Ricarda Nauenburg

Automatisiertes Fernrechnen mit amtlichen Mikrodaten – aktuelle Entwicklungen

Das Interesse der Wissenschaft an der Nutzung amtlicher Mikrodaten wächst ständig. In Deutschland betreuen die neu eingerichteten Forschungsdatenzentren der amtlichen Datenanbieter Auswertungsprojekte der unabhängigen wissenschaftlichen Forschung. Um den gesetzlichen Datenschutzbestimmungen Genüge zu tun, werden die Daten vor der Nutzung anonymisiert bzw. die Analyseergebnisse, bevor sie freigegeben werden, auf Geheimhaltung geprüft. Diese Prüfungen finden in Deutschland ausschließlich ohne maschinelle Hilfe statt. Im Zuge der verstärkten Öffnung amtlicher Mikrodaten für die Wissenschaft kann die Geheimhaltungsprüfung der Ergebnisse jedoch zu einem Engpass bei der Bearbeitung führen. Das Problem wird sich noch verschärfen, wenn es auch in Deutschland Möglichkeiten zum internetbasierten Fernrechnen geben wird. Die Einrichtung dieser bei Nutzern stark nachgefragten Datenzugangsform macht im Vorhinein Überlegungen zur Automatisierung der Geheimhaltungsprüfung nötig. Die zu realisierenden Maßnahmen beinhalten aber nicht nur die Output-Kontrolle, sondern auch die Verhinderung von sonstigen geplanten und ungeplanten Datenangriffen. Die folgenden Ausführungen geben einen Überblick über die international vorhandenen Implementierungen zum automatisierten Fernrechnen mit amtlichen Daten und die dabei angewandten Geheimhaltungsmaßnahmen. Sowohl vorhandene Probleme als auch „best practice" werden aufgezeigt und es wird deutlich, dass es ein ideales System nicht gibt. Die gezogenen Schlussfolgerungen verstehen sich als ein Beitrag zur Diskussion eines analogen Konzeptes in Deutschland.

1 Datenzugangsformen in Deutschland

Aktuell kann die unabhängige Wissenschaft[1] in Deutschland Sekundäranalysen amtlicher Mikrodaten auf drei verschiedene Arten durchführen: Die Forschungsdatenzentren der amtlichen Statistik bieten Scientific-Use-Files für die Nutzung am eigenen Arbeitsplatz (Off-Site), Scientific-Use-Files für die Nutzung am Gastwissenschaftlerarbeitsplatz im Statistischen Amt (On-Site) und Originaldaten für die kontrollierte Datenfernverarbeitung an. Nach §16 Abs. 6 des Bundesstatistikgesetzes müssen die Daten faktisch anonym sein, bevor sie zur Verfügung gestellt werden zu können[2]. Sie werden deshalb teilweise vor der Nutzung vergröbert. Scientific-Use-Files für die Off-Site-Nutzung sind stärker anonymisiert als Datenfiles für Gastwissenschaftlerarbeitsplätze. Letztere erfordern einen geringeren Anonymisierungsgrad, da die Nutzung unter anderen Rahmenbedingungen an abgeschotteten PCs in den Räumen der amtlichen Statistik erfolgt. Für die kontrollierte Datenfernverarbeitung werden die Daten so zur Verfügung gestellt, wie sie auch im Amt gespeichert sind: ohne direkte Identifikatoren wie Namen und Adresse. Der Anonymisierungsgrad der Daten unterscheidet sich also nach der Umgebung, in der diese Daten genutzt werden dürfen. Durch diese Kombination, ergänzt mit einer eingeschränkten Zugangsberechtigung, ist die faktische Anonymität der Daten sichergestellt.

Die kontrollierte Datenfernverarbeitung ist die einzige Möglichkeit, nicht veränderte Originaldaten der amtlichen Statistik zu analysieren. Der Nutzer hat – anders als beim Scientific-Use-File und am Gastwissenschaftlerarbeitsplatz – keinen direkten Zugang zu den Daten, sondern schickt anhand von Dummy-Daten selbst erstellte Auswertungsprogramme (Syntax) zum Forschungsdatenzentrum. Dort werden mit Hilfe dieser Syntax und den Originaldaten Ergebnisse produziert, die auf statistische Geheimhaltung geprüft und danach dem Nutzer zur Verfügung gestellt werden. Die manuelle Prüfung kann – besonders in explorativen Projektphasen – sehr aufwändig sein, da jeder Output geprüft werden muss. Bei der Datennutzung am Gastwissenschaftlerarbeitsplatz ist der Aufwand geringer, geprüft wird nur der Output, der das Amt verlassen soll. Hier ist dem Datennutzer jedoch die Unbequemlichkeit der Anfahrt aufgebürdet. Der oben erwähnte Off-Site-Scientific-Use-File erfordert keine Geheimhaltungsprüfung der Ergebnisse, der Nutzer muss jedoch mit relativ stark veränderten Daten und vermindertem Analysepotenzial vorlieb nehmen. Aus den angebotenen Datenzugangsformen kann der Nutzer die für ihn optimale wählen bzw. kombinieren: Sinnvoll wären z. B. Exploration und Modellentwicklung mit Hilfe des Scientific-Use-Files oder am Gastwissenschaftlerarbeitsplatz und finale Arbeiten mittels der Kontrollierten Datenfernverarbeitung.

2 Automatisiertes Fernrechnen mit amtlichen Mikrodaten

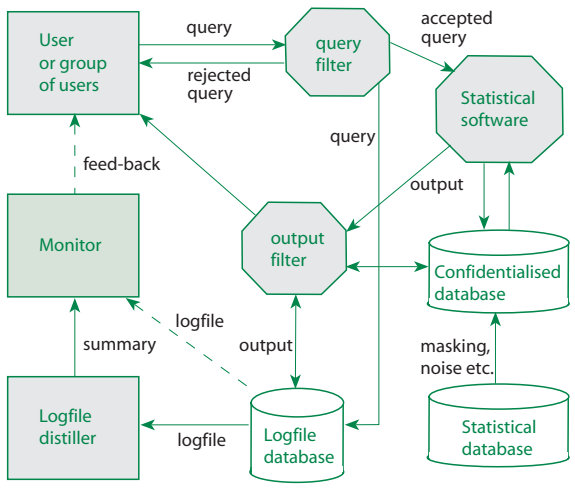
Das Prinzip dieser neuen Datenzugangsform ist einfach: Datennutzer schicken vom heimischen PC aus ihre Auswertungswünsche an den Datenanbieter und erhalten von dort innerhalb kürzester Zeit die automatisch anonymisierten Ergebnisse zurück. Automatisiertes Fernrechnen hat für beide Seiten Vorteile: Der Datenanbieter behält bei geringst möglichem eigenen Aufwand die vollständige Kontrolle über die Daten, der Nutzer muss keine Anfahrten auf sich nehmen und erhält die Analyseergebnisse idealerweise sofort. Die Möglichkeit des Fernrechnens mit integrierter automatischer Geheimhaltungsprüfung könnte also eine sinnvolle Ergänzung für die in Deutschland bisher angebotenen Datenzugangswege sein.

International gibt es bereits verschiedene Systeme für das automatisierte Fernrechnen mit sensiblen Daten.

1 §16 Abs. 6 Bundesstatistikgesetz (BstatG): „Für die Durchführung wissenschaftlicher Vorhaben dürfen vom Statistischen Bundesamt und den statistischen Ämtern der Länder Einzelangaben an Hochschulen oder sonstige Einrichtungen mit der Aufgabe unabhängiger wissenschaftlicher Forschung übermittelt werden …"

2 Faktische Anonymität ist erreicht, wenn die Einzelangaben nur mit einem unverhältnismäßig hohen Aufwand an Zeit, Kosten und Arbeitskraft zugeordnet werden können (§16 Abs. 6 BStatG).

Abb. 1 **Schematische Übersicht eines Remote Access Systems**



Übernommen aus: Schouten/ Cigrang (2003), S.17

Zwei Prinzipien können unterschieden werden:

1. Systeme, die auf E-Mail-Kontakt basieren (Remote Jobs Execution Systems, Remote Access Systems, Offline Systems, siehe Abbildung 1): Der Nutzer schickt seine Auswertungsprogramme als E-Mail oder E-Mail-Attachment. Die Syntax kommt nach einer Prüfung auf unerlaubte Inhalte in eine Warteschlange und wird nacheinander abgearbeitet. Die Geheimhaltungsprüfung der Ergebnisse läuft in zwei Stufen, die erste Stufe prüft automatisch. Werden Verletzungen der Regeln festgestellt, erfolgt in der zweiten Stufe die manuelle Prüfung. Der Nutzer erhält die Ergebnisse innerhalb von Minuten bis Tagen per E-Mail zurück, abhängig von der Größe des Programms und der Notwendigkeit manueller Nachprüfung. Die bereitgestellten Daten sind üblicherweise anonymisiert. Alle Programme, Protokolle und Outputs werden archiviert. In Zukunft sollen diese archivierten Unterlagen automatisch nach bestimmten Regeln zusammengefasst werden, um Datenangriffe zu identifizieren. Off-line-Systeme nutzen gängige Statistik-Softwarepakete. Diese Systeme können mit einer automatisiert kontrollierten Datenfernverarbeitung gleichgesetzt werden.
2. Die zweite Form des automatisierten Fernrechnens basiert auf dem Direktkontakt zwischen Nutzer und System über ein Web-Interface (On-line Systems, siehe Abbildung 2). Die Datenverarbeitung und Geheimhaltungsprüfung findet wärend der On-line-Sitzung statt und ist voll automatisiert, manuelle Einflussnahme gibt es nicht. Diese Systeme benutzen eigens kreierte Auswertungssoftware, da die gängigen Statistik-Pakete nicht web-interface-tauglich sind. On-line-Systeme sind für anspruchsvolle wissenschaftliche Auswertungen nicht geeignet. Sie bieten nur eingeschränkte Analysemöglichkeiten: Es werden Tabellen mit Prozentwerten oder Mittelwerten, allenfalls noch Varianzen oder Korrelationen berechnet und die Kombinationsmöglichkeiten von Merkmalen in diesen Tabellen ist meist eingeschränkt. Nichtversierte Nutzer können das Web-Interface leicht lernen und beherrschen, professionelle Nutzer müssen sich umstellen und bleiben trotzdem unbefriedigt.

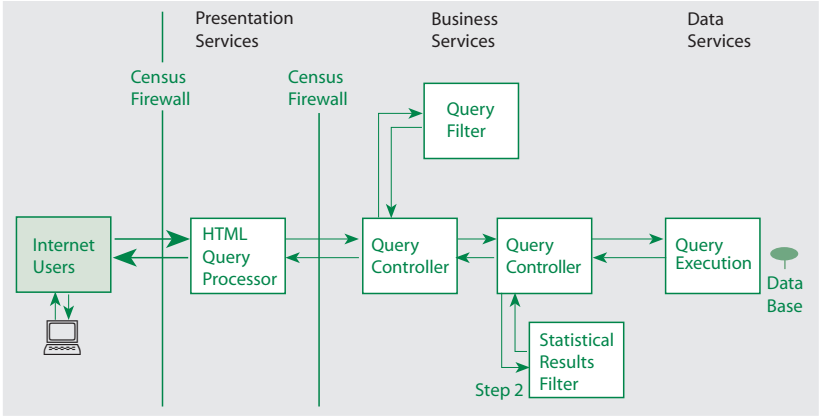
Beide Hauptsysteme des Fernrechnens sind mit einer automatisierten Geheimhaltungsprüfung kombiniert, die primäre und sekundäre Geheimhaltung sicherstellen soll. Primäre Geheimhaltung verhindert Deanonymisierungen, die unmittelbar aus einzelnen Elementen des Ergebnisses hervorgehen. Die sekundäre Geheimhaltung ergänzt die primäre Geheimhaltung. Sie verhindert, dass die zunächst gelöschten Elemente, z. B. Tabellenzellen, aus anderen Ergebnisbestandteilen wie Randverteilungen in der gleichen Tabelle oder ähnlichen anderen Tabellen rekonstruiert werden können. Dies funktioniert für einen begrenzten, einmaligen Output sehr gut. Fordert der Nutzer jedoch über längere Zeit immer wieder Auswertungen derselben Daten an, ist es durchaus möglich, dass er dann genügend Informationen gesammelt hat, um ursprünglich geheimgehaltene Output-Elemente zurückrechnen zu können. Dieses Problem ist bisher in keinem automatischen System vollständig gelöst.

Primäre und sekundäre Geheimhaltung werden durch drei generelle Methoden erreicht:

1. Veränderung (Anonymisierung) der Daten, z. B. durch Mittelwertbildung;
2. Veränderung einzelner Ergebnisse, z. B. Runden von Zellen in Tabellen;
3. Unterdrückung einzelner Ergebnisse, z. B. Löschen bzw. „Sperren" von Zellen in Tabellen.

Bei den vorhandenen Fernrechensystemen werden im Allgemeinen die erste und die dritte Möglichkeit kombiniert und eher rigide angewandt. Die erste Möglichkeit ist zwar relativ einfach zu realisieren – und stellt auch die primäre und sekundäre Geheimhaltung sicher, allerdings vermindert sie empfindlich den Informationsgehalt der Daten und ist deshalb bei den Datennutzern nicht sehr beliebt. Die beiden anderen Möglichkeiten sind bei einer Prüfung größerer Output-Mengen nicht machbar. Die vorhandenen automatischen Systeme, die die statistische Geheimhaltung bei geringstem Informationsverlust

Abb. 2 **Schematische Übersicht eines On-line-Systems (hier: Advanced Query System)**



Übernommen aus: Hawala u. a. (2004), S. 211

sicherstellen können, z. B. τ-ARGUS aus dem EU-CASC-Projekt (siehe unten), sind im Moment noch zu rechenzeitaufwändig und nur Prototypen. Die gegenwärtigen Fernrechensysteme löschen deswegen nicht einzelne Zellen, sondern ganze Tabellen, was gleichzeitig das Problem der sekundären Geheimhaltung in verbundenen Tabellen entschärft.

Der Kontrollaufwand bei den gegenwärtigen Fernrechensystemen wird aber auch im Vorhinein schon begrenzt: Es wird eine Liste erlaubter Anfragen definiert, in der z. B. auch die Anzahl der Tabellen, die maximale Größe der Tabellen und die erlaubten Kombinationen von Variablen festgelegt werden. Auch kann die maximale Anzahl von Anfragen pro Nutzer festgesetzt oder die Überlappung von Anfragen eingeschränkt werden. Die Prüfung dieser Regeln vermeidet Kontrollaufwand, bevor er überhaupt entsteht.

Die existierenden Fernrechensysteme zeigen auch, dass für jede Statistik maßgeschneiderte Kombinationen aus Anonymisierung und Geheimhaltungsprüfung notwendig sind, je nach Informationsgehalt und kombinierbarem, bereits veröffentlichtem Wissen. Datenfiles für das automatisierte Fernrechnen sind generell nicht mehr Originaldaten, wie z. B. für die kontrollierte Datenfernverarbeitung. Sie sind stärker anonymisiert, als es Files für Gastwissenschaftlerarbeitsplätze wären, aber nicht so stark wie Off-Site-Scientific-Use-Files.

Diese datenseitigen Maßnahmen werden in jedem Fall von anderen Sicherheitsmaßnahmen flankiert. Für das automatisierte Fernrechnen sind das beispielsweise ein kontrollierter Webzugriff über Nutzerregistrierung, Passwortschutz oder Output-Verschickung durch E-Mail statt auf den PC-Bildschirm. Dazu kommt die vertragliche Verpflichtung des Nutzers zur Geheimhaltung unter Strafandrohung. Besonders effektiv, aber aufwändig ist die nochmalige Prüfung der zur Veröffentlichung vorgesehenen Ergebnisse durch den Dateneigner.

3 Bestehende Systeme für automatisiertes Fernrechnen

Im Folgenden wird eine Auswahl internationaler Fernrechensysteme vorgestellt.

3.1 LISSY Remote-Access-System

Das LIS-Projekt (LIS – Luxembourg Income Study) sammelt seit 1983 vergleichbare Haushaltseinkommensbefragungen aus inzwischen 25 hauptsächlich europäischen Ländern. Es wird aus öffentlichen Geldern seiner Mitgliedsländer finanziert. Für den Datenzugriff wurde das Remote-Access-System LISSY[3] entwickelt, das für die Nutzer der Mitgliedsländer kostenfreien automatisierten Zugriff auf die Mikrodaten ermöglicht. LISSY ist die älteste Fernrechenmöglichkeit und Vorbild für viele Nachfolger. Es funktioniert offline über E-Mail-Kontakt. Durch die Programmierung mit JAVA ist LISSY auf fast allen Arbeitsplattformen einsetzbar. Die technischen Anforderungen sind: Pop3/SMTP kompatibler Mail-Server, Database System: Oracle (DB2 oder JDBC kompatibel). Es werden die Statistik-Pakete SPSS, STATA und SAS unterstützt. Potenzielle Nutzer müssen ihr Forschungsvorhaben skizzieren und eine Geheimhaltungsvereinbarung unterzeichnen. Sie erhalten im Internet ein umfangreiches Metadatenangebot einschließlich Dummy-Datenfiles. Die LISSY-Komponenten sind in Abbildung 3 dargestellt.

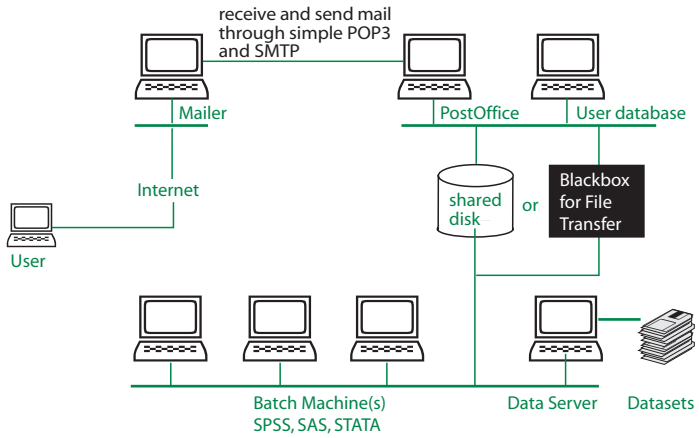
Die Post Office-Komponente holt alle fünf Sekunden neu eingegangene E-Mails vom Mail-Server ab. Sie überprüft die Nutzeridentität und sucht in der übermittelten Syntax nach unerlaubten Befehlen, Befehlsfolgen, Variablen oder nicht erlaubten Kombinationen. Es wird nur reines Textformat akzeptiert, Einbettungen oder E-Mail-Anhänge sind nicht gestattet. Bei Verletzung der Konventionen wird die E-Mail mit einer Fehlermeldung an den Absender zurückgeschickt. Die geprüfte Syntax gelangt zu den Batch-Computern, die den Output erzeugen. Dieser wird automatisch auf Größe und Inhalt geprüft. Wenn Sicherheitskonventionen verletzt sind, erfolgt eine Weiterleitung zur manuellen Prüfung. Der Output wird nur an die registrierte E-Mail-Adresse zurückgesandt, nicht an den Absender der Syntax. Die Datenbasen sind natürlich schreibgeschützt, so dass kein übermitteltes Programm sie verändern kann. Alle Jobs werden gespeichert, um gegebenenfalls Datenangriffe rekonstruieren zu können. Die angedrohten Sanktionen reichen vom Nutzungsausschluss bis zum Freiheitsentzug.

Das LISSY-System wurde für eine spezielle Datenbasis entwickelt, die Personendaten aus Stichprobenerhebungen mit freiwilliger Teilnahme enthält. Es ist deswegen nicht ohne weiteres auf andere Statistiken übertragbar. Vom Statistischen Bundesamt (Heitzig 2003) wird es als relativ unkomfortabel mit mangelhafter Geheimhaltung eingeschätzt.

3.2 ANDRE – Analytic Data Research by Email (USA)

ANDRE wird seit 1998 vom US National Center for Health Statistics betrieben. Es ist wie LISSY ein automatisches Remote-Access-System, erlaubt aber nur SAS-Jobs, da SAS-Jobs und -Outputs leichter automatisch auf Deanonymisierungen zu scannen sind. SAS-Kommandos wie ADD, PRINT, OBS werden unterdrückt, PROC MEANS, NMEAN STD werden modifiziert. Befehle, die unstrukturierten Output erzeugen, der nicht automatisch zu scannen ist, werden ebenfalls nicht zugelassen. Die automatische Output-Prüfung identifiziert Extremwerte und Werte, die auf zu geringen Fallzahlen beruhen, als kritisch und leitet diesen Output an die manuelle Prüfung weiter. Syntax, Protokolle und Outputs

Abb. 3 **Das LISSY Remote-Access-System**



Übernommen aus: Coder/ Cigrang (2003), S.2

3 http://www.lisproject.org

werden gespeichert. Zur Nutzung des Systems ist ein Projektantrag notwendig und es wird ein Vertrag mit Geheimhaltungsklauseln abgeschlossen. Der Nutzer erhält eine User ID und ein Passwort. Innerhalb des Systems sind das Abspeichern eigener Datensets und das Mergen von Dateien möglich. Dem Datennutzer werden, wie bei der kontrollierten Datenfernverarbeitung, Dummy-Daten zur Verfügung gestellt. Die Kosten für Nutzer sind bei diesem Komfort erheblich: 500 $ pro Monat. An einem internetbasierten, nutzerfreundlicheren System wird zur Zeit gearbeitet (ANDREW: Analytical Data Research by E-Mail and Web, siehe Gambhir/Harris 2005).

3.3 DAS – Data Analysis System (USA)

Vom National Center for Educational Statistics wird DAS[4] als web-interface-basiertes Direktzugriffs-System zur Generierung von Tabellen und Korrelationsmatrizen angeboten. Es hat wie alle On-line-Systeme eine eigene Programmiersprache. Die angebotenen Daten sind anonymisiert, Zellen und Randverteilungen unter 30 Fällen in Output-Tabellen werden unterdrückt. Der Zugriff ist kostenlos, Hilfe ist über E-Mail oder online erreichbar.

3.4 AQ – Advanced Query System (USA)

Das US Census Bureau bietet das Advanced Query System für die Mikrodaten des Zensus an (vgl. Hawala u.a. 2004). Das System funktioniert online und voll automatisiert auf der Basis einer kommerziellen Business Intelligence Software. Den Nutzern entstehen keine Kosten, wobei der Nutzerkreis allerdings auf staatliche Stellen eingeschränkt ist.

Die Zensusdaten durchlaufen vor der Nutzung eine Anonymisierung. Die Ausprägungen verschiedener Merkmale wie Alter und Rasse werden zu Kategorien zusammengefasst und bei metrischen Merkmalen wie Einkommen und Ausgaben die höchsten Werte abgeschnitten. Außerdem kommt die Technik des „data swapping" zum Einsatz: Haushalte mit dem höchsten Reidentifizierungsrisiko erhalten eine Markierung und werden mit ähnlichen aus anderen geographischen Einheiten paarweise getauscht. Die Geheimhaltung wird zusätzlich über Nutzerregistrierung bzw. -autorisierung sichergestellt.

Die Funktionskomponenten des Advanced Query Systems sind in Abbildung 2 dargestellt. Zwei Firewalls schützen die Mikrodaten vor unerlaubten Zugriffen aus dem Internet. Der Query-Filter identifiziert Anfragen, die von vornherein nicht die Geheimhaltungskontrolle passieren würden. Damit werden die Systemressourcen geschont und der Nutzer erhält in relativ kurzer Zeit Nachricht darüber, ob seine Tabellenanforderung den Geheimhaltungsregeln entspricht. Dieser Vorfilter kontrolliert vor allem die regionale Tiefe der eingehenden Auswertungswünsche. Abgesehen von der Regionalvariablen ist die Anzahl der Variablen in einer Tabelle auf drei beschränkt. Der Results-Filter überprüft die Zellen der berechneten Tabellen. Enthält zum Beispiel die Mehrzahl der Tabellenzellen keine oder nur einen Fall, wird die Tabelle komplett gesperrt. Der Median und der Mittelwert der Zellenbesetzungen einer Tabelle müssen über einem – geheimgehaltenen – Grenzwert liegen. Hawala u.a. (2004) fanden heraus, dass die Forderung nach einer minimalen Zellenbesetzung überflüssig ist, da in diesem Fall andere Bedingungen ebenfalls nicht erfüllt sind. Deshalb wird die Prüfung der minimalen Zellenbesetzung künftig entfallen. Alle Anfragen werden gespeichert, um Informationen über Nutzerprioritäten zu erhalten und mögliche Denanonymisierungsrisiken zu identifizieren.

Dem Geheimhaltungsmanagement des Advanced Query System wird von externen Experten bescheinigt, dass es sinnvoll, umfassend und substantiell „best practice" ist (Duncan 2000, zit. nach Hawala u.a. 2004, S. 123, eigene Übersetzung). Die Datennutzer sind ebenfalls sehr zufrieden. Bei einer Öffnung des Systems für die Allgemeinheit gibt es nicht etwa sicherheitsrelevante Befürchtungen, sondern Bedenken, dass die technischen Ressourcen den Anfragen nicht gewachsen sein könnten.

3.5 American FactFinder (USA)

Für die Allgemeinheit und ohne Nutzerregistrierung zugänglich ist der American FactFinder[5], der aggregierte Zensus-Daten verarbeitet und ebenfalls online vordefinierte Tabellen für ausgewählte geografische Einheiten ausgibt.

3.6 Das dänische Fernrechensystem

2001 von Danmarks Statistik eingerichtet, sollte das Fernrechensystem als Ersatz für die kontrollierte Datenfernverarbeitung dienen, stellt aber inzwischen auch die vorhandenen Gastwissenschaftlerarbeitsplätze zur Disposition. Das dänische System wird hier ausführlicher dargestellt, da es vom Statistischen Bundesamt als Vorbild für eine mögliche deutsche Lösung angesehen wird (vgl. auch Rowland 2003, Heitzig 2003, Statistics Sweden 2003, Wende u.a. 2004, Borchsenius 2005).

Danmarks Statistik hat eine Kombination aus On-Line- und Off-Line-System geschaffen, vorstellbar als eine Art Gastwissenschaftlerarbeitsplatz, der sich jedoch nicht im Statistischen Amt befindet. Der Nutzer startet von einem entfernten PC aus eine Terminalsitzung auf dem Server im Amt. Der Programmcode wird online übermittelt, der Output kommt via E-Mail. Die Daten werden abgetrennt von den normalen Arbeitsservern des Amtes auf SUN/UNIX Servern gespeichert. Der Zugriff findet über Citrix Server und verschlüsselte Internetverbindungen statt und ist passwortgeschützt. Nur die Client-Software, die Danmarks Statistik zur Verfügung stellt, kann für den Datenzugriff genutzt werden. Inzwischen kann der Terminal-PC unter bestimmten Bedingungen sogar an der privaten Adresse stehen. Die Daten bleiben immer auf den Servern, sie werden niemals auf die Terminal-PCs übertragen. Sie können auf dem Bildschirm angelistet, jedoch nicht heruntergeladen oder ausgedruckt werden. Der Nutzer kann die Daten verändern und diese Veränderungen auf den Servern auch abspeichern. Danmarks Statistik führt unangemeldete Kontrollen der Terminal-PCs durch. Bei Verletzung der Regeln droht der Ausschluss von weiterer Datennutzung – bis jetzt ist noch kein Fall bekannt geworden.

Der Zugang zu den Daten ist Wissenschaftlern anerkannter Institute mit Sitz in Dänemark vorbehalten. Institute müssen sich bewerben und werden autorisiert. Auch private Einrichtungen wie Nichtregierungsorganisationen, Beratungsfirmen und Unternehmen können eine Autorisation erhalten. Einzelne Wissenschaftler ohne institutionellen Hintergrund, Einzelpersonen und die Massenmedien erhalten keinen Datenzugang. Diese

4 http://nces.ed.gov/dasol/
5 http://factfinder.census.gov/home/saff/main.html

Einschränkungen werden unter anderem gemacht, um im Fall einer Datenschutzverletzung eine Adresse für Sanktionen zu haben. Forscher, deren Institute die Berechtigung zum Fernrechnen (noch) nicht erhalten haben, können immer noch die On-Site-Möglichkeiten, d. h. Gastwissenschaftlerarbeitsplätze an zwei verschiedenen Standorten nutzen. Wissenschaftler an berechtigten Instituten müssen einen projektbezogenen Antrag auf Datennutzung stellen. Über den Datenzugang im Einzelnen entscheiden die Direktoren von Danmarks Statistik und die dänische Datenschutzbehörde. Gestattet wird nur die Auswertung der im Rahmen des Forschungsvorhabens unbedingt nötigen Merkmale.

Nicht alle vorhandenen Statistiken sind für die wissenschaftliche Analyse zugänglich, z. B. nicht die Kriminalitätsstatistik. Unternehmensdaten werden vor der Freigabe sorgfältig geprüft und grundsätzlich nur als Stichprobe zur Verfügung gestellt. Wenn ein Forscher die Totalerhebung von Unternehmensdaten auswerten möchte, wird die Anzahl der Variablen beschränkt. Dafür sind alle Daten nur formal anonymisiert, d. h. es werden nur die direkten Identifikatoren entfernt. Die in Dänemark verfügbaren amtlichen Mikrodaten bieten mehr Analysemöglichkeiten, als in Deutschland nach dem BStatG rechtlich zulässig wäre, da z. B. anhand einer einheitlichen Personennummer mehrere Datensätze zusammengefügt werden können. Die Personennummer wird danach ersetzt.

Der produzierte Output wird den Forschern innerhalb von 5-10 Minuten per E-Mail zugesandt und in einem Protokoll gespeichert. Eine Geheimhaltungskontrolle erfolgt nur stichprobenartig im Nachhinein (!) und nicht automatisiert. Falls sich dann Geheimhaltungsprobleme ergeben, wird der Forscher darauf hingewiesen, wie er mit der Veröffentlichung verfahren soll und wie solche Fälle in Zukunft zu vermeiden sind. Ein solches Vorgehen ist nur möglich, weil die Wissenschaftler der autorisierten Institute den Status eines Geheimnisträgers, vergleichbar mit der ärztlichen oder anwaltlichen Schweigepflicht erhalten. So übernehmen sie ebenfalls Verantwortung für die Geheimhaltung der Individualdaten.

Die Bereitstellung der Daten erfolgt nach Vertragsabschluss innerhalb von ein bis drei Monaten. Metadaten sind über das Internet verfügbar. Nutzer des dänischen Remote-Access-Systems beschreiben einen Komfort ähnlich einer On-Site-Nutzung. Das Statistische Amt ist mit der Lösung ebenfalls zufrieden, da sie nach eigener Aussage bessere Kontrollmöglichkeiten und einen Sicherheitsgewinn gegenüber der üblichen Gastwissenschaftlerlösung bietet. Die Nutzungsintensität ist seit der Einrichtung erheblich angestiegen. Im Jahr 2003 wurden 250 Projekte bearbeitet, pro Arbeitstag arbeiteten ca. 20-50 Forscher im System; dreizehn Mitarbeiter betreuen die Wissenschaftler.

Das dänische System wurde von Mitarbeitern des Statistischen Bundesamtes ebenfalls sehr positiv bewertet: „Das dänische System erscheint geeignet, Wissenschaftlern auf komfortable Weise Analysen mit Einzeldaten zu ermöglichen, ohne dass diese Einzeldaten als Dateien den geschützten Bereich des Servers verlassen. Die Umsetzbarkeit dieses Weges für Deutschland bleibt zu prüfen." (Wende/Heitzig 2004). Leitprinzip des dänischen Systems ist, dass amtliche Daten in keiner Form aus dem Statistischen Amt gelangen dürfen. Das bedeutet, dass es nicht – wie z. B. in Deutschland – absolut anonymisierte Public-Use-Files oder faktisch anonymisierte Scientific-Use-Files gibt, Mikrodaten also in anonymisierter Form herausgegeben werden. In der Summe bietet Danmarks Statistik deshalb, verglichen mit Deutschland, nicht unbedingt erleichterte Zugangswege. Ein direkter Kontakt externer Forscher mit Originaldaten, verbunden mit einer nur punktuellen Geheimhaltungskontrolle im Nachhinein ist jedenfalls in Deutschland mit der Gesetzeslage kaum vereinbar.

3.7 Die Position von Eurostat

Die Statistikbehörde der EU Eurostat, ist beiden Grundkonzepten des automatisierten Fernrechnens (E-Mail und Web-Interface) gegenüber skeptisch. Dies ist sehr verständlich, muss doch Eurostat die Datenschutzansprüche aller Mitgliedsländer auf den kleinsten gemeinsamen Nenner bringen und hat dabei keine rechtlichen Möglichkeiten, Sanktionen bei Datenschutzverletzungen europaweit durchzusetzen (Eurostat 2004). Eurostat sieht deshalb für seine Datenbestände in naher Zukunft keine Möglichkeiten für einen automatisierten Fernrechenzugriff, räumt aber dem kombinierten On-line-Off-line-System, so wie es in Dänemark angewandt wird, die größten Chancen ein. Eurostat will sich dem internationalen Trend jedoch insofern nicht verschließen, als es jeden Fortschritt im Hinblick auf die Fernrechenmöglichkeiten begrüßen wird und sich auch an Machbarkeitsstudien und Tool-Entwicklungen (wie beim CASC-Projekt) beteiligen will.

3.8 Das CASC - Projekt

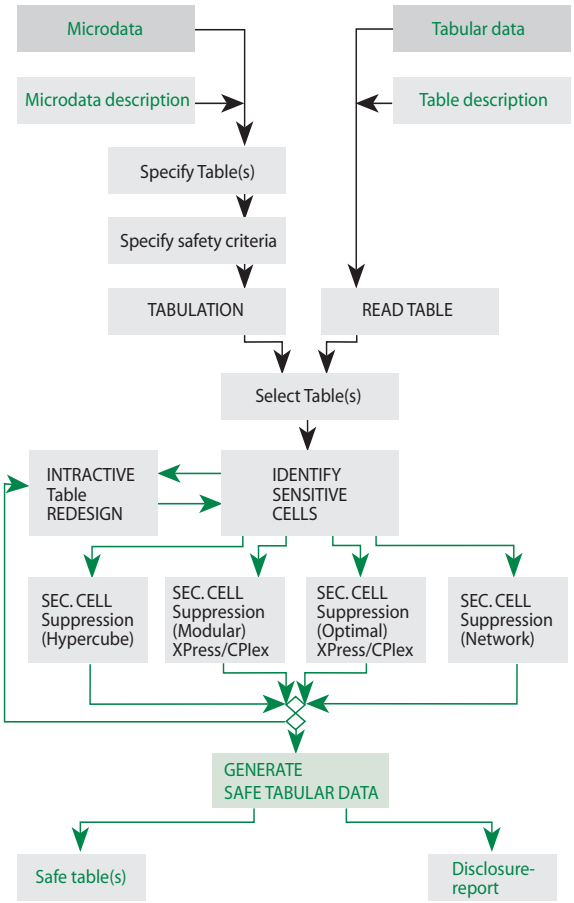
Das EU-CASC-Projekt[6] hat keine eigene Fernrechenmöglichkeit entwickelt, aber zwei dafür unabdingbare Bausteine: die Softwarepakete μ-ARGUS und τ-ARGUS zur mehr oder weniger automatischen Gewährleistung der statistischen Geheimhaltung. μ-ARGUS anonymisiert Daten und τ-ARGUS prüft Tabellen auf Geheimhaltung. Beide Softwareprodukte sind kostenlos im Internet erhältlich.

μ-ARGUS ist demnach prinzipiell zur Vorbereitung der Daten für das automatisierte Fernrechnen geeignet, τ-ARGUS für die automatisierte Geheimhaltungsprüfung. τ-ARGUS liest sowohl Mikrodaten als auch Tabellen ein und berechnet selbst Tabellen mit spezifizierten Geheimhaltungsparametern. Es werden Mikrodaten im ASCII-Format mit fester Feldlänge oder mit Separator verarbeitet. Die meisten amtlichen Daten sind unkompliziert in diese Formate zu überführen. Wenn τ-ARGUS fertige Tabellen einliest, sind Tabellenrestrukturierungen (d. h. Zusammenlegen von Zeilen oder Spalten) jedoch nicht mehr möglich. τ-ARGUS bietet verschiedene Methoden der Optimierung der Geheimhaltung an, bei manchen ist ein kommerzieller, d. h. kostenpflichtiger so genannter „LP-Solver" nötig (vgl. Abbildung 4).

Das Arbeiten mit τ-ARGUS erscheint für die automatisierte Anwendung, bei der große Mengen verschiedenster Tabellen geprüft werden, etwas mühevoll. Jede Tabelle wird manuell am Bildschirm zusammengefügt, das Ergebnis wird wieder manuell modifiziert. Jede Tabellenzelle kann einzeln angesprochen bzw. auch gegen Veränderung geschützt werden. Dieses aufwändiges Unterfangen ist der Verkürzung der Prüfzeiten nicht dienlich. Mit

6 http://neon.vb.cbs.nl/casc/

Abb. 4 **Die Funktionskomponenten von t-ARGUS**



Übernommen aus: Statistic Netherlands (2004), S. 19

Hilfe der eigenen (umständlichen) Programmiersprache bzw. dem Ausschreiben der zusammengeklickten Kommandos als Programmcode ist jedoch ein „Batch-mode" möglich, der wiederum nur für die wiederholte Erstellung identischer Tabellensätze interessant ist.

Diese Einschränkungen und die langen Rechenzeiten führen dazu, dass τ-ARGUS bisher nicht großflächig im Einsatz ist.

3.9 Die Entwicklung in Deutschland

Auch in Deutschland gibt es Diskussionen, den Zugriff auf amtliche Daten komfortabler zu gestalten und durch ein Fernrechensystem zu automatisieren. Das Statistische Bundesamt ist dabei konzeptioneller Vorreiter, wird aber ohne die Zusammenarbeit mit den Statistischen Landesämtern keine befriedigende Lösung anbieten können. Bisher haben die Statistischen Landesämter berechtigte Bedenken wegen des unbefriedigenden Datenschutzes beim automatisierten Fernrechnen. Da sie Dateneigner bei den meisten Statistiken sind, also deren Verwendung kontrollieren, ist hier eine enge Zusammenarbeit von Statistischem Bundesamt und Statistischen Landesämtern unumgänglich.

Das Konzept des Statistischen Bundesamtes für eine deutsche Fernrechenmöglichkeit, der so genannte Wissenschaftsserver oder auch SAM (Server for Access to Microdata, siehe auch Wende u.a. 2005, Zühlke 2005) ist eng an das dänische Vorbild angelehnt. Die Titelgrafik zeigt eine mögliche Lösung mit Einbindung der Anonymisierungssoftware μ-ARGUS und der Geheimhaltungssoftware τ-ARGUS. Es handelt sich hier ebenfalls um ein Mischsystem aus On-line- und Off-line-System. Der Nutzer soll den Eindruck haben, er arbeite an einem gewöhnlichen Gastwissenschaftlerarbeitsplatz. Er hat wie dort die Wahl zwischen den Statistikpaketen SAS, SPSS und Stata. Die Jobs oder auch die Menüeingaben werden vom Nutzer von einem Terminal-Rechner aus an einen entfernten Server im Statistischen Bundesamt abgeschickt, das Ergebnis kommt als E-Mail an ihn zurück. Durch verschiedene Sicherheitsmaßnahmen werden die zur Verfügung stehenden Daten als faktisch anonym angesehen. Dazu gehört, neben dem technischen und vertraglichen Schutz der Daten und anders als in Dänemark, auch die Geheimhaltungsprüfung der Ergebnisse. Dadurch wäre hier in jedem Fall mit längeren Verzögerungen zu rechnen als im dänischen System.

Für die Sicherstellung der Geheimhaltung wird unter anderem die Jack-Knife-Methode diskutiert (Heitzig 2005). Diese Technik ist als vollautomatische Ad-hoc-Anonymisierung der Daten während der Analyse zu verstehen, die eine vorherige Anonymisierung und auch eine Geheimhaltungsprüfung der Ergebnisse überflüssig machen soll. Bei der Jack-Knife-Methode enthält der Output keine Punktwerte, sondern ein Intervall, in dem sich der wahre Wert befindet. Das Intervall entsteht durch das mehrmalige Ersetzen eines oder mehrerer Einzeldaten im Mikrodatenfile durch Zufallswerte und anschließend durchgeführter Analysen. Ein Datenangreifer soll so keine Chance mehr haben, aus den berechneten Tabellenzellen, Koeffizienten oder Testwerten das wahre Mikrodatum für den interessierenden Fall über Gleichungssysteme zurückzurechnen. Prototypische Anwendungen bietet im Moment das Statistikpaket SAS, die Methode ist allerdings noch in einem frühen Entwicklungsstadium. Bei der Anwendung müsste außerdem sichergestellt sein, dass der Nutzer die Daten nicht auf dem Schirm sieht, was aber alle gebräuchlichen Statistik-Pakete gestatten und am herkömmlichen Gastwissenschaftlerarbeitsplatz kein Problem ist.

Eine Punktlösung in Hinblick auf die automatisierte Geheimhaltungskontrolle wird im Landesamt für Datenverarbeitung und Statistik Nordrhein-Westfalen angewandt. Für die Tabellierung der Beherbergungsstatistik wird ein EDV-gestütztes System zur Identifizierung und Sperrung kritischer Zellen in Tabellen eingesetzt (vgl. Radmacher-Nottelmann 2004). Dieses Verfahren kombiniert die p%-Regel[7] für die primäre Geheimhaltung mit dem so genannten Quaderverfahren für die sekundäre Geheimhaltung (Repsilber 1999), mit dem die Zahl der zu sperrenden Tabellenzellen minimiert werden kann. Es ist durchaus denkbar, dieses Verfahren in die automatisierte Output-Prüfung beim Fernrechnen einzubinden.

4 Fazit

Auch in Deutschland wird zur Zeit die Möglichkeit geprüft, automatisiertes Fernrechnen als weiteren Zugangsweg zu amtlichen Daten anzubieten. Zwar gehört Deutschland wegen der strengen Datenschutzgesetze beim automatisierten Fernrechnen nicht zu den Vorreitern, das hat für die amtliche Statistik jedoch den Vorteil, dass sie aus den internationalen Erfahrungen lernen kann. Dabei wären die folgenden Punkte besonders beachtenswert.

7 „p%-Regel": Geheimhaltungsregel der amtlichen Statistik für Summen in Tabellenzellen: Geheimhaltungsbedarf ist gegeben, wenn nach Abzug des zweitgrößten Anteils von der Gesamtmenge eine Schätzung für das größte Objekt entsteht, deren Fehler weniger als p% (in der Regel 5%, statistikabhängig) beträgt

Das Ziel von automatisierten Fernrechungssystemen muss sein, dass sie so komfortabel wie möglich sind und dem Nutzer den Eindruck vermitteln, er arbeite an seinem eigenen PC (vgl. Desai 2003). Das bedeutet, dass die Ergebnisse dem Nutzer schnell zugestellt werden, der Nutzer nicht erst neue Auswertungsprogramme lernen muss und die Restriktionen der Datenbearbeitung auf ein Minimum reduziert sind. Es sollte u.a. die Möglichkeit geben, den Datensatz zu verändern und die Veränderungen auch abzuspeichern. Gleichzeitig muss die statistische Geheimhaltung gewährleistet sein.

Reine On-line-Systeme, wie der American FactFinder, können anspruchsvolle Analysewünsche nicht befriedigen, reine Off-line-Systeme, wie z. B. LISSY, sind fast als überholt anzusehen, deswegen ist ein kombiniertes Online-Off-line-System, wie es von Danmarks Statistik eingerichtet wurde, das System der Wahl für Deutschland. Hier ist jedoch zu überlegen, ob die Zugriffsrechte dieses verlagerten Gastwissenschaftlerarbeitsplatzes so weit gehen sollten, dass der Nutzer wie beim Gastwissenschaftlerarbeitsplatz im Statistischen Amt Einblick in die Mikrodaten nehmen kann.

Was die zu beschaffende Hardware betrifft, beziehen sich notwendige Überlegungen bei der Einrichtung eines automatisierten Fernrechensystems auf den benötigten Speicherplatz und die Entscheidung, ob das System auf PC- oder UNIX-Basis laufen soll. Im Bereich der Software muss entschieden werden, ob diese neu entwickelt oder gängige Software eingesetzt wird und was im letzteren Fall die Lizenzen kosten. On-line-Systeme sind der Trend, es bleibt deshalb zu hoffen, dass die üblichen Statistikpakete internetfähig werden und eine Geheimhaltungskontrolle erlauben. Zudem gilt es einigen Aufwand für die Datenvorbereitung zu kalkulieren.

Nützliche Ergänzungen sind eine User-Group-Mail-List und eine Website. Sie entlasten Mitarbeiter, da sich viele Fragen von alleine oder zwischen den Usern klären. Im Falle eines Remote-Access-Systems ist das Angebot von Datenstrukturfiles sehr hilfreich und ressourcenschonend, da der potenzielle Nutzer mit deren Hilfe die Daten kennen lernen und seine Auswertungsprogramme ausprobieren kann.

Noch nicht gelöst ist das Problem komplementärer Outputs, die kombiniert zur Deanonymisierung genutzt werden können. Hier existiert auch international noch keine zuverlässige automatisierte Kontrolle. Deshalb müssen Restrisiken vertraglich abgesichert werden. Zu den üblichen Sicherheitsmaßnahmen gehören auch ein Passwortschutz, die vertragliche Verpflichtung der Datennutzer auf statistische Geheimhaltung und eine generelle Zugangsbeschränkung – in Deutschland nach §16 BstatG auf Wissenschaftler an Universitäten und anderen Einrichtungen unabhängiger wissenschaftlicher Forschung.

Die Ergebnisprüfung bleibt sehr aufwändig, deshalb gilt es, kritischen Output von vornherein zu verhindern, indem bestimmte Befehle und Prozeduren durch eine Textsuche blockiert werden.

Zu guter Letzt ist zu bedenken, dass eine voll automatisierte Geheimhaltungsprüfung immer restriktiver sein wird als eine manuelle. Deshalb bleibt für das nutzerorientierte Funktionieren des Systems die zweistufige Prüfung aktuell. In der automatisierten ersten Stufe wird geheimhaltungskritischer Output identifiziert, der in der zweiten Stufe durch einen Mitarbeiter noch einmal geprüft wird. Dieses stufenweise Vorgehen wird dem Zielkonflikt zwischen Datensicherheit und Ausbeutung des Informationsgehaltes der Daten am besten gerecht. Der Auswertungsmöglichkeit verschiedenster Statistiken (und nicht nur einer wie beim LISSY-System) über das Fernrechnen wird eine umfangreiche individuelle Vorbereitung dieser Daten vorausgehen, und die Geheimhaltungsprüfung der jeweiligen Ergebnisse muss ebenfalls mit einer zwar automatisierten, aber an die Statistik individuell angepassten Geheimhaltungskontrolle sichergestellt werden. Eine Aufwandsersparnis durch die automatisierte Fernrechenmöglichkeit ist deswegen vorerst und bis auf weiteres nur für den Datennutzer zu erwarten.

Quellennachweis

- Coder, J./Cigrang, M. (2003): LISSY Remote Access System. Joint ECE/ Eurostat work session on statistical data confidentiality Luxembourg, 7 - 9 April 2003, Working Paper No. 7.
- Borchsenius, L., Statistics Denmark (2005): New Developments in the Danish System for Access to Microdata. Joint ECE/Eurostat work session on statistical data confidentiality Geneva, 9 - 11 November 2005, Invited Paper.
- Desai, T. (2003): Providing Remote Access to Data: The Academic Perspective. Joint ECE/Eurostat work session on statistical data confidentiality Luxembourg, 7 - 9 April 2003, Working Paper No. 9.
- Duncan, G. (2000): Final Report on the American FactFinder Disclosure Audit Project for the U.S. Census Bureau. Prepared under contract to the U.S. Census Bureau.
- Eurostat (2004): Remote Access to Confidential Data. Information Technologies Directors Meeting, Doc. ITDG/October 2004/3.5.
- Franconi, L./Merola, G. (2003): Implementing Statistical Disclosure Control for Aggregated Data Released via Remote Access. Joint ECE/Eurostat work session on statistical data confidentiality Luxembourg, 7 - 9 April 2003, Working Paper No. 30.
- Gambhir, V./Harris, K. W. (2005): Analytical Atat Research by Email and Web (ANDREW). Joint ECE/ Eurostat work session on statistical data confidentiality Geneva, 9 - 11 November 2005, Supporting Paper.
- Hawala, S./Zayatz, L./Rowland, S. (2004): American FactFinder: Disclosure Limitation for the Advanced Query System, in: Journal of Official Statistics, Vol. 20, Nr. 1, S. 115-124.
- Heitzig, J., Statistisches Bundesamt (2003): Fernrechnen mit Mikrodaten, unveröffentlichtes Papier.
- Heitzig, J. (2005): The „Jackknife" Method: Confidentiality Protection for Complex Statistical Analyses. Joint ECE/Eurostat work session on statistical data confidentiality Geneva, 9-11 November 2005, Invited Paper.
- Heitzig, J./Wende, T. (2004): Bericht über den Forschungsaufenthalt in Dänemark zur Evaluierung des Dänischen Systems zum Online-Zugriff auf amtliche Mikrodaten vom 4. bis zum 6. Februar 2004, vorgelegt auf der 6. Sitzung der Bund-Länder AG am 4. März 2004.
- Radmacher-Nottelmann, N. (2004): Geheimhaltung mit Makrodaten. Das Beispiel der Beherbergungsstatistik in: Statistische Analysen und Studien NRW, Band 19, S. 19 - 37.
- Repsilber, R. D. (1999): Das Quader-Verfahren, in: Methoden zur Sicherung der statistischen Geheimhaltung, hrsg. vom Statistischen Bundesamt, Wiesbaden, S. 27 - 64.
- Rowland, S. (2003): An Examination of Monitored, Remote Microdata Access Systems. Presented at the National Academy of Sciences Workshop on Access to Research Data, October 16 - 16, 2003.
- Schouten, B./Cigrang M., Statistics Netherlands (2003): Remote Access Systems for Statistical Analysis of Microdata, Discussion paper 03004.
- Statistics Netherlands (2004): τ-ARGUS, User's Manual.
- Statistics Sweden (2003): Access to Microdata in the Nordic Countries.
- Wende, T./Heitzig, J./Erdmann, K. (2005): Konzept zur Wahrung der faktischen Anonymität (§16 Abs. 6 BstatG) im Rahmen eines Zugangs zu amtlichen Einzeldaten über einen Wissenschaftsserver im Statistischen Bundesamt, unveröffentlichtes Papier.
- Zühlke, S./Zwick, M./Scharnhorst, S./Wende T. (2005): The research data centres of the Federal Statistical Office and the statistical offices of the Länder, Statistische Ämter des Bundes und der Länder, Forschungsdatenzentren, FDZ-Arbeitspapier Nr. 3.