RatSWD Working Paper Series

www.ratswd.de

RatSWD.
Rat für Sozial- und WirtschaftsDaten

261

Standards des sicheren Datenzugangs in den Sozial- und Wirtschaftswissenschaften: Überblick über verschiedene Remote-Access-Verfahren

David H. Schiller, Johanna Eberle, Daniel Fuß, Jan Goebel, Jörg Heining, Tatjana Mika, Dana Müller, Frank Röder, Michael Stegmann und Karsten Stephan

März 2017

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Working Paper Series des Rates für Sozial- und Wirtschaftsdaten (RatSWD)

Die *RatSWD Working Papers* Reihe startete Ende 2007. Seit 2009 werden in dieser Publikationsreihe nur noch konzeptionelle und historische Arbeiten, die sich mit der Gestaltung der statistischen Infrastruktur und der Forschungsinfrastruktur in den Sozial-, Verhaltens- und Wirtschaftswissenschaften beschäftigen, publiziert. Dies sind insbesondere Papiere zur Gestaltung der Amtlichen Statistik, der Ressortforschung und der akademisch getragenen Forschungsinfrastruktur sowie Beiträge, die Arbeit des RatSWD selbst betreffend. Auch Papiere, die sich auf die oben genannten Bereiche außerhalb Deutschlands und auf supranationale Aspekte beziehen, sind besonders willkommen.

RatSWD Working Papers sind nicht-exklusiv, d. h. einer Veröffentlichung an anderen Orten steht nichts im Wege. Alle Arbeiten können und sollen auch in fachlich, institutionell und örtlich spezialisierten Reihen erscheinen. Die *RatSWD Working Papers* können nicht über den Buchhandel, sondern nur online über den RatSWD bezogen werden.

Um nicht deutsch sprechenden Nutzer/innen die Arbeit mit der Reihe zu erleichtern, sind auf den englischen Internetseiten der *RatSWD Working Papers* nur die englischsprachigen Papers zu finden, auf den deutschen Seiten werden alle Nummern der Reihe chronologisch geordnet aufgelistet.

Einige ursprünglich in der *RatSWD Working Papers* Reihe erschienenen empirischen Forschungsarbeiten sind ab 2009 in der RatSWD Research Notes Reihe zu finden.

Die Inhalte der *RatSWD Working Papers* stellen ausdrücklich die Meinung der jeweiligen Autor/innen dar und nicht die des RatSWD. Das Bundesministerium für Bildung und Forschung hat die Publikationen nicht beeinflusst.

Standards des sicheren Datenzugangs in den Sozial- und Wirtschaftswissenschaften: Überblick über verschiedene Remote-Access-Verfahren

Autorinnen und Autoren: David H. Schiller (Freiberufler, ehemaliges Mitglied des RatSWD), Johanna Eberle (IAB), Daniel Fuß (LIfBi), Jan Goebel (SOEP), Jörg Heining (IAB), Tatjana Mika (DRV), Dana Müller (IAB), Frank Röder (DRV), Michael Stegmann (DRV) und Karsten Stephan (DZHW).

Inhalt

Abstract 1
Einleitung 2
Begriffsklärungen und Rahmenbedingungen 2
Remote-Execution-Verfahren 8
Remote-Desktop-Verfahren 17
Remote-Access-Verfahren weltweit 27
Zusammenfassung und Ausblick 30
Abbildungsverzeichnis 32
Literaturverzeichnis 33

Abstract

Die Forschung in den Sozial- und Wirtschaftswissenschaften ist immer öfter auf einen abgesicherten Zugang zu Forschungsdaten angewiesen, da ansonsten die Anforderungen des Datenschutzes nicht erfüllt werden können. Remote-Access-Lösungen bieten hier einen komfortablen Datenzugang und ermöglichen gleichzeitig einen hohen Sicherheitsstandard. Der Text klärt kurz Begrifflichkeiten und Rahmenbedingungen des Datenzugangs in den Sozial- und Wirtschaftswissenschaften und beschreibt daraufhin Funktionsweisen und Ausformungen des Remote Access. Als exemplarische Lösungen werden Remote-Access-Verfahren von fünf deutschen Forschungsdatenzentren (FDZ), die im „Rat für Sozial- und Wirtschaftsdaten“ (RatSWD) organisiert sind, dargestellt. Dabei werden für jedes FDZ die folgenden Themenkomplexe erörtert: „Organisation und Motivation für die Einrichtung des Verfahrens“, „Beschreibung des Verfahrens“, „Vorteile und Nachteile des Verfahrens“, „Rechtliches und Datensicherheit“ und „Aufwand für den Betrieb“. Nach der Darstellung der deutschen Remote-Access-Lösungen folgt ein Überblick zu Remote-Access-Verfahren in anderen Ländern. Der Text schließt mit einer Zusammenfassung bezüglich der vorgestellten Remote-Access-Lösungen und einem Ausblick auf weitere Entwicklungen.

Einleitung

Moderne Analysemethoden in den Sozial- und Wirtschaftswissenschaften (z. B. multivariate Verfahren) sind oft auf Mikrodaten angewiesen. Diese Mikrodaten beziehen sich auf einzelne Untersuchungseinheiten (z. B. Personen, Betriebe, Haushalte). Das Risiko einer Zuordnung dieser Mikrodaten zu realen Individuen ist entsprechend hoch, insbesondere bei Panelstudien. Zur Einhaltung der notwendigen Datenschutzmaßnahmen konnte Forschung mit diesen Daten nur in abgesicherten Räumen der jeweiligen datenhaltenden Institutionen durchgeführt werden. Um Forschung auch am eigenen Arbeitsplatz zu ermöglichen, wurden Scientific Use Files (SUF) erstellt, die weniger Detailinformationen enthalten und aufgrund dessen an die Forscherinnen und Forscher, unter Einhaltung von Sicherheitsstandards, übermittelt werden dürfen. Die moderne Forschung befand sich somit in dem Dilemma, mit SUF zu arbeiten, die teils nicht genug Detailinformationen enthalten, um aussagekräftige Ergebnisse zu erzielen, oder zu den datenhaltenden Institutionen reisen zu müssen, was einen hohen finanziellen und zeitlichen Aufwand zur Folge hatte. Remote-Access-Verfahren eröffnen neue Flexibilität bei der Forschung mit sensiblen Mikrodaten, indem sie einen nutzungsfreundlichen Datenzugang ermöglichen. Die Folge ist eine Erhöhung der Nutzungszahlen und dadurch eine effizientere Analyse der vorliegenden Daten.

Eine genauere Betrachtung von Remote-Access-Verfahren macht daher für datenhaltende Institutionen, aber auch für Forscherinnen und Forscher durchaus Sinn. Der folgende Text richtet sich an beide Gruppen und vermittelt ein fundiertes Verständnis der Möglichkeiten des Remote Access. Dadurch soll Datenhaltern die Entscheidung erleichtert werden, ob Remote-Access-Verfahren den Zugang zu ihren spezifischen Datenbeständen verbessern können. Für die Forschung wiederum wird dargestellt, wie verschiedene Datenzugangsverfahren die Analyse von sensiblen Daten erleichtern oder sogar erst möglich machen. Hierfür werden zunächst Begrifflichkeiten und Rahmenbedingungen des Datenzugangs in den Sozial- und Wirtschaftswissenschaften im Allgemeinen geklärt und eine Einführung in die Funktionsweisen und Ausformungen des Remote Access zu sensiblen Forschungsdaten gegeben. Für eine spezifischere Betrachtung werden Remote-Access-Verfahren von fünf deutschen Forschungsdatenzentren (FDZ), aufgeteilt nach Remote-Execution- und Remote-Desktop-Verfahren, erläutert und ergänzend ein Überblick zu Remote-Access-Verfahren in anderen Ländern gegeben. Der Text schließt mit einer Zusammenfassung bezüglich der vorgestellten Remote-Access-Lösungen und einem Ausblick auf weitere Entwicklungen.

Begriffsklärungen und Rahmenbedingungen

Die folgenden Abschnitte beschreiben Begrifflichkeiten und Rahmenbedingungen, die für das Verständnis und die Einordnung von Remote-Access-Verfahren relevant sind. Dabei wird zunächst umrissen, um welche Art von Daten es sich in den Sozial- und Wirtschaftswissenschaften handelt. Die folgende Einführung in den Themenkomplex „Anonymisierungsgrade von Forschungsdaten“ bildet die Basis für die Erläuterung des Portfolioansatzes. Dieser zeigt, dass die Gewährleistung des Datenschutzes aus einem Portfolio an Sicherungsmaßnahmen besteht. Auf Basis des Portfolioansatzes können Remote-Access-Verfahren eine höhere Flexibilität beim Zugang zu sensiblen Daten ermöglichen. Eine Kurzdarstellung der Funktionsweisen und Ausformungen des Remote Access zeigt, dass Remote-Access-Verfahren die Lücke zwischen dem Datenzugang durch den Download von Forschungsdaten und der Forschung an Gastwissenschaftlerarbeitsplätzen in den Räumen der Datenhalter schließen und einen nutzerfreundlichen Zugang zu sensiblen Forschungsdaten bei der gleichzeitigen Einhaltung hoher Sicherheitsstandards ermöglichen.

Daten für die Forschung

Forschungsdaten sind Daten, die für die wissenschaftliche Nutzung aufbereitet wurden und für mehr als eine Analyse (Sekundäranalysen) zur Verfügung stehen. Dabei handelt es sich meist um Mikrodaten; d. h. Daten, die sich auf einzelne Untersuchungseinheiten (z. B. Personen, Betriebe, Haushalte) und nicht auf Aggregate beziehen. Diese können aus Befragungen und Testungen stammen und direkt für die Forschung erhoben werden oder ihm Rahmen von administrativen

Prozessen entstanden sein und danach einer wissenschaftlichen Analyse zugeführt werden. Bei der Erstellung von Forschungsdaten durchlaufen die Originaldaten (auch Roh- bzw. Primärdaten) einen Aufbereitungsprozess (Forschungsdatenmanagement (Jensen, 2012)). In den Sozial- und Wirtschaftswissenschaften werden sie meist als Datenmatrix bzw. zweidimensionale Datendateien aus Spalten und Zeilen (Merkmale und Fälle) gespeichert.

Anonymisierungsgrade von Forschungsdaten

Beinhalten Forschungsdaten personenbezogene Daten im Sinne des Bundesdatenschutzgesetzes (BDSG) § 3 (1), definiert als: „Einzelangaben über persönliche oder sachliche Verhältnisse einer bestimmten oder bestimmbaren natürlichen Person", sind besondere Vorkehrungen zu treffen. Ziel ist es, die Identität der Individuen hinter den Forschungsdaten zu schützen. Hierzu sind Techniken der Anonymisierung (Statistical Disclosure Control, SDC (Hundepool et al., 2012)) anzuwenden, die das Risiko einer De-Anonymisierung bzw. einer Re-Identifikation kontrollieren (Höhne, 2010). Daten, die auf solche Art und Weise zu schützen sind, werden auch als sensible (von: besonders viel Sorgfalt, Umsicht, Fingerspitzengefühl o. Ä. erfordernd) Daten bezeichnet. Individuen können entweder direkt (bestimmt) oder indirekt (bestimmbar) identifiziert werden.[1] Für eine direkte Identifizierung sind eindeutige Merkmale (wie Name, Adresse, Sozialversicherungsnummer) nötig. Bei einer indirekten Identifizierung wird auf Basis mehrerer Merkmale, die für sich genommen keine eindeutige Zuordnung zulassen, auf das Individuum geschlossen.[2] Im Folgenden werden die Schritte der Anonymisierung skizziert. Dabei werden auf Grundlage der Gesetzestexte des BDSG und des Bundesstatistikgesetzes (BStatG) Anonymisierungsgrade und korrespondierende Bezeichnungen für Forschungsdaten herausgearbeitet. Im Anschluss wird die Definition der Europäischen Kommission erörtert und in Einklang mit den deutschen Gesetzestexten gebracht.

Ein erster Schritt zur Einhaltung des Datenschutzes ist das Pseudonymisieren der Originaldaten. Damit ist laut BDSG § 3 (6a) das: „Ersetzen des Namens und anderer Identifikationsmerkmale durch ein Kennzeichen zu dem Zweck, die Bestimmung des Betroffenen auszuschließen oder wesentlich zu erschweren" gemeint. Das BStatG spricht hier von „formal anonymisierten Einzelangaben" (§5a (3) sowie §16 (6) 2). Anonymisierung wird im BDSG § 3 (6) als: „das Verändern personenbezogener Daten derart, dass die Einzelangaben über persönliche oder sachliche Verhältnisse nicht mehr oder nur mit einem unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft einer bestimmten oder bestimmbaren natürlichen Person zugeordnet werden können" definiert. Das BStatG verwendet für diesen Zustand der Daten den Begriff „faktisch anonymisierte Einzelangaben" (§16 (6) 1).[3] Des Weiteren ist noch die Formulierung „absolute Anonymisierung"[4] in Verwendung. Sie bezeichnet Daten, die so stark anonymisiert wurden, dass eine Identifizierung von Individuen de facto ausgeschlossen ist.

Daten, die als absolut anonym angesehen werden, tragen auch den Namen Public Use Files (PUF). Den gleichen Anonymisierungsgrad erreichen Campus Files bzw. Campus Use Files (CUF), die als

[1] Für indirekt identifizierbar hat sich der Begriff der „personenbeziehbaren Daten" etabliert. Dieser mag zwar im Alltag hilfreich sein, als Kontrast zu personenbezogenen Daten ist er jedoch nicht korrekt verwendet, da „personenbezogene Daten" laut BDSG bereits die Dimensionen „bestimmte" und „bestimmbare" enthalten. Mit anderen Worten: Laut BDSG handelt es sich um personenbezogene Daten, wenn eine direkte oder eine indirekte Identifizierung möglich ist.

[2] Z. B. kann eine genaue Berufsbezeichnung in Zusammenhang mit einer genauen Ortsangabe zu einer Identifikation eines spezifischen Individuums führen.

[3] Einzelangaben können bereits nach der Pseudonymisierung und ohne Anwendung von weiteren Anonymisierungsmethoden als absolut anonym gelten, wenn die Informationen in den Forschungsdaten keine indirekte Identifikation ermöglichen. Somit kann auch faktische Anonymität bereits nach der Pseudonymisierung erreicht sein. Nämlich wenn Einzelangaben „nur mit einem unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft einer bestimmten oder bestimmbaren natürlichen Person zugeordnet werden können" (BDSG § 3 (6)). Entscheidend ist jeweils die spezifische Möglichkeit der indirekten Zuordnung.

[4] Vergleiche: http://www.forschungsdatenzentrum.de/anonymisierung.asp

Übungsdaten für die Lehre Verwendung finden. Strukturdateien (bzw. Testdaten), die lediglich die Struktur des Datensatzes und der enthaltenen Merkmale wiedergeben und zur Vorbereitung von Analyseprogrammen dienen, sind ebenfalls absolut anonym. Diese Daten (PUF o. Ä.) sind als Informationsquelle und zum Erlernen erster Fähigkeiten hilfreich, enthalten aber in der Regel für Analysen zur Beantwortung von Forschungsfragen zu wenige oder sogar keine Detailinformationen. Daher sind in der Wissenschaft meist SUF im Einsatz. Diese verfügen über mehr Detailinformationen, sind somit näher an den Originaldaten, wurden aber so erstellt, dass sie als faktisch anonym angesehen werden können. Moderne Methoden der Datenanalyse und komplexere Datenbestände machen immer öfter den Zugang zu weniger stark anonymisierten Daten notwendig, um relevante Forschungsergebnisse zu erzielen.

Auf EU-Ebene regelt die COMMISSION REGULATION (EU) No 557/2013 den "access to confidential data for scientific purposes". Sie beschreibt SUF als "confidential data for scientific purposes to which methods of statistical disclosure control have been applied to reduce to an appropriate level and in accordance with current best practice the risk of identification of the statistical unit". Confidential data for scientific purposes meint wiederum: "data which only allow for indirect identification of the statistical units, taking the form of either secure-use files or scientific-use files"; und "statistical disclosure control" (SDC) Methoden sind hier "methods to reduce the risk of disclosing information on the statistical units, usually based on restricting the amount of, or modifying, the data released". Ein Secure Use File (SecUF) besteht demnach aus "confidential data for scientific purposes to which no further methods of statistical disclosure control have been applied".[5] Die Definition für SUF ist der deutschen sehr ähnlich. Das gleiche gilt für formal anonymisierte Daten, für die der Begriff des SecUF eingeführt wird.

An dieser Stelle erscheint es sinnvoll, die verschiedenen Begriffe einzuordnen. Da sich Datenbestände und Anonymisierungsverfahren im Detail unterscheiden, handelt es sich sicherlich nur um eine annähernde Kategorisierung, die im speziellen Fall einer genaueren Betrachtung bedarf.[6] PUF sind absolut anonymisierte Daten, die auch als CUF oder als Strukturdateien eingesetzt werden. Diese Daten können ohne Restriktionen weitergegeben (heruntergeladen) werden. Sie sind somit als Open Data anzusehen. SUF sind faktisch anonymisierte Daten, für die eine kontrollierte Weitergabe notwendig ist (abgesicherter Download). Zur Kontrolle gehören Verträge, abgesicherte Übertragung und Verarbeitung. Schließlich gibt es noch den Bereich der formal anonymisierten Daten. Diese sind pseudonymisierten Daten sehr nahe. Wobei festgehalten werden kann, dass neben der Entfernung von direkten Identifikatoren (wie z. B. Name, Anschrift und Steueridentifikationsnummer) meist noch weitere Modifikationen (Datenaufbereitungsprozess) an den Daten vorgenommen werden. Diese werden aber nicht aus Gründen des Datenschutzes durchgeführt. Die EU kennt für solche Daten die Bezeichnung SecUse. SecUF dürfen nur in dafür vorgesehenen Räumen, die unter der Kontrolle der Datenanbieter stehen, analysiert werden.[7]

Für den eigentlichen Prozess der Anonymisierung von Daten gibt es eine Vielzahl von Möglichkeiten (Müller et al., 1991); eine der gebräuchlichsten ist die Aggregation. So werden z. B. genaue Gehaltsangaben zu Gehaltsgruppen vergröbert (z. B.: 1.000-1.500 Euro). Im Grundsatz gilt: Je stärker z. B. aggregiert wird, desto stärker die Anonymisierung und desto geringer das Risiko der Re-Identifikation. Gleichzeitig sinkt jedoch auch der Informationsgehalt der Daten für die wissenschaftliche Nutzung.

[5] Die neue „Verordnung (EU) 2016/679 des Europäischen Parlaments und des Rates vom 27. April 2016 zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten, zum freien Datenverkehr und zur Aufhebung der Richtlinie 95/46/EG (Datenschutz-Grundverordnung)" und deren Auswirkungen werden hier außer Acht gelassen, da eine Einordnung zum jetzigen Zeitpunkt noch kaum möglich ist.

[6] So bergen z. B. Teilerhebungen eine geringere Gefahr der Re-Identifikation in sich als Vollerhebungen.

[7] Das IAB bietet sogenannte „schwach-anonymisierte" Daten an. Diese enthalten weniger Anonymisierungen als SUF und werden daher ausschließlich im Gastaufenthalt oder über Remote Execution bereitgestellt.

Portfolioansatz

Der Schutz von sensiblen Forschungsdaten wird nicht nur durch die Anonymisierung der Daten gewährleistet. Vielmehr ist eine Vielzahl von Sicherungsmechanismen – ein Portfolioansatz wie bei Desai et al. (2016) beschrieben - anzuwenden. Dieser Portfolioansatz setzt sich aus den sogenannten Five Safes zusammen. Diese sind: Safe People (geschulte Forscherinnen und Forscher), Safe Projects (geprüfte Projekte), Safe Settings (technisches Umfeld), Safe Outputs (kontrollierte Ergebnisse) und Safe Data (anonymisierte Daten).

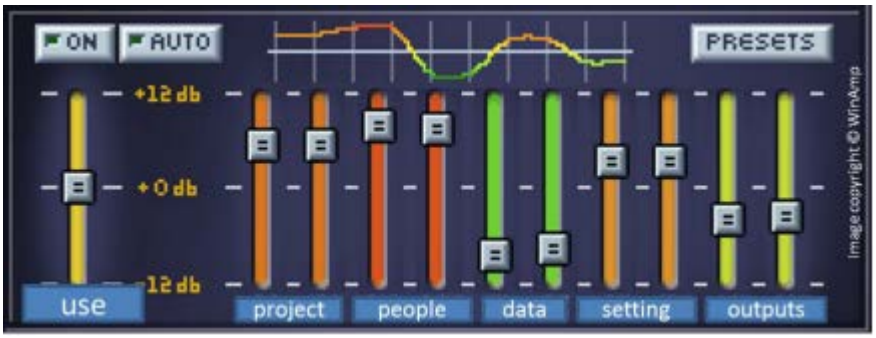

Abbildung 1 - Portfolioansatz als "equaliser" (Desai et al., 2016).

Für das hier behandelte Thema des sicheren Datenzugangs via Remote-Access-Verfahren genügt eine Betrachtung der Dimensionen Safe Setting (das Remote-Access-Verfahren selbst) und Safe Data (Forschungsdaten mit unterschiedlichem Anonymisierungsgrad).[8]

Remote-Access-Verfahren eröffnen weitere Variationen beim Zugang zu sensiblen Forschungsdaten. Im Vergleich zu SUF im kontrollierten und verschlüsselten Download bietet Remote Access einen höheren Sicherheitsstandard. Somit können SUF im Remote Access mehr Detailinformationen beinhalten (d. h. weniger stark anonymisiert sein). Faktische Anonymität ist durch die Einhaltung der Regelung, dass Einzelangaben „nur mit einem unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft […] zugeordnet werden können“ (BDSG §3 (6)) weiterhin gewährleistet. Somit ist die faktische Anonymität von Forschungsdaten stets in Bezug auf Safe Data (Anonymisierung) und Safe Setting (Datenzugangsweg) zu sehen.

Die Abschnitte zur Definition von Forschungsdaten sowie zur Erläuterung von Anonymisierungsgraden von Forschungsdaten und zur Notwendigkeit eines Portfolioansatzes bilden die Basis für die Einordnung und das Verständnis von Remote-Access-Verfahren. Diese werden nun im nächsten Abschnitt beschrieben werden.

Remote Access – Kurzbeschreibung

Remote Access kann als ein Überbegriff verstanden werden, der zunächst nichts anderes aussagt, als dass ein Zugriff, z. B. auf Server für Forschungsdaten, aus der Ferne vorgenommen wird (Schiller & Welpton, 2014). Geht es um die Analyse von sensiblen Forschungsdaten, muss es sich um einen abgesicherten Remote Access handeln. Es ist somit von Secure Remote Access zu sprechen. Dieser kann als Remote Execution oder Remote Desktop durchgeführt werden. Bei Remote Execution werden Dateien mit Syntax (Programmcodes) zur Modifizierung und Analyse der Datenbestände an die datenhaltende Organisation gesendet. Dort werden dann die Auswertungen auf den Servern mit den Forschungsdaten durchgeführt und Ergebnisdateien (Output) erzeugt. Diese Dateien werden nach einer Datenschutzprüfung, meist muss faktische oder absolute Anonymität erreicht werden, an

[8] Safe Output, die Notwendigkeit der Kontrolle von Ergebnissen, ist auch relevant, wird hier jedoch nicht weiter behandelt, da für alle Remote-Access-Verfahren gilt, dass nur datenschutztechnisch unbedenkliche Ergebnisse an die Forscherinnen und Forscher übermittelt werden.

die Forscherinnen und Forscher weitergeleitet. Die Forscherinnen und Forscher bekommen somit die eigentlichen Forschungsdaten nie zu Gesicht. Sie können Analysen nur über Anfragen und Ergebnisdateien durchführen. Ein Browsen der Forschungsdaten (z. B. um Extremwerte zu finden, die eventuell Analysen verzerren können) ist nicht möglich. Neben der Bezeichnung Remote Execution ist auch der Begriff „Job Submission“ in Verwendung. Die Begriffe werden oft als Synonyme verwendet. Hält man sich an die genaue Begriffsbedeutung, könnte jedoch durchaus eine inhaltliche Unterscheidung spezifiziert werden. Job Submission würde hierbei die einfache Übermittlung von Syntaxdateien an eine datenhaltende Organisation bedeuten. Im Kontrast dazu würde bei Remote Execution der Analyseprozess aus der Ferne auch gleich angestoßen werden. D. h. bei der ersten Variante löst die externe Forscherin bzw. der externe Forscher keine Funktion auf den Servern des Datenhalters aus; bei der zweiten Variante wird eine Funktion auf den Servern angestoßen, nämlich die Analyse der geschützten Forschungsdaten. Diese Definition wird jedoch nicht immer angewendet. Die Begriffe „Fernrechnen“ bzw. „Datenfernverarbeitung“ schließlich können im hier behandelten Zusammenhang als Synonyme für Remote-Execution-Verfahren im Allgemeinen verstanden werden.

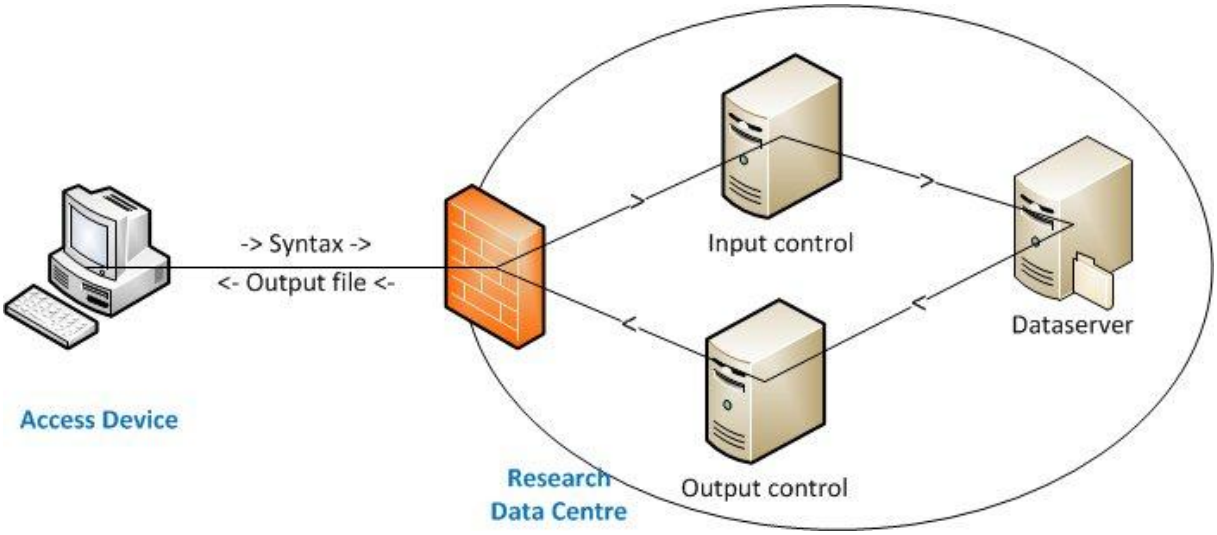

Abbildung 2 - Remote Execution (Schiller & Welpton, 2014).

Bei Remote Desktop erfolgt die Datenspeicherung und Datenverarbeitung ebenfalls auf zentralen Servern, meist innerhalb der Gebäude der Datenhalter. Über eine abgesicherte Verbindung wird das User Interface (UI), also die Benutzeroberfläche, auf einen entfernten Bildschirm übertragen. Das Access Device besteht dabei, neben dem Bildschirm, meist nur aus Eingabeinstrumenten (Tastatur, Maus) und einem Thin Client (dieser ist lediglich dafür ausgelegt, mit dem entfernten Server zu kommunizieren, nicht um selbst als Speicher oder Rechner zu agieren).[9] Schließlich bleibt noch

[9] Eine Datenübermittlung findet somit nur in einer sehr beschränkten Weise statt. Übertragen werden nicht die Forschungsdaten. Diese verbleiben auf den Servern des Datenhalters in einer sogenannten „Datenenklave“. Übertragen werden Daten in einem informatischen Sinn. D. h. es werden Datenströme zwischen dem Access Device und der Datenenklave übermittelt, die es erlauben, das UI auf dem Access Device und die Analysetools auf dem Server zu betreiben. Datenströme heißt hier nicht gleich Forschungsdaten. Dieser Unterschied in der „Übermittlung“ von Einzelangaben ist bei der Beurteilung von Remote-Desktop-Verfahren zu beachten. Remote Desktop bietet hier einen bei Weitem höheren Schutz als der sichere Download von Daten. Dies hat im Rahmen eines Portfolioansatzes Auswirkungen auf das Erreichen der faktischen Anonymität. Die im UI dargestellten Ausschnitte der Datenmatrix und visualisierten Ergebnisse können durchaus datenschutzrechtlich relevante Informationen enthalten. Daher muss abgesichert sein, dass nur berechtige Forscherinnen und Forscher (Safe People) den Bildschirm einsehen können (Teil des Safe Setting) und keine Informationen aus der Datenenklave entnommen werden können (Thema Safe Setting, wie Safe Output). Ersteres wird durch Verträge

festzuhalten, dass sich das Access Device in verschiedenen Umgebungen befinden kann. Grob bestehen die drei folgenden Möglichkeiten: (1) irgendein Ort, indem die Forscherinnen und Forscher sich z. B. von ihren eigenen Notebooks aus anmelden dürfen; (2) das Büro der Forscherin bzw. des Forschers, gewährleistet durch Verträge und technische Kontrollen (IP-Adresse etc.); (3) ein vom Datenhalter bereitgestellter Raum (Safe Room), der einer Zugangskontrolle durch geschultes Personal unterliegt. Der Grad der Sicherheit des Safe Settings nimmt dabei von (1) zu (3) zu, da die Kontrollmöglichkeiten des Datenanbieters größer werden.

Im Detail sollte bei der Betrachtung von Remote-Access-Verfahren stets geprüft werden, ob es sich um Remote Desktop (mit oder ohne Safe Room) oder um Remote Execution handelt.

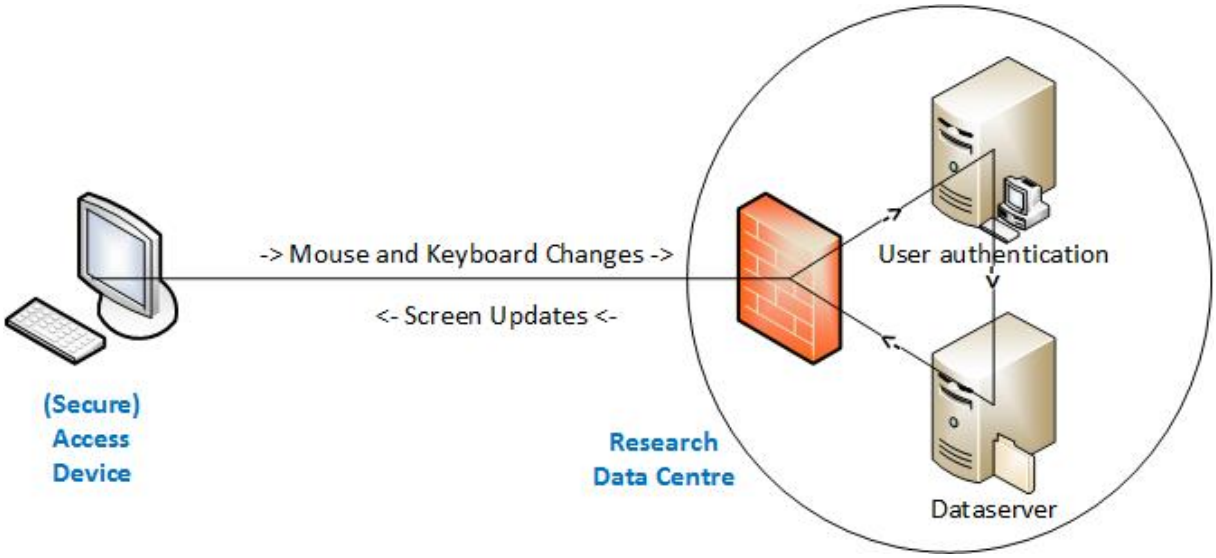

Abbildung 3 - Remote Desktop (Schiller & Welpton, 2014).

In Bezug auf die gesetzlichen Regelungen zum Arbeiten mit sensiblen Forschungsdaten ist die Unterscheidung nach faktisch und formal anonymen Daten wichtig. Das BStatG stellt in §16 (6) fest, dass für „die Durchführung wissenschaftlicher Vorhaben“ faktisch anonymisierte Einzelangaben übermittelt und formal anonymisierte Einzelangaben „innerhalb speziell abgesicherter Bereiche des Statistischen Bundesamtes und der statistischen Ämter der Länder […], wenn wirksame Vorkehrungen zur Wahrung der Geheimhaltung getroffen werden,“ (BStatG §16 (6) 1 und 2) analysiert werden können. Der Unterschied liegt somit in der Übermittlung bei faktisch anonymen und der Nutzung in gesicherten Räumen des Datenhalters bei formal anonymen Daten. Festzuhalten ist hier noch, dass das BStatG für den Bereich der Statistik und dessen spezifische Anforderungen und Datenangebote gilt. Für Befragungsdaten beispielsweise können andere Definitionen oder Einordnungen sinnvoll sein.

Remote-Access-Verfahren verbessern somit den Zugang zu Forschungsdaten auf drei Ebenen:

(1) Durch eine höhere Informationsdichte bei SUF bei der Nutzung von Remote Desktop im Rahmen eines Portfolioansatzes,

(2) durch die Schaffung von Zugangspunkten zu SecUF bei der Nutzung von Remote Desktop unter Verwendung eines Safe Room und

---

und die Aufsichtspflicht der Forscherinnen und Forscher gewährleistet bzw. von geschultem Personal abgesichert. Letzteres durch technische Maßnahmen, die eine Entnahme von Daten verhindern und durch eine Überprüfung von Ergebnis- und Ausgabedateien auf deren Anonymisierungsniveau.

(3) durch die Ermöglichung der Forschung mit SecUF über Remote Execution ohne Restriktionen beim Zugangspunkt – jedoch mit der Einschränkung, dass die Forschungsdaten nicht direkt eingesehen werden können.

Die folgende Tabelle zeigt die möglichen Kombinationen zwischen Safe Data / Anonymisierungsgrad und Safe Setting / Zugangsweg.

	Safe Data / Anonymisierungsgrad		**Safe Setting / Zugangsweg**
1	Public Use File (PUF)	Absolut anonym	Download/CD/DVD
2	Scientific Use File (SUF)	Faktisch anonym	Gesicherter Download/CD/DVD
3	Scientific Use File (SUF)	Faktisch anonym	Remote Access (RA) als Remote Desktop
4	Secure Use File (SecUF)	Formal anonym	RA als Remote Desktop mit Safe Room
5	Secure Use File (SecUF)	Formal anonym	RA als Remote Execution
6	Secure Use File (SecUF)	Formal anonym	Gastaufenthalt

Zusammenfassend existieren sechs Kombinationen aus Safe Data und Safe Setting. (1) PUF können über freien Download weitergegeben werden, da sie als Open Data anzusehen sind. (2) SUF können faktisch anonym über abgesicherten und vertraglich festgelegten Download und (3) im Rahmen von sicheren Remote-Desktop-Zugängen analysiert werden. (6) SecUF wiederum können aufgrund ihres hohen Informationsgehalts lediglich im Gastaufenthalt oder (5) über Remote Execution bearbeitet werden. (4) Eine Weiterentwicklung stellt hier die Verwendung eines Safe Rooms unter Kontrolle des Datenanbieters im FDZ-im-FDZ Ansatz dar (Bender & Heining, 2011). Hierdurch wird ein Remote-Desktop-Zugriff auf SecUF ermöglicht.

Zur besseren Veranschaulichung werden in den folgenden beiden Abschnitten Remote-Access-Verfahren deutscher FDZ dargestellt. Dadurch kann gezeigt werden, wie Portfolioansätze im Alltagsbetrieb funktionieren. Es handelt sich um die FDZ der Deutschen Rentenversicherung (DRV), des Deutschen Zentrums für Hochschul- und Wissenschaftsforschung (DZHW), der Bundesagentur für Arbeit (BA) im Institut für Arbeitsmarkt- und Berufsforschung (IAB), des Leibniz-Instituts für Bildungsverläufe e.V. (LIfBi) und des Sozio-Oekonomischen Panels (SOEP). Zunächst werden Remote-Execution-Lösungen und daraufhin Remote-Desktop-Lösungen vorgestellt. Dabei werden für eine leichtere Vergleichbarkeit jeweils die folgenden fünf Kurzbeschreibungen geliefert: „Organisation und Motivation für die Einrichtung des Verfahrens“, „Beschreibung des Verfahrens“, „Vorteile und Nachteile des Verfahrens“, „Rechtliches und Datensicherheit“ und „Aufwand für den Betrieb“. Da das IAB und das SOEP sowohl Remote-Execution- als auch Remote-Desktop-Verfahren anbieten, kommen die beiden FDZ auch in beiden Abschnitten vor.

Remote-Execution-Verfahren

Mit der DRV, dem IAB und dem SOEP bieten drei der oben genannten FDZ ihre sensiblen Forschungsdaten über Remote-Execution-Verfahren an.

Deutsche Rentenversicherung (DRV)[10]

Kurzbeschreibung der Organisation und Motivation für die Einrichtung des Verfahrens

Das Forschungsdatenzentrum der Rentenversicherung (FDZ-RV) stellt der hauptamtlich wissenschaftlichen und sonstigen Forschung anonymisierte Daten aus dem Bestand der prozessproduzierten Statistikdaten der deutschen Rentenversicherung aus den Bereichen Versicherung, Renten und Rehabilitation zur Verfügung (Stegmann, 2008). Das FDZ-RV bietet Daten grundsätzlich als SUF an, die in den Institutionen der Wissenschaftlerinnen und Wissenschaftler

[10] Verantwortlich: Michael Stegmann, Frank Röder und Tatjana Mika.

genutzt werden können. Die SUF sollen eine Stichprobengröße und Merkmalstiefe bieten, die für die meisten empirischen Fragestellungen ausreicht. Lediglich 3% der genutzten Datensatzzugänge werden im Fernrechenverfahren bereitgestellt. Zudem wird angeboten, im Gastaufenthalt einen umfassenderen und/oder detaillierteren Datensatz zu nutzen. Interessierten Wissenschaftlerinnen und Wissenschaftlern bietet sich außerdem die Möglichkeit, zumeist im Anschluss an einen Gastwissenschaftleraufenthalt in Berlin oder Würzburg, ergänzende Auswertungen im Rahmen des online-gestützten kontrollierten Fernrechnens durchzuführen. Hierbei werden die Daten wie bei einem Gastwissenschaftleraufenthalt nicht nach außen gegeben, sondern lediglich die geprüften Auswertungsergebnisse (z. B. in Form von Auswertungstabellen oder Koeffizienten) zur Verfügung gestellt. Wurde im Gastaufenthalt ein Datensatz bearbeitet, so kann an diesem im Fernrechenverfahren die Analyse fortgesetzt werden.

Beschreibung des Verfahrens

Das Datenmaterial wird in ein persönliches Verzeichnis gespeichert, welches dann im Rahmen des Fernrechenverfahrens von außen zugänglich gemacht wird. Dazu müssen sich die Nutzerinnen und Nutzer mit einer vom FDZ-RV auf Antrag zur Verfügung gestellten Zugangskennung und einem Kennwort authentifizieren und können dann den personalisierten Login-Bereich nutzen. Derzeit können die Analyseprogramme SPSS in der Version 22 und Stata in der Version 13 genutzt werden (Stand Januar 2016).

Folgende Regeln werden hierbei beachtet:

- Kontrollierter Zugriff mit Analyseprogrammen auf anonymisierte Statistikdaten. Die Analysen dürfen nicht dazu führen, dass Informationen über eine konkrete Einzelperson offenbart werden (Vertraulichkeit).
- Die Datenbestände müssen außerdem vor Manipulation geschützt werden (Integrität).
- Es muss sichergestellt werden, dass der Forschungsserver nicht mit Fernrechenanfragen überlastet wird (Verfügbarkeit).
- Die Sicherheit der persönlichen Anmeldedaten der Nutzerinnen und Nutzer muss gewährleistet werden (Vertraulichkeit).
- Verhinderung des Einschleusens von Schad-Code durch Fremdnutzung.

a.) Datenzugriff und Datenhaltung

Die Daten des Fernrechenprojekts werden in ein separates Verzeichnis auf dem File-Server gelegt. Hier können die Nutzerinnen und Nutzer indirekt zugreifen. Dazu wird ein projektbezogenes Arbeitsverzeichnis zur Verfügung gestellt, der erzeugte Output (i.d.R. Tabellen, Statistische Kennzahlen etc.) wird in die Datenbank geschrieben und steht dort den FDZ-Mitarbeiterinnen und Mitarbeitern zur Prüfung bereit.

b.) Verfahrensablauf

Damit potentielle Datennutzerinnen und Datennutzer Auswertungen über die spezifische Internet-Plattform „online-gestütztes kontrolliertes Fernrechnen“ durchführen können, müssen diese einen Nutzungsantrag über das bereits etablierte Antragswesen im FDZ-RV-Webauftritt stellen (Abb. 4: Serviceangebot und Antragswesen).

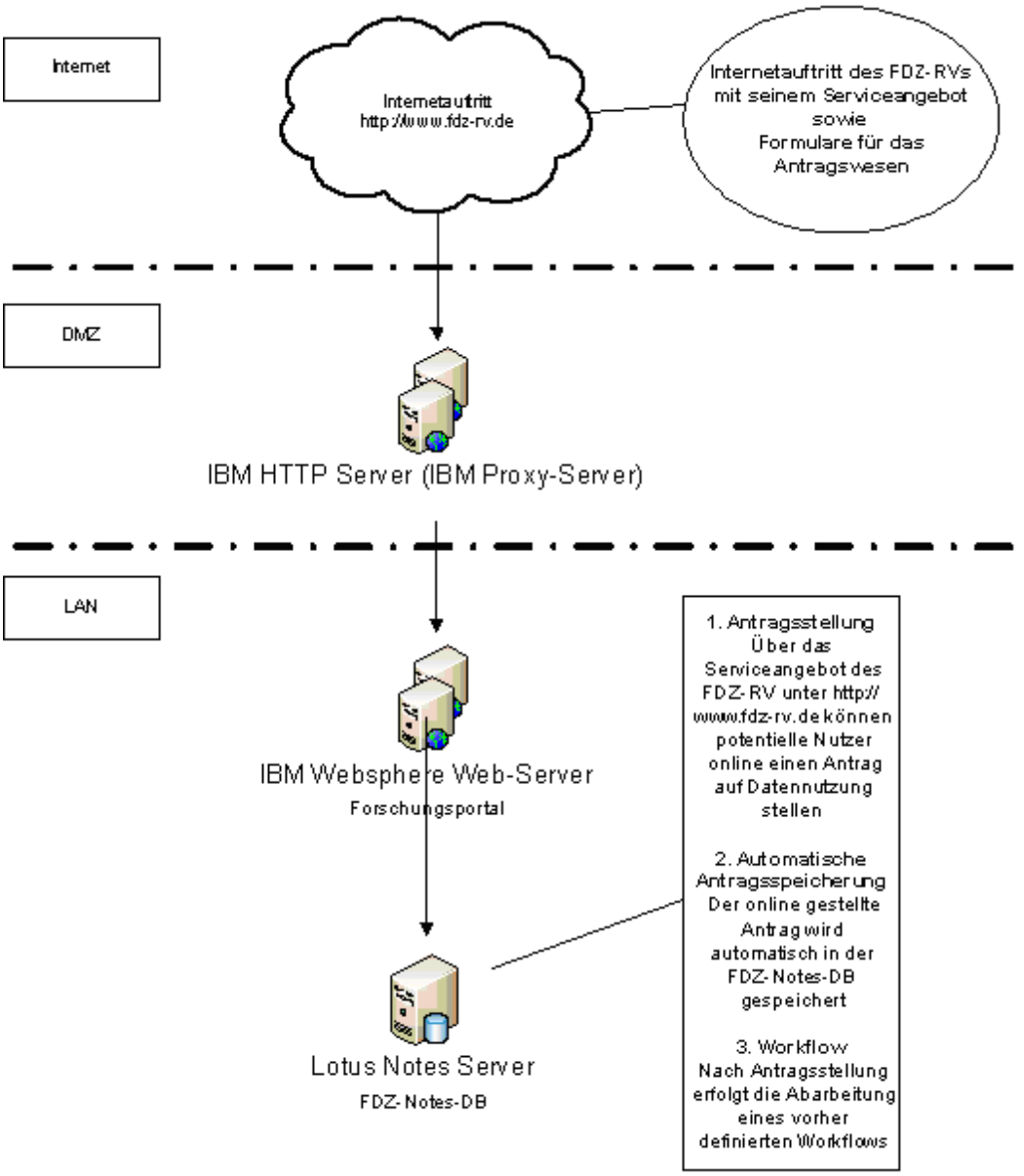

Abbildung 4 - Serviceangebot und Antragswesen unter http://www.fdz-rv.de (Quelle: FDZ-RV).

Die Forscherinnen und Forscher erhalten eine persönliche Kennung, mit welcher der Zugang über die Website des FDZ-RV auf die Fernrechenplattform möglich ist. Zur Datenauswertung setzen sie dort SPSS- oder Stata-Syntaxen ab, die auf den für sie zugeschnittenen Datenbestand zugreifen, der in einem abgeschotteten Bereich des File-Servers abgelegt wird. Die Abarbeitung der Syntaxdateien erfolgt in einer Batch-Verarbeitung. Die Forscherinnen und Forscher können jedoch über das Portal den Status des Jobs kontrollieren. Sobald die Ergebnisdateien durch das FDZ-RV freigegeben sind, können die Wissenschaftlerinnen und Wissenschaftler diese Dateien in ihrem persönlichen Login-Bereich abholen.

c.) EDV-technische Umsetzung

Der Webauftritt für das Fernrechnen ist über einen Link auf der FDZ-RV-Homepage sowie unter http://www.fernrechnen.de erreichbar. Nach der erfolgreichen Antragstellung werden individuelle Zugangsdaten für die Datennutzerin bzw. den Datennutzer vergeben. Die Zugangsdaten sind nur für den beantragten und genehmigten Zeitraum gültig und laufen automatisch ab. Das Verfahren

„online-gestütztes kontrolliertes Fernrechnen“ kann erst nach erfolgreicher Authentifizierung erreicht werden.

d.) Ablauf einer Auswertung

Nach Eingabe der Zugangsdaten im Login-Bereich werden die Nutzerinnen und Nutzer automatisch auf das Fernrechnenportal weitergeleitet. Aktuell ist der folgende Funktionsumfang nutzbar:

- Start-Übersichtsseite (Anzeige von: Ablaufdatum der Kennung, maximale Jobs etc.)
- Anzeige der Pfade zu den freigegebenen Datenbeständen und zum persönlichen Ordner, um temporär Dateien anlegen zu können.
- Formularmaske zur Eingabe der abzuschickenden Syntax im SPSS- oder Stata-Format. Hier kann auch das Format der Ausgabe-Log-Datei gewählt werden (Text oder HTML).
- Job-Übersichtsseite: Anzeige der einzelnen Statusmeldungen für alle angelegten User-Jobs. Weiterhin können hier die Jobs „verwaltet“ werden. Es können hier jedoch keine Jobs freigegeben bzw. gesperrt werden.

Nach dem Anlegen einer Auswertungssyntax wird diese noch nicht sofort zur Verarbeitung abgeschickt. Somit erhalten die Nutzerinnen und Nutzer einerseits die Möglichkeit, mehrere Auswertungsjobs anlegen zu können. Soll eine Auswertung durchgeführt werden, muss diese durch die Nutzerinnen und Nutzer explizit abgeschickt werden. Da die Verarbeitung asynchron läuft, erhalten die Nutzerinnen und Nutzer in der Job-Übersichtsseite Statusmeldungen der aktuellen Verarbeitung der einzelnen Jobs. Auch das vorzeitige Abbrechen und Löschen eines bereits abgeschickten Jobs ist durch die Nutzerinnen und Nutzer möglich.

Die Ergebnisse dieser Auswertungen liegen nach erfolgreicher Verarbeitung als Log-Dateien im Text- oder HTML-Format vor. Fehlermeldungen werden unmittelbar angezeigt, so dass die Syntaxen korrigiert und wieder abgeschickt werden können. Solche Meldungen werden nicht von den Mitarbeiterinnen und Mitarbeitern des FDZ geprüft. Enthalten die Log-Dateien hingegen Ergebnisse der Analyse, dann werden sie durch Mitarbeiterinnen und Mitarbeiter des FDZ-RV geprüft und anschließend freigeschaltet, wenn sie den oben genannten Kriterien entsprechen, also keine Deanonymisierung von Personen zulassen. Die freigegebenen Ergebnisse können die Nutzerinnen und Nutzer dann abrufen und lokal auf dem eigenen PC speichern. Auf nicht freigegebene Log-Dateien erhalten die Nutzerinnen und Nutzer keinen Zugriff. Über den aktuellen Status der Jobs (freigegeben, nicht freigegeben, Job läuft derzeit, es ist ein Fehler aufgetreten etc.) können sich die Datennutzerinnen und -nutzer auf einer Übersichtsseite informieren. Darüber hinaus ist noch eine E-Mail-Funktionalität implementiert, so dass die Datennutzerinnen und -nutzer automatisch per E-Mail über die Status ihrer Jobs informiert werden.

Vorteile und Nachteile des Verfahrens

Anders als bei der Datenweitergabe erhalten die Forscherinnen und Forscher keine Mikrodaten auf Datenträgern, sondern die Möglichkeit, Auswertungsjobs zu erstellen, die via Fernrechenplattform über einen vertraglich vereinbarten Datenbestand geschickt werden können. Als Ergebnis dieser Verarbeitung erhalten die Datennutzerinnen und -nutzer durch SPSS bzw. Stata generierte Log-Dateien. Die Nutzerinnen und Nutzer erhalten keinen direkten Zugriff auf die für sie bereitgestellten anonymisierten Daten und können die Daten nicht betrachten.

Damit die Datennutzerinnen und -nutzer nicht unbegrenzt Jobs abschicken können und den SPSS-Server völlig blockieren, darf nur eine bestimmte Anzahl an Verarbeitungen pro Tag abgeschickt werden. Diese Anzahl ist für jede Nutzerin und jeden Nutzer frei wählbar und wird von den Mitarbeiterinnen und Mitarbeitern des FDZ-RV vergeben. Ist die maximale Anzahl überschritten, können an diesem Tag keine weiteren Verarbeitungen mehr gestartet werden. Dies kann die Zeit für die Analysen in die Länge ziehen.

Rechtliches und Datensicherheit

Das Fernrechenverfahren steht nicht jedem offen, sondern ist wie das gesamte FDZ-RV Angebot auf wissenschaftliche Forschung beschränkt. Die Nutzung setzt voraus, dass im Vorfeld ein SUF genutzt wird und dazu ein Vertrag besteht. Die Nutzerinnen und Nutzer sollen damit bereits über eine gewisse Datenkompetenz verfügen. Bei den extern bereitgestellten Daten handelt es sich nicht um personenbezogene, sondern um faktisch anonymisierte Statistikdaten.

Aufwand für den Betrieb

Nach Beendigung des Auswertungslaufs müssen die Ergebnisdateien (Syntax und erzeugte Tabellen) von einer Mitarbeiterin oder einem Mitarbeiter des FDZ-RV dahingehend geprüft werden, ob die Vertraulichkeit von Einzelinformationen gewahrt ist. Nur wenn keine datenschutzrechtlich bedenklichen Inhalte zu erkennen sind, werden die Ergebnisdateien vom FDZ-RV freigegeben. Dies bindet Personal. Außerdem kommt es zu Rückfragen, wenn Überlastungen auf dem Server vorliegen oder sonstige technische Schwierigkeiten auftreten.

Institut für Arbeitsmarkt- und Berufsforschung (IAB)[11]

Kurzbeschreibung der Organisation und Motivation für die Einrichtung des Verfahrens

Das Forschungsdatenzentrum (FDZ) der Bundesagentur für Arbeit (BA) im Institut für Arbeitsmarkt- und Berufsforschung (IAB) ermöglicht externen Forschenden Zugang zu Mikrodaten für die nicht-kommerzielle Forschung im Bereich der Sozialversicherung und der Arbeitsmarkt- und Berufsforschung. Zum Datenangebot gehören zum einen administrative Daten, die aus den Arbeitgebermeldungen zur Sozialversicherung und aus den Geschäftsprozessen der BA sowie kommunaler Träger gewonnen werden. Dabei handelt es sich um Angaben, die zum Zwecke der Aufgabenerfüllung staatlicher Träger verpflichtend erhoben werden und für die der Sozialdatenschutz nach § 75 SGB X gilt. Zum anderen werden Befragungsdaten und kombinierte administrative Daten und Befragungsdaten angeboten. Befragungsdaten des IAB werden den administrativen Daten datenschutzrechtlich gleichgestellt.

Zugang zu schwach-anonymisierten Datensätzen ist aus rechtlichen Gründen nur per Gastaufenthalt in geschützten Räumlichkeiten des FDZ oder den externen Standorten sowie per Datenfernverarbeitung möglich.[12] Bei der Datenfernverarbeitung bereiten Nutzerinnen und Nutzer ihre Programme mit Testdaten vor und übermitteln diese ans FDZ. Dort werden sie ausgeführt und die Ergebnisse nach einer Datenschutzprüfung rückübermittelt. Bei der Datenfernverarbeitung greifen Forschende nicht direkt auf die schwach-anonymisierten Daten zu. Eine Beschreibung der Datenschutzprüfungen am FDZ findet sich bei Hochfellner et al. (2012).

Die Datenfernverarbeitung am FDZ wurde vor 2015 größtenteils manuell durchgeführt und die geprüften Ergebnisse per E-Mail an die Wissenschaftlerinnen und Wissenschaftler versandt. In den letzten Jahren hat die Anzahl an Projekten und Forschenden kontinuierlich zugenommen, wodurch der Aufwand für die technische Ausführung der Aufträge und für die Datenschutzprüfung deutlich gestiegen ist. In 2014 wurden beispielsweise etwa 1.800 Fernverarbeitungsaufträge bearbeitet.

Um den steigenden Nutzungszahlen gerecht zu werden, wurde im April 2015 die Anwendung JoSuA (Job Submission Application) am FDZ eingeführt. Die Anwendung wurde vom Institut zur Zukunft der Arbeit (IZA) entwickelt. Das IZA unterstützt im Rahmen eines Pflege- und Wartungsvertrags laufend die technische Stabilität der Software. Um den spezifischen rechtlichen und organisatorischen

[11] Verantwortlich: Johanna Eberle und Dana Müller.

[12] Für schwach-anonymisierte Daten gilt der Sozialdatenschutz nach § 75 SGB X. Unter bestimmten Voraussetzungen ist nach § 75 SGB X auch eine Übermittlung von Sozialdaten zulässig. Am FDZ wurde auf Basis des § 75 SGB X jedoch ein standardisiertes Verfahren etabliert, das externen Wissenschaftlerinnen und Wissenschaftlern Zugang zu den schwach-anonymisierten Daten des IAB per Gastaufenthalt ermöglicht.

Erfordernissen am FDZ gerecht zu werden, wurden einige Bestandteile von JoSuA modifiziert bzw. um weitere Funktionen erweitert. Durch die Software JoSuA läuft die Datenfernverarbeitung am FDZ nahezu komplett automatisiert ab. Es verbleibt die manuelle Datenschutzprüfung von Seiten des FDZ.

Der Datenzugang am FDZ erfolgt standardisiert und hängt vom Anonymitätsgrad der Datensätze ab. Bei Gastaufenthalten und bei der Datenfernverarbeitung mit JoSuA führen Forschende Auswertungen auf Basis der schwach-anonymisierten Daten des IAB durch. Diese Daten haben den geringsten Grad an Anonymität und gelten daher als besonders schützenswert. Neben schwach-anonymisierten Daten, die im Gastaufenthalt und per Datenfernverarbeitung mit JoSuA genutzt werden können, bietet das FDZ einige SUFs an. Diese können nach Abschluss eines Datennutzungsvertrags gemäß § 282 Abs. 7 SGB III an wissenschaftliche Einrichtungen übermittelt werden. Eine weitere Form des Datenzugangs sind absolut anonymisierte CUF, die zum Zwecke der wissenschaftlichen Ausbildung von registrierten Nutzerinnen und Nutzern von der Homepage des FDZ heruntergeladen werden können.

Beschreibung des Verfahrens

Aus Sicht der Forschenden wird die Datenfernverarbeitung vollständig über eine Webschnittstelle (https://josua.iab.de) gesteuert. Die Nutzerinnen und Nutzer loggen sich über ihren Webbrowser auf der Seite ein und laden ihre vorbereiteten Programme in ihrem Projektzugang hoch. Die Programme werden dann automatisch an die FDZ-Rechenserver weitergeleitet. Diese Rechenserver befinden sich in einem sicheren Netzwerk, das bis auf wenige Schnittstellen abgeschottet ist. Dort werden die Programme mit Stata ausgeführt. Nach einer Datenschutzprüfung werden die Ergebnisdateien über die Webschnittstelle zugänglich gemacht.

Die Mitarbeiterinnen und Mitarbeiter des FDZ nutzen eine administrative Schnittstelle, die aus dem lokalen Netzwerk über einen Webbrowser aufgerufen wird. Mit dieser werden verschiedene administrative Tätigkeiten durchgeführt wie die Kontrolle laufender Aufträge, das Starten und Beenden von Prozessen sowie die Nutzer- und Projektverwaltung. Auch die manuelle Datenschutzprüfung erfolgt über diese Schnittstelle.

Eine wichtige Neuerung, die für JoSuA konzipiert wurde, ist die Unterscheidung zwischen internen und externen Ergebnissen. Ausgehend von der Beobachtung, dass ein Großteil der erzeugten Ergebnisse der Datenaufbereitung, Auszählung und Vorbereitung auf die relevanten Analysen dient und nicht veröffentlicht wird, können Forschende beim Hochladen ihrer Programme zwischen zwei Modi wählen. Im Modus „Presentation / Publication“ werden Ergebnisdateien manuell auf Einhaltung des Datenschutzes geprüft und zum Download verfügbar gemacht. Diese Ergebnisse können für Vorträge und Publikationen genutzt werden. Im Modus „Internal Use“ können Ergebnisse hingegen nicht heruntergeladen, sondern nur als Vorschaubilder innerhalb der Webschnittstelle von JoSuA betrachtet werden. Die Ergebnisse dürfen nur von Projektmitarbeiterinnen und -mitarbeitern eingesehen werden, die einen JoSuA-Zugang besitzen und dienen allein der Entwicklung der Auswertungsprogramme. Ein Wasserzeichen in den dargestellten Ergebnissen weist die Forschenden darauf hin, dass ihr Datennutzungsvertrag es ihnen verbietet, aus den Bildern Ergebnisse abzuschreiben oder Screenshots zu erstellen. Weil keine Ergebnisübertragung stattfindet, erfolgt eine skriptbasierte Datenschutzprüfung und die Ergebnisse stehen unmittelbar nach dem Programmdurchlauf zur Einsicht bereit. Sobald ein Job im Modus „Presentation / Publication“ hochgeladen wird, sind die vorherigen Ergebnisse im Modus „Internal Use“ aus Datenschutzgründen nicht mehr einsehbar.

Vorteile und Nachteile des Verfahrens

Die Einführung der Anwendung JoSuA war für die Forschenden und für das FDZ selbst mit einigen Änderungen bestehender Arbeitsabläufe verbunden. Es wurden neue Verträge erstellt, die Projekte und Forschenden in die Nutzerverwaltung von JoSuA überführt und Arbeitshilfen für die Nutzerinnen und Nutzer aufbereitet. Die Forschenden mussten alle notwendigen Unterschriften für die Verträge einsammeln und wenige Veränderungen in ihren Programmen vornehmen. Dennoch kann die

Einführung von JoSuA als Erfolg gewertet werden, da die Vorteile des neuen Systems die Anpassungskosten bei Weitem überwiegen. Einer der zentralen Vorteile von JoSuA ist die schnellere Verfügbarkeit der Ergebnisse. Ergebnisse des Modus „Internal Use“ können unmittelbar nach Programmdurchlauf eingesehen werden. Damit wird der Forschungsprozess erheblich beschleunigt. Darüber hinaus sind die Programme und Ergebnisdateien für die Forscherinnen und Forscher archiviert und jederzeit einsehbar.

Ein weiterer Vorteil besteht darin, dass nur noch diejenigen Ergebnisse nach außen übermittelt werden, die tatsächlich in einen Vortrag oder eine Veröffentlichung münden. Indem die Menge des transferierten Outputs sinkt, wird auch die Gefahr der Offenlegung von schützenswerten Daten minimiert. Angesichts steigender Nutzungszahlen ist eine Fokussierung auf den tatsächlich publizierten Output essentiell, um den Prüfaufwand für die manuelle Datenschutzprüfung in einem Rahmen zu halten, der gründlich und zeitnah bearbeitet werden kann. Mittlerweile werden 85 % aller Aufträge im Modus „Internal Use“ eingeschickt.

Durch die Effizienzsteigerung hat sich nach Einführung von JoSuA die Anzahl der durchgeführten Fernverarbeitungsaufträge pro Monat in etwa vervierfacht. Gleichzeitig hat sich der Prüfaufwand für die manuelle Datenschutzprüfung der Aufträge im Modus „Publication / Presentation“ gegenüber dem Prüfaufwand vor Einführung von JoSuA um etwa 25 % reduziert.

Rechtliches und Datensicherheit

Ein direkter Zugriff auf die schwach-anonymisierten Daten des IAB ist aus rechtlichen Gründen nur in den geschützten Räumlichkeiten des FDZ oder an den externen Standorten möglich. Durch die Datenfernverarbeitung mit JoSuA können externe Forschende jedoch auch außerhalb des FDZ Auswertungen mit Daten des IAB durchführen. Dabei erfolgt kein direkter Datenzugriff, sondern es werden nur anonyme Ergebnisse übermittelt. Durch den Modus „Internal Use“ ist die zeitliche Verzögerung bei der Ergebnisbereitstellung im Gegensatz zu einem direkten Datenzugriff sehr gering. Durch eine automatisierte Löschung von Ergebnissen, die auf einer zu geringen Fallzahl beruhen, können die dargestellten Ergebnisse als faktisch anonym betrachtet werden.

Die sensiblen Datensätze des IAB befinden sich auf Servern in einem abgeschotteten Netzwerk des FDZ. Zugriff haben nur Forschende mit gültigem Datennutzungsvertrag innerhalb zugewiesener Projektordner. Zur Einrichtung der Anwendung JoSuA wurde eine Schnittstelle zu diesem Netzwerk geschaffen. Eine geeignete technische Umgebung sorgt hierbei dafür, dass nur legitime Anforderungen von außen weitergeleitet werden. Der Zugriff auf die JoSuA-Plattform erfolgt kennwortgeschützt. Jede Nutzerin und jeder Nutzer erhält pro beantragtem Datensatz einen personalisierten Account, der für die Vertragslaufzeit die Möglichkeit zur Datenfernverarbeitung bietet.

Aufwand für Implementierung und Betrieb

Im Vorfeld der Einführung von JoSuA wurden verschiedene Abteilungen im IAB und der BA in die Planung einbezogen und Abstimmungs- und Genehmigungsprozesse durchlaufen. Das FDZ hat in Kooperation mit dem IZA einige Modifikationen von JoSuA festgelegt und den Einsatz der Anwendung am FDZ umfangreich getestet. Der Personalaufwand wurde aus Eigenmitteln finanziert. Während der Einführungsphase der Anwendung war der zeitliche Aufwand für die Mitarbeiterinnen und Mitarbeiter des FDZ wegen der Häufung von Nutzeranfragen und einigen notwendigen organisatorischen Anpassungen erhöht.

Für den Einkauf der Anwendung JoSuA fielen einmalige Kosten an, daneben gibt es monatliche Kosten im Rahmen eines Pflege- und Wartungsvertrags. Weitere Lizenzen mussten nicht angeschafft werden, da JoSuA auf quelloffenen Paketen beruht. Da größtenteils die bestehende Infrastruktur genutzt wird, waren bis auf einen zusätzlichen Webserver keine größeren Hardwarebeschaffungen nötig. Wichtig ist eine ausreichende Speicherkapazität des Servers, um die Vielzahl an Ergebnisdateien zu speichern, sowie die Bereitstellung ausreichenden Arbeitsspeichers für die Programmläufe der Nutzerinnen und Nutzer.

Sozio-Oekonomisches Panel (SOEP)[13]

Kurzbeschreibung der Organisation und Motivation für die Einrichtung des Verfahrens

Die Daten des SOEP werden der Wissenschaft und Forschung im Rahmen eines Vertragsverhältnisses zweckgebunden zur Nutzung bereitgestellt. Dabei ist der Datenzugang in einer hierarchischen Abstufung entlang des Potentials zur Deanonymisierung gestaltet. Im Standard SUF werden daher neben allen direkten Identifikatoren ebenfalls alle Klartexte, sowie alle Regionalzuordnungen unterhalb der Bundesländer entfernt. Raumordnungsregionen sind die einzigen zusätzlichen Daten, die bei Vorlage eines erweiterten Datenschutzkonzeptes noch direkt am Arbeitsplatz der einzelnen Forscherin bzw. des einzelnen Forschers genutzt werden können. Ist eine Nutzung der Kreise, Gemeinden oder Postleitzahlen für die wissenschaftliche Fragestellung notwendig, muss die Forscherin bzw. der Forscher einen Aufenthalt am FDZ einplanen. Für die zusätzliche Information der Kreiszugehörigkeit der Befragten bietet das FDZ SOEP den Forscherinnen und Forschern seit 2006 mit der Einführung von SOEPremote (Remote Execution) einen Fernzugang an. Es kann mit den Daten gearbeitet werden ohne zu einem Gastaufenthalt gezwungen zu sein, aber unter Einhaltung zusätzlicher Kontrollmöglichkeiten der spezifischen Nutzung der Daten.

Beschreibung des Verfahrens

Für SOEPremote wird seit Beginn das bereits seit 1990 von der Luxembourg Income Study entwickelte Programm LISSY (Coder & Cigrang, 2003) genutzt. Dabei wird entweder per E-Mail oder per Webclient die Syntax zur Auswertung für ein spezifisches Set an Daten an einen internen Server geschickt (derzeit unterstützt LISSY die Syntax von Stata, SPSS, SAS und R). Der die Anfrage von außen annehmende Server (die sogenannte Post-Office) prüft zum einen, ob die formalen Bedingungen zutreffen, also ob diese Nutzerin bzw. dieser Nutzer registriert ist, stimmt das Passwort und auf welche Daten darf sie bzw. er zugreifen. Danach erfolgt eine inhaltliche Prüfung, die je nach Einstellungen für jede Nutzerin und jeden Nutzer individuell ausgelegt sein kann. Die restriktivste Einstellung ist eine manuelle Prüfung jeglichen Syntaxinputs und Ergebnisoutputs. Es können aber auch Regeln definiert werden, wie die eingehende Syntax oder der an die Nutzerinnen und Nutzer geschickte Output geprüft werden soll, also z. B. nach dem Vorliegen von nicht erlaubten Befehlen oder nach der Länge des Outputs. Wird nach der Prüfung ein Output versendet, erfolgt dies immer auch parallel zur Webschnittstelle an die registrierte E-Mail Adresse.

Ist die Prüfung der eingehenden Syntax erfolgreich (entweder per manueller oder per automatischer Prüfung), wird die auszuführende Syntax in einer internen Datenbank abgelegt. Die im Intranet stehenden und speziell abgeschotteten Recheninstanzen fragen regelmäßig diese interne Datenbank ab, um noch nicht abgearbeitete Jobs abzuarbeiten. Wurde eine Syntax abgearbeitet, wird der entsprechende entstandene Ergebnisfile ebenfalls in die interne Datenbank geschrieben. Der Eingangsserver kann nun den vorliegenden Ergebnisfile wieder prüfen (ebenfalls je nach Einstellung manuell oder nach definierten automatischen Regeln) und sendet nach erfolgreicher Prüfung den Output an die jeweilige Nutzerin bzw. an den jeweiligen Nutzer.

Damit werden jegliche Anfragen und generierte Outputs für jede Nutzerin und jeden Nutzer auf individueller Ebene nachvollziehbar gespeichert, egal ob die jeweilige Anfrage erfolgreich abgearbeitet oder versendet wurde. Die Dauer für den zusätzlichen Rechenaufwand von Empfang, Prüfung, Verteilung und Versendung der Jobs ist dabei zu vernachlässigen und würde theoretisch, bei entsprechender Ressourcenausstattung auf der Ebene der Recheninstanzen, eine dem nicht-interaktiven lokalen Rechnen ähnliche Arbeitsweise ermöglichen.

[13] Verantwortlich: Jan Goebel.

Vorteile und Nachteile des Verfahrens

Die Vorteile des Verfahrens sind, dass jegliche Syntax und alle übermittelten Ergebnisse protokolliert und nachvollziehbar sind. Anders als bei der üblichen Datenweitergabe erhalten die Forscherinnen und Forscher keinen direkten Zugriff auf die Mikrodaten und müssen daher auch keine eigenen Anstrengungen zur Absicherung auf ihren lokalen Rechnern unternehmen. Ebenfalls vorteilhaft ist die hervorragende Skalierbarkeit eines solchen Systems, da über zusätzliche modulare Recheninstanzen einfach auf zusätzliche Nutzerinnen und Nutzer reagiert werden kann.

Ein Nachteil eines solchen Systems ist, dass Nutzerinnen und Nutzer die Daten als solche nicht sehen können, da sie ja keinen direkten Zugriff auf die Mikrodaten haben. D. h., eine interaktive Arbeitsweise bei der Erstellung eines Auswertungsskripts ist zumindest inhaltlich nicht möglich. Teilweise kann dies umgangen werden, durch die Bereitstellung sogenannter Strukturfiles, die aber lediglich eine syntaktische Prüfung erlauben. Im Fall der Regionaldaten des SOEP ist dieser Nachteil jedoch weniger gravierend, da SOEP Nutzerinnen und Nutzer den normalen SUF lokal verfügbar haben, lediglich die zusätzliche Information zur Verortung der Haushalte in den Kreisen ist als Zusatzinformation im Remote-Verfahren für die Nutzerinnen und Nutzer nicht sichtbar.

Rechtliches und Datensicherheit

Auch das Remote-Execution-Verfahren des SOEP (SOEPremote) steht nicht jedem offen, sondern ist wie das gesamte FDZ SOEP Angebot auf wissenschaftliche Forschung beschränkt. Die Nutzung setzt voraus, dass im Vorfeld ein SUF genutzt wird und dazu ein Vertrag besteht. Bei den bereitgestellten Daten handelt es sich nicht um personenbezogene, sondern um faktisch anonymisierte Mikrodaten.

Aufwand für Implementierung und Betrieb

Das LISSY System ist komplett in Java implementiert und kann daher quasi auf allen Plattformen genutzt werden. Eine Nutzung verlangt eine kostenpflichtige jährliche Nutzungslizenz. Der Eingangs- und Prüfserver braucht kaum Ressourcen und kann daher auf einem sehr klein bemessenen Rechner laufen (oder auf einem entsprechenden Server als Dienst mitlaufen). Die Ablage der Datenbank ist davon unabhängig in jeder SQL fähigen Datenbank möglich, am FDZ SOEP wird hierbei eine frei verfügbare Postgres Installation genutzt. Die Recheninstanzen können und müssen je nach Datenmenge und Nutzungsaufkommen skaliert werden. Am FDZ SOEP werden drei parallele Recheninstanzen genutzt, wobei pro Monat ca. 1.000 Jobs gerechnet werden.

Die Installation des Systems ist unproblematisch und die benötigten Abhängigkeiten sind Open Source Programme.

Die benötigten Ressourcen zur Outputprüfung sind insbesondere abhängig von der Nutzung der automatischen Prüfungsmöglichkeiten. Werden bei der automatischen Prüfung Auffälligkeiten gefunden, so muss eine Mitarbeiterin bzw. ein Mitarbeiter in einer manuellen Prüfung dem nachgehen. Die Erfahrung zeigt, dass insbesondere bei neuen Nutzerinnen und Nutzern der Prüfbedarf höher ausfällt als bei erfahrenen und mit den Restriktionen vertrauten Nutzerinnen und Nutzern.

Zusammenfassung

Alle drei Verfahren zeichnen sich durch ein hohes Sicherheitsniveau aus (Safe Setting). Daten, die sonst nicht verfügbar gemacht werden könnten, stehen für die Analyse zur Verfügung. Dabei müssen die Nutzerinnen und Nutzer nicht reisen; sie können Analysen von ihrem Arbeitsplatz aus anstoßen und dort auch die Ergebnisse entgegennehmen. Nachteilig für den Forschungsprozess ist das Fehlen der Möglichkeit, die Datensätze zu browsen. Dies ist vor allem bei der Phase der ersten Datenanalyse, bei der es um die Vorbereitung auf die eigentliche Analyse geht und z. B. Extremwerte betrachtet werden müssen, ein Manko. Gerade bei einer starken Nutzung durch die Forschung entsteht der Bedarf, Prozesse zu automatisieren. Damit sind zunächst Aufwände auf Seiten des Datenhalters verbunden, die jedoch in der Folge zu einer Einsparung von Ressourcen führen. Für

weitere Entwicklungen im Bereich Remote Execution (neue Implementationen oder Weiterentwicklungen) sollten Kooperationen und Harmonisierungen der Verfahren erörtert werden.

Remote-Desktop-Verfahren

Zugriff auf sensible Daten über Remote-Desktop-Verfahren bieten vier FDZ, wobei sich hier im Hinblick auf den Access Point zwei Untergruppen bilden lassen. Das DZHW und das LIfBi ermöglichen den Zugriff über die Rechner der Forscherinnen und Forscher und bieten somit eine räumlich flexible Nutzung der Forschungsdaten. Das IAB und das SOEP erlauben den Zugriff lediglich über Thin Clients, die in speziellen und kontrollierten Räumen stehen. Dadurch ist weniger räumliche Flexibilität gegeben, es können aber Forschungsdaten mit einer höheren Detailvielfalt angeboten werden.

Deutsches Zentrum für Hochschul- und Wissenschaftsforschung (DZHW)[14]

Kurzbeschreibung der Organisation und Motivation für die Einrichtung des Verfahrens

Das Deutsche Zentrum für Hochschul- und Wissenschaftsforschung (DZHW) betreibt anwendungsorientierte Forschung im Bereich des Hochschul- und Wissenschaftssystems. Am DZHW werden Befragungen verschiedener Reihen (z. B. Studienberechtigtenpanel, Sozialerhebung, Absolventenpanel, Wissenschaftsbefragung) sowie unterschiedliche Querschnittserhebungen des Forschungsfeldes durchgeführt. Die Daten werden datenschutzkonform für Forschungszwecke zur Verfügung gestellt. Distributionswege sind die Übermittlung von SUF (aktuell teilweise auch über ein GESIS-FDZ), Remote Desktop, On-site Nutzung und in seltenen Fällen auch Job Submission. Zudem werden absolut anonymisierte CUF zu Lehrzwecken herausgegeben. Der überwiegende Anteil der Forschungsdaten wird über das Remote-Desktop-Verfahren zur Verfügung gestellt.

Das Remote-Desktop-System wurde am DZHW (vormals Abteilung Hochschulforschung der Hochschul-Informations-System GmbH) im Jahr 2006 eingeführt. Ziel war die Einrichtung eines datenschutzkonformen sicheren Datenzugangswegs, der die (kollaborative) Nutzung auch sensiblerer Daten des DZHW von einem selbstgewählten Arbeitsort aus erlaubt. Solche Daten waren vorher nur im Rahmen eines Gastaufenthalts am DZHW (On-site) oder über das Einreichen von Analyseskripten (Job Submission) zugänglich.

Seit 2015 wird am DZHW ein forschungsfeldbezogenes Forschungsdatenzentrum für Hochschul- und Wissenschaftsforschung (FDZ) aufgebaut, welches im Juni 2017 seinen Betrieb aufnehmen wird. Sowohl die Daten der im DZHW durchgeführten Studien als auch ins FDZ übernommene Daten anderer Datenproduzenten werden je nach Sensibilitätsgrad weiterhin über die genannten Distributionswege zugänglich gemacht. Durch das FDZ wird jedoch die bisherige Praxis der auf individuelle Datennutzungsanfragen ausgerichteten Aufbereitung von Mikrodaten zugunsten der Herausgabe stärker standardisierter Datenprodukte wie SUF und CUF abgelöst werden. Zudem werden – in Anlehnung an die Praxis anderer Forschungsdateneinrichtungen (wie z. B. LIfBi) – Befragungsdaten derselben Erhebung über mehrere Distributionswege zur Verfügung gestellt werden, indem je nach Distributionsweg unterschiedlich stark anonymisiert wird. Auch im FDZ wird der Datenzugang über das Remote-Desktop-Verfahren in einer technisch modernisierten Variante der zentrale Distributionsweg bleiben.

Beschreibung des Verfahrens

Am DZHW ist der Datenzugang über das Remote-Desktop-Verfahren folgendermaßen organisiert: Nach Einreichung eines Antrags auf Datennutzung zu wissenschaftlichen Zwecken wird mit jeder

[14] Verantwortlich: Karsten Stephan.

nutzenden Person ein Mikrodatennutzungsvertrag geschlossen. Wahlweise kann der Datenzugang individuell oder für eine (verteilt arbeitende) Gruppe eingerichtet werden. Auf den Computern der nutzenden Personen wird (mit Unterstützung durch das DZHW) als plattformunabhängiges Programm ein sogenannter Terminalserverclient installiert. Dieser ermöglicht nach Eingabe eines individuellen Passworts den sicheren Zugang zu einer individuellen Arbeitsumgebung auf dem Terminalserver des DZHW. Dort stehen die Daten, entsprechende (Analyse-)Software und Dateispeicherplatz zur Verfügung. Auf den lokalen Computern der Datennutzerinnen und -nutzer wird die Arbeitsumgebung des im DZHW befindlichen Servers angezeigt. Tastatur- und Mauseingaben am lokalen Computer werden an den entfernten Server weitergereicht, so dass die Nutzerinnen und Nutzer in der Arbeitsumgebung des Terminalservers arbeiten können. Dabei können Datenanalysen mit der auf dem Server vorhandenen Software durchgeführt und Analyseergebnisse direkt eingesehen werden. Ein Kopieren von Dateien aus der Arbeitsumgebung des Servers in die Arbeitsumgebung des lokalen Computers ist nicht möglich. Analyseergebnisse können aber nach einer Prüfung auf datenschutzrechtliche Unbedenklichkeit durch das DZHW an die Datennutzerinnen und -nutzer übermittelt werden (z. B. per E-Mail).

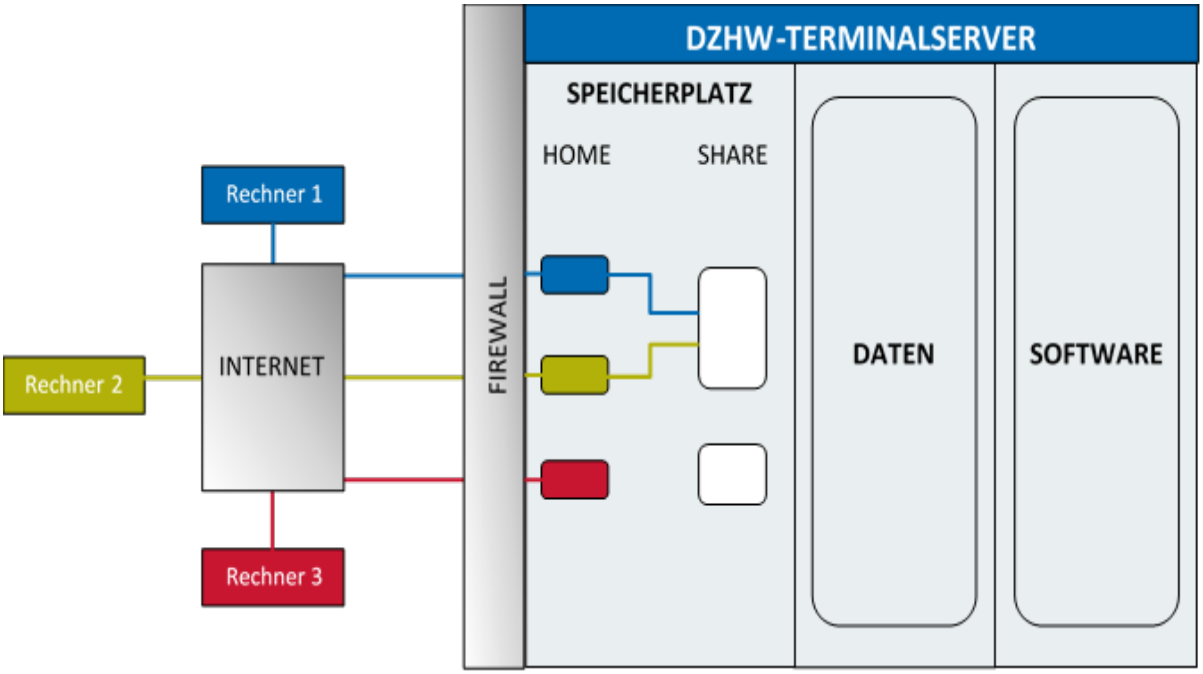

Abbildung 5 - Prinzip des Remote-Desktop-Systems am DZHW (Quelle: DZHW).

Vorteile und Nachteile des Verfahrens

Vorteile des Verfahrens sind die Möglichkeit zur Nutzung auch sensiblerer Daten an einem selbstgewählten Arbeitsort, die Möglichkeit zur räumlich verteilten kollaborativen Datennutzung sowie die Möglichkeit zur kostenlosen Nutzung von Software, die zentral durch das DZHW bereitgestellt wird. Nachteilig gegenüber der Nutzung stärker anonymisierter Daten auf dem eigenen Computer sind Wartezeiten, die durch die datenschutzrechtliche Prüfung der Analyseergebnisse entstehen. Zudem besteht die Notwendigkeit, die vordefinierte Arbeitsumgebung auf dem Terminalserver zu verwenden, die nur in Grenzen durch die nutzende Person konfigurierbar ist.

Rechtliches und Datensicherheit

Die Datensicherheit des Verfahrens wird durch verschiedene Maßnahmen gewährleistet. Der Datennutzungsvertrag regelt den möglichen Umfang der Nutzung und die Verpflichtungen der Datennutzerinnen und -nutzer. Die Verbindung zwischen dem Computer der nutzenden Person und dem Terminalserver ist verschlüsselt. Die Datenverarbeitung findet ausschließlich auf dem

Terminalserver statt, insbesondere verlassen die Daten niemals den Server des DZHW. Ein Kopieren von Dateien wird in beide Richtungen technisch unterbunden. Alle Dateien mit Analyseergebnissen werden vor der Herausgabe auf datenschutzrechtliche Unbedenklichkeit geprüft.

Aufwand für Implementierung und Betrieb

Für die Implementierung und den Betrieb des verwendeten Remote-Desktop-Systems werden im DZHW entsprechend leistungsfähige Server, eine spezielle zusätzliche Firewall, Terminalserversoftware sowie verschiedene Datenanalyseprogramme eingesetzt. Die Betreuung des Remote-Desktop-Verfahrens übernehmen Mitarbeiterinnen und Mitarbeiter, die die Server technisch warten, die Benutzerkonten und Arbeitsumgebungen einrichten, die benötigten Forschungsdaten bereitstellen, Analyseergebnisse auf datenschutzrechtliche Unbedenklichkeit prüfen und mit den Nutzerinnen und Nutzern kommunizieren.

Im Rahmen des FDZ ist zukünftig mit einem wesentlich höheren Datennutzungsaufkommen zu rechnen. Daher wird die Remote-Desktop-Infrastruktur im DZHW aktuell modernisiert und erweitert.

Leibniz-Institut für Bildungsverläufe e.V. (LIfBi)[15]

Kurzbeschreibung der Organisation und Motivation für die Einrichtung des Verfahrens

Angesichts der komplexen und teils sensiblen Befragungs-, Kompetenz- und Kontextdaten des vom LIfBi verantworteten Nationalen Bildungspanels (NEPS) bestand eine wesentliche Herausforderung beim Aufbau der Forschungsdateninfrastruktur darin, der wissenschaftlichen Gemeinschaft einen modernen und flexiblen Zugang zu den aufbereiteten Daten einzurichten. Dieser Zugang sollte Forschenden ein Maximum an Informationsverfügbarkeit und Komfort bei der Arbeit mit den Daten bieten, musste zugleich aber auch den hohen Standards der Datensicherheit und des Datenschutzes gerecht werden. Mit RemoteNEPS als einem System des Ferndatenzugriffs per Remote Desktop wurde am FDZ des LIfBi eine Lösung implementiert, die die Lücke zwischen den Möglichkeiten des Downloads von stärker anonymisierten Datenversionen eines SUF einerseits und des Zugangs zu sensiblen Daten an geschützten On-site-Arbeitsplätzen für Gastwissenschaftlerinnen und Gastwissenschaftler (GWAP) am Standort Bamberg andererseits schließt (Koberg & Stark, 2016). Neben der Gewährleistung eines sicheren Zugangs zu sensiblen Mikrodaten erfüllt RemoteNEPS eine weitere wichtige Funktion, indem es Datennutzerinnen und Datennutzern eine leistungsstarke Forschungsumgebung zur Verfügung stellt.

Beschreibung des Verfahrens

RemoteNEPS ist seit 2009 für externe Datennutzerinnen und Datennutzer in Betrieb, was dem NEPS innerhalb der Forschungsdateninfrastruktur in Deutschland eine Vorreiterrolle bei der Entwicklung und Etablierung von Remote-Desktop-Systemen zuweist.

Die Funktionsweise von RemoteNEPS ist so konzipiert, dass berechtigte Personen ohne größeren Aufwand bzgl. Hardware- und Softwarekonfiguration zu jeder Zeit und von überall mit den sensiblen Daten des NEPS arbeiten können. Nutzerseitig beschränken sich die technischen Voraussetzungen auf einen aktuellen Webbrowser, eine aktuelle Java-Version und eine stabile Internetverbindung. Die Anmeldung am RemoteNEPS-Server erfolgt dreistufig mittels Login, tippbiometrischer Authentifizierung und Passwort. Über eine sichere Verbindung (TLS) wird nach erfolgreicher Anmeldung in einem separaten Fenster eine Arbeitsoberfläche mit Icons zu allen verfügbaren Softwareanwendungen und Verzeichnissen auf den Bildschirm der Nutzerin bzw. des Nutzers projiziert. Jede Datennutzerin und jeder Datennutzer verfügt in RemoteNEPS über ein persönliches Arbeitsverzeichnis; darüber hinaus haben sie Zugriff auf das Datenverzeichnis mit allen veröffentlichten NEPS-SUF sowie auf ein Import/Export-Verzeichnis und auf gemeinsame

[15] Verantwortlich: Daniel Fuß.

Projektverzeichnisse, sofern diese für kollaborative Datenanalysen beantragt wurden. Die Software-Ausstattung von RemoteNEPS erlaubt neben statistischen Analysen die Anfertigung von Texten, Tabellen und Grafiken. Möchten die Forschenden bestimmte Dateien aus der geschützten Serverumgebung auf ein lokales Medium übertragen bzw. vice versa, so bedarf es einer entsprechenden Export- bzw. Importanfrage. Sobald die jeweiligen Dateien von den Mitarbeiterinnen und Mitarbeitern des FDZ-LIfBi geprüft und freigegeben sind, werden sie den Nutzerinnen und Nutzern als Download im NEPS-Webportal bzw. im Import-Verzeichnis in RemoteNEPS zur Verfügung gestellt. Ein direkter Austausch von Dateien zwischen Serverumgebung und lokalem Rechner per „Drag & Drop“ oder „Copy & Paste“ ist nicht möglich. Eine ausführliche Beschreibung von RemoteNEPS findet sich bei Skopek et al. (2016); Hinweise zum Umgang mit der Forschungsumgebung sowie technische Unterstützung bei Problemen bieten die NEPS-Webseite, eine Sammlung von FAQ sowie die Nutzerbetreuung durch das FDZ.

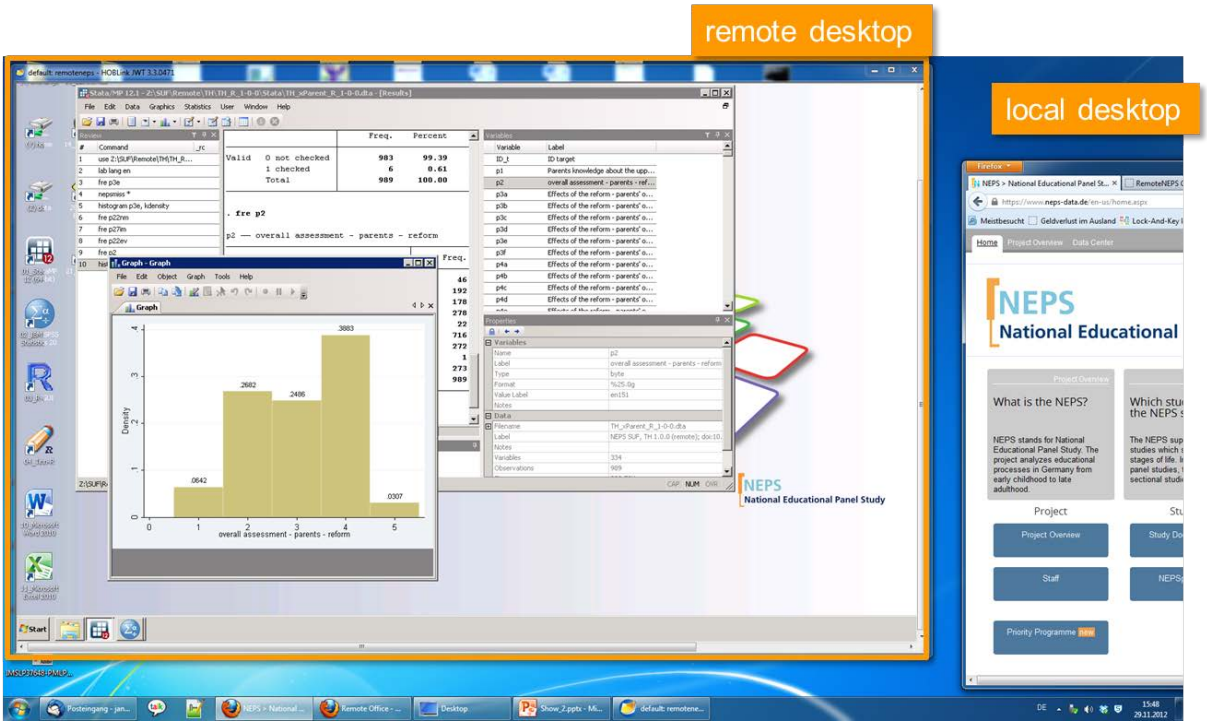

Abbildung 6 - LIfBi Remote Desktop – Screenshot (Quelle: LIfBi).

Vorteile und Nachteile des Verfahrens

Der wesentliche Vorteil von Remote-Desktop-Systemen wie RemoteNEPS besteht in der Unmittelbarkeit der Datenanalyse seitens der Datennutzerinnen und Datennutzer, ohne die Nachteile des Download-Zugangs (eingeschränkte Informationsverfügbarkeit) oder des Zugangs über einen On-site-Gastwissenschaftlerarbeitsplatz (Zeit und Kosten) in Kauf nehmen zu müssen. Die Forschenden können ihre Auswertungen im direkten Zugriff auf die sensiblen Daten und ohne Zeitverzögerung von ihren Rechnern aus durchführen. Dies fördert nicht nur eine intensivere Auseinandersetzung mit den Daten und damit die Qualität der wissenschaftlichen Resultate. Es vermindert zugleich den Aufwand für die Ergebniskontrolle seitens des Datenanbieters, da nur „finale“ Ergebnisse auf Anfrage hin geprüft und exportiert werden müssen. Ein weiterer Vorteil von RemoteNEPS liegt in der Ausstattung mit aktueller und frei verwendbarer Statistiksoftware (Stata, R, SPSS) sowie weiteren Anwendungen (Microsoft Office, Libre Office, Notepad++ etc.). Die RemoteNEPS-Umgebung ermöglicht zudem eine effiziente Zusammenarbeit von mehreren Personen, indem für gemeinsame Projekte spezielle Verzeichnisse mit individuellen Zugriffsrechten für die jeweiligen Kooperationspartner definiert werden.

Rechtliches und Datensicherheit

RemoteNEPS ist eingebettet in das allgemeine Datenschutzkonzept des Nationalen Bildungspanels. Zusätzlich zur Unterzeichnung eines NEPS-Datennutzungsvertrags, in dem die wissenschaftliche Anbindung sowie die Dauer und der Zweck der Datennutzung anzugeben sind, bedarf es für den Zugang zu RemoteNEPS einer Vereinbarung, die unter anderem die Verwendung von Geräten zur Bildaufnahme (Kameras, Fotohandys) und die Anfertigung von Screenshots bei der Arbeit mit den NEPS-Daten untersagt. Für die Anmeldung am Serversystem ist eine biometrische Authentifizierung erforderlich. Mit diesem Vorgang wird sichergestellt, dass nur vertraglich gebundene Personen einen Zugang zu den sensiblen Daten erhalten, da biometrische Anmeldeinformationen im Unterschied zur Zugangsberechtigung mittels Login und Passwort nicht weitergegeben werden können. Im Falle von RemoteNEPS erfolgt die biometrische Erkennung über das individuelle Tippverhalten bei der Eingabe einer bestimmten Phrase. Voraussetzung für die Registrierung des tippbiometrischen Profils ist ein schriftliches Einverständnis zur Speicherung dieser Informationen und der Besuch einer NEPS-Nutzerschulung, in deren Verlauf die zukünftigen Nutzerinnen und Nutzer von RemoteNEPS explizit über relevante Aspekte des Datenschutzes und der Datensicherheit aufgeklärt werden. Auch ein verbreitetes Problem im Kontext klassischer Datenzugriffe über Download - die unzulässige Verwendung von Forschungsdaten nach Ablauf der im Datennutzungsvertrag angegebenen Projektdauer - entfällt bei RemoteNEPS, da die Fristen für den Datenzugang automatisiert über ein datenbankgesteuertes Rechtemanagement umgesetzt werden.

Die Gewährleistung von Datensicherheit erfolgt bei RemoteNEPS im Wesentlichen durch den hohen Grad an Kontrolle bzgl. des Zugangs und der Nutzung der Daten sowie durch das Prinzip der „Datenenklave“. Letzteres meint, dass die sensiblen NEPS-Forschungsdaten zu keinem Zeitpunkt das geschützte Serversystem des LIfBi als Datenhalter verlassen. Die bzw. der Forschende arbeitet letztlich mit einer visuellen Repräsentation der Daten auf ihrem bzw. seinem Rechner, die über eine gesicherte Verbindung (HTTPS) zwischen Client und Server übertragen wird. Dabei ist RemoteNEPS so konfiguriert, dass keinerlei Austausch zwischen der Remote-Desktop-Oberfläche und dem lokalen Rechner stattfinden kann. Jedweder Export von Dateien aus der Enklave oder Import von Dateien in die Enklave durchläuft zwingend eine manuelle Kontrolle durch die Mitarbeiterinnen und Mitarbeiter des FDZ-LIfBi. Werden dabei unzulässige Datenoperationen - seien sie von den Nutzerinnen und Nutzern beabsichtigt oder nicht - identifiziert, können entsprechende Maßnahmen in die Wege geleitet werden. Anders als bei der physischen Bereitstellung von Daten mittels Download wird bei Systemen wie RemoteNEPS eine direkte Kontrolle der Einhaltung von Datensicherheitsbestimmungen durch den Datenhalter selbst wirksam. Hinzu kommt die bereits erwähnte Absicherung des personalisierten Zugangs zu RemoteNEPS über die tippbiometrische Authentifizierung, die gegenüber dem Datenzugriff mittels Download einen zusätzlichen Schutz vor der unberechtigten Weitergabe von Forschungsdaten an Dritte bietet.

Um maximale Flexibilität beim Zugang zu den Daten des Nationalen Bildungspanels zu ermöglichen, stehen den Nutzerinnen und Nutzern drei Optionen zur Auswahl: (1) Die Daten können als klassischer Download nach persönlicher Anmeldung vom NEPS-Webportal heruntergeladen werden, wobei die relativ geringen Kontrollmöglichkeiten durch eine Reihe von Anonymisierungsmaßnahmen kompensiert werden. (2) Die Daten können an speziell gesicherten Arbeitsplätzen am LIfBi in Bamberg analysiert werden, wobei der hohe Grad an Kontrolle bei der On-site-Nutzung keine bzw. nur geringfügige Datenmodifikationen erforderlich macht. (3) Die Daten sind über die Remote-Desktop-Anwendung von RemoteNEPS zugänglich, was im Hinblick auf Informationsverfügbarkeit und Sicherheitsvorkehrungen eine Zwischenstellung zu den beiden erstgenannten Optionen bedeutet.

Die bei der Aufbereitung der NEPS-Daten angewandten Maßnahmen der statistischen Anonymisierung beschränken sich auf Aggregation (z. B. Geburtsland in „Deutschland vs. Ausland“), Top-Coding (z. B. Anzahl der Angestellten mit der obersten Kategorie „mehr als 50“), Bottom-Coding (z. B. Geburtsjahr mit der niedrigsten Kategorie „vor 1950“), Prozentuierung (z. B. Anzahl der

Mädchen in einer Klasse zu „Anteil Mädchen“) und Entfernen von Variablen (z. B. Kreiskennziffer des Wohnorts). Zur besseren Orientierung für die Nutzerinnen und Nutzer von NEPS-Daten enthalten alle drei Versionen eines SUF - Download, Remote Desktop, On-site - das vollständige Set an Variablen; anonymisierte Variablen werden in der jeweiligen Datenversion mit einem speziellen Missing-Code befüllt und mittels Suffix im Variablennamen gekennzeichnet. Außerdem veröffentlicht das FDZ-LIfBi zu jedem NEPS-Datenprodukt einen Report, der die durchgeführten Anonymisierungsmaßnahmen beschreibt und die betroffenen Variablen dokumentiert. Um einen Eindruck vom Umfang der Datenmodifikation bei den NEPS-Forschungsdaten zu gewinnen, haben Koberg et al. (2016) auf der Basis verschiedener Berechnungen eine „Informationseinbuße“ von durchschnittlich ca. 20 % zwischen On-site- und Downloadversion ermittelt. Dagegen unterscheidet sich der Informationsumfang zwischen On-site- und Remoteversion im Mittel um ca. zwei Prozentpunkte. Mithin kann über RemoteNEPS auf nahezu alle verfügbaren Informationen zugegriffen werden.

Aufwand für Implementierung und Betrieb

Der Betrieb von RemoteNEPS bzw. die Bereitstellung dieser Forschungsumgebung für eine Vielzahl von Datennutzerinnen und Datennutzern stellt hohe Anforderungen an die technische Ausstattung des Serversystems. Für RemoteNEPS werden am LIfBi dedizierte Hardware-Server mit jeweils 4 CPU-Sockeln betrieben. Die Systeme sind so ausgelegt, dass für jede und jeden der maximal 50 gleichzeitigen Nutzerinnen und Nutzer mindestens eine „desktopadäquate" Rechenleistung zur Verfügung steht. Da im Hintergrund jedoch Rechnersysteme mit einer höheren Ausstattung laufen, können für die RemoteNEPS-Arbeitsprozesse einer einzelnen Nutzerin bzw. eines einzelnen Nutzers bis zu 16 CPU-Kerne mit jeweils 2 GHz und 64 GB RAM Arbeitsspeicher zur Verfügung gestellt werden. Wie viele Ressourcen letztlich in einer RemoteNEPS-Session nutzbar sind, hängt dabei sowohl vom jeweiligen Bedarf als auch von der jeweiligen Auslastung des Systems ab.

Die speziell für RemoteNEPS betriebenen Systeme sind unmittelbar in die IT-Infrastruktur des LIfBi eingebunden. Dadurch profitieren sie von einer redundanten Servervirtualisierungs-, Speicher- und Netzwerkarchitektur, deren Kosten- und Wartungsaufwand sich nicht direkt auf einzelne Dienste umlegen lässt. In jedem Fall sind für den Betrieb der Gesamtinfrastruktur, die Pflege der beteiligten Systemkomponenten und den technischen Support mindestens zwei Vollzeitstellen im technischen Personal einzuplanen.

Der Aufbau und die Einrichtung von RemoteNEPS erfolgte mit Beratung durch die Betreiber von „MONA – Microdata Online Access" am schwedischen Amt für Statistik (Statistics Sweden). Deren langjährige Erfahrungen bildeten die Grundlage für RemoteNEPS und ermöglichten den Aufbau eines Grundsystems innerhalb von zwei Wochen. Für die Anpassungen an die eigenen Notwendigkeiten und Bedürfnisse des NEPS folgte zunächst eine mehrmonatige Einführungsphase. Für RemoteNEPS wurden in dieser Zeit hausinterne Tests durchgeführt, eine detaillierte Rechtesteuerung u. a. für den Zugriff auf einzelne Programme realisiert, Erfahrungen mit der neu integrierten Tippbiometrie-Authentifizierung gesammelt sowie ein Import-/Exportmechanismus programmiert und in die Webseiten des NEPS integriert.

Ein wesentlicher Aspekt ist die Ausstattung von RemoteNEPS mit Software. Die hierfür erforderlichen Lizenzen bilden einen erheblichen Kostenfaktor, der sich nicht allein auf die „sichtbaren“ Anwendungen wie z. B. Stata, SPSS und Microsoft Office beschränkt, sondern auch die für den Betrieb der Server-Infrastruktur notwendigen Basisprogramme bzw. Betriebssysteme umfasst. Trotz der regelmäßig durch das LIfBi zu tätigenden Investitionen für den Erwerb bzw. die Aktualisierung von Software-Lizenzen ist die Nutzung von RemoteNEPS für Mitglieder der Wissenschaftsgemeinschaft, die die eingangs erwähnten Voraussetzungen erfüllen, kostenfrei möglich.

Institut für Arbeitsmarkt- und Berufsforschung (IAB)[16]

Kurzbeschreibung der Organisation und Motivation für die Einrichtung des Verfahrens

Das Forschungsdatenzentrum (FDZ) der Bundesagentur für Arbeit (BA) im Institut für Arbeitsmarkt- und Berufsforschung (IAB) in Nürnberg bietet einen Fernzugriff auf datenschutzrechtlich sensible Mikrodaten an. Datennutzerinnen und -nutzer können an sechs Standorten der Forschungsdatenzentren der Statistischen Ämter der Länder sowie an der Hochschule der Bundesagentur für Arbeit in Mannheim auf die Daten des FDZ der BA im IAB zugreifen. Zusätzlich besteht diese Möglichkeit an sechs Universitäten in den USA und an zwei Standorten in England, dem UK Data Archive in Essex und dem University College London. Dieser Fernzugriff zu den Daten des FDZ der BA im IAB wurde im Rahmen des FDZ-im-FDZ Ansatzes (Bender & Heining, 2011) etabliert. Die Grundidee bei der Implementierung dieses Fernzugriffs besteht darin, Zugriff auf die Mikrodaten des FDZ aus den Räumen eines anderen Forschungsdatenzentrums (unterschiedliche Institution und Verortung) zu ermöglichen (siehe Abbildung 7).

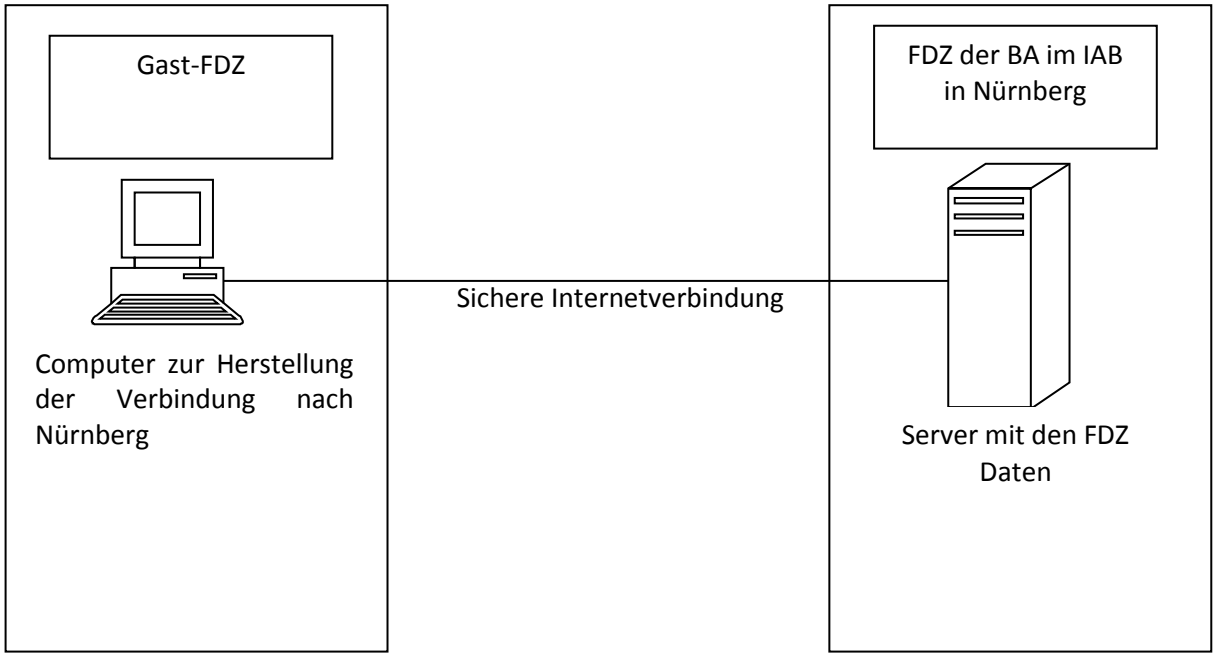

Abbildung 7 - Grundidee des FDZ-im-FDZ Ansatzes (Bender & Heining, 2011).

Motiviert wurde die Entwicklung und Etablierung dieses Verfahrens zunächst durch einen, sich auch international, kontinuierlich ausweitenden Kreis an Nutzerinnen und Nutzern der Daten des FDZ der BA im IAB, wodurch eine stärkere Dezentralisierung des Datenzugangs notwendig wurde. Dem übergeordnet war das Ziel, ein Verfahren zu entwickeln, bei dem die Verteilung von Zugängen anstelle lokaler Datenkopien im Vordergrund steht und damit den Datenzugang in Deutschland näher an einen umfassenden Remote Access bringt.[17]

Beschreibung des Verfahrens

Die technische Umsetzung des Fernzugriffs auf die Daten des FDZ der BA im IAB erfolgt durch sogenannte Thin Client Rechner. Diese Geräte erlauben es den Datennutzerinnen und –nutzern, sich über eine sichere Internetverbindung in das Netzwerk des FDZ in Nürnberg einzuwählen. Die Datenverarbeitung erfolgt nicht auf den Thin Client Rechnern, sondern auf einem Server im abgeschotteten Netzwerk des FDZ.

[16] Verantwortlich: Jörg Heining.

[17] Die Ausführungen in diesem Abschnitt entsprechen in großen Teilen der Darstellung in Heining & Bender (2012).

Der Verbindungsaufbau zwischen Thin Client Rechner am externen Standort und einem Server im Netzwerk des FDZ der BA im IAB erfolgt über die Access Gateway Software und einem dazugehörigen Server der Firma Citrix. Hiermit wird eine verschlüsselte Verbindung zwischen dem Thin Client und dem Netzwerk des FDZ der BA im IAB über das Internet aufgebaut.

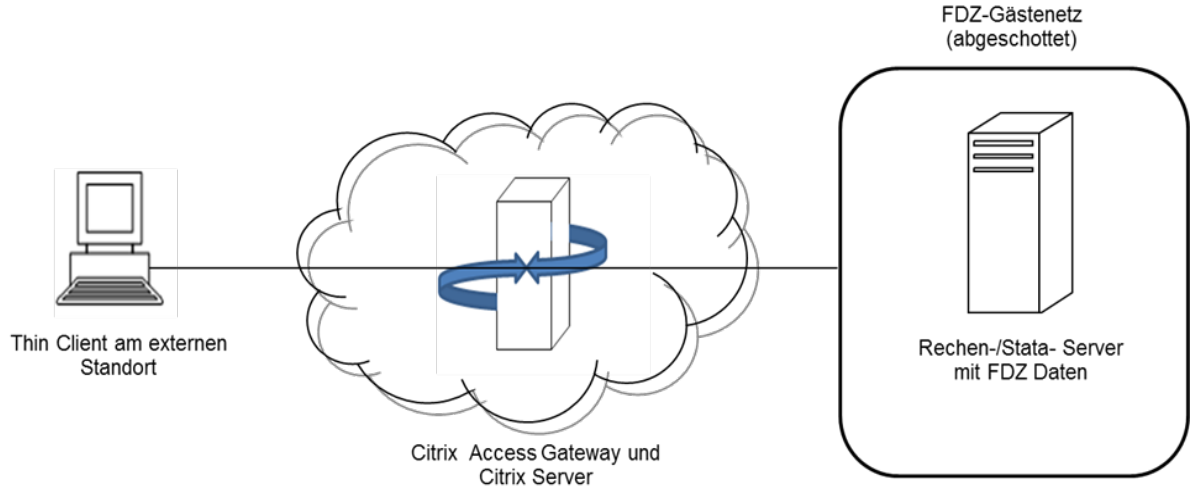

Abbildung 8 - Technische Umsetzung des Fernzugriffs (Bender & Heining, 2011).

Zum Aufbau einer sicheren Verbindung vom Thin Client Rechner zum Server im Netzwerk des FDZ startet die Datennutzerin oder der Datennutzer am externen Standort den Windows Internet Explorer am Thin Client. Die Nutzerin bzw. der Nutzer gelangt zur Citrix Eingabemaske, in der der personalisierte Benutzername und das Kennwort für das Projekt von der Nutzerin bzw. vom Nutzer einzugeben sind. Der Verbindungsaufbau ist zudem durch ein weiteres Kennwort gesichert. Dieses ist nur den vor Ort betreuenden Mitarbeiterinnen und Mitarbeitern bekannt und wird nicht an die externen Datennutzerinnen und -nutzer weitergegeben.

Vorteile und Nachteile des Verfahrens

Durch die Etablierung dieses Verfahrens wurde eine Dezentralisierung des Zugangs zu den Daten des FDZ der BA im IAB entlang der geltenden datenschutzrechtlichen Regelungen und Auflagen erreicht. Datennutzerinnen und –nutzer können nun effizienter und kostengünstiger on-site mit den Daten des FDZ der BA im IAB arbeiten.

Der Erfolg des etablierten Verfahrens wird eindrucksvoll durch die Auslastung an den Standorten im In- und Ausland dokumentiert. Jedoch zeigte sich ein deutlicher Mehraufwand hinsichtlich der Beratung zu den Zugängen und den Daten bei Datennutzerinnen und –nutzern in den USA. Insbesondere müssen hier institutionelle Zusammenhänge erklärt und für die Datennutzerinnen und –nutzer in den USA dokumentiert werden.

Rechtliches und Datensicherheit

Die Datensicherheit wird durch verschiedenste technische und organisatorische Maßnahmen gewährleistet. Wie oben ausgeführt erfolgt die eigentliche Datenverarbeitung auf den Servern im abgeschotteten Netzwerk des FDZ der BA im IAB. Weiterhin verfügen die Thin Client Rechner über keinerlei Anschlussmöglichkeiten für Wechselmedien oder externe Peripheriegeräte, wodurch eine Entnahme von Daten erfolgen könnte. Die Zugangskontrolle und die Einhaltung der Vorgaben des FDZ der BA im IAB während der Arbeit mit den Daten wird durch die Standortbetreuung sichergestellt. Die Datennutzerin bzw. der Datennutzer erhalten lediglich von den Mitarbeiterinnen und Mitarbeitern des FDZ der BA im IAB geprüfte Protokoll- und Ergebnisdateien, die absolut anonym sind (Hochfellner et al., 2012).

Im Hinblick auf den Anonymisierungsgrad der über diesen Zugangsweg verfügbaren Daten gibt es keinen Unterschied zwischen einem Gastaufenthalt in Nürnberg oder an einem der externen Standorte in Deutschland. Hier können berechtigte Datennutzerinnen und -nutzer auf schwach-anonymisierte Daten zugreifen. Anders dagegen an den Standorten in den USA und England. Hier werden die Datensätze durch die Mitarbeiterinnen und Mitarbeiter des FDZ der BA im IAB projektspezifisch zu faktisch anonymen Datensätzen aufbereitet.

Aufwand für Implementierung und Betrieb

Die Entwicklung und Etablierung dieses Verfahrens an zunächst vier Standorten in Deutschland und einem Standort in den USA wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 01UW1002 gefördert. Das IAB erhielt dabei Fördergelder zur Finanzierung von entsprechendem Personal. Hardware und Software wurden dagegen durch das IAB finanziert. Der laufende Betrieb wurde nach Ende der Förderung durch das IAB bzw. IT-Systemhaus der BA übernommen.

Sozio-Oekonomisches Panel (SOEP)[18]

Kurzbeschreibung der Organisation und Motivation für die Einrichtung des Verfahrens

Um die Kreiskennziffern der SOEP Haushalte nutzen zu können, mussten die Nutzerinnen und Nutzer entweder nach Berlin an das FDZ kommen, oder eine sogenannte Remote-Execution-Schnittstelle nutzen (siehe Beschreibung SOEPremote).

Beide Zugangswege haben jeweils spezifische Vor- und Nachteile. Per Remote Execution kann die Datennutzerin bzw. der Datennutzer nicht interaktiv mit den Daten arbeiten und keine Einzelfallprüfungen vornehmen, was den normalen Arbeitsprozess insbesondere bei der Entwicklung einer Auswertungssyntax erschwert. Vorteilhaft für die Nutzerinnen und Nutzer ist jedoch, dass sie für die Auswertung der Daten nicht nach Berlin reisen müssen. Bei einem Aufenthalt am FDZ SOEP ist die interaktive Nutzung der Daten zwar ohne Einschränkung möglich, dafür muss aber die Forscherin bzw. der Forscher nach Berlin reisen.

Ziel des in einem Pilotprojekt etablierten Fernzugriffs ist es, die beiden Vorteile dieser Zugangswege zu kombinieren und die Nachteile aufzuheben. Forscherinnen und Forscher können über eine sichere Internetverbindung von einem externen FDZ aus Zugang zu den Daten des FDZ SOEP bekommen, die Daten selbst aber verbleiben im Intranet des DIW Berlin.

Die Grundidee bei der Implementierung dieses Fernzugriffs besteht darin, Zugriff auf die Mikrodaten des FDZ SOEP aus den Räumen eines anderen Forschungsdatenzentrums (unterschiedliche Institution und Verortung) zu ermöglichen, analog dem FDZ-im-FDZ Ansatz des IAB. Der Datenzugang erfolgt analog zur bisherigen Praxis der Gastaufenthalte im FDZ des SOEP. Der einzige Unterschied besteht darin, dass sich der Gastraum, in dem die Forscherin bzw. der Forscher vor Tastatur und Bildschirm sitzt, nicht im FDZ des SOEP in Berlin, sondern in einem anderen FDZ (Gast-FDZ) befindet.

Beschreibung des Verfahrens

Nach Abschluss eines Nutzungsvertrages mit dem FDZ SOEP können die Forscherinnen und Forscher sich mit Hilfe eines sogenannten Thin Client Terminals auf einem Server im abgeschotteten Netzwerk des DIW Berlin verbinden. Auf diesem Server erfolgt die eigentliche Datenverarbeitung, das Terminal dient somit lediglich dem Verbindungsaufbau. Die Verschlüsselung der Verbindung zwischen Thin Client Rechner und dem Server in Berlin erfolgt unter Verwendung eines restriktiv konfigurierten OpenVPN Tunnels, durch den eine abgesicherte NX-Sitzung geführt wird und ist somit komplett End-to-End verschlüsselt. Ist die Verbindung zur NX-Sitzung aufgebaut, kann die Nutzerin bzw. der Nutzer

[18] Verantwortlich: Jan Goebel.

einen üblichen Linux Desktop nutzen, d. h. Statistikprogramme wie R und Stata oder OpenOffice sind nutzbar als wären sie lokal installiert.

Vorteile und Nachteile des Verfahrens

Durch diese Art des Remote Access können die Wissenschaftlerinnen und Wissenschaftler interaktiv mit den Daten arbeiten und sind nicht den stärkeren Restriktionen einer Remote-Execution-Schnittstelle ausgesetzt. Die Arbeitsweise entspricht daher nahezu komplett der üblichen Arbeitsweise am eigenen Arbeitsplatz. Der Nachteil eines solchen Verfahrens ist jedoch die weiterhin gegebene örtliche Bindung des Zugangs zu einem Thin Client. Es können zwar durch dieses Verfahren einfacher zusätzliche Zugänge in verschiedenen anderen FDZ erstellt werden, jedoch ist dies kein grundsätzlich ortsunabhängiger Zugang.

Rechtliches und Datensicherheit

Unter einem Thin Client Terminal versteht man einen Rechner, der im Gegensatz zu herkömmlichen Rechnern mit weniger Hardware(-leistung) ausgestattet ist. Ein Thin Client stellt lediglich die Benutzerschnittstelle zu einem Server dar. Die eigentliche Kommunikation, Datenverarbeitung und die Speicherung der Daten erfolgen auf dem Server.

Die Thin Clients dieser Lösung verfügen über keine konfigurierte Anschlussmöglichkeit für Wechselmedien (USB Stick, externe Festplatte, usw.) oder externe Peripheriegeräte (Drucker, externe optische Laufwerke usw.). Einer externen Datennutzerin bzw. einem externen Datennutzer ist es daher neben der fehlenden Zugriffsberechtigung auf dem Thin Client nicht möglich, Software oder Dateien auf diesen aufzuspielen oder zu entnehmen. Darüber hinaus ist die für eine externe Datennutzerin bzw. einen externen Datennutzer verfügbare Software auf dem Thin Client Rechner auf den vorkonfigurierten NX-Client als einziger Anwendung beschränkt.

Zur Gewährleistung des Datenschutzes sind bei diesem Zugang zwei Dinge zentral. Zum einen, dass der Zugang nur von einem räumlich kontrollierten Punkt aus möglich ist (kein allgemeiner Zugang durch alle internetfähigen Rechner). Und zum anderen, dass nur absolut anonymisierte Ergebnisse und Dateien den abgeschotteten Bereich des Servers verlassen. Die Mitarbeiterinnen und Mitarbeiter des FDZ des SOEP stellen durch Prüfung der von den Nutzerinnen und Nutzern erzeugten Ergebnisdateien sicher, dass keine Einzelangaben (z. B. Haushalte, Individuen) identifiziert werden können. Diese Grundsätze gelten auch für Datennutzungen an den externen Standorten. Auch bei einer Datennutzung außerhalb des Standortes Berlin wird der Datennutzerin und dem Datennutzer durch das FDZ des SOEP nur absolut anonymer Output übermittelt. Eine selbständige Entnahme von Daten- oder Ergebnisdateien durch die Nutzerinnen und Nutzer am externen Standort ist durch die technische Umsetzung nicht möglich. Es erfolgt eine identische Abgrenzung an nutzbaren Daten wie in SOEPremote.

Aufwand für Implementierung und Betrieb

Für eine Erstimplementierung ist die Anschaffung und entsprechende Konfiguration eines ausreichend performanten Servers unumgänglich, inklusive der auch im normalen Betrieb eventuell notwendigen Lizenzen (eventuelle Statistikprogramme u. Ä.). Die gesicherte Verbindung ist komplett mit Open Source Software umgesetzt. Grundsätzlich ist die Architektur des eingerichteten Fernzugriffs darauf ausgelegt, dass weitere Standorte ohne grundsätzliche Probleme aufgenommen werden können.

Zusammenfassung

Remote-Desktop-Verfahren schließen die Lücke zwischen der Datenanalyse im Rahmen eines Gastaufenthalts und des Downloads von Forschungsdaten. Dabei wird es ermöglicht, die Daten direkt einzusehen und Zwischenergebnisse in Echtzeit am Bildschirm zu betrachten. Durch die unterschiedliche Kontrolle der Zugangspunkte werden neue Zwischenstufen für die Analyse erschaffen. Secure Remote Desktop ermöglicht die Analyse von faktisch anonymen Daten, die mehr

Informationen enthalten als SUF im sicheren Download. Durch die Verwendung eines Safe Rooms, eines Raums, der durch den Datenhalter kontrolliert werden kann, wird sogar die Nutzung von SecUF möglich. Dabei sind allerdings Reisen zu den Safe Rooms nötig. Alle Lösungen benötigen zusätzlichen Aufwand bei der Implementierung und auch im Regelbetrieb. Sie ermöglichen aber auch eine bessere Nutzung von Forschungsdaten. Eine klarere rechtliche Einordnung von Remote-Desktop-Verfahren ist wünschenswert. Dabei sind sowohl die aktuellen Gesetze als auch die technische wie organisatorische Ausformung von Remote Desktop genau und ergebnisoffen zu evaluieren.

Remote-Access-Verfahren weltweit

David H. Schiller

Der folgende Abschnitt stellt exemplarisch und kurz einige Remote-Desktop- und Remote-Execution-Verfahren vor, die außerhalb Deutschlands im Einsatz sind. Danach werden anhand weiterer Beispiele mögliche Zukunftsszenarien erläutert.

Bei einer internationalen Betrachtung von Remote-Desktop-Lösungen[19] muss zunächst NORC at the Universtiy of Chicago erwähnt werden. Die dort im Einsatz befindliche Datenenklave (Lane et al., 2008) war das Vorbild für einige weitere Lösungen und steht in den USA sozusagen als ein Synonym für Remote Desktop. Auf den Erfahrungen von NORC aufgebaut ist z. B. der MONA (Microdata ONline Access) in Schweden, welcher wiederum Vorbild für das Remote-Desktop-Verfahren des deutschen Bildungspanels (NEPS) war. Da sich die Lösungen im Grundsatz ähnlich sind – sicherer Fernzugriff auf geschützte Daten von einem definierten Zugangspunkt – soll in der Folge auf spezifische Unterschiede eingegangen werden. Dabei handelt es sich um die Ausgestaltung des Zugangspunktes, die verfügbaren Datenbestände und die Übermittlung von Ergebnissen.

Um auf die abgesicherten Datenbestände zugreifen zu können, ist stets ein Access Device notwendig. Lösungen wie MONA in Schweden oder das Secure Lab des UK Datenarchives (UKDA) erlauben es den Nutzerinnen und Nutzern ihren eigenen PC zu verwenden. Zur Absicherung werden bestimmte Sicherheitsmaßnahmen, wie die Verwendung eines One-Time-Token und/oder die Kontrolle der IP-Adressen eingesetzt. Bei anderen Lösungen, z. B. bei dem Remote Access Centre Frankreichs CASD (centre d'accès sécurisé aux données), wird den Nutzerinnen und Nutzern ein Thin Client zur Verfügung gestellt. Bei CASD ist dies die sogenannte SD-Box[20], die nur genutzt werden kann, um auf die geschützten Datenbestände zuzugreifen. Meist müssen sich die Geräte in bestimmten Räumen befinden. Das kann das Büro der Forscherin oder des Forschers sein, wobei diese bzw. dieser unterzeichnet, dass gewährleistet wird, dass nur berechtigte Personen Zugang zu diesem Raum haben. Es kann sich aber auch um spezifische Räume handeln, die nur für die Datennutzung reserviert sind, so z. B. beim RDC (Research Data Center) Netzwerk von Statistics Canada, welches derartige Räume an Universitäten einrichten ließ. Bei den verfügbar gemachten Datenbeständen gibt es ebenfalls verschiedene Herangehensweisen. So werden in den skandinavischen Ländern Datensätze nach dem Need-to-know-Prinzip spezifisch für jedes Projekt zur Verfügung gestellt. Need-to-know bedeutet, dass nur die für das Analysevorhaben des spezifischen Projekts notwendigen Informationen aus dem Datenbestand extrahiert und in den Projektdatensatz integriert werden. Dies stellt einen hohen Aufwand bei der Vorbereitung von Projekten dar. Daher werden in anderen Ländern standardisierte Datensätze angeboten, die so gebaut sind, dass sie auf möglichst viele Projekte anwendbar sind. Ein nicht unbedeutender Aspekt bei den verschiedenen über Remote Desktop angebotenen Datenbeständen sind die anfallenden Gebühren. Hier werden stark

[19] Im Jahr 2011 wurde im Rahmen des Projekts „Data without Boundaries“ (DwB, 7 Framework Program Project N°: 262608) eine Befragung europäischer Remote-Desktop-Verfahren bei NSI (National Statistical Institutes) und CESSDA Datenarchiven durchgeführt (Data without Boundaries (DwB), Deliverable 4.1, 2012).
[20] https://casd.eu/en/on-site-integration-sd-box

unterschiedliche Finanzierungskulturen auffällig. Ist es in den skandinavischen Ländern selbstverständlich, dass die Kosten für die Erstellung der spezifischen Datensätze durch die Forschung zu tragen sind, gibt es in Ländern wie z. B. Frankreich den Ansatz, dass die Daten kostenlos sind, da sie ein öffentliches Gut darstellen, die Infrastruktur für den Zugang zu den Daten aber der Forschung in Rechnung gestellt wird. Schließlich existiert noch ein dritter Ansatz, der, wie in Deutschland, die Forschungsfreiheit mit einem kostenlosen Datenzugang gleichsetzt. Die verschiedenen Ansätze spiegeln lediglich unterschiedliche Finanzierungstrukturen und -kulturen wieder, können im Einzelfall für die Forschung respektive für den Datenanbieter aber tiefgreifende Folgen haben.

Bei der Nutzung von Remote-Desktop-Verfahren können die Forscherinnen und Forscher die Mikrodaten „live" sehen und generierte Ergebnisse sofort am Bildschirm betrachten. Bevor die Ergebnisdateien jedoch direkt, z. B. per E-Mail, an die Forscherinnen und Forscher übermittelt werden (d. h. die Ergebnisse aus dem abgesicherten Remote-Desktop-System herausgegeben werden), muss eine Kontrolle auf Konsistenz mit Datenschutzregelungen erfolgen (Safe Output). Die Methoden hierzu unterscheiden sich. In Skandinavien erfolgt oft keine sofortige Kontrolle der Ergebnisse, bevor diese versendet werden. Es wird jedoch jedes übermittelte Ergebnis gespeichert. Auf Basis einer Zufallsstichprobe werden dann im Nachhinein einzelne Ergebnisse geprüft. Außerdem können alle gespeicherten Ergebnisse bei gegebenem Anlass auch nachträglich begutachtet werden. Die Forscherinnen und Forscher sind über dieses Verfahren informiert. Des Weiteren sind es in Skandinavien die Forscherinnen und Forscher, die auf die Einhaltung der Datenschutzregeln zu achten haben. Das UKDA (Datenarchiv des Vereinigten Königreichs) legt einen hohen Wert auf die Schulung der Nutzerinnen und Nutzer. Diese müssen einen Trainingskurs bestehen, der unter anderem auch über den Datenschutz informiert, bevor sie mit den Mikrodaten arbeiten dürfen. In Deutschland wiederum muss jedes Ergebnis vor dem Versand geprüft werden, damit festgestellt werden kann, ob bei den Ergebnissen der Analysen SUF- oder teils auch PUF-Niveau erreicht ist. Es ist offensichtlich, dass der Prozess in Deutschland ressourcen- und zeitaufwändiger ist als der in Skandinavien. Bei den im Folgenden beschriebenen Remote-Execution-Verfahren ist ebenfalls eine Ergebniskontrolle von Nöten. Diese wird hier oft stark automatisiert durchgeführt.

Für den Bereich Remote Execution steht das LISSY System der Luxembourg Income Study[21] als Vorreiter. Die am weitesten entwickelte Variante findet sich in Australien. Das Australian Bureau of Statistics (ABS) bietet mit den Werkzeugen[22] TableBuilder und RADL einen vielfältigen und schnellen Zugriff auf ABS Mikrodaten. Dabei dient TableBuilder zum Erstellen von Tabellen und Grafiken auf Basis der ABS Daten. RADL (Remote Access Data Laboratory) erlaubt vielfältigere Analysen mit den ABS CURF (Confidentialised Unit Record Files). Mit CURF beschreibt ABS seine Mikrodatensätze; dabei entsprechen (im Groben) Basic CURF den oben erwähnten SUF und Expanded CURF den oben beschriebenen SecUF. Das RADL ist das eigentliche Job-Submission-System von ABS und dient der Arbeit mit Expanded CURF. Es ist ein von ABS komplett selbst erstelltes Werkzeug zur statistischen Analyse, das Anfragen sehr schnell beantworten kann (real time), da auf die möglichen Anfragen angepasste SDC Methoden angewendet werden können. Zwar ist dadurch die Variantenbreite möglicher Anfragen beschränkt, doch ermöglicht ABS eine hohe Zahl verschiedener Analyseformen und schafft im Resultat Ergebnisse nahezu ohne Zeitverzug. Neben solchen hochautomatisierten Remote-Execution-Lösungen werden auch Varianten angeboten, die über eine Teilautomatisierung verfügen. Dabei erfolgt z. B. noch eine manuelle Kontrolle der Ergebnisse auf deren Einklang mit Datenschutzverordnungen.

[21] http://www.lisdatacenter.org/data-access/lissy/
[22] http://www.abs.gov.au/websitedbs/D3310114.nsf/home/Microdata+entry+page

Neuere Entwicklungen im Bereich Remote Execution stellen sich neuen Herausforderungen. So fand das DataSHIELD Projekt[23] eine Lösung für Analysen an harmonisierten aber physikalisch an unterschiedlichen Orten gespeicherten Mikrodaten (Jones et al., 2012). Basierend auf Analysen in R[24] können Auswertungen von einem zentralen Punkt über einen zentralen Server ausgeführt werden. Der zentrale Server verteilt die Anfrage auf die betroffenen Datenbestände (Datenserver an verschiedenen Orten), lässt dort Teilkalkulationen durchführen und liefert den Nutzerinnen und Nutzern ein zusammengefasstes Ergebnis zurück. Dabei können automatisierte SDC Methoden durchgeführt werden, um den Datenschutz zu gewährleisten. Für die Nutzerinnen und Nutzer scheint es, als ob alle Daten auf einem zentralen Rechner abgelegt wären, wobei sie in Realität an verschiedenen Orten in der ganzen Welt gespeichert sein können. Neben der technischen Lösung ist die statistische Methode und die Harmonisierung der Datenbestände grundlegend für diese Remote-Execution-Lösung.

Das RAIRD (Remote Access Infrastructure for Register Data)[25] Projekt stellt einen neuen Ansatz des norwegischen Datenarchives (NSD) und von Statistics Norway dar (Heldal et al., 2015). Sein Ziel ist es, möglichst viele der Registerdaten in Norwegen für die Forschung verfügbar zu machen. Das RAIRD System besteht aus drei Hauptkomponenten: Dem RAIRD Data Store, der die Rohdaten vorhält; der RAIRD Remote Storage and Statistical Execution Platform, einer sicheren Umgebung, in der alle Datenmanipulationen (editieren, analysieren etc.) durchgeführt werden; und der RAIRD Online Statistical Environment, welche die gleichen Funktionalitäten wie andere Statistiksoftware (z. B. Stata) zur Verfügung stellt. Im Zusammenspiel bietet RAIRD eine integrierte Infrastruktur zur sicheren Datenanalyse, die basierend auf Rohdaten eine Vielzahl von Analysen auf sichere Weise ermöglicht.

Diese kurze Darstellung zeigt, dass beide Ansätze – Remote Desktop wie Remote Execution – weltweit weiterentwickelt werden. Die Ansätze unterscheiden sich dabei vor allem in drei Dimensionen. Bei der Frage, ob die Mikrodaten direkt, also sozusagen live, gesehen werden können (Remote Desktop) oder ob sie über Anfragen analysiert werden (Remote Execution); und, als Folge dieser Unterscheidung, in der Kontrolle der Zugangspunkte – bei Remote Execution unerheblich; bei Remote Desktop immanent. Die dritte Dimension liegt schließlich im Umgang mit der Weitergabe von Ergebnissen.

Die Entwicklungen des internationalen Remote Access gehen weiter. Neue Systeme werden z. B. in Osteuropa[26] oder von Eurostat[27] entwickelt. Durch das Administrative Data Research Centre Scotland wurden standarisierte safePODS (Dibben, 2015) konzipiert. Dabei handelt es sich um Boxen, die mit einem Thin Client, Schreibtisch und Stuhl, sowie einem Fingerabdruckscanner zur Zugangskontrolle und einer Videoüberwachung ausgestattet werden. Die Boxen sollen an verschiedenen Orten aufgestellt werden und sicheren Remote-Desktop-Zugriff auf die Daten des Administrative Data Research Centre Scotland ermöglichen. Die skandinavischen Länder verfügen mit den Remote-Desktop-Systemen in Dänemark, Finnland und Schweden bereits über moderne Zugangswege zu den Mikrodaten der jeweiligen Statistischen Institute. Im Rahmen eines neuen Projekts kooperieren nun sechs skandinavische Länder (Dänemark, Finnland, Grönland, Island, Norwegen und Schweden), um den Zugang zu ihren Registerdaten zu verbessern (Thaulow & Nielsen, 2015). Dabei werden die drei bestehenden Infrastrukturen genutzt, um Projekten Datenzugang zu ermöglichen, die mit Daten aus mehr als einem skandinavischen Land arbeiten wollen. Grundlage hierfür sind Vereinbarungen, die es erlauben, Daten aus einem Land in ein anderes zu kopieren und dabei die sichere Remote-Desktop-

[23] http://www.datashield.ac.uk/ DataSHIELD kommt in den Projekten BioSHaRE-EU Healthy Obese Project und BioSHaRE-EU Environmental Core Project für die gleichzeitige Analyse von Datensätzen, die in acht verschiedenen europäischen Ländern gespeichert sind zum Einsatz.
[24] https://www.r-project.org/
[25] http://raird.no/about/
[26] http://dwbproject.org/events/workshop_bucharest.html
[27] http://www.safe-centre.info/

Infrastruktur des datenempfangenden Landes zu nutzen. Dieser Ansatz widmet sich einer bedeutenden Limitierung der bestehenden Lösungen, sei es Remote Desktop oder Remote Execution. All diese Lösungen ermöglichen den Zugang zu den Daten eines Landes, noch spezifischer den Daten einer Organisation. Dabei können zwar unter einer Organisation eine Vielzahl von Datenbeständen verfügbar gemacht werden, spätestens bei Projekten, die Ländergrenzen überschreitend arbeiten wollen, sind die aktuellen Lösungen jedoch an ihren Grenzen. Der Ansatz der skandinavischen Länder versucht dies durch die Möglichkeit, Daten zu einem vertrauten Partner zu kopieren aufzulösen. In vielen anderen Konstellationen ist eine solche Übertragung aber aus legalen und organisatorischen Gründen nicht durchführbar. Seit einiger Zeit werden daher Konzepte von Remote-Access-Netzwerken diskutiert (Schiller & Welpton, 2014). Diese würden den Nutzerinnen und Nutzern einen Einstieg über ein Interface (Webseite) ermöglichen und im Hintergrund Zugang zu Datenbeständen verschiedener Länder ermöglichen. Dabei können sowohl Remote-Desktop- als auch Remote-Execution-Lösungen integriert werden. Neben der technischen Herausforderung eines solchen Ansatzes sind Aspekte wie die Harmonisierung von Datenbeständen, Zugangsregeln etc. von hoher Bedeutung. Durch eine solche Infrastruktur kann jedoch eine Weiterentwicklung von einer nationalen und organisationisspezifischen zu einer internationalen und forschungsfragengeleiteten Nutzung der verfügbaren Datenbestände vollbracht werden. Auch auf nationaler Ebene bietet ein solcher Ansatz Vorteile für Forschung und Datenanbieter. Zwar geht es dann nicht um die Harmonisierung über nationale Grenzen hinweg, jedoch können über den Einstieg über ein zentrales Nutzerinterface Dienste wie Datendokumentation, Suchfunktionen und Antragswesen bereitgehalten werden, wodurch die Forscherinnen und Forscher leichter an die gewünschten Daten gelangen und Datenanbieter an zentralen Infrastrukturen partizipieren können. Durch die Implementierung von Remote-Desktop- und Remote-Execution-Verfahren in ein solches Forschungsdatenzentren-Netzwerk kann für die Forschung ein einheitlicher und überschaubarer nationaler Datenbestand definiert und nutzbar gemacht werden (Schiller, 2015).

Zusammenfassung und Ausblick

Ziel des RatSWD, des Ständigen Ausschusses Forschungsdateninfrastruktur (FDI) und der beteiligten FDZ ist ein verantwortungsvoller Umgang mit Forschungsdaten, die Einhaltung von Datenschutzstandards und die bestmögliche Unterstützung der Forschung. Dies ist nur durch die Nutzung moderner, angepasster und nachhaltiger Infrastrukturen möglich. Die Verwendung von Remote-Access-Verfahren ist dabei ein wichtiger Schritt für eine sichere, komfortable und zukunftsfähige Analyse von Forschungsdaten. Die verschiedenen Ausformungen des Remote Access – Remote Execution sowie Remote Desktop mit oder ohne Safe Room – erlauben eine hohe Variabilität im Angebot passender Datenzugangswege. So können unterschiedliche Datenbestände (z. B. administrative oder Befragungsdaten) mit unterschiedlichem Anonymisierungsgrad (z. B. SUF und SecUF) in einer angepassten Umgebung verfügbar gemacht werden. Von den FDZ sind dabei nicht unwesentliche Ressourcen aufzubringen, um Remote-Access-Verfahren aufzubauen und zu betreiben. Eine klare Einordnung der Verfahren in die aktuelle Gesetzeslage ist dabei oft nicht allzu leicht. Umso stärker ist es zu würdigen, dass Mitglieder des Ausschusses FDI neue Wege begehen, um bestmöglichen Datenzugang in Einklang mit Datenschutzregeln und dem generellen Schutz der Persönlichkeitsrechte zu bringen.

Weiterentwicklungen des Remote Access sind auf verschiedenen Ebenen von Nöten. Eine valide Einordnung in gesetzliche Bestimmungen (sei es das BDSG, die jeweiligen Gesetze der Länder, das BStatG oder europäische Regelungen) ist genauso wichtig wie organisatorische und technische Verbesserungen. Hierzu zählen Harmonisierungen von Abläufen, best practice Lösungen und die kontinuierliche Anpassung an moderne IT Standards. Diese Aufgaben sind von einzelnen Institutionen nur schwerlich zu leisten. Entwicklungen müssen in Expertengremien wie dem Ausschuss FDI, flankiert durch Beratungsorganisationen wie dem RatSWD, diskutiert sowie durch nachhaltige Förderung und den Austausch mit verschiedenen Disziplinen ermöglicht werden.

Ein weiterer Aspekt für die Weiterentwicklung von Datenzugangswegen zu sensiblen Forschungsdaten liegt in einer Kultur der Forschung und dem Zusammenspiel zwischen Forscherinnen und Forschern und Datenanbietern. Durch Remote-Access-Verfahren werden Arbeitsumgebungen vom Arbeitsplatz der Forscherinnen und Forscher zu den Datenanbietern verlagert. Datenanbieter werden die Cloudservices der wissenschaftlichen Forschung. Was zunächst als Einschränkung der Freiheiten bei der Forschung erscheinen mag, eröffnet eine Vielzahl neuer Möglichkeiten und Verbesserungen in der Analyse von Individualdaten. Neben besserem Service und besseren Informationen zu den Datenbeständen können durch derartige Infrastrukturen in der Summe Ressourcen eingespart und das Ziel der guten wissenschaftlichen Arbeit, z. B. durch die Verbesserung von Replikationsstudien und Sekundäranalysen, unterstützt werden.

Abbildungsverzeichnis

Abbildung 1 - Portfolioansatz als "equaliser" (Desai et al., 2016). 5
Abbildung 2 - Remote Execution (Schiller & Welpton, 2014). 6
Abbildung 3 - Remote Desktop (Schiller & Welpton, 2014). 7
Abbildung 4 - Serviceangebot und Antragswesen unter http://www.fdz-rv.de (Quelle: FDZ-RV). 10
Abbildung 5 - Prinzip des Remote-Desktop-Systems am DZHW (Quelle: DZHW). 18
Abbildung 6 - LIfBi Remote Desktop – Screenshot (Quelle: LIfBi). 20
Abbildung 7 - Grundidee des FDZ-im-FDZ Ansatzes (Bender & Heining, 2011). 23
Abbildung 8 - Technische Umsetzung des Fernzugriffs (Bender & Heining, 2011). 24

Literaturverzeichnis

Bender, S. & Heining, J., 2011. The Research-Data-Centre in Research-Data-Centre approach: A first step towards decentralised international data sharing. *IASSIST Quarterly*, 35(3), pp. 10-16.

Coder, J. & Cigrang, M., 2003. LISSY Remote Access System. *Joint Eurostat UNECE Work Session on Statistical Data Confidentiality*.

Data without Boundaries (DwB), Deliverable 4.1, 2012. *www.dwbproject.org.* [Online].

Desai, T., Ritchie, F. & Welpton, R., 2016. Five Safes: Designing data access for research. *Working Paper; University of the West of England*, Issue 1601.

Dibben, C., 2015. Micro, remote, safe settings (safePODS) – extending a safe setting network across a country. *Joint Eurostat UNECE Work Session on Statistical Data Confidentiality*.

Heining, J. & Bender, S., 2012. Technische und organisatorische Maßnahmen für den Fernzugriff auf die Mikrodaten des Forschungsdatenzentrums der Bundesagentur für Arbeit. *FDZ-Methodenreport*, Issue 8.

Heldal, J., Monstad, E., Risberg, T. & Risnes, O., 2015. The RAIRD Project: Remote Access Infrastructure for Register. *Joint Eurostat UNECE Work Session on Statistical Data Confidentiality*.

Hochfellner, D., Müller, D., Schmucker, A. & Roß, E., 2012. Datenschutz am Forschungsdatenzentrum. *FDZ-Methodenreport*, Issue 6.

Höhne, J., 2010. *Verfahren zur Anonymisierung von Einzeldaten. Statistik und Wissenschaft.* Wiesbaden: s.n.

Hundepool, A. et al., 2012. *Statistical Disclosure Control.* Wiley Series in Survey Methodology: JOHN WILEY & SONS INC.

Jensen, U., 2012. Leitlinien zum Management von Forschungsdaten. *GESIS-Technical Reports*, Issue 07.

Jones, E. M. et al., 2012. DataSHIELD – shared individual-level analysis without sharing data: a biostatistical perspective. *Norwegian Journal of Epidemiology*, 21 (2), pp. 231-239.

Koberg, T. & Stark, K., 2016. Measuring Information Reduction caused by Anonymization Methods in NEPS Scientific Use Files. *NEPS Working Paper*, Issue 65.

Lane, J., Heus, P. & Mulcahy, T., 2008. Data Access in a Cyber World: Making Use of Cyberinfrastructure. *TRANSACTIONS ON DATA PRIVACY* , Issue 1, pp. 2-16.

Müller, W., Blien, U., Knoche, P. & Wirth, H., 1991. *Die faktische Anonymität von Mikrodaten.* Stuttgart: Metzler-Poeschel.

Schiller, D., 2015. Virtual Research Environments (VREs) to enable access to confidential data for scientific purposes. *Joint Eurostat UNECE Work Session on Statistical Data Confidentiality*.

Schiller, D. & Welpton, R., 2014. Distributing Access to Data, not Data. *IASSIST Quarterly*, 38 (3), pp. 6-14.

Skopek, J., Koberg, T. & Blossfeld, H.-P., 2016. RemoteNEPS—An Innovative Research Environment. In: H. Blossfeld, J. v. Maurice, M. Bayer & J. Skopek, Hrsg. *Methodological Issues of Longitudinal Surveys.* s.l.:VS Verlag für Sozialwissenschaften, pp. 611-626.

Stegmann, M., 2008. Das aktuelle Datenangebot des FDZ-RV: Erschließung der Längsschnittdaten der Rentenversicherung. *DRV-Schriftenreihe, Fünf Jahre FDZ-RV, DRV-Schriften Band*, 55, pp. 27-35.

Thaulow, I. & Nielsen, C., 2015. New Nordic model for researchers joint access to data from the Nordic Statistical Institution. *Joint Eurostat UNECE Work Session on Statistical Data Confidentiality*.