issue of confidentiality is not confined to information collected through personal interviews nor is it limited to elite populations or even individuals. Data collected through means other than personal interviews - through laboratory experiments, for example - also present the issue, and data collected form public record sources, auch as court records, can be seen as unnecessarily and unjustifiably damaging to individuals unless anonymity is provided. Data on organizations or particular groups can also present confidentiality issues as can data relevant to deviant populations and the issue is also raised by mass surveys. While the following policies are intended to apply specifically to individual level data for elite populations, they are also seen as applying in more general terms to other categories of data such as those suggested above.

To protect the identity of respondents, the following guidelines will be applied :

1. The Archive will not accept any documentation, list or data files that explicitly identify respondents by name except in the case of data that are in publicly available sources and which do not present problems of confidentiality.

2) In cooperation and consultation with principal investigator(s) the Consortium staff will assist in the identification of "sensitive" variables but reserves the right to extend the list of variables to be deleted, aggregated or otherwise masked beyond those identified by the principal investigator(s).

3) Variables will be flagged when a) they allow identification of particular respondents, and b) when used in combination with other variables, they also identify particular respondents.

4) The list of sensitive variables will be presented to the principal investigator(s) with recommendations for which variables should be deleted, col-

lapsed or grouped (such as age) or have the code category descriptions changed to generalize the category.

5) A version of the dataset which incorporates the above recommendations will be produced for general distribution. 1a

6) The codebook for the version of the data to be distributed will include documentation for the masked variables, a description of the nature of the deletions and maskings and frequencies for the masked variables.

7) Requests for data reductions (for example, cross tabulations) or aggregations involving the masked variables will be accepted, but the Consortium staff will judge whether the requested data reductions or aggregations will preclude identification of individuals and will supply only those results that do so. Decisions will be made in consultation with principal investigators in order to insure that protection of confidentiality does not unnecessarily limit researchers' access to analytic infromation.

8) The original version of the data will be maintained under security (see document entitled "Physical File Security").

9) Public record data will not be added or merged into a survey data file if that data allows identification of individuals. If public record data will not be added to the survey data file and identifiaction of individuals as a consequence is possible, these data will be removed from the survey data file and retained separately without a linking variable that would allow the two files to be merged again.

The proposed method for handling confidentiality attempts to protect the respondents without destroying the meaning of the data. If this process renders the data meaningless in any given case, the data could be maintained under security by the Consortium and not be generally distributed. The documentation could be distributed upon request and only data reduction requests would be accepted.

If none of the above is acceptable, the data cannot be archived.

ad 1.

Die Frage ist : Was heißt "explicitely identify respondent"? Personenbezogene Umfragedaten werden in der Regel gar nicht archiviert und anonymisierte Daten sind schlecht zu deano-

nymisieren. Die oft heraufbeschworenen Deanonymisierungsverfahren (auch von uns) haben sich bis heute für Umfragedaten mehr als theoretische Möglichkeiten für Personenbestimmung erwiesen. In der Praxis sind solche Verfahren 1) sehr teuer und 2) ohne Zuhilfenahme von Zusatzinformationen gar nicht durchführbar.

Als vorbeugende Maßnahmen werden vorgeschlagen :

1. Nur anonymisierte Daten zu archivieren bzw. nur solche weiterzugeben;
2. Demographievariablen wie z.B. Berufsbezeichnung und Geburtstage kategorial zu vercoden (Nachvercodung wenn nötig!);
3. Umfragebeschreibungen und Daten getrennt aufzubewahren;
4. Dateinutzung nur durch Personal des Archivs zulassen;
5. Weitergabe der Originaldaten ausschließen. Nur vorgeprüfte (regecodete) Dateien werden der Nutzung überlassen.

ad 2.
Ohne eine Liste was "sensitive Variablen" sind, kann ein Archiv keine Handlungsanweisung für geeignete Maßnahmen zu dieser an sich begrüßenswerten Norm ziehen.

Zur Verdeutlichung :

Variable : "Einstellung zur Todesstrafe". Für Arbeiter und Angestellte kann eine Einstellung dafür oder dagegen weniger "sensibel" sein als etwa für einen Richter.

Variable : "Einstellung zur konzertierten Aktion". Für Arbeitgebervertreter und Gewerkschaftsfunktionäre sind Aussagen zu dieser Variablen sicherlich sensibler als für "Otto Normalverbraucher".

Variable : "Wahlentscheidung". Aussagen zu dieser Variable können für jeden sensibel sein.

Facit : Aus Sicht des Archivs selbst in Abstimmung mit dem "principal investigator" sind keine Entscheidungen, ob Variablen sensibel sind, zu treffen, sondern das muß in das Belieben des Befragten gestellt bleiben

Es ergibt sich folgendes Problem :

Dem Befragten ist ja auf Verlangen der Zweck der Befragung offen zu legen. Wenn der Befragte antwortet, ist damit das implizite Einverständnis für diese Zweckverfolgung gegeben.

Gerade in der Feldphase sind Stimuli soweit wie möglich - wenn nicht anders gewünscht - auszuschalten. Von dort her ist es üblich, den Befragungszweck mit "statistischer Auswertung" und ähnlichem anzugeben. Dies schließt auch die Sekundäranalyse mit ein. Ein anderes Problem ergibt sich, wenn zum Zeitpunkt der Befragung ein bestimmter Befragungszweck angegeben wird, der eine weitere Verwendung als die angegebene ausschließt. Wenn nun die Daten anonymisiert worden sind, ist die Weitergabe nach BDSG-Regeln unkritisch, die zugrundeliegende Information jedoch unter Ausschluß einer weiteren Verwendung gewonnen worden. Daraus folgt, daß in der Feldphase der Befragte möglichst nicht durch die Ausschließlichkeit einer nur einmaligen Auswertung zur Informationsabgabe provoziert werden sollte.

ad 3.

Der Forscher sollte bei einer weitergehenden Analyse die Möglichkeit haben, sich den Fragebogenaufbau - insbesondere den unmittelbaren Kontext der Variablen, die er in die Analyse einbeziehen will, anzusehen. Von daher ist es sinnvoll, sensitive Variablen zu maskieren und nicht zu löschen. Zwar wird durch die Maskierung aktiv informiert, daß eine sensitive Variable vorliegt. Damit wird den von Datenschützern latent unterstellten unlauteren Absichten Vorschub geleistet, doch bleibt die Kontrolle über die das Individuum betreffende Information erhalten, da die Daten selbst nicht freigegeben werden. Es ist bei diesem Vorgehen lediglich darauf zu achten, daß in den Randverteilungen keine Besetzungen mit auffälligen Extremwerten ausgewiesen sind.

ad 4.
Zwar ist der Primärforscher in der Regel derjenige, der die Daten am besten kennt. Von daher ist eine gewisse Insiderkenntnis bei der Identifizierung möglicherweise brisanter Variablen zu erwarten.

Dennoch ist nicht einzusehen, warum der Primärforscher die Verantwortung für die vom Archiv vorgeschlagenen Maßnahmen übernehmen soll. Hier ist ihm allenfalls eine beratende- nicht eine legitimierende Rolle - zuzugestehen.

Vielmehr sollten Archive als letztlich verantwortlich für die weitere Distribution der Daten sicherstellen, daß

1. alle als kritisch bekanntgewordene Variablen gesperrt oder transformiert werden
2. die Verwendungsabsichten der Benutzer legitim sind, d.h. den Vorbehalten des potentiell Betroffenen bei der Informationsabgabe nicht entgegenstehen. Insoweit können die Archive Kontrollfunktionen übernehmen. Die tatsächliche Auswertung unterliegt jedoch der Kontrolle der weiteren Benutzung, Datenmißbrauch in dieser Phase muß entsprechend sanktioniert werden.

ad 5.
Soweit die genannten Transformationen bzw. Sperrungen realisiert sind, ist eine mißbräuchliche Verwendung des vom Archiv freigegebenen Datensatzes unwahrscheinlich. Dennoch ist auch bei größter Sorgfalt ein "Harrisburg Effekt" nicht auszuschließen. Es können nicht vorherzusehende Informationskonstellationen auftreten. Ließe sich jeder mögliche Risikofall vorausberechnen, dann wäre perfekter Datenschutz möglich. Da dies jedoch nicht zu garantieren ist, besteht perfekter Datenschutz nur bei totaler Informationssperre. Damit führt sich diese Überlegung aber selbst ad absurdum. Es kann also auch für die Archive letzlich nur um abgestufte Maßnahmen gehen. Der Versuch, diese Maßnahmen im Rahmen der gesetzlichen Rege-

lungen auszugestalten, wird von den Mitgleidsinstitutionen der International Federation of Data Organizations (IFDO) weiterverfolgt. Besondere Probleme ergeben sich dabei auch aus dem internationalen Datentransfer. Soweit ausländische Archive als Mittler für den Datentransfer auftreten, können bei fortbestehenden Diskrepanzen der Datenschutzgesetze in den verschiedenen Ländern gesetzliche Probleme auftauchen, die bei der direkten Übermittlung der Daten an den ausländischen Benutzer umgangen werden können.

Tatsächlich ist hier noch ein Markt für "Informationsflucht". Es steht einem Forscher nichts im Wege, seine (brisanten) Ergebnisse in einem Lande zu veröffentlichen, das noch keine oder eine sehr liberale Datenschutzgesetzgebung hat.

ad 6 - 9

Es ist vorzuschlagen, daß die Datenarchive aus ihren Erfahrungen eine Liste sensitiver Variablen aufstellen (in anonymisierter Form) und sie international zirkulieren lassen.

Alle Maßnahmen, auf die wir uns beziehen, gehen davon aus, daß man mit sensitiven Variablen irgendetwas macht, ohne zu sagen, welche Variablen das im einzelnen sind bzw. sein können. Desweiteren wäre vorzuschlagen, ein Forschungsprogramm zu etablieren, in dem Befragungsinhalte der üblichen Untersuchungen einmal empirisch auf die einzelnen Sensibilisierungsgrade analysiert werden. Es kann nämlich nicht angehen, daß Archive erst dann über die sensitiven Variablen Kenntnis erhalten, wenn entsprechender Mißbrauch eingetreten ist.

"Data Archives for Social Research, Towards Ethical Standards to Ensure Confidentiality Preserving Modes of Access"

(Auszug aus einem paper, präsentiert auf der IFDO-Konferenz in Köln)(Anlage 2)

"Procedural techniques which can be applied to reduce confidentiality - related problems include :

1. collecting anonymous information (for example, no names on completed face sheets or questionnaires, anonymous telephoning),

2. destruction of questionnaires and face sheets once data have been processed,

3. separation of data collection staffs from those responsible for processing and analysis,

4. security devices for protecting access to the data once they are in machine readable form; and various methods of code linkage and models of broker systems which demonstrate the feasibility of linking records without disclosure by data collection and dissemination agents.

5. Identifiers can be deleted.

6. Randomizing response methods can be used so that the respondent does not reveal the question he answered.

7. Random error can be injected in the data.

8. Data can be averaged for groups (microaggregation).

9. Various data can be suppressed and classifications can be collapsed, thus limiting the detail of small samples and preventing the disclosure of individual identities by way of cross classifications.

In dem Beitrag von Alice Robbin werden technische Prozeduren vorgeschlagen, deren Zielrichtung begrüßenswert ist, aber ebenfalls in der praktischen Umsetzung Probleme schafft.

ad 1.

"Collecting anonymous information..."

Die Weitergabe von anonymisierten Daten ist nach BDSG erlaubt. Eine Akquisition solcher Daten ist damit problemlos. Selbst die Akquisition nicht anonymisierter Daten ist nicht

ein Problem für Archive, sondern ein Problem des principal investigator (Datenherr). Für den letzten sind Regeln der Weitergabe an Dritte zu beachten! Bei Nichtbeachtung dieser Regeln durch den Datenherrn macht sich der Empfänger von Daten (hier Archiv) nur dann strafbar, wenn er nicht in gutem Glauben handelt.

Das Problem für Archive besteht darin, akquirierte personenbezogene Daten gar nicht erst in Dateien zu speichern, sondern einen Anonymisierungsprozeß der Speicherung vorzuschalten, um dem Gesetz Genüge zu tun.

ad 2.

"Destruction of questionnaires..."
Archive sind in der Regel (wenn sie keine eigene Forschung treiben!) gar nicht an einer Akquisition bzw. Speicherung von Fragebögen gehalten. Insofern ist eine Vernichtung nur ein Problem des principal investigators. Davon unabhängig - und nicht datenschutz-relevant - ist die Akquisition eines Musters des Fragebogens (unausgefüllt!), um Fragetechnik und sonstige Erkenntnisse für die Umfragebeschreibung zu gewinnen.

ad 3.

"Separation..."
Eine Trennung der Mitarbeiter der Feldphase von denen, die anschließend analysieren, ist ebenfalls für Datenarchive natürlicherweise nicht relevant, sondern nur für den principal investigator. Eine Bemerkung jedoch dazu am Rand : Eine Trennung der Verantwortlichkeiten ist aus unserer Sicht zwar organisatorsich möglich, doch wohl nicht ernst gemeint. Wer Feldarbeit kennt und die nachfolgende Primäranalyse einmal durchgezogen hat, wird zugeben müssen, daß eine strigente Befolgung dieser Maßnahme einem Sprechverbot gleichkommt. Rückschlüsse in der Analysephase auf die Feldarbeit ohne Kommunikation ist nicht sinnvoll.

Daneben wären alle die ad hoc Forschergruppen an Universitäten und ähnlichen Institutionen, die schon aus Kostengründen und Interessenlagen beide Phasen in einer Gesamtverantwortlichkeit bearbeiten, ex tunc vernichtet !

ad 4.
"Security devices..."
Diese Maßnahmen sind begrüßenswert, bdürfen jedoch weiterer Konkretisierung, um in Handlungsanweisungen für Archive umgesetzt werden zu können. Bisher sind z.B. passwords, file-trennung und entsprechende Dateiensicherung (labels) bekannt.

ad 5.
"Identifiers..."
Es ist zu prüfen, was darunter zu verstehen ist. Name und Adresse der Befragten liegen bei anonymisierten Daten sowieso nicht vor. Sonstige Identifizierungsmerkmale (Berufscodes, Geburtstage usw.) sind durch eine kategoriale Vercodung abzufangen. Damit ist keine "deletion", sondern "aggregation" vorzuschlagen. Wenn hingegen alle Variablen, die möglicherweise identifiers sein können, herausgelassen würden, wird der Informationsgehalt (gerade wenn es Demographievariablen sind) der Umfrage stark reduziert.

ad 6.
"Randomizing..."
Diese Maßnahme ist in dem Bereich des Fragebogendesigns nicht in das Belieben des Archivs gestellt, sondern als Maßnahme für den principal investigator interessant, es sei denn, daß Archive aus Datenschutzerwägungen Umfragestrukturen verändern (z.B. Antworten werden nicht objektbezogen kummuliert, sondern per Zufallsgenerator).

ad 7.
"Random error..."
Diese Maßnahme erscheint zunächst unverständlich, da bisher Zufallsfehler mit erheblichem Aufwand in der Analysephase

festzustellen waren, um Validität und Reliabilität der Daten nachzuweisen. Eine "künstliche" Einfügung eines solchen Fehlers setzt für Dritte voraus, daß der Umfang dieses Fehlers nicht bekannt wird. Damit wären die Einzelmaßnahmen der Fehlereinfügung selbst sensibel und damit als schtuzwürdig zu betrachten.

ad 8.

"Micro aggregation..." "limiting..."
Diese Maßnahmen wurden oben schon als geeignete Datenschutzmaßnahmen vorgeschlagen.