Peter Bartelheimer
Tanja Schmidt

Modellprojekt „Kollaborative Datenauswertung und Virtuelle Arbeitsumgebung“ – VirtAug – Abschlußbericht

soeb-Arbeitspapier 2011-1

Forschungsverbund sozioökonomische Berichterstattung (Hrsg.)
Internet: http://www.soeb.de
Koordination: Soziologisches Forschungsinstitut (SOFI)
Friedländer Weg 31
D-37085 Göttingen
Projektleitung: Dr. Peter Bartelheimer

Inhalt

1. Das Projekt 4

2. Virtuelle Forschungsumgebungen als Teil der „E-Infrastruktur“ für die Sozialwissenschaften 8

2.1 E-Science“ – Innovationspotenzial für die wissenschaftliche „Leistungskette“ 8

2.2 Was eine virtuelle Forschungsumgebung leisten soll 11

2.3 Derzeitiger Entwicklungsstand virtueller Forschungsumgebungen 13

2.4 Zugang zu Forschungsprimärdaten als Voraussetzung 16

3. Kooperation an Sozial- und Wirtschaftsdaten 20

3.1 Arbeitsabläufe 20

3.2 Anforderungen von Forschungs- und Dateneinrichtungen 23

Datenzugang (Datenbereitstellung) 24

„Syntax Sharing“ und Metadaten 26

Archivierung 29

Datenverarbeitung 30

Arbeit im Verbund – nicht nur ein technisches Problem 30

4. Funktionen einer virtuellen Forschungsumgebung 33

4.1 Datensicherheit 33

4.2 Zwei Schnittstellen zu Forschungsdaten 35

4.3 Datenverwaltung 36

4.4 Datenbearbeitung 37

4.5 Datenvergleich 39

4.6 Datenverarbeitung 39

4.7 Kollaboration und Kommunikation 41

4.8 Konfiguration und Verwaltung 41

4.9 Publikation 42

5. Schlussfolgerungen und Empfehlungen 43

5.1 Forschungsnahe und projektbegleitende Entwicklung 43

5.1 Schwerpunkt und Kernelemente einer VFU 44

5.3 Kooperation mit Einrichtungen der Dateninfrastruktur 44

5.4 Datenzugang und Datenschutz 45

5.5 Arbeitsumgebung und Arbeitsprozess 45

6. Literaturverzeichnis 47

Anhang 1 Erster Workshop „Fachwissenschaftliche Anforderungen an eine virtuelle Arbeitsumgebung für soeb“ 50
Anhang 2 Zweiter Workshop „Virtuelle Arbeitsumgebung für sozioökonomische Forschung und Berichterstattung“ 55
Anhang 3 Dritter Workshop „Rechtliche Aspekte der Nutzung von Forschungsdaten“ 61
Anhang 4 Vierter Workshop „Auswirkungen einer Virtuellen Arbeitsumgebung auf Arbeitsabläufe in der sozioökonomischen Forschung und Berichterstattung“.....65

1. Das Projekt

Von 2000 bis 2004 und von 2005 bis 2009 förderte das Bundesministerium für Bildung und Forschung (BMBF) zwei Verbundvorhaben zur Erstellung eines ersten und zweiten „Berichts zur sozioökonomischen Entwicklung Deutschlands (*soeb* 1 und *soeb* 2[1], mehr Information auf der Projekt-Website http://www.soeb.de). Beide Verbundvorhaben wurden vom Soziologischen Forschungsinstitut (SOFI e.V.) an der Georg-August-Universität Göttingen koordiniert. Ziel des sozioökonomischen Berichtsansatzes ist es, wissenschaftsgestützte Sozialberichterstattung problemorientiert weiter zu entwickeln und Brücken von empirischer Forschung zu regelmäßiger Berichterstattung zu schlagen. Die empirische Arbeit der beteiligten Institute soll darauf ausgerichtet werden, neue oder verbesserte Zugänge zu Sozial- und Wirtschaftsdaten für die Weiterentwicklung von Beobachtungskonzepten und Indikatoren zur gesellschaftlichen Entwicklung zu nutzen.

Von August 2009 bis Dezember 2010 moderierte das SOFI – wieder mit Förderung des BMBF – eine fachöffentliche Konzeptphase für ein drittes Verbundvorhaben (soeb 3). Während in einer „soeb-Werkstatt“[2] Fragestellungen und Forschungsstand zu möglichen Themen eines dritten Berichts diskutiert wurden, sollte ein Modellprojekt im Rahmen der Konzeptphase auch klären, welche IT-Infrastruktur die Arbeit eines dritten Forschungsverbunds besser unterstützen könnte.

Das Modellprojekt „Kollaborative Datenauswertung und Virtuelle Arbeitsumgebung“ (VirtAug) sollte am „Anwendungsfall“ der sozioökonomischen Berichterstattung untersuchen und dokumentieren, wie der gemeinsame Datenzugang und die datenbezogene Kooperation von Sozialwissenschaftler/innen an verschiedenen quantitativ-empirisch orientierten Forschungseinrichtungen und insbesondere eine kollaborative Auswertung der Mikrodaten von Forschungsdatenzentren künftig besser organisiert und technisch unterstützt werden könnten. Dabei sollten insbesondere Erfahrungen und IT-Lösungen aus der D-Grid-Initiative (vgl. unten: 2.1) berücksichtigt werden. Das Projekt sollte fachdisziplinäre Anforderungen der empirischen Sozialwissenschaften an virtuelle Forschungsumgebungen

---

[1] Der erste Bericht zur sozioökonomischen Entwicklung Deutschlands erschien 2005 mit dem Untertitel „Arbeit und Lebensweisen“ (SOFI u.a. 2005); der zweite Bericht mit dem Untertitel „Teilhabe im Umbruch“ erscheint im 4. Quartal 2011 im gleichen Verlag (Forschungsverbund Sozioökonomische Berichterstattung 2012).

[2] Die Ergebnisse der fünf Werkstattgespräche, die 2010 in Göttingen stattfanden, sind auf der Projektwebsite (http://www.soeb.de) dokumentiert.

(VFU)[3] klären und deren mögliche Komponenten skizzieren. Ziel war eine Konzeptentwicklung, also weder ein detailliertes Systemdesign noch eine Systementwicklung. Dr. Peter Bartelheimer und Tanja Schmidt bearbeiteten das Teilprojekt VirtAUG am SOFI von August 2009 bis Dezember 2010.

Um bisherige Arbeitserfahrungen in den Verbundprojekten zur sozioökonomischen Berichterstattung zu evaluieren, führte Tanja Schmidt in der zweiten Jahreshälfte 2009 leitfadengestützte Interviews mit acht Projektbeteiligten aus soeb 2; drei der Befragten waren nicht nur Datennutzer/innen, sondern vertraten zugleich datenhaltende Einrichtungen. (Zu den Ergebnissen vgl. unten: 3.1.)

Gleichzeitig nahm das SOFI Kontakt zum Projekt WissGrid in der D-Grid-Initiative auf (vgl. unten: 2.3), die zum Ziel hat, verschiedenen wissenschaftlichen Disziplinen konzeptionelle Grundlagen und technische Lösungen für die kooperative Nutzung von IT-Ressourcen in Netzen (Grids) zu vermitteln. Überlegungen zu einer virtuellen Forschungsumgebung (VFU) für die sozioökonomische Forschung wurden im Rahmen des Review-Workshops des WissGrid-Arbeitspakets Langzeitarchivierung am 28. am 28. Januar 2010 im Astrophysikalischen Institut Potsdam präsentiert.[4] Im April 2010 schloss das SOFI mit der D-Grid Entwicklungs- und Betriebsgesellschaft mbH einen Forschungs- und Entwicklungsvertrag über eine technische Evaluation verfügbarer Grid-Technologie zur Umsetzung einer virtuellen sozialwissenschaftlichen Forschungsumgebung. An der Expertise arbeiteten Harry Enke (Astrophysikalisches Institut Potsdam), Patrick Harms (Abteilung Forschung und Entwicklung der Niedersächsischen Staats- und Universitätsbibliothek Göttingen) und Frank Dickmann (Abteilung Medizinische Informatik der Universitätsmedizin Göttingen). Ihr Bericht (Dickmann u.a. 2010, im Folgenden zitiert als Technische Evaluation), der im Sommer 2010 vorlag, behandelt Anforderungen an eine virtuelle Arbeitsumgebung, deren Komponenten und technische Umsetzungsmöglichkeiten mit Hilfe der D-Grid-Infrastruktur (vgl. unten: 4.1).

Im Verlauf des Jahres 2010 fanden vier Workshops mit Projektbeteiligten der sozioökonomischen Berichterstattung, Expert/inn/en aus dem Projekt WissGrid und Vertreter/innen

---

[3] Im Berichtstext wird entsprechend dem inzwischen üblichen Sprachgebrauch durchgängig diese Bezeichnung verwendet.

[4] Peter Bartelheimer / Tanja Schmidt: Anwendungsfall Sozialwissenschaften: Kollaborative Datenauswertung in virtueller Arbeitsumgebung, Beitrag auf dem Workshop zur Begutachtung des WissGrid AP 3, 28. Januar 2010, AIP Potsdam, URL: http://www.wissgrid.de/workgroups/ap3/workshop-2010-01-28.html.

von Forschungsdatenzentren und Datenserviceeinrichtungen statt, deren Ergebnisse im Anhang dokumentiert sind:

- Workshop 1 „Fachwissenschaftliche Anforderungen an eine virtuelle Arbeitsumgebung für soeb“, Göttingen, 9. Februar 2010 (vgl. Anhang 1).
- Workshop 2 „Virtuelle Arbeitsumgebung für sozioökonomische Forschung und Berichterstattung“, Göttingen, 19. Juli. 2010 (vgl. Anhang 2).
- Workshop 3 „Rechtliche Aspekte der Nutzung von Forschungsdaten“, Göttingen, 05.11. November 2010 (vgl. Anhang 3),
- Workshop 4: „Auswirkungen einer Virtuellen Arbeitsumgebung auf Arbeitsabläufe in der sozioökonomischen Forschung und Berichterstattung“, Göttingen, 15. Dezember 2010 (vgl. Anhang 4).

Überlegungen zu einer virtuellen Umgebung für die Arbeit mit Forschungsdaten stellten Peter Bartelheimer und Tanja Schmidt im Forum „Future Data Access“ der Fünften Konferenz für Sozial- und Wirtschaftsdaten (5|KSWD) am 13. und 14. Januar 2011 in Wiesbaden vor.[5]

Der vorliegende Bericht fasst die Ergebnisse des Modellprojekts zusammen. Er macht Vorschläge zu den Hauptfunktionen einer VFU für die sozioökonomische Berichterstattung und schlägt eine Vorgehensweise zu ihrer Entwicklung vor. Diese Vorschläge orientieren sich weiter am „Anwendungsfall“ der sozioökonomischen Berichterstattung, da diese Verbundvorhaben typischerweise viele verschiedene Mikrodatensätze nutzen, eine Verständigung über Indikatoren und deren Replikation anstreben und dabei verlangen, dass verschiedene Arbeitspakete Analyseverfahren und Ergebnisse untereinander kommunizieren und austauschen. Jedoch sind die grundlegenden Arbeitsabläufe, die im Modellprojekt berücksichtigt wurden, durchaus charakteristisch für die Nutzung von Mikrodaten in der quantitativ-empirischen sozialwissenschaftlichen Forschung. Daher ist zu erwarten, dass die Ergebnisse des Modellprojekts und die skizzierte virtuelle Forschungsumgebung in den Sozialwissenschaften[6] breiter genutzt werden können.

Der nachfolgende zweite Abschnitt stellt das Konzept der Virtuellen Forschungsumgebung in den Zusammenhang der Entwicklung verbesserter digitaler Infrastrukturen für die Forschung. Der dritte Abschnitt skizziert die zu unterstützenden Arbeitsprozesse in Ver-

[5] http://www.ratswd.de/5kswd/5KSWD_Praesentationen/5KSWD_Bartelheimer_Schmidt_Forum4.pdf

[6] Die Sozial-, Wirtschafts- und Verhaltenswissenschaften werden im Folgenden summarisch als Sozialwissenschaften bezeichnet.

bundvorhaben, die Sozial- und Wirtschaftsdaten nutzen, und geht auf Nutzungsinteressen aus Sicht von Projektbeteiligten der sozioökonomischen Berichterstattung ein. Im vierten Abschnitt werden auf der Grundlage der Technischen Evaluation und der Projekt-Workshops Funktionen und Werkzeuge einer VFU für kollaborative Datenauswertung dargestellt. Schlussfolgerungen für Datenschutzbelange, Nutzungsinteressen und für die Entwicklung einer VFU die den dritten Bericht zur sozioökonomischen Entwicklung Deutschlands zieht der fünfte Abschnitt.

2. Virtuelle Forschungsumgebungen als Teil der „E-Infrastruktur“ für die Sozialwissenschaften

2.1 E-Science“ – Innovationspotenzial für die wissenschaftliche „Leistungskette“

Zwei Entwicklungen haben die Forschungsinfrastruktur für die Sozialwissenschaften in den letzten zehn Jahren rasch und grundlegend verändert. Zum einen entstanden in Umsetzung der 2001 vorgelegten Empfehlungen der Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik (KVI) Forschungsdatenzentren und Datenserviceeinrichtungen, die immer mehr amtliche Daten und öffentlich finanzierte Umfragedaten als Mikrodaten für Forschungszwecke zugänglich machen. Zum anderen verändern die technische Entwicklung und zunehmende Verfügbarkeit netzbasierter IT-Ressourcen und IT-Dienste den wissenschaftlichen Zugang zu Daten, die Datenbereitstellung und die datenbasierte Forschungsarbeit.

„e-Science“ oder „e-Infrastruktur“ sind gängige Bezeichnungen für diese sowohl daten- als auch technikgetriebenen Veränderungen. Grundsätzlich haben die damit etikettierten technischen Innovationen das Potenzial, die gesamte „Leistungskette“[7] der wissenschaftlichen Arbeit zu verändern: Die digitale Vernetzung räumlich verteilter Arbeitsplätze, von denen aus über „Middleware“ (d.h. Programme, die zwischen Anwendungen vermitteln) und Dienstprogramme auf Rechnerkapazitäten und Daten zugegriffen werden kann, soll neue standortunabhängige Formen der Zusammenarbeit in der Forschung ermöglichen und unterstützen.

Häufig wird der Aufbau einer eigenen Netz-Infrastruktur als eine Voraussetzung für „e-Science“ angesehen. Seit 2004 fördert das BMBF mit der deutschen D-Grid-Initiative“ [8] eine Erweiterung des Internets durch Hochleistungsnetzwerke („Grids“), die der Wissenschaft „nahezu unendlich große Rechen- und Speicherkapazität, Flexibilität und automatische Anpassung von komplexen Rechenprozessen durch dynamischen und konzertierten Betrieb der vernetzten Ressourcen, höhere Qualität der Ergebnisse durch Grid-unterstützte Entwicklung und schließlich Einsparungen durch eine verbrauchsorientierte Abrechnung“ versprechen (Gentzsch 2007: 9). Zugleich sollen die Potentiale des „Höchstleistungsrechnens“

[7] Die Kommission “Zukunft der Informationsinfrastruktur” (2011: 25) spricht von der „wissenschaftlichen Wertschöpfungskette“.

auch für Unternehmen erschlossen werden. Solche „On-Demand-Infrastrukturen“ werden in den letzten Jahren auch unter der Bezeichnung „Cloud Computing“ entwickelt und diskutiert.

In der D-Grid-Initiative sind bislang überwiegend naturwissenschaftliche Fachdisziplinen vertreten, die einen hohen Bedarf an großen Rechnerleistungen haben, sowie mit TextGrid (vgl. unten: 2.3) ein geisteswissenschaftlicher Forschungsverbund. Dagegen haben die sozialwissenschaftlichen Disziplinen bislang in der Regel keine Notwendigkeit gesehen, bei der Entwicklung von „E-Infrastrukturen“ auf Grid- oder Cloud-Computing zurückzugreifen, da ihre Daten- und Rechneranforderungen für virtuell vernetztes, standortunabhängiges Arbeiten auch im Rahmen der bestehenden Internet-Infrastruktur (dem sogenannten „Web 2.0“) erfüllt werden können. „Results of these developments are possibly less perfect than those designed for Grid applications, but they are facilitated by cooperative approaches within the science community and they take usually much less time to implement.“ (Mochmann 2010: 270.) Daher wird die Weiterentwicklung der Informationsinfrastruktur für die Sozialwissenschaften in den letzten Jahren unabhängig von den technischen Lösungen für digitale Netze, also von der Entwicklung des Grid- oder Cloud-Computings diskutiert.

Ein Leitbild für eine „nachhaltige integrierte digitale Forschungsumgebung“ hat zunächst 2008 die gemeinsame Initiative „Digitale Information“ der Allianz der deutschen Wissenschaftsorganisationen[9] (Alexander von Humboldt-Stiftung u.a. 2008; im Folgenden: Allianz-Initiative) formuliert. Im Jahr darauf hat das Präsidium der Wissenschaftsgemeinschaft Gottfried Wilhelm Leibniz (WGL) im Auftrag der Gemeinsamen Wissenschaftskonferenz des Bundes und der Länder (GWK) 2009 die Kommission „Zukunft der Informationsinfrastruktur“ (im Folgenden KZII) ins Leben gerufen. Diese Kommission unter dem Vorsitz von Sabine Brünger-Weilandt legte im April 2011 ein Gesamtkonzept für die Informationsinfrastruktur in Deutschland vor (KZII 2011). Ebenfalls im Jahr 2009 bat die GWK den Wissenschaftsrat (WR), Empfehlungen für die Weiterentwicklung des Gesamtsystems der Informationsinfrastruktur bis zum Jahr 2020 zu erarbeiten und dabei zu dem Gesamtkonzept der WGL Stellung zu nehmen; diese Stellungnahme wird für Mitte 2012 erwartet (WR 2011: 27). Eine weitere grundsätzliche Stellungnahme stammt von der Kommission für IT-Infrastruktur der Deutschen Forschungsgemeinschaft (DFG 2010).

---

[8] http://www.d-grid.de.
[9] http://www.allianzinitiative.de

Das Gesamtkonzept der Kommission „Zukunft der Informationsinfrastruktur“ behandelt acht Themenfelder:

- nationale Lizensierungen für die Informationsversorgung,
- Hosting (d.h. Bereithalten von Inhalten für den unmittelbaren, sofortigen Zugriff) und Langzeitarchivierung,
- Verbesserung der Zugangs- und Nutzungsbedingungen für nichttextuelle Materialien,
- Retrodigitalisierung des schriftlichen Kulturerbes in Deutschland,
- die Entwicklung *virtueller Forschungsumgebungen* als flexible IT-Infrastrukturen für kollaboratives Arbeiten,
- Open Access (d.h. Zugänglichkeit und Nachnutzbarkeit von Forschungsergebnissen),
- nachhaltige Sicherung, Erschließung/Bereitstellung, Nachnutzung und langfristige Bewahrung von *Forschungsprimärdaten,*
- Verbesserung der Informationskompetenz durch veränderte Ausbildung.

Jedes Themenfeld wurde durch eine Arbeitsgruppe untersucht, die Handlungsbedarfe identifizierte und detaillierte Empfehlungen abgab. Fünf der acht Handlungsfelder, darunter auch das der virtuellen Forschungsumgebungen, wurden in Zusammenarbeit mit Arbeitsgruppen der Allianz „Digitale Information“ bearbeitet.

In der Begriffsabgrenzung der KZII umfasst „Informationsinfrastruktur“

> „die Erwerbung, Aufbereitung, Erschließung, der Nachweis, die Bereitstellung und Archivierung von Information (‚klassische’ Aufgaben), die Sicherstellung von nachhaltiger Retrieval- und Analysefähigkeit von relevanter Information, das Management von Information aller Art (Daten, textuelle und nichttextuelle Objekte, Medien) einschl. der Bereitstellung von Werkzeugen zur Bearbeitung, die Sicherstellung des dauerhaften Zugriffs (Langzeitverfügbarkeit), die Gewährleistung von Sicherheit, Vertraulichkeit und Vertrauenswürdigkeit, die Bereitstellung von Möglichkeiten der kollaborativen Nutzung (z. B. data sharing) und der virtuellen Kommunikation, die Unterstützung dieser neuen Prozesse und Arbeitsgebiete durch adäquate Methoden in der Lehre und Ausbildung“ (2011: 14 f).

Dabei wird in den programmatischen Texten gelegentlich auch die gesamte Informationsinfrastruktur als „virtuelle Forschungsumgebung“[10] bezeichnet. Im Folgenden soll dieser Begriff jedoch nur im engeren Sinn für spezielle Lösungen zur IT-Unterstützung standortunabhängiger Zusammenarbeit in der Forschung verwendet werden. Virtuelle Forschungsumgebungen (VFU) sind demnach Teil einer digitalisierten, webbasierten Informationsinfrastruktur und müssen auf andere Infrastrukturkomponenten zugreifen.

---

[10] Das BMBF spricht im Zusammenhang mit Grid-Computing auch von „virtuellen Wissensumgebungen“; vgl. http://www.bmbf.de/de/298.php.

2.2 Was eine virtuelle Forschungsumgebung leisten soll

Als virtuelle Forschungsumgebungen (VFU) werden „flexible Infrastrukturen“ bezeichnet, „die es Forschern erlauben, die Potenziale elektronischer Medien und Technologien für das kollaborative Arbeiten zu nutzen und daraus auch neue Forschungsmethoden und -gegenstände zu entwickeln“ (KZII 2011: 28). Sie sollen „alle nötigen Instrumente, Daten, Informationen und Werkzeuge zur Verfügung stellen, so dass der Wissenschaftler losgelöst von Ressourcen-und Zugangsproblemen (Speicher, Rechenzeit, Log-In etc.) in einem virtuellen Netzwerk seiner Forschungstätigkeit nachgehen kann“ (Neuroth u.a. 2007: 273).

> Eine virtuelle Forschungsumgebung ist „eine Arbeitsplattform, die eine kooperative Forschungstätigkeit durch mehrere Wissenschaftler an unterschiedlichen Orten zu gleicher Zeit ohne Einschränkungen ermöglicht. Inhaltlich unterstützt sie potentiell den gesamten Forschungsprozess – von der Erhebung, der Diskussion und weiteren Bearbeitung der Daten bis zur Publikation der Ergebnisse - während sie technologisch vor allem auf Softwarediensten und Kommunikationsnetzwerken basiert.“ (Allianz-Initiative)[11]
>
> VFU „ermöglichen vernetztes, zeitlich und räumlich unabhängiges Arbeiten in Gruppen und stellen die dafür benötigte IT-Infrastruktur, Informationsressourcen, Werkzeuge zur Datenproduktion und –weiterverarbeitung sowie Kommunikations- und Publikationsmittel zur Verfügung“ (DFG 2010: 25).
>
> „Technisch betrachtet bestehen sie (meist) aus einer Kernarchitektur mit allgemeinen Dienstleistungen und Werkzeugen (Medienspeicher, Rechenressourcen, Kommunikationsmittel etc.), an die Umgebungen und Module für einzelne Forschungsgruppen mit spezifischen Konfigurationen und Erweiterungen angeschlossen werden können. Virtuelle Forschungsumgebungen decken in der Regel die wesentlichen Arbeitsprozesse in der Forschung ab: Diese umfassen die Recherche und Informationsbeschaffung, die Literatur- und Forschungsdatenverwaltung, die Bearbeitung von Literatur- und Forschungsdaten durch Annotieren, Sequenzierung, Analysieren etc., die Kommunikation von Forschungszwischenständen und die Produktion wissenschaftlichen Outputs bis hin zur Publikation von Forschungsergebnissen.“ (KZII, 2011, S. B74)

Unter den verschiedenen Komponenten der Informationsinfrastruktur nehmen VFU aus zwei Gründen eine Sonderstellung ein. *Erstens* lassen sie sich nicht standardisieren. Sie sollen ausdrücklich die fachlich-inhaltliche Vielfalt unterstützen[12] und „je nach fachspezifischen und individuellen Charakteristika strukturell sehr weit ausdifferenziert sein“ (KZII 2011: B74). So geht die KZII von einem Bedarf von wenigstens 5 VFU für jedes der 48 Fachkollegien der DFG aus. *Zweitens* greifen sie stärker in die Arbeitspraxis der Forschung ein als die anderen „eher technologiebezogenen bzw. der Grundversorgung dienenden Infrastrukturen“ (ebd.). Sicher gilt generell, dass die Arbeitsweisen der Wissenschaftler/innen Grund-

[11] http://www.allianzinitiative.de/de/handlungsfelder/virtuelle_forschungsumgebungen/definition/

[12] Bei VFU ist „der Erhalt der fachlich-inhaltlichen Diversivität ausdrücklich gewünscht, um der Kreativität und Innovation in der Forschung keine unnötigen Beschränkungen aufzuerlegen.“ (Allianz-Initiative; http://www.allianzinitiative.de/de/handlungsfelder/virtuelle_forschungsumgebungen/definition/.)

lage für die zu entwickelnden Informationsangebote und Dienstleistungen sein sollen (Winkler-Nees 2011). Doch VFU „nehmen stärkeren Bezug auf den tatsächlichen Gebrauch von IT-Diensten in den jeweiligen Disziplinen oder interdisziplinären Feldern, den dort eingesetzten Methoden und zu bearbeitenden Forschungsfragen“ (DFG 2010: 25). Aus beiden Gründen können VFU nur „forschungsnah“ entstehen – ihre Entwicklung bedeutet zugleich eine technische und eine soziale Innovation.

> „Virtuelle Forschungsumgebungen werden nur dann intensiv genutzt, wenn sie die Anforderungen der jeweiligen Forschungsprozesse möglichst gut abdecken. Deshalb können die Umgebungen nicht mehr wie bisher oftmals für die Fachwissenschaftler, sondern nur mit ihnen gemeinsam entwickelt werden. Das heißt, dass Forschende den Aufbau und die Weiterentwicklung von virtuellen Forschungsumgebungen aktiv mitgestalten und mit steuern: Die kooperative Entwicklung unter Beteiligung von Forschern, Informationsspezialisten und IT-Fachleuten gehört zu den entscheidenden Erfolgsfaktoren einer virtuellen Forschungsumgebung. Die Fachwissenschaftler formulieren ihre Anforderungen und begleiten die informationelle und softwaretechnische Entwicklung der virtuellen Forschungsumgebungen in einem iterativen Prozess. Bibliotheken und Rechenzentren sorgen für wissenschaftsnahe Standardisierungen, die Unterstützung von wissenschaftlichen Fachdiensten durch Basisdienste sowie die Sicherung der Nachhaltigkeit von Services und Forschungsergebnissen.“ (KZII 2011; S. B87f)

Die AG KZII zum *Themenfeld Virtuelle Forschungsumgebungen* benennt bereits einige Erwartungen an Funktionalität und Struktur von VFU (KZII 2011: B80), wobei zu berücksichtigen ist, dass die an solchen Systemen beteiligten Wissenschaftler/innen „zugleich Nutzer und Anbieter sein“ können, also nicht nur die Nutzerperspektive einnehmen.

- Unterstützung des kompletten Forschungszyklus (z. B. von der Datenerhebung durch Messinstrumente über die Analyse und Sequenzierung bis zur Publikation und Archivierung der Forschungsdaten),
- Unterstützung von Projektvorbereitung und -management unter Einbeziehung der lokalen administrativen Systeme (z. B. Finanz- und Personalmanagement, E-Mail-Dienst, Zeit- und Terminplanung),
- Umfassender Informationszugriff unter Einbeziehung unterschiedlicher Informationstypen (z. B. Literatur, Forschungsdaten, Simulationen sowie Forschungsprozessinformationen wie Projekte, Forscher u. a.) aus verteilten Quellen,
- Einfache und benutzerfreundliche Nutzbarkeit virtueller Forschungsumgebungen,
- Modularer Aufbau und flexible Konfigurierbarkeit hinsichtlich beteiligter Forscher/innen, verfügbarer Funktionen und integrierter Informationen mit entsprechender Zugangsrechteverwaltung,
- Unterstützung des Datenaustausches zwischen den Forscher/innen und den Modulen der Forschungsumgebung sowie mit externen Quellen und Systemen (Bibliothekskatalogen, sozialen Netzwerken, Publikationsplattformen und Open Access-Repositorien) und Standardsoftware (z. B. Office- oder Statistikpaketen),
- Nachhaltige Verfügbarkeit von Informationen und Funktionen, gewährleistet durch wissenschaftliche Infrastruktureinrichtungen.

Für die Aufbauphase von VFU liegen insbesondere von der KZII bereits eine Reihe von Empfehlungen vor:

- „Community- oder projektspezifische Entwicklungen“ sollen „in übergeordnete Strukturen eingebettet“ sein, „auf bestehenden Systemen aufbauen und die Neuentwicklungen auf die fachspezifischen Services fokussieren“ (KZII 2011: B83),
- Infrastrukturdienstleister, die den späteren Betrieb dauerhaft übernehmen können, sollen bereits während der Aufbauphase einbezogen werden.

- Als Träger von VFU sieht die KZII (2011: B84) Forschungsorganisationen und Hochschulen. Die DFG (2010: 25) sieht die Ressourcenorganisation und das Identitätsmanagement der Hochschulen gefordert, solche Kooperationsplattformen zu unterstützen.
- Die KZII sieht die Notwendigkeit, „Förderprogramme für virtuelle Forschungsumgebungen aus(zu)weiten und zusätzliche Finanzmittel für den dauerhaften Betrieb bereit(zu)stellen". Das erforderliche Fördervolumen schätzt die Arbeitsgruppe der Kommission aufgrund der bisherigen Förderaktivitäten von BMBF und DFG für einen Zeitraum von mindestens zehn Jahren auf „eine Größenordnung von mehreren 100 Mio. EUR" (KZII 2011: B87).[13]
- Geeignete Koordinations- und Unterstützungsstrukturen sollen Redundanzen vermeiden, Förderaktivitäten „kartieren" und bestehende VFU nachweisen. Wissenschaftliche Netzwerke sollen Fachwissenschaftler und Infrastruktureinrichtungen beim Aufbau von VFU methodisch unterstützen. (KZII 2011: B87)
- Vorhandene Infrastruktureinrichtungen sollten in den Aufbau von VFU integriert werden; dass Fachwissenschaftler/innen mit ihnen zusammenarbeiten, sollte ein Prüfkriterium bei der Förderung von VFU sein (ebd.).
- VFU sollten kompatible Mindeststandards erfüllen und nachnutzbare Strukturen und Daten bereitstellen (ebd.)
- Qualifizierungsmaßnahmen werden vor allem für den postgradualen Bereich, für Informationsexpert/innen, Bibliothekar/innen aber auch Informatiker/innen und Fachwissenschaftler/innen empfohlen (ebd.).

2.3 Derzeitiger Entwicklungsstand virtueller Forschungsumgebungen

Der Aufbau virtueller Forschungsumgebungen wird in Deutschland zum einen von der Deutschen Forschungsgemeinschaft (DFG)[14] und zum anderen vom Bundesministerium für Bildung und Forschung (BMBF) gefördert. Das BMBF fördert insbesondere die D-Grid-

[13] Die von der Arbeitsgruppe der KZII errechneten Durchschnittswerte für die Projektförderung von VFU liegen jedoch weit auseinander. Für die D-Grid-Initiative ergibt sich ein Durchschnittswert von ca. 5 Mio. EUR, für 22 bisher geförderte DFG-Projekte ein Wert von ca. 340.000 EUR je Projekt (KZII 2011: B86).

[14] http://www.dfg.de; Infrastruktur, Themenschwerpunkt digitale Information, Förderbereich Informationsmanagement, Programm „Virtuelle Forschungsumgebungen".

Initiative[15] und das Projekt WissGrid (vgl. dazu unten). Auf europäischer Ebene erging 2009 im Rahmen des 7. Forschungsrahmenprogramms (FP7) ein Call zu „Virtual Research Communities“[16]. Auch die Allianz der Wissenschaftsorganisationen (vgl. oben) und die Deutsche Initiative für Netzwerkinformation (DINI)[17] beschäftigen sich thematisch mit VFU. Wie die Arbeitsgruppe der KZII (KZII 2011: B76 ff.) feststellt, fehlt ein systematischer Überblick über den Stand der Entwicklung. Weder habe sich eine beherrschende Technologie etabliert, noch gebe es Referenzarchitekturen.

In dieser Situation kommt dem vom BMBF geförderten Projekt WissGrid[18] im Rahmen der D-Grid-Initiative besondere Bedeutung zu. Das WissGrid-Konsortium, das an der Niedersächsischen Staats- und Universitätsbibliothek (SUB) Göttingen koordiniert wird, bietet grundlegende Unterstützung „für die langfristige Nutzung der deutschen Grid-Infrastruktur sowie IT-technische Lösungen“ und will „die organisatorische Zusammenarbeit der Wissenschaftsdisziplinen im Grid fördern und die Eintrittsschwellen für neue Community-Grids senken“ (Neuroth 2010, S. 18). Auch wenn VFU für die Sozialwissenschaften nicht zwingend das Grid voraussetzen bietet WissGrid einen Überblick über generische Basisinfrastrukturen, die für die Entwicklung spezifischer, auch internetbasierter Lösungen genutzt werden können. Zu den Leistungen von WissGrid gehört die Erstellung von „Blaupausen für Community-Grids“. „Community Grid“ steht dabei nicht nur für Hochleistungsnetzwerke, sondern für die gesamte technische Infrastruktur an Hardware, Netzwerken, Sicherheitsmechanismen und höherwertigen Diensten, die eine Nutzung verteilt vorliegender Datenmengen und Rechenressourcen ermöglicht. Fachberater sollen die verschiedensten Fachdisziplinen beim Aufbau virtueller Umgebungen für ihre Forschungsaufgaben unterstützen. Ein eigenes Arbeitspaket in WissGrid ist die Langzeitarchivierung digitaler Forschungsdaten.

Als Beispiel für eine VFU, die im Rahmen der D-Grid-Initiative entwickelt wurde, kann TextGrid[19] (Herpay u.a. 2009: 38 f.) stehen. Ziel ist eine für Grid Computing geeignete Forschungsinfrastruktur für die Geisteswissenschaften. TextGrid bietet ein fachwissenschaftliches Langzeitarchiv (Repository) für geisteswissenschaftliche Forschungsdaten und einen Einstiegspunkt (Laboratory) für eine ständig erweiterbare Zahl von Diensten, mittels derer

---

[15] Die Grid-Technologie vernetzt geographisch verteilte Hardware, also Rechner- und Speicherkapazitäten, und stellt sie allen Beteiligten zur Verfügung. Derzeit sind 35 Projekte mit über 140 Partnern an D-Grid beteiligt (vgl. Gemeinnützige D-Grid Entwicklungs- und Betriebsgesellschaft mbH, 2010).

[16] ftp://ftp.cordis.europa.eu/pub/fp7/ict/docs/e-infrastructure/e-infrastructures-in-fp7-call7_en.pdf.

[17] http://www.dini.de/ag/vforum.

[18] www.wissgrid.de.

[19] http://www.textgrid.de.

Textressourcen kollaborativ editiert, annotiert, analysiert und veröffentlicht werden können. [20]

Auf europäischer Ebene werden in einer Reihe von Projekten Komponenten einer Informationsinfrastruktur entwickelt, auf die VFU zugreifen können. Um eine neue europäische Forschungsinfrastruktur zu planen und zu implementieren, wurde 2002 das „European Strategy Forum on Research Infrastructures“[21] (ESFRI) ins Leben gerufen (Publications Office of the European Union 2011). Eine erste Roadmap aus dem Jahr 2006 wurde 2010 aktualisiert. Derzeit werden zehn von insgesamt 44 geplanten Projekten bereits implementiert. Als Dateninfrastruktureinrichtung für die Human- und Sozialwissenschaften integriert das „Council of European Social Science Data Archive“ (CESSDA)[22] über 20 Sozialdatenarchive aus 20 europäischen Ländern. Auch zwei sozialwissenschaftliche Bevölkerungsumfragen, der „European Social Servey“ (ESS)[23] und der „Survey of Health, Aging and Retirement in Europe“ (SHARE)[24] werden durch ESFRI gefördert. Die Survey-Daten aus ESS und SHARE können relativ unproblematisch direkt aus dem Internet auf den lokalen Rechner geladen und dort mit den üblichen Statistikprogrammen und Verfahren lokal ausgewertet werden. Der ESS stellt zusätzlich die Möglichkeit zur Verfügung, einfache Analysen direkt mittels Nesstar, einer speziellen Software[25], online durchzuführen. Auch die ESFRI-Projekte „Common Language Resources and Technology Infrastructure” (CLARIN)[26] und “Digital Research Infrastructure for the Arts and Humanities (DARIAH)[27] zielen darauf ab, Komponenten für virtuelle Forschungsumgebungen zu entwickeln.

Das im Rahmen des 7. EU-Forschungsrahmenprogramms (FP7) geförderte Projekt „Permanent Access to the Records of Science in Europe“ (PARSE.insight)[28] evaluiert Anforderungen insbesondere an die nachnutzbare Langzeitarchivierung europäischer Forschungsdaten und entwickelt dazu Handlungsempfehlungen (vgl. PARSE.insight 2009).

---

[20] Ein Beispiel für eine internetbasierte VFU außerhalb der D-Grid-Initiative ist Edumeres.net, ein Web-Portal des Georg Ecker Instituts für Internationale Schulbuchforschung (www.gei.de) mit einer integrierten, teilweise öffentlich zugänglichen virtuellen Forschungsumgebung.
http://www.edumeres.net/virtuelle-forschungsumgebung.html

[21] http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=esfri

[22] http://www.cessda.org

[23] http://www.europeansocialsurvey.org

[24] http://www.share-project.org

[25] Vgl. http://www.nesstar.com/; für den ESS: http://ess.nsd.uib.no/ess/

[26] http://www.clarin.eu

[27] http://www.dariah.eu

[28] http//www.Parse-insight.eu

2.4 Zugang zu Forschungsprimärdaten als Voraussetzung

Komponenten der Informationsinfrastruktur für die sozialwissenschaftliche Forschung können grob danach unterschieden werden, ob sie an der Datenproduktion und -bereitstellung, an der Schnittstelle zwischen Datenhaltung und Datennutzung oder am datenbezogenen Forschungsprozess ansetzen. Die Empfehlungen der Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik (KVI 2001) und die Tätigkeit des 2004 eingerichteten Rats für Sozial- und Wirtschaftsdaten[29] (RatSWD 2011) konzentrierten sich auf die Datenschnittstelle: also darauf, den Zugang der empirisch arbeitenden Sozial-, Verhaltens und Wirtschaftswissenschaften zu Mikrodaten aus öffentlich finanzierten Erhebungen zu verbessern und zu erleichtern.

Der Zugriff auf Forschungsdaten bzw. das data-sharing wird durch verschiedene Forschungsdatenzentren (FDZ) und durch Datenservicezentren (DSZ) ermöglicht. Derzeit sind beim Rat für Sozial- und Wirtschaftsdaten sechzehn Forschungsdatenzentren und drei Serviceeinrichtungen akkreditiert.

> „Die FDZ und DSZ als institutionalisierte Orte des data sharing ermöglichen nicht nur den Zugang zu Daten, sondern bieten darüber hinaus einen Service um die Daten herum an. Ein solcher Service ist wegen der komplexen Strukturen vieler Datensätze, und der jeweils beschränkten Aussagekraft der Daten (Reichweite, Validität und Reliabilität), welche durch die Operationalisierungen der Erhebungen bedingt sind, nötig und kann am besten von denen geleistet werden, die die Daten produzieren." (Huschka et al 2011, S. 2).

Als „Vision" der künftigen Forschungsdateninfrastruktur formuliert ein Bericht an die EU-Kommission den „seamless access, use, re-use, and trust of data":

> "In a sense, the physical and technical infrastructure becomes invisible and the data themselves become the infrastructure – a valuable asset on which science, technology, the economy and society can advance." (High Level Expert Group 2010: 4, vgl. Kommission Zukunft der Informationsinfrastruktur, 2011: 16 f).

Virtuelle Forschungsumgebungen sollen sowohl den Datenzugang als auch die Arbeit mit Daten unterstützen. Daher bleibt die Datenschnittstelle die wesentliche Komponente der Informationsinfrastruktur, auf die VFU zugreifen müssen: „Integraler Bestandteil sind Forschungsdaten; Open Access[30] ist eine wesentliche Voraussetzung für die volle Entfaltung des Innovationspotenzials von virtuellen Forschungsumgebungen." (KZII 2001: B73)

Die Empfehlungen der AG der KZII zum *Themenfeld Forschungsprimärdaten* sind daher für die Entwicklung von VFU gleichfalls von großer Bedeutung. Die Kommission macht vier

---

[29] http://www.ratswd.de.

[30] Der Zusammenhang zwischen VFU und Open Access zu digitalen Publikationen konnte im Rahmen des Modellprojekts noch nicht eingehend berücksichtigt werden.

Handlungsfelder aus: die nachhaltige Sicherung, die Erschließung und Bereitstellung, die Nachnutzung und die langfristige Bewahrung von Forschungsdaten (KZII 2011: 111). Für die Sozialwissenschaften sollen dazu vor allem bereits bestehende Infrastrukturen des GESIS – Leibniz-Instituts für Sozialwissenschaften (GESIS)[31] und des RatSWD genutzt werden.

Im Bereich der Sicherung verweist die KZII (2011: B115) auf das Eigeninteresse der Nutzer/innen an Datensicherheit und auf die Anforderung mehrjähriger Datensicherung als Teil „guter wissenschaftlicher Praxis". Wissenschaftliche Nutzer/innen erwarten von der Technik „Standards in Bezug auf Verfügbarkeit, Integrität und Vertraulichkeit der Daten" sowie schnellen, orts- und zeitunabhängigen Zugriff. Sie müssen sich „auf vertrauenswürdige, langfristig stabile Einrichtungen verlassen können", und sie wollen „weitestgehende Kontrolle über die eigenen Daten behalten". Zugriffsrechte (Schutz der Urheberrechte) müssen klar geregelt sein.

Für die Erschließung und Bereitstellung von Forschungsdaten sieht die KZII vor allem zwei Hürden: Dem hohen Ressourceneinsatz für Dokumentation stehe „derzeit kein wissenschaftlicher Reputationsgewinn gegenüber". Die Veröffentlichung berge Risiken, etwa der Verwertung durch andere, der Kritik oder der Verletzung des Datenschutzes. „Aus Forschersicht überwiegen die wahrgenommenen Kosten bzw. Risiken des ‚data sharing' daher oftmals noch den antizipierten individuellen Nutzen." (KZII 2011: B115 f.)

Bei der Nutzung von Forschungsdaten konzentriert sich der Bericht der KZII auf Suchdienste, Metadaten und Tools sowie auf die Kosten der Datennutzung. Handlungsbedarf sieht sie vor allem bei Richtlinien zum Datenmanagement und bei der Finanzierung. Die für die Sozialwissenschaften zentralen Belange des Datenschutzes behandelt sie dagegen nur am Rande.

Zur langfristigen Bewahrung von Forschungsdaten empfiehlt die KZII, „stärker auf zukunftsfeste Datenformate/Standards sowie Hard- und Softwareunabhängigkeit zu achten" und Daten durch „Disziplin-spezifisch begründete Metadatenschemata" zu erschließen (KZII 2011: 116).

Für das Datenumfeld, in dem VFU für die Sozialwissenschaften entstehen, ist vor allem die Arbeit der Arbeitsgruppe „Future Data Access"[32] von Bedeutung. Unter Leitung von Prof.

---

[31] GESIS ist als größte deutsche Infrastruktureinrichtung für die Sozialwissenschaften zugleich Träger von zwei Forschungsdatenzentren (FDZ ALLBUS und FDZ Wahlen) und des Datenservicezentrums German Microdata Lab (GML). http://www.gesis.org/.

[32] http://www.ratswd.de/Future_Data_Access/index.php

Dr. Ulrich Rendtel (FU Berlin) soll sie den derzeitigen Stand an Datenzugangsmöglichkeiten evaluieren und weitere Handlungsempfehlungen erarbeiten. Eine weitere Verbesserung der Informationsinfrastruktur wird vor allem von neuen Möglichkeiten des Datenfernzugriffs (Remote Data Access) auf Forschungsdaten erwartet.

Für viele Forschungszwecke ist eine Arbeit mit den Originaldatensätzen der Auswertung faktisch anonymisierter Scientific Use Files (SUF) vorzuziehen. Zudem lässt sich eine wachsende Zahl komplexer Datensätze nicht ohne wesentlichen Informationsverlust in ein faktisch anonymisiertes Format bringen, das die datenschutzrechtliche Voraussetzung für eine Weitergabe als SUF erfüllt. Für die Arbeit mit Originaldaten gibt es zwei Nutzungswege: On-Site-Nutzung an Arbeitsplätzen für Gastwissenschaftler/innen und kontrolliertes Fernrechnen. Das Fernrechnen verursacht den FDZ hohen Arbeitsaufwand und den Nutzer/innen lange Wartezeiten, da alle Ergebnisse zunächst manuell auf ihre datenschutzrechtliche Unbedenklichkeit geprüft werden müssen. Daher beschäftigt sich das Projekt „An informational infrastructure for the E-Science-Age“ (infinitE) beim Forschungsdatenzentrum des Bundes und der Länder (Brandt/Zwick 2011) mit der Entwicklung anonymisierter Strukturdatenfiles und neuer Verfahren einer automatischen Ergebniskontrolle, die das Fernrechnen erleichtern sollen.

Eine weitere absehbare Entwicklung sind „FDZ-in-FDZ“-Nutzungen: Eine Kooperation zwischen Forschungsdatenzentren kann Nutzer/innen die Möglichkeit von kontrolliertem Fernrechnen mit den Daten aller beteiligten Einrichtungen eröffnen.

Schließlich können bei der Entwicklung neuer Bevölkerungsumfragen neue Lösungen für Datenfernzugriff von vornherein umgesetzt werden. So sieht das Nationale Bildungspanel (NEPS) künftig neben der herkömmlichen Nutzung von SUF und Gastarbeitsplätzen den Fernzugriff auf Forschungsdaten über „RemoteNEPS“ vor.

> „Über einen „virtuellen Desktop“ greift der Nutzer innerhalb einer kontrollierten Umgebung auf Mikrodaten zu. Der Zugang zu dieser Datenenklave ist unabhängig vom Betriebssystem und erfordert keine Softwareinstallation, sondern lediglich einen Interzugang und einen aktuellen Browser. Die Kommunikation zwischen Forscher und RemoteNEPS erfolgt über eine verschlüsselte Verbindung. Nutzer melden sich über ein biometrisches Authentifizierungssystem an, das sie eindeutig identifiziert (keystroke biometrics, TÜV-geprüft). Anschließend kann jeder Nutzer innerhalb einer individuellen Arbeitsumgebung (Windows-Desktop) auf die beantragten Daten zugreifen. Nach Abschluss der Analysen können Outputs angefordert werden, die das NEPS auf Einhaltung des Datenschutzes prüft und zeitnah versendet.“[33]

Micro Data Access steht auch im Mittelpunkt des diesjährigen Wissenschaftlichen Kolloquiums von Statistischem Bundesamt und Deutscher Statistischen Gesellschaft am 10./11. No-

[33] http://www.uni-bamberg.de/?id=43660

vember 2011 in Wiesbaden. Die Veranstaltung gibt einen Überblick über innovative Methoden zur kontrollierten netzbasierten Weitergabe von Forschungsdaten (statistical disclosure).

Innovationen *an* und *hinter* der Datenschnittstelle beeinflussen einander. Einerseits kann Remote Data Access die Zusammenarbeit an Daten im Rahmen einer VFU erleichtern, etwa die gemeinsame Verwaltung von Syntax und Metadaten, die Dokumentation des Forschungsprozesses und die Langzeitarchivierung von Ergebnissen. Andererseits werfen „virtuelle Organisationen" von Datennutzer/inne/n neue datenschutzrechtliche Probleme auf: Wie können virtuell vernetzte Forschungseinrichtungen die personellen, inhaltlichen und zeitlichen Nutzungsbeschränkungen garantieren, die Grundlage jeder wissenschaftlichen Nutzung von Forschungsdaten bleiben?

3. Kooperation an Sozial- und Wirtschaftsdaten

Die Entwicklung einer virtuellen Forschungsumgebung (VFU), die eine kollaborative Nutzung der Infrastruktur an Wirtschafts- und Sozialdaten unterstützt, muss an praktischen Problemen im Forschungsprozess ansetzen. Das zweite Verbundvorhaben „Berichterstattung zur sozioökonomischen Entwicklung Deutschlands“ (soeb 2) diente für die Evaluation der fachwissenschaftlichen Anforderungen als Anwendungsfall. Viele Erfahrungen aus diesem Verbundvorhaben dürften jedoch für die quantitativ-empirisch arbeitenden Sozialwissenschaften von allgemeiner Bedeutung sein.

3.1 Arbeitsabläufe

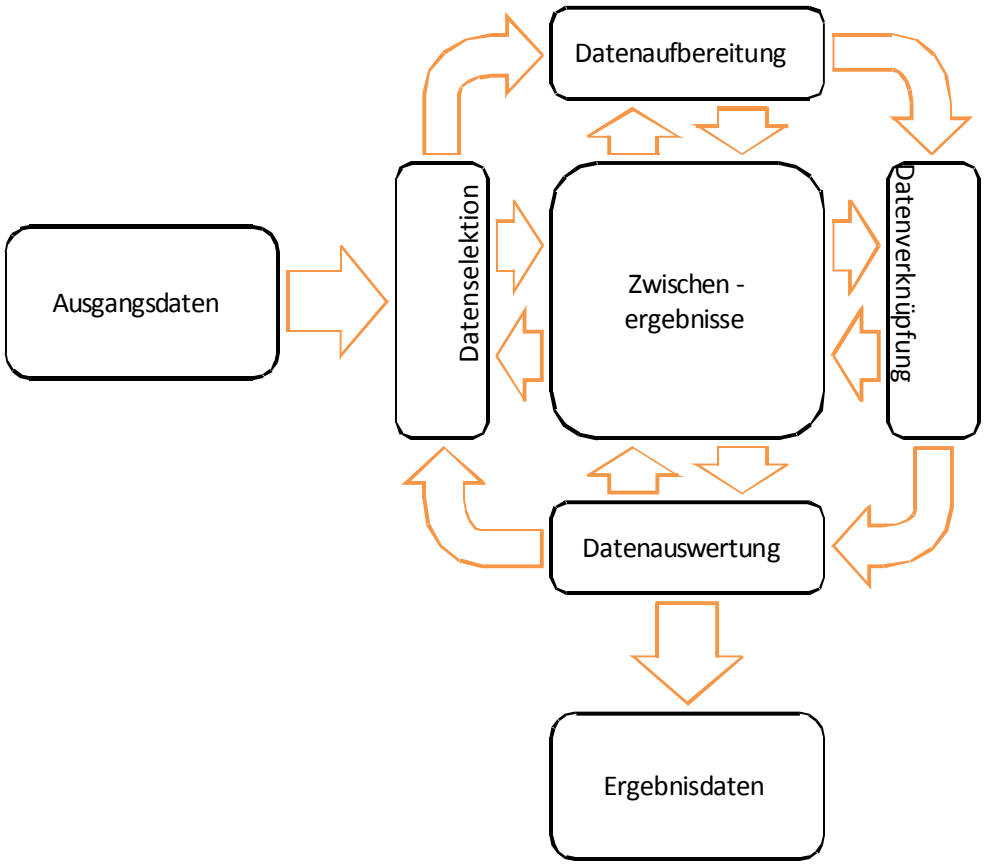
Eines der Ziele des Verbundvorhabens zur sozioökonomischen Berichterstattung besteht darin, neue Datenzugänge in der Forschungsdateninfrastruktur intensiv für einen Typ empirischer Forschung zu nutzen, der auf regelmäßige und wiederholbare Beschreibung gesellschaftlicher Entwicklungen angelegt ist. Die 23 empirischen Arbeitspakete von soeb 2 griffen auf eine Vielzahl verschiedenartiger Datensätze zu, und die zahlreichen Einzelentscheidungen, die in der Arbeit mit den jeweils genutzten Daten zu treffen sind, müssen im Verbund konzeptionell abgestimmt werden. Etwa sind Stichproben und Kohorten, Beobachtungszeiträume und –zeitpunkte nach einheitlichen Gesichtspunkten abzugrenzen, Mess- und Beobachtungskonzepte (z.B. Einkommensbegriffe, Haushaltstypen) festzulegen und Operationalisierungen für Indikatoren auf unterschiedliche Datensätze zu übertragen.

Abbildung 1 zeigt den datenbezogenen Ablauf des Forschungsprozesses in der vereinfachten Darstellung, die im Projekt auch den Ausgangspunkt für verschiedene Anwendungsszenarien in der Anforderungsanalyse für die VFU bildete. Ausgangsdaten können sowohl aggregierte Datensätze sein, in denen keine Individuen zu identifizieren sind, wie etwa Länderergebnisse der volkswirtschaftlichen Gesamtrechnung oder regionalisierte Indikatoren auf Kreisebene, als auch Individualdaten (Mikrodaten) aus Querschnitts- oder Längsschnittdatensätzen, etwa aus dem Mikrozensus oder dem SOEP, deren Nutzung datenschutzrechtlichen Anforderungen genügen muss.

Zu Beginn des Forschungsprozesses oder eines Arbeitsschritts werden aus den Ausgangsdaten die benötigten Variablen und Fälle selektiert und als Arbeitsdateien zwischengespeichert. Bei aggregierten Datensätzen können beispielsweise für internationale Vergleiche bestimmte Länder ausgewählt werden, im SOEP Subpopulationen verschiedener Beobachtungsjahre und verschiedene Variablen. Diese zwischengespeicherten Daten werden aufbereitet, z.B. indem Variablen generiert und recodiert werden, und anschließend wieder zwischengespeichert. Zudem können ggf. Verknüpfungen mit anderen Daten des gleichen Datensatzes nötig sein: etwa werden im Mikrozensus oder im SOEP Haushaltsinformationen jeder Person des Haushaltes zugeordnet. Erneut wird dieses Zwischenergebnis

gespeichert. Auf diese Arbeiten des Datenmanagements folgen dann einfache Analysen wie uni- oder bivariate Auszählungen für alle Datentypen oder komplexe Verfahren, etwa multivariate Querschnittanalysen, multivariate oder Sequenzenanalysen mit Längsschnittdaten.

Abbildung 1: Grundform des Workflows in der Arbeit mit Forschungsdaten



Quelle: Dickmann u. a. 2010: 9.

Alle oder wenigstens einige dieser Aktivitäten werden sich im empirischen Forschungsprozess wiederholen. Denn im Zuge der Forschungsarbeiten an den Daten müssen erste Ergebnisse betrachtet, Schlüsse gezogen und dann Veränderungen an verschiedenen Stellen im Analyseprogramm vorgenommen werden. Beispielsweise stellt sich im Laufe der Arbeiten heraus, dass mit der anfangs gewählten Stichprobenabgrenzung ein als wichtig erachtetes Merkmal aufgrund zu niedriger Fallzahlen nicht adäquat abgebildet werden kann. Auch komplexe Verfahren bedürfen häufig wiederholter Analyse- und Prüfdurchläufe mit Anpassungen des Analyseprogramms. Daher werden die abgebildeten Schritte häufig solange iterativ durchgeführt, bis am Ende der Arbeiten die endgültigen Ergebnisdaten in Form von Datensätzen, Output- und anderen Ausgabedateien abgespeichert werden. Jeder dieser Arbeitsschritte wird durch die Syntax des jeweiligen Statistikprogramms (z.B. SPSS, Stata, TDA)) gesteuert und dokumentiert. Die Ausgestaltung dieser Abläufe im Forschungs-

verbund kann danach unterschieden werden, an welchen Punkten des Ablaufs und wie intensiv „kollaborativ“ an Daten gearbeitet wird.

Im *ersten* Fall nutzen einzelne Projektpartner/innen aufgrund spezieller Datenkenntnisse oder aufgrund der speziellen Fragestellung in ihrem Arbeitspaket einzelne Datensätze allein. Die datenbezogene Zusammenarbeit beschränkt sich in diesem Fall auf Abstimmungen zu vereinbarten Standards und Operationalisierungen, die Vergleichbarkeit mit Ergebnissen anderer Arbeitspakete sicherstellen. Die Umsetzung dieser Entscheidungen bei der Datenselektion, Datenaufbereitung und Datenverknüpfung sowie alle weiteren iterativen Schritte im Rahmen der empirischen Arbeit bleiben einer Projektpartnerin oder einem Projektpartner überlassen. Ein Austausch findet erst wieder über Zwischenergebnisse oder Ergebnisse am Ende eines Arbeitsschritts statt, die als Outputs, als Excel-Dateien oder als fertige Textdateien bilateral auf Nachfrage ausgetauscht werden.

Ein *zweites* Anwendungsszenario entsteht, wenn mehrere Projektpartner/innen den gleichen Datensatz, wie z.B. den Mikrozensus, für verschiedene Arbeitspakete nutzen. Dabei arbeiten die Nutzungsberechtigten insofern zusammen, als sie einzelne generierte Variablen oder auch Teile der Syntax zur Erstellung spezieller gemeinsamer Variablen, beispielsweise einer gemeinsamen Haushaltstypologie, untereinander austauschen („Syntax Sharing“). Auch diese Vergleichs- und Tauschaktivitäten finden iterativ statt, jedoch überwiegend zwischen Projektbeteiligten, die sich in der Phase der Datenselektion, der Datenaufbereitung und der Datenverknüpfung befinden. Beispielsweise generiert eine Projektbeteiligte eine Syntax, in der eine bestimmte Altersgruppe ausgewählt wird und für sie ein kombinierter Elternschaft- und Erwerbsstatus anhand eines bestimmten Kindesalters festgelegt wird. Diese Syntax schickt sie dann an eine andere Projektbeteiligte, die diese Syntax in ihre integriert, um mit derselben Abgrenzung zu arbeiten. Auch hier bleibt die eigentliche Auswertungsarbeit, die Wahl des Statistikprogramms und der Analyseverfahren Sache jedes Projektbeteiligten.

Beim *dritten* Typ der Kollaboration ist die datenbezogene Arbeit am stärksten integriert. Dabei wenden mehrere Nutzungsberechtigte auf den gleichen Ausgangsdatensatz, wie z.B. das SOEP, auch die gleichen Auswertungsmethoden an. Etwa wurden für die Abteilung „Lebensverläufe im Umbruch“ von soeb 2 in drei Arbeitspaketen Verlaufsmuster des Jugendalters, der Haupterwerbsphase und des Altersübergangs mittels Sequenzanalyse deskriptiv analysiert und anschließend auf der Grundlage der gleichen Variante eines Optimal-Matching-Verfahrens (OMA) mittels Clusteranalysen statistisch gruppiert. Darüber hinaus war die Berechnung vergleichbarer Turbulenzkennziffern mit dem gleichen Software-Tool vereinbart. Bei einer solchen Arbeitsweise bezieht sich die Kollaboration iterativ auf alle Phasen des Ablaufs: auf harmonisierte Datenselektionen, Variablenrecodierungen und –generierungen, Datenverknüpfungen und insbesondere harmonisierte Analysen. Im Detail werden die verwendeten Verfahren einschließlich der festgelegten Parameter und der Hochrechnungsfaktoren harmonisiert und festgelegt und von den einbezogenen Projektbe-

teiligten mit Blick auf ihre jeweilige Unterstichprobe und ihre Forschungsfrage umgesetzt. Die Beteiligten entwickeln und nutzen Syntaxbestandteile, speziell angepasste Softwarekomponenten oder Spezialprogramme[34] gemeinsam und diskutierten Syntaxdateien und Teilarbeitsdatensätze mit ausgewählten Variablen und Matchingvariablen bilaterial oder in persönlichen Arbeitstreffen.

In allen verschiedenen Kollaborationsszenarien, welche von der virtuellen Forschungsumgebung zu unterstützen sind, müssen Mikrodatensätze aus Personen- und Haushaltssurveys, Verwaltungsregistern oder Unternehmensbefragungen durch Forschungsdatenzentren (FDZ) oder durch Daten haltende statistische Ämter, Forschungseinrichtungen oder Verwaltungen zur Verfügung gestellt werden. Im ersten Fall würde eine VFU vor allem die Generierung und Verwaltung von Meta- und Paradaten[35] zu allen Arbeitsdatensätzen und den Austausch darüber unterstützen. Dabei geht es nicht nur um technische Daten, wie beispielsweise das Entstehungsdatum, sondern auch um inhaltliche Datenbeschreibungen, die von den Nutzer/innen eigene Einträge erfordern. In den beiden anderen, stärker kollaborativen Arbeitsformen sind diese Hilfsmittel für spätere Re-Analysen ebenfalls erforderlich. Jedoch steht hier die Unterstützung der gemeinsamen Arbeit an und mit Syntax im Vordergrund.

3.2 Anforderungen von Forschungs- und Dateneinrichtungen

Anforderungen an eine Virtuelle Forschungsumgebung lassen sich aus zwei Perspektiven formulieren, aus Sicht von Forschungseinrichtungen und aus Sicht von Einrichtungen der Forschungsdateninfrastruktur. Projektbeteiligte können fallweise beide Perspektiven einnehmen, da Wissenschaftler/innen aus Dateneinrichtungen auch mit eigenen Forschungsbeiträgen in die sozioökonomische Berichterstattung oder andere Verbundvorhaben einbezogen sind.

Als Ausgangspunkt für die Evaluation von Grid-Technologien und für die Verständigung über Funktionen einer VFU war im Rahmen des Projekts eine Anforderungsanalyse zu erstellen (vgl. Technische Evaluation: 16 ff.). Grundlage hierfür waren zum einen Interviews zu Arbeitserfahrungen von Projektbeteiligten aus Forschungseinrichtungen, die für soeb 2 Forschungsdaten genutzt hatten (sieben Interviews mit Datennutzer/innen[36], zwei Gesprä-

[34] Beispielsweise spezielle Ado-files aus Stata oder das Spezialprogramm CHESA.
[35] Daten über den Prozess der Datenerhebung, vgl. Kreuter/Casas-Cordero2010.
[36] DatennutzerInnen wurden zu folgenden Themenbereichen befragt: Nutzungswege für Datensätzen, Programme, Analysedesign und Methoden, Probleme mit Rechnerleistung, Erfahrungen mit Dokumentation,

che mit Projektbeteiligten aus Dateneinrichtungen und ein weiteres Interview mit einem FDZ-Mitarbeiter[37]), zum anderen die vier Workshops zu VFU (vgl. oben: 1.), die in 2010 im Rahmen des Projekts stattfanden.

Datenzugang (Datenbereitstellung)

In der Arbeit an soeb2 wurden alle derzeit bestehenden Wege des Zugriffs auf Forschungsdaten genutzt, wenn auch unterschiedlich häufig. Dabei wurden ausschließlich individuelle Nutzungsverträge zwischen Projektbeteiligten und den Dateneinrichtungen geschlossen.

Die Arbeit mit Scientific Use Files (SUF) auf den lokalen Rechnern der Projektbeteiligten war der häufigste Nutzungsweg. Datenbezogene Kooperation war zwischen den Projektbeteiligten am intensivsten, die den gleichen SUF nutzten. Denn an solchen SUF kann ausgetauschte oder gemeinsam entwickelte Syntax immer wieder getestet werden. Da alle Beteiligten die Auswertungen ihrer Projektpartner/innen reproduzieren können, ist eine gemeinsame Validierung und Qualitätskontrolle bis zu den letzten Auswertungsschritten möglich.

Datensätze, für die SUF nicht verfügbar waren oder aus inhaltlichen Gründen nicht ausreichten, wurden nur von einzelnen, spezialisierten Projektbeteiligten für ihr jeweiliges Arbeitspaket genutzt. Die meisten Projektbeteiligten vermieden die Arbeit mit Onsite-Ausgangsdaten: zum einen wegen des damit verbundenen zusätzlichen Zeitaufwands (terminlich abzustimmende Fahrten zu Gastarbeitsplätzen, zeitverzögerte Aushändigung kontrollierter Outputs), zum anderen, da iterative Arbeitsabläufe bei einer explorativen Syntaxentwicklung durch die Bedingungen der Nutzung verlangsamt wurden. Erst wenn nach einer Wartezeit die kontrollierten Outputs ausgehändigt werden, können Ergebnisse und Probleme (etwa abweichende Randverteilungen oder unplausible Resultate) wieder mit anderen Projektbeteiligten diskutiert werden. Auch dies erschwerte eine kollaborative Arbeitsweise, etwa die gemeinsame Fehlersuche. Zudem entstehen bei den Projektbeteiligten keine Arbeitsdateien, die ausgetauscht werden könnten.

Datenfernanalysen mit Onsite-Ausgangsdaten fanden also nur in dem ersten der drei oben beschriebenen Anwendungsszenarien statt, womit sich die Kooperation zu diesen Datensätzen auf konzeptionelle Absprachen beschränkte. Beim Versuch, den Mikrozensus an

---

Output, Ergebnisverwendung, Bewertung der Zusammenarbeit im Verbund und der Arbeitsabläufe, andere Erfahrungen in virtueller Zusammenarbeit, Einschätzung der Möglichkeiten von VFU.

[37] Mitarbeiter/innen von Dateneinrichtungen wurden zu folgenden Themenbereichen befragt: Struktur des FDZ, Datenangebot, Nutzungsberechtigungen, Formen der Datenbereitstellung, Datenpflege / Betreuung, Hardwareausstattung, beabsichtigte weitere Entwicklung der Datenangebote und der Datenzugangswege.

einem FDZ-Gastarbeitsplatz kollaborativ zu nutzen, zeigte sich, dass für die erforderlichen Qualitätskontrollen, Korrekturen und Anpassungen bei der Übertragung von Auswertungssyntax von SUF auf Onsite-Files deutlich mehr Zeit einzuplanen ist. (Diese Abläufe verursachen auch in den FDZ erheblichen Arbeitsaufwand, da die Outputs immer wieder kontrolliert werden müssen.) Aus Zeitgründen wurde die kollaborativ erstellte Auswertungssyntax für soeb 2 schließlich allein mit den verfügbaren SUF erstellt.

Aus Sicht der Datennutzer/innen ergeben sich vor allem zwei Anforderungen an VFU: Scientific Use Files und daraus erzeugte Arbeitsdateien sollten für Nutzergruppen in die gemeinsame Arbeitsumgebung (Archivstruktur) integriert sein. Und das kontrollierte Fernrechnen mit FDZ-Datenbeständen sollte in diese Arbeitsumgebung integriert werden, um eine kollaborative Arbeit an Onsite-Ausgangsdaten besser zu unterstützen.

Auch die Vertreter/innen der an der Diskussion beteiligten Dateneinrichtungen melden das Interesse an, die Datenbereitstellung und die damit verbundenen Datendienstleistungen zu verbessern. Dabei stehen für sie aber die Belange des Datenschutzes und die rechtlichen Nutzungsbeschränkungen im Vordergrund. Für alle Forschungseinrichtungen, die in einer VFU kooperieren, müssen Einzelnutzungsverträge mit den (berechtigten) Dateneinrichtungen bestehen. Die Vertragsgestaltung setzt verantwortliche (aufsichtspflichtige) Antragsteller voraus, die in einer hierarchischen Arbeitsorganisation unterbinden können, dass Mitarbeiter/innen Datenschutzanforderungen verletzen, und die eine Nutzung ausschließlich für den vertraglich bestimmten Forschungszweck an lokalen Standorten sicherstellen können. Die Stellen und Personen, die personenbezogene Daten bearbeiten, sind in den Nutzungsvereinbarungen zu nennen. Für die Frage, wie personelle Nutzungsbeschränkung, Befristung, Zweckbindung und Genehmigungsvorbehalt in einer VFU gewährleistet werden können, unterscheiden die Dateneinrichtungen zwei Szenarien.

Szenario 1: Nutzungsberechtigte greifen auf SUF und daraus erzeugte Arbeitsdateien in einer gemeinsamen Arbeitsumgebung zu. Diese Lösung, die für viele Nutzer/innen die praktischste wäre, ist datenschutzrechtlich die problematischere. Nur eines der im Projekt einbezogenen FDZ hält dies für möglich, sofern der Zugriff gegen nicht nutzungsberechtigte Dritte abgegrenzt ist. Ein anderes FDZ wägt ab, dass eine VFU zwar die in Nutzungsverträgen unterstellte Hierarchie schwächt, andererseits aber einen weniger privaten Raum für die Datennutzung schafft und damit die Einhaltung der Nutzungsbeschränkungen eher unterstützt. Bei der Bewertung einer solchen Lösung bleiben für die FDZ jedoch zu viele Fragen offen.

- Kann man einen gemeinsamen Ort des Datenaustauschs[38] für SUF und Arbeitsdateien in individuelle Datennutzungsverträge mit Partnereinrichtungen einer VFU aufnehmen?
- Wird mit den Forschungsdaten in der gemeinsamen Arbeitsumgebung gerechnet, etwa um Rechnerkapazität gemeinsam zu nutzen?
- Sind bestehende Datenschutzkonzepte beteiligter Forschungseinrichtungen auf eine VFU übertragbar? (Ist dies nicht der Fall, müsste das Sicherheitskonzept für die Datenaustauschplattform von den FDZ, die Daten bereitstellen sollen, zeitaufwändig und mit unsicherem Ergebnis geprüft werden.)
- Handelt es sich bei dem VFU-Provider um eine beteiligte Forschungseinrichtung? (Mit dem Provider einer VFU, der nicht zur scientific community gehört, hätten die FDZ kein Vertragsverhältnis. Lägen die Daten bei ihm, betriebe er eine rechtlich nicht zulässige Datenvorratshaltung.)
- Wann dürfen Mikrodaten aus der VFU auf die Workstations der Beteiligten gespeichert werden? Lässt sich abgleichen, was auf dem Repository, was auf Workstations vorhanden ist?

Beim bisherigen Diskussionsstand zeichnet sich ab, dass für die meisten Forschungsdaten eine solche Lösung kaum zu verwirklichen sein wird.

Szenario 2: Die VFU unterstützt Datenfernverarbeitung (remote access processing) über eine remote-access-Plattform. Bei dieser Lösung würden die Daten das FDZ gar nicht verlassen, damit ist auch die Gefahr durch unsichere Leitungen gebannt. Eine gemeinsame Remote-Access-Architektur wäre aus Sicht der Dateneinrichtung die ideale Lösung; sie träfe sich auch am ehesten mit ihren eigenen Vorstellungen über die Verbesserung der Datenbereitstellung (vgl. oben: 2.4).

„Syntax Sharing“ und Metadaten

Aus Sicht der einbezogenen Forschungseinrichtungen könnte die Datenverwaltung in einer VFU vor allem die gemeinsame Nutzung von Syntax für Statistikprogramme („Syntax Sharing“), die Dokumentation von Syntax und Forschungsdaten (Metadaten und Paradaten) und die nachnutzbare Archivierung unterstützen.

Einerseits liegt der wichtigste potenzielle Vorteil, den Arbeit im Verbund bietet, in der gemeinsamen Arbeit an Auswertungssyntax, etwa für generierte Variablen und in deren

[38] Unproblematisch wäre lediglich eine sichere Online-Bereitstellung von Scientific Use Files an Nutzungsberechtigte. Sie wäre einem Postversand sogar vorzuziehen.

Austausch und Nachnutzung. Andererseits setzt ein solches „Syntax Sharing“ gemeinsame Dokumentationsstandards voraus. Das wichtigste Dokumentationsformat bilden Syntaxdateien (z.B. Stata-do-files) mit Kommentaren zu Arbeitsschritten im ASCII-Format. Damit Syntaxdateien für andere nachnutzbar sind, müssen in ihnen die Arbeitsschritte erläutert und die dafür genutzten Variablen nachgewiesen werden. Doch während die FDZ für Syntax im Datenfernzugriff oder an Gastarbeitsplätzen Mindeststandards vorgeben, hing es in der Verbundarbeit vom persönlichen Arbeitsstil ab, wie dokumentiert wurde. Die individuell sehr unterschiedliche Dokumentationspraxis der Beteiligten begrenzte die Nachnutzbarkeit zwischen den beteiligten Einrichtungen und im Fall des Weggangs von Mitarbeiter/inne/n selbst innerhalb einer Forschungseinrichtung. Projektbeteiligte, die Syntax nachnutzen wollten, konnten nicht sicher sein, die jeweils aktuellste Version zu nutzen oder über Updates informiert zu werden. Beschränkt wird die Nachnutzung von Syntax auch durch Übersetzungsprobleme, die sich durch Nutzung verschiedener statistischer Software, wie SPSS, Stata, TDAund Newspell ergeben. Probleme bei der Migration in andere Formate wurden verbundintern durch bilateralen Austausch, mündliche Syntaxerläuterungen sowie manuelle Konvertierung gelöst.

Einige Verbundpartner/innen verfassten ergänzend zur Syntax Protokolle und kurze Arbeitspapiere in Word oder Excel zu Operationalisierungen. Eher selten wurden diese Produkte im internen Bereich auf der soeb-Website für die anderen Projektbeteiligten hinterlegt, überwiegend wurden sie bilateral auf Nachfrage per Email ausgetauscht.

Auch die Dateneinrichtungen haben ein Interesse an der Entwicklung von Dokumentationsstandards, die eine problembezogene Recherche ermöglichen. Je mehr generierte Variablen auf ihren Web-Portalen hinterlegt werden, desto wichtiger wird für die wissenschaftlichen Nutzer/innen, dass die in ihnen hinterlegten Vorannahmen dokumentiert sind und nachvollzogen werden können. Auch Imputationen sollten gut dokumentiert sein. Die Dateneinrichtungen verweisen auch darauf, dass Peer Reviewer wissenschaftlicher Beiträge künftig ebenfalls den Nachweis der verwendeten Daten und der Syntaxen verlangen werden. Einige FDZ beabsichtigten daher ihrerseits eine Verbesserung bestehender Serviceleistungen (z.B. SoepInfo, PanelWhiz[39]) und deren Erweiterung um „Paradaten“ (z.B. Referenz-

[39] Das Tool „Panelwhiz“ soll für das SOEP und Paneldatensätze des FDZ-IAB die Stichprobenzusammenstellung und −Abgrenzung mit gleichzeitiger Syntaxgenerierung unterstützen. Voraussetzung für die Nutzung ist, dass ein Datennutzungsvertrag geschlossen wurde und die Daten selbst auf der lokalen Festplatte vorliegen. Es soll zukünftig auch die Option bieten, selbst generierte Syntax online einzustellen, mit der Zusicherung, dass die Urheber/innen bei Verwendung durch Dritte zitiert werden. (http://www.panelwhiz.eu/)

werte, Fragenkontext und Referenzstudien). Eine weitere wichtige Aufgabe sehen die Dateneinrichtungen darin, die explorative Syntaxentwicklung für Originaldaten, die nur onsite zugänglich sind, durch verteilungstreue Strukturdatensätze zu unterstützen (vgl. hierzu Brandt/Zwick 2011).

Für eine zukünftige kollaborative Arbeitsweise ergibt sich als Anforderung, Standards für eine detaillierte Dokumentation aller Arbeitsschritte im empirischen Forschungsprozess, von der Operationalisierung der zu analysierenden Konstrukte über Entscheidungen zum Analysedesign bis zur Umsetzung in Syntax zu verabreden und diese Dokumentationen verbundintern zu archivieren. Dies dient auch der Qualitätssicherung: Die Regeln guter wissenschaftlicher Praxis verlangen nicht nur die Sicherung und Aufbewahrung von Primärdaten, sondern auch die Dokumentation von Resultaten und die Reproduzierbarkeit des Wegs zum Ergebnis.[40] Eine VFU kann die Einhaltung solcher Standards dadurch unterstützen, dass diese in einer durchsuchbaren gemeinsamen Arbeitsplattform hinterlegt sind, welche die notwendigen Angaben extrahiert oder von den Nutzer/innen abfragt. Da in der gleichen Arbeitsumgebung auch Arbeitsdateien und andere Dokumente abgelegt werden können, könnte Syntax ggf. auch zu den Daten verlinkt sein, auf die sie angewendet wurde. Eine Dateiverwaltung, die dies unterstützt, muss jedoch ein intuitiv nachvollziehbares Ordnungssystem verwenden. Nutzer/innen müssen sicher sein, dass die angebotene Syntax aktuell ist, und sie müssen leicht erkennen können, welchen Status sie hat (z.B „work in progress“) und ob eine Qualitätsprüfung stattgefunden hat.

Allerdings verweisen sowohl die einbezogenen Projektbeteiligten als auch die Dateneinrichtungen darauf, dass dem „Syntax Sharing“ und der Arbeit mit verbindlichen Dokumentationsstandards nicht nur technische Probleme entgegenstehen. Gute Dokumentation ist zeitaufwändig, und wird dieser Aufwand bei der Projektplanung und -finanzierung nicht angemessen berücksichtigt, wird sie in der Praxis am ehesten eingespart. Arbeitsunterlagen, die allgemein zugänglich eingestellt werden, müssen auch für Projektbeteiligte verständlich sein, die nicht direkt an einer Datenauswertung kooperieren, möglichst sogar unabhängig von bestimmten Statistikprogrammen, und für solche Aufbereitungen fehlt meist die Zeit. Für die beteiligten Wissenschaftler/innen besteht ein erheblicher Teil ihrer Leistung in Syntax, und sie behalten sich nicht nur die Urheberschaft daran vor, sondern auch

[40] „Der primäre Test eines wissenschaftlichen Ergebnisses ist seine Reproduzierbarkeit. Je überraschender, aber auch je erwünschter (im Sinne der Bestätigung einer liebgewordenen Hypothese) ein Ergebnis ist, um so wichtiger ist die unabhängige Wiederholung des Weges zu ihm in der Gruppe, ehe es außerhalb der Gruppe weitergegeben wird.“ (DFG 1998: 8.)

die Entscheidung, zu welchem Zeitpunkt sie diese wem zugänglich machen. Viele Beteiligte äußern die Sorge, aus ihrer Sicht unfertige oder nicht hinreichend geprüfte Syntax könnte ohne ihr Wissen genutzt werden, sie könnten für fehlerhafte Ergebnisse anderer verantwortlich gemacht werden, oder die Bilanz von Leistung und Gegenleistung könne für sie nicht aufgehen. Dass es eine wechselseitige Qualitätskontrolle geben müsse, äußern viele Befragte, aber sie wollen nur Arbeitsschritte offen legen, die aus ihrer Sicht „als fertig abgehakt“ sind. Das Hochladen von Syntax oder von Zwischenergebnissen setze Vertrauen voraus, und die Projektbeteiligten müssten die Kontrolle über ihre Arbeitsergebnisse behalten.

Auch die Dateneinrichtungen schätzen bisherige Erfahrungen mit „Syntax-Sharing“ als nicht ermutigend ein: Mit Syntax zitiert zu werden, sei hierfür kein hinreichend starker Anreiz. Die FDZ beklagen, dass sie bislang von den Nutzer/innen kaum Informationen über generierte Variablen oder Erkenntnisse über Datenmängel und Datenqualität zurückbekommen; vielfach erhielten sie selbst die vertraglich zustehenden Publikationen nicht. Metadatensysteme mit Wiki-Elementen setzen aus ihrer Sicht generell eine andere Bereitschaft der Nutzer/innen voraus, sich an „kollektiven Gütern“ zu beteiligen.

Zusammengefasst ergibt sich als Anforderung an eine VFU, dass sie Syntax, Arbeitsunterlagen und Arbeitspapiere sowie – soweit dies mit den Datenschutzanforderungen vereinbar ist – Arbeitsdatensätze und Outputdateien für die gemeinsame Arbeit im Verbund mittels durchsuchbarer Metadaten verwaltet und teilweise editierbar und nachnutzbar verfügbar macht. Dabei müssen jedoch die Projektbeteiligten in einem abgestuften System projektinterner Öffentlichkeit die Kontrolle über ihre Arbeitsergebnisse behalten und festlegen können, für wen diese wann einsehbar und editierbar sind.

Archivierung

Die Langzeitarchivierung von Arbeitsdateien und Outputs ist für die kollaborativen Arbeitabläufe im Verbund von geringerer unmittelbarer Bedeutung als die Dokumentation und Archivierung von Syntax, Paradaten und Arbeitsdateien. Die Datenverwaltungsfunktionen einer VFU könnten die beteiligten Forschungseinrichtungen dabei unterstützen, ihren Aufbewahrungspflichten zum Nachweis guter wissenschaftlicher Praxis nachzukommen. Für einen erheblichen Teil der zu sichernden Daten käme jedoch die Datenverwaltung einer VFU nicht in Betracht: nämlich immer dann, wenn die Nutzungsbedingungen für den jeweiligen Forschungsdatensatz die Speicherung in einem gemeinsamen Repository außerhalb der nutzungsberechtigten Einrichtung ausschließen. (Diese Beschränkung gilt dann nicht nur für die Ausgangsdaten, sondern auch für daraus erzeugte Arbeitsdateien.) Ein weiteres Problem ergibt sich daraus, dass für einen großen Teil der wissenschaftlich genutzten Forschungsdaten Löschfristen vereinbart sind. Aus beiden Gründen kann eine VFU die Archivierung nach den Regeln guter wissenschaftlicher Praxis (DFG 1998) nur für Syntaxdateien in vollem Umfang übernehmen.

Für die FDZ ist die Befristung ein entscheidender Teil der Nutzungsbedingungen. In einem größeren Projekt sollte daher die Nutzung für vier oder fünf Jahre beantragt werden. Jedoch weisen sie auch darauf hin, dass bei der bisherigen Vertragsgestaltung die Löschung nur durch Eigenerklärungen der Nutzungsberechtigten gesichert ist. In einer VFU wäre die fristgemäße Löschung dagegen auch technisch gesichert und nachvollziehbar.

Einige FDZ bieten den Nutzer/inn/en nach Ablauf der Nutzungsfrist für Arbeitsdateien an, ihre Daten wiederaufrufbar und langfristig (zehn bis 20 Jahre) zu archivieren. Sie gewährleisten, dass die von den Nutzer/innen entwickelte Syntax auf den Ursprungsdatensätzen lauffähig bleibt. Die FDZ verweisen hierzu auf ihre Kompetenz bei Datenmanagement, Versionierung und Versionskontrolle. Eine solche Archivierung bei den FDZ dürfte daher für einen Teil der zu archivierenden Daten eine bessere Lösung versprechen als die Langzeitachivierung in einer VFU.

Datenverarbeitung

Der größte Teil der Rechenoperationen mit den Forschungsdaten, die in der Arbeit an *soeb 2* genutzt wurden, ließ sich mit den lokalen Rechnern an den Arbeitsplätzen der Projektbeteiligten durchführen. Ergaben sich bei großen Arbeitsdateien und rechenaufwändigen Analyseverfahren (etwa bei Sequenzanalysen) zu lange Rechenzeiten, wurden die lokalen Kapazitäten durch Upgrades sowie durch Neuanschaffungen erweitert, oder es wurde auf mehreren Arbeitsplatzrechnern der Einrichtung parallel gearbeitet. Bei sehr großen Ausgangsdatensätzen wurde die Auswertung von vornherein über Datenfernzugriff geplant und ausgeführt. Ein gleichzeitiger Zugriff mehrerer Projektbeteiligter auf Ausgangsdatensätze oder Arbeitsdateien kommt in der Arbeit bislang praktisch nicht vor.

Eine VFU sollte die Möglichkeit, aufwändige Operationen in der Grid-Umgebung durchzuführen, technisch unterstützen. Es ist jedoch zu erwarten, dass der größte Teil der Rechenoperationen auch künftig vor allem an lokalen Workstations und über Fernrechnen in den Dateneinrichtungen durchgeführt wird.

Arbeit im Verbund – nicht nur ein technisches Problem

In den Gesprächen über Anforderungen an eine gemeinsame VFU wurde von Projektbeteiligten und Expert/inn/en immer wieder angesprochen, dass abgestimmte Datenanalysen verschiedener Projektbeteiligter in Verbundvorhaben zwar technisch besser unterstützt

werden sollten, jedoch vernetztes Arbeiten auch eine ganze Reihe nichttechnischer Voraussetzungen hat.

In einer VFU werden bestimmte Verhaltensanforderungen technisch „hinterlegt". Die Arbeitsumgebung muss aber weiterhin individuell verschiedene Arbeitsweisen zulassen und unterstützen. Die wissenschaftliche Freiheit der Beteiligten und ihr Interesse an selbständiger kreativer Forschung sind produktiv und daher zu erhalten. Das Ziel kann nicht sein, Datenanalysen und Vorgehensweisen zu „zentralisieren".[41]

> „Ich glaube, dass jeder eine gewisse Idee und Fragestellung mit seinen Sachen verbindet und es wichtig ist, um gute Arbeit zu machen, dass man die auch hat und dass man die verfolgt. Ansonsten könnte man eigentlich auch arbeiten wie [Institut]. Da hast Du einen Projektleiter, der sagt Dir genau, also der sagt dir, das und das will ich wissen, ganz genau, und schickt es unten in den Keller, und irgendein Datenknecht schickt ihm das dann hoch. Aber so verstehen wir uns als Verbund nicht. Deswegen finde ich es wichtig: Man muss, wenn man an Daten arbeitet, wenn man sie interpretiert und wenn man Wissenschaft macht, ein Erkenntnisinteresse daran haben und muss eine eigene Fragestellung haben, und die muss man auch entwickeln und hinter der muss man auch stehen, ansonsten macht man keine gute Wissenschaft. Deswegen glaube ich, kann man viele Sachen diskutieren, man kann sich nicht immer auf alles einigen und dann ist irgendwann auch mal die Frage, wer es entscheidet." (Interview 4.)

Es ist anzunehmen, dass vor allem solche Teilarbeitsprozesse durch eine VFU verbessert werden können, die heute schon einen gewissen Grad an Standardisierung aufweisen. Andere Qualitäten verbundener Forschungsprozesse hängen stärker von Koordination und Kommunikation ab als von Technik. Die das Gesamtvorhaben betreffenden konzeptionelle Abstimmung über Messkonzepte im weitesten Sinne, also über Begriffe, Definitionen, Abgrenzungen und Operationalisierungen, und die Verständigung über themenübergreifende Hypothesen, die inhaltliche Beziehungen zwischen Auswertungsschritten in Arbeitspaketen herstellen, muss zu gemeinsamen Zeitpunkten im Arbeitsprozess mit ausreichend Zeit stattfinden. Anschließend müssen datenbezogene Kooperationen zwischen Arbeitspaketen in Arbeitstreffen der jeweils unmittelbar Beteiligten persönlich abgestimmt werden, und Verabredungen über Zulieferungen zwischen Arbeitspaketen sind verbindlich einzuhalten.

Vernetzungsangebote werden nur genutzt, wenn sie praktische Bedürfnisse treffen und die individuelle Arbeit unterstützen. So stand den Projektbeteiligten für die Arbeit an soeb 2 bereits eine Austauschplattform in einem passwortgeschützten Bereich der Projekt-Website zur Verfügung. Diese wurde uneinheitlich und insgesamt zu wenig genutzt. Als Gründe hierfür wurden neben dem vergleichsweise schwierigen Upload-Verfahren im ver-

[41] Sennett (2011: 102f.) führt das Scheitern der Plattform GoogleWave, die das gemeinsame Arbeit im Netz gestalten sollte, darauf zurück, dass die Entwickler von einer zu einfachen, linearen Vorstellung von praktischer Zusammenarbeit ausgegangen seien und ihre Benutzeroberfläche „laterales Denken" nicht unterstützte.

wendeten Content-Management-System auch Vorbehalte genannt, Zwischenergebnisse und unfertige Arbeitsunterlagen verbundintern zugänglich zu machen, ferner die mangelnde Verbindlichkeit dieses Kommunikationswegs und eine unzureichende Moderation der Plattform.

Erst wenn eine bessere technische Unterstützung mit einer stärkeren Organisation und Steuerung arbeitsteiliger Forschungsprozesse einhergeht, kann sie ihr Potenzial, eine stärker kollaborative Nutzung von Forschungsdaten zu unterstützen, wirkungsvoll entfalten. Umgekehrt ist zu erwarten, dass die neue Qualität der IT-technischen Vernetzung Aufgaben der Organisation und Koordination bewusster macht, indem es sie stärker formalisiert, und ihnen mehr Aufmerksamkeit sichert. Schließlich ist zu berücksichtigen, dass neue Technik auch neue Probleme mit sich bringt. Viele Funktionen einer VFU bedürfen einer Moderation, und auch hierfür sind ausreichend Personalressourcen zu veranschlagen.

4. Funktionen einer virtuellen Forschungsumgebung

Auf Basis der Technischen Evaluation zur GRID-Technologie (Dickmann, Enke, Harms 2010) und der daran anschließenden Diskussionen in den Workshops des Projekts werden im Folgenden die Hauptfunktionen einer Virtuellen Forschungsumgebung (VFU) für die sozioökonomische Berichterstattung als exemplarischer Anwendungsfall dargestellt. Die Technische Evaluation unterscheidet zwischen Funktionen der VFU und den technischen Komponenten des Systemkerns und der grafischen Nutzerschnittstelle, wobei die Komponenten jeweils mehrere zusammengehörige Funktionalitäten umfassen. In der nachfolgenden nichttechnischen Darstellung liegt der Schwerpunkt auf den Funktionen selbst, nicht auf ihrer Umsetzung in „Werkzeuge".

Aufgabe der Technischen Evaluation war es, einen Überblick über bestehende Grid-Technologien zu geben, die für die Entwicklung einer VFU zur Nutzung von Forschungsdaten geeignet sind. Um dies leisten zu können, skizzierte die Technische Evaluation zugleich bereits die mögliche Architektur einer solchen Arbeitsumgebung. Unter Grid-Technologie werden dabei alle Komponenten verstanden, die IT-Ressourcen, wie Rechner- und Speicherkapazität, aber auch Anwendungen und Daten der im Grid befindlichen Rechner vernetzen. Als Grid-Middleware wird eine bestimmte Kombination von Softwarekomponenten bezeichnet, „mit der die transparente Nutzung von Computern und der von ihnen bereitgestellten Ressourcen ermöglicht wird" (Technische Evaluation: 39). Wenn im Folgenden von einer Grid-Umgebung die Rede ist, wird also weder die Nutzung eines besonderen Hochleistungsnetzes noch eines bestimmten Providers aus der D-Grid-Initiative unterstellt.

4.1 Datensicherheit

Die Sicherheitsarchitektur der VFU muss allen Datenschutzanforderungen der Dateneinrichtungen Rechnung tragen. Um diese Anforderung auch für off-site-SUF-Files und daraus generierte Arbeitsdaten zu erfüllen, die nicht per Remote-Onsite-Rechnen (Fernrechnen), sondern in der VFU verwaltet und bearbeitet werden können (vgl. hierzu 4.2), wird die IT-Infrastruktur auf der Grundlage eines Service-Level-Agreements auf dem Server eines möglichst forschungsnahen Providers betrieben.[42]

[42] Z.B. die Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG), die als Hochschulrechenzentrum für die Georg-August-Universität Göttingen und als Rechen- und IT-Kompetenzzentrum für die Max-Planck-Gesellschaft arbeitet.

Mit der Grid Security Infrastructure (GSI) ist ein umfassendes Sicherheitssystem für das Arbeiten im Grid und für den Zugang zum Grid und seinem sicheren Server verfügbar. Innerhalb des Grids wird durch die GSI jeder Datentransfer verschlüsselt. Der Zugang zum Grid wird über die sog. Public Key Infrastructure (PKI) geregelt, ein persönliches Zertifizierungssystem, das durch die EUGrid/PMA[43] und damit EU-weit akzeptiert ist und der ISO-Norm und den Vorgaben des Bundesamtes für Sicherheit in der Informationstechnik (BSI) entspricht.

Mit der GSI sollte sich auch ein Datenfernzugriff und ein Remote-Rechnen (Fernrechnen) in den FDZ in die Nutzerschnittstelle der VFU integrieren lassen. Dabei könnte es jedoch zusätzlich erforderlich sein, durch die physische Prüfung am Standort oder durch andere Authentifizierungsverfahren (zu „keystroke biometrics“ vgl. oben: 2.4) sicherzustellen, dass die Person, die das Zertifikat nutzt, auch die berechtigte Person ist.[44]

Eine wesentliche Funktion der Sicherheitskomponenten besteht darin, Virtuelle Organisationen (VO) im Grid zu definieren, für die abgestufte Zugangsrechte festgelegt und verwaltet werden. Der Forschungsverbund als ganzer bildet eine solche Organisation, innerhalb derer die Beteiligten unterschiedlichste Rollen mit entsprechenden Rechten einnehmen. Vor allem aber können VO für einzelne Arbeitspakete oder bestimmte Forschungsdaten definiert werden. Der Zugriff auf einzelne Forschungsdatensätze kann so auf die VO der Projektbeteiligten beschränkt werden, die mit der jeweiligen Dateneinrichtung einen individuellen Nutzungsvertrag haben. Ein VO- bzw. Account-Management stattet die verschiedenen VO sowie die einzelnen Mitarbeiter/innen im Grid über einen Virtual Organization Management Registration Service (VOMRS) mit den für sie zutreffenden Rollen und Rechten aus.[45] Alle Aktivitäten von Nutzer/innen im Grid sind nachvollziehbar, und einzelne Aktionen können gegenüber den Aktivitäten anderer Nutzer/innen abgeschottet werden. So ist gewährleistet, dass nur nutzungsberechtigte Projektbeteiligte Lese-, und / oder Schreibrechte zu den jeweiligen Ausgangsdaten (z.B. Offsite-SUF-Dateien) und den daraus entstandenen und verknüpften Arbeitsdatensätzen erhalten. Zusätzlich kann durch die vollständige Protokollierung von Änderungen an Daten in der virtuellen Arbeitsumgebung sicherge-

[43] European Grid Policy Management Authority - www.eugridpma. org

[44] Im Projekt „FDZ in FDZ“ geschieht dies in den Räumen des jeweils nächstgelegenen FDZ.

[45] „Die Anbieter beziehen von den VOMRS die Informationen über die zugelassenen Mitglieder und Gruppen und bilden diese in den lokalen Nutzerberechtigungen ab. Diese Informationen werden täglich neu vom VOMRS bezogen.“ (Technische Evaluation: 57)

stellt und verifiziert werden, dass die maximalen Aufbewahrungszeiten für die Ausgangsdaten (z.B. SUF) nicht überschritten werden.

4.2 Zwei Schnittstellen zu Forschungsdaten

Wie die Datenschnittstelle der VFU zu Forschungsdaten zu gestalten ist, richtet sich nach den Bedingungen der Datenbereitstellung. Hier sind (vgl. auch 3.2: Datenbereitstellung) zwei Nutzungsszenarien zu unterscheiden.

Im *ersten Fall* können Forschungsdaten in der gemeinsamen Arbeitsumgebung gehalten und verarbeitet werden. Dies gilt zunächst für *Daten, die für registrierte Nutzer/innen online frei verfügbar sind*, wie z.B. für den European Social Survey (ESS). Für diese Datensätze können Ausgangsdaten und entstehende Arbeitsdatensätze direkt in der VFU abgelegt und verwaltet werden. Darüber hinaus könnte es auch für einige faktisch anonymisierte *Offsite-Scientific Use Files (SUF),* z.B. für das SOEP, zulässig sein, dass nutzungsberechtigte Projektbeteiligte auf sie in der VFU zugreifen, etwa um sie mittels einer dort integrierten Statistiksoftware (R, vgl. unten: 4.6) weiter zu verarbeiten und zu analysieren. Auch diese Offsite-SUF würden in der VFU auf einem sicheren Server des Providers (vgl. dazu oben: 4.1) unter Berücksichtigung der entsprechenden Rechte- und Fristenbeschränkungen abgelegt. Nach Beendigung eines Arbeitsschritts können alle Arbeitsdaten in der VFU archiviert werden. Gilt für solche Daten eine Löschfrist und wird diese erreicht, kann die VFU das Löschen nachfragen und Daten vor der Löschung zur weiteren Archivierung an das FDZ übertragen. Bei Bedarf könnten nutzungsberechtigte Projektbeteiligte diese Offsite-SUFs aber auch auf ihre individuelle Workstation laden und dort mit eigenen oder speziellen Statistikprogrammen analysieren, um anschließend Arbeitsdatensätze wieder in die Forschungsumgebung zurückzuspeichern.

Im *zweiten Fall* soll für Forschungsdaten, die aus datenschutzrechtlichen Gründen nicht in die Forschungsumgebung integriert werden können, in der VFU eine Schnittstelle zum jeweiligen FDZ eingerichtet werden, die den Datenfernzugriff und das Fernrechnen unterstützt. Dieser Zugangsweg kommt zum einen für *Offsite-SUF* in Betracht, deren Nutzungsbedingungen einen gemeinsamen Zugriff im Grid ausschließen (z.B. für den Mikrozensus-SUF); zum anderen nach erfolgreicher Erprobung auch für *Onsite-Datensätze* (beispielsweise die Versichertenkontenstichprobe der Rentenversicherung), die nur in der Infrastruktur des FDZ genutzt werden dürfen. In beiden Fällen werden beim „Remote-Rechnen“ (Fernrechnen), das nur Nutzungsberechtigten zur Verfügung steht, aus der Benutzeroberfläche der VFU heraus Auswertungsschritte im FDZ angestoßen, die mit den dortigen Statistikprogrammen auf einem FDZ-Rechner ausgeführt werden. Werden Offsite-SUF auf diese Weise genutzt, können die Ergebnisse ohne Outputkontrolle online übermittelt werden. Wird auf onsite-Datensätze zugegriffen, unterliegen die Ergebnisse der Outputkontrolle durch das FDZ. Alle entstandenen Zwischen- und Enddatenprodukte verbleiben auf dem Rechner des FDZ.

Auf diese Weise lässt sich „Remote Data Access“ in die VFU integrieren, und die gemeinsame Metadatenverwaltung und sowie andere Funktionen der VFU (vgl. unten: 4.3 bi 4.5) stehen für beide Zugangswege in gleicher Weise zur Verfügung. Daten, die den FDZ zur Langzeitarchivierung übermittelt wurden oder die von vornherein auf den Rechnern des FDZ verbleiben, werden in der VFU „weiterhin als virtuelle Datensätze, mit einem Verweis auf den Archivierungsstandort, erhalten bleiben“ (Technische Evaluation: 83).

Bei allen Forschungsdaten, die Nutzungsbeschränkungen unterliegen, ist zu berücksichtigen, ab welchem Zeitpunkt oder Aggregationsgrad der Daten die Datenschutzbestimmungen der Ausgangsdaten nicht mehr gelten. So ist z.B. bei Organisationsdaten zu klären, bei welchem Grad der Verdichtung Arbeits- oder Ergebnisdateien in die Datenverwaltung der VFU integriert werden können.

4.3 Datenverwaltung

Die Datenverwaltung stellt verschiedene Funktionen zu Datenorganisation und Datenablage in der VFU bereit. Sie wird durch eine Schnittstelle genutzt, die dem Dateibrowser eines Betriebssystems mit den üblichen bekannten Dateiverwaltungsfunktionen (Datei laden, Datei speichern usw.) sehr ähnlich sein soll. Welche dieser Funktionen für wen verfügbar sind, hängt vom jeweiligen Dateityp ab. Etwa können Arbeitsdateien zu *Onsite-SUF* s und Remote-Processing-Dateien gar nicht innerhalb der VFU gespeichert und geladen werden. Sie werden, wie oben unter 4.2 beschrieben, über eine Schnittstelle im jeweiligen FDZ verwaltet. In der VFU verbleiben dazu jedoch Metadaten, die auf den Datensatz beim FDZ verweisen. Daten aus Offsite-SUF Dateien, die das jeweilige FDZ für die Grid-Umgebung freigegeben hat, und Daten aus frei verfügbaren Datensätzen wie beispielsweise dem ESS können von allen Projektbeteiligten entsprechend ihren Nutzungsrechten mit den Datenverwaltungswerkzeugen in der VFU abgelegt und verwaltet werden.

Die Datenverwaltung soll schwerpunktmäßig die Archivierung, Dokumentation und Nachnutzung bzw. kollaborative Nutzung von Auswertungssyntax unterstützen. Neben Syntaxdateien und (freigegebenen) Forschungsdaten werden über sie auch alle anderen Dateiarten, etwa Outputdateien, Tabellen- oder Textdokumente, formatunabhängig abgelegt und mittels Inhaltssuche wie über Metadaten auffindbar verwaltet werden. Es soll möglich sein, im Rahmen der bestehenden Nutzungsberechtigungen logische Verknüpfungen zwischen den verschiedenen Datenarten (z.B. zwischen Syntax und Forschungsdaten) herzustellen.

Eine entscheidende Voraussetzung für die Nutzung einer gemeinsamen Arbeitsumgebung besteht darin, dass die Daten anhand von Metadaten gut beschrieben werden. Die Metadatenverwaltung der VFU soll an den Standard der Data Documentation Initiative (DDI, aktuell: DDI 3.0) angepasst sein, der sich für Forschungsdaten international etabliert hat (vgl. Gregory u.a. 2010: 500); sie soll aber auch flexibel erweitert werden können. Das Schema muss technische und fachliche Metadaten, zusätzliche Versionsinformationen, In-

formationen über Verknüpfungen mit Dateien in den FDZ, über Nutzungszeiträume und über virtuelle (z.B. in den FDZ abgelegte) Datensätze berücksichtigen. Zur technischen Umsetzung kommen die Grid-Komponenten Fedora und iRODS in Betracht (Technische Evaluation: 69 ff.).

Die eigentliche Herausforderung für die Entwicklung der Metadatenverwaltung liegt jedoch eindeutig nicht auf der technischen Seite, sondern in der fachlichen Aufgabe, ein geeignetes Kategoriensystem zu entwickeln und die Arbeit damit zu moderieren (vgl. dazu auch unten: 4.4). Die Technische Evaluation (72; 96) empfiehlt, die aus Standardanwendungen bekannten Verzeichnissysteme so zu erweitern, dass eine Darstellung der Inhalte sowohl nach Dateien als auch nach inhaltlichen Kategorien möglich ist. So lassen sich Dateien in mehrere Kategorien gleichzeitig einordnen und in verschiedenen Kategorienhierarchien anzeigen. Metadaten und Kategorien können über Kontextmenüs erstellt und definiert werden.[46]

Eine weitere wichtige Funktion der Datenverwaltung besteht darin, Dateien zu „versionieren“, d.h. verschiedene Dateiversionen anzuzeigen, zu vergleichen und zu löschen“, was insbesondere das „Syntax Sharing“ unterstützen würde.

4.4 Datenbearbeitung

Zu den Funktionen der Datenbearbeitung zählen zum einen die Konvertierung und Validierung, zum anderen das Editieren von Syntax und Metadaten.

Mit Konvertierungs- und Validierungsfunktionen können Dateien von einem alten in ein neues Format konvertiert werden. Solche Funktionen sind auch für die langfristige Nachnutzung aller vorhandenen Ausgangs- und Arbeitsdaten notwendig. Gleichzeitig können die jeweils verwendeten Formate automatisch in der VFU validiert werden.

Von zentraler Bedeutung sind Editoren für Syntax- und Metadaten aller verfügbaren Datenformate. Anders als Forschungsdaten, die Nutzungsbeschränkungen unterliegen, können diese Daten ohne Einschränkung (also für frei verfügbare Datensätze und Offsite-SUF ebenso wie für Onsite-Datensätze der FDZ und für Remote-Processing-Dateien) abgelegt und bearbeitet werden. Denn Syntax zur Analyse einer Remote-Processing-Ausgangsdatei sowie die zugehörigen Metadaten kann anhand eines Datenstrukturfiles innerhalb der ge-

[46] „Kernpunkte sind hierbei die erweiterten Möglichkeiten zur Arbeit mit Metadaten, die Verwendung mehrerer paralleler Hierarchien für einen Satz von Dateien sowie entsprechenden Werkzeugen zum effektiven Umgang und der Pflege dieser Hierarchien. “ (Technische Evaluation: 73).

meinsamen Arbeitsumgebung erstellt werden, auch wenn anschließend die entsprechende Syntaxdatei via Schnittstelle auf Daten im FDZ zugreift.

Die technische Evaluation empfiehlt, bei den Editoren für Syntax und für Metadaten möglichst auf bereits existierende und bekannte Software zurückzugreifen. Der Syntax-Editor soll „Unterstützung für die korrekte Formulierung von Befehlen bieten und gleichzeitig eine Validierung von Syntaxdateien erlauben“ (ebd.: 74). Z.B. lässt sich TextPad[47] so erweitern, dass die Syntax-Formate von SPSS, Stata und R unterstützt werden.

Der Metadaten-Editor soll Metadaten sowohl anzeigen als auch bearbeiten. Die Datei, zu der die jeweils bearbeiteten Metadaten gehören, „soll gleichzeitig im entsprechenden Editor mit angezeigt oder in der Dateiverwaltung markiert“ werden (ebd.). Einige technische und inhaltliche Metadaten (Entstehung einer Datei, in einer Syntax angesprochene Input-Dateien oder vorhandene Kommentarzeilen) können durch eine „Erkennung für Dateiformate und eine im Hintergrund arbeitende Informationsextraktion“ automatisch aus den entsprechenden Dateien extrahiert werden (ebd.: 75). Weitere Metadaten können bei der Datensicherung standardmässig, z.B. nach einem Schlagwortsystem, abgefragt werden. Der Editor könnte hier auch die Funktion bieten, „die von den FDZ erwarteten Kommentarstrukturen auf Basis der Metadaten zu erzeugen“ (ebd.: 75). Oder es könnten evtl. bereits aufbereitete Metadaten von definierten Datensätzen der FDZ (z.B. Mikrozensus, auch über das Microdata Lab von GESIS) auf der Suche nach geeigneten Hinweisen gezielt durchgearbeitet werden.

Ein großer Teil der Metadaten wird jedoch individuell einzugeben und zu bearbeiten sein. Beispielsweise bedürfen Aggregatdaten einer detaillierten Beschreibung. Wie bereits unter 4.4 angesprochen, muss die Erfassung von Metadaten im Verbund so moderiert werden, dass den Beteiligten klar ist, welche Informationen in die Metadaten eingehen sollen, und dass dabei ihr fachwissenschaftlicher Erfassungsaufwand vertretbar bleibt. Metadatensysteme sind nie fertig, und Felder können „missbraucht“ werden, indem mehrere Informationen in ein Feld eingetragen werden. Zudem ist im Arbeitsprozess abzustimmen, in welcher Reihenfolge Metadaten für kollaborativ erstellte Syntaxdateien eingegeben oder bearbeitet werden.

Der Editor soll es ermöglichen, standardisierte und frei eingegebene Metadaten in Metadaten anderer Dateien zu importieren. Der DDI-Standard zur Erfassung von Metadaten sollte berücksichtigt werden. So könnten sowohl DDI-Datensätze von externen Daten in die

[47] http://www.textpad.com

VFU importiert und nachgenutzt werden, als auch selbst erstellte Metadaten im DDI-Format für andere Anwendungen exportiert werden.

Zudem ist es „für Syntaxdateien sinnvoll, die durch Befehle referenzierten Eingangs- und Ausgangsdaten mit den Syntaxdateien zu verknüpfen“ (ebd.: 75).

4.5 Datenvergleich

Für einen schnellen Überblick über den Arbeitsstand aller Beteiligten beispielsweise zu einem bestimmten Ausgangsdatensatz oder zu einem bestimmten Variablenkonstrukt bieten die Datenvergleichswerkzeuge wichtige Unterstützung. Denn alle in der VFU vorgenommenen Änderungen sind technisch und durch ausreichende Metadatenbeschreibung dokumentiert und nachvollziehbar. Z.B. lassen sich Dateien verschiedener Formate auf bestimmte Indikatoren hin vergleichen und Arbeitsprozesse dokumentieren.

Dateien können auf zwei Wegen verglichen werden. SubVersionControl Systeme vergleichen zwei Dateien (Bit für Bit) auf Identität hin. Falls Dateien in diesem technischen Sinne identisch, jedoch inhaltlich verschieden sind, wird dieser Unterschied im Vergleich der jeweiligen Metadaten oder Kategorien festzustellen sein. Bei Syntaxdateien ist ein zeilenweiser Vergleich sinnvoll:

> „So können zum Beispiel unterschiedliche Versionen der gleichen Syntaxdatei mit einander verglichen und deren Entwicklung nachvollzogen werden. Aber auch parallele Entwicklungen an Syntaxdateien könnten damit effektiv wieder zusammengeführt werden. In einer erweiterten Variante sollten hier auch einzelne Wort- oder Buchstabenänderungen sichtbar gemacht werden können. Diverse Tools, u.a. aus der Softwareentwicklung (z.B. Eclipse IDE, TextPad, Emacs, etc.), liefern Teile der entsprechenden Funktionalitäten mit. Diese sollten wenn möglich wiederverwendet werden.“ (Technische Evaluation: 77).

4.6 Datenverarbeitung

Unter Datenverarbeitung wird hier die statistische Datenanalyse von Forschungsdaten (Ausgangs- und Arbeitsdaten) verstanden. Nach den beiden unter 4.2 eingeführten Zugriffsszenarien findet dies teils an Daten innerhalb der gemeinsamen Arbeitsumgebung[48], teils durch Ferndatenzugriff in den Dateneinrichtungen statt. Der zweite Fall (das Rechnen mit Daten in den FDZ) wurde oben behandelt und muss hier nicht näher betrachtet werden.[49]

---

[48] Forschungsdaten, die innerhalb der VFU verwaltet werden, können natürlich von Nutzungsberechtigten, verarbeitet werden. Dieser Fall ist hier nicht besonders zu behandeln.

[49] Die Schnittstelle zur Bearbeitung von Forschungsdaten in den FDZ soll in die Sicherheitsarchitektur der VFU integriert sein. Die Datenverarbeitung wird in der VFU lediglich angestoßen. Gerechnet wird – mit oder ohne

An Forschungsdaten, die für Nutzungsberechtigte in der VFU verwaltet werden dürfen (bestimmte offsite-SUF oder andere, frei verfügbare Mikrodaten, vgl. 4.2) sollen Berechnungen in der gemeinsamen Arbeitsumgebung, auf dem gesicherten Server, möglich sein. (Sofern sie zum Download der Daten berechtigt sind, haben Nutzungsberechtigte die Alternative, diese mit individueller Software auf der individuellen Workstation zu verarbeiten. Im Falle von SUF-Dateien erleichtert ein von vielen Arbeitsplätzen zugängliches Daten-Repository auch das lokale Ausführen von Berechnungen durch einfachen Zugriff auf den gewünschten Datensatz.) Um diese Option zu ermöglichen, soll in der VFU zunächst das Statistikprogramm R (eine Open-Source-Version des kommerziellen Statistikprogramms S-PLUS) verfügbar sein. (Eine Lizenzierung der VFU für andere Statistikprogramme wie SPSS und Stata erscheint derzeit nicht als wirtschaftlich.)

Ein Rechnen mit R in der gemeinsamen Arbeitsumgebung bietet sich für sehr rechenintensive Operationen an, die unter Nutzung der Speicher- und Prozessorkapazitäten des Grid schneller erledigt werden können. Zudem bietet R inzwischen einige Rechenoperationen für Längsschnittanalysen, die nur mit großen Rechnerkapazitäten zu bewältigen sind. Vor der Durchführung der Analysen mit R prüft die Sicherheitsarchitektur der VFU die Berechtigung zur Nutzung der ausgewählten Datensätze. Der Ablauf der Datenverarbeitung im Grid wird in der Technischen Evaluation (ebd.: 79) genauer beschrieben.

> „Um statistische Berechnungen im Grid aus einer virtuellen Arbeitsumgebung heraus anstoßen zu können, müssen entsprechende Jobs an die Grid-Middleware geschickt werden. Dies sollte für die Anwender/innen transparent passieren, da die Spezifikation solcher Jobs bestimmtes Fachwissen erfordert und in einem abgeschlossenen Kontext, wie dem der virtuellen Arbeitsumgebung, sich je Job nur geringfügig unterscheiden wird. Die virtuelle Arbeitsumgebung muss demnach in einem Teil ihrer Oberfläche eine einfach zu nutzende Möglichkeit bieten, um eine solche Verarbeitung anzustoßen. Im Hintergrund wird dabei eine Jobbeschreibung erstellt und an die Grid-Middleware gesendet. Die graphische Oberfläche muss danach eine Überwachung der Verarbeitung ermöglichen und den Benutzer über die Fertigstellung der Berechnung, z.B. per Anzeige auf dem Bildschirm oder per E-Mail, informieren. Außerdem sollen die Ergebnisse direkt zugreifbar zur Verfügung gestellt werden. Die Erstellung einer Jobbeschreibung muss gegebenenfalls auf Inhalte der Syntaxdatei zurückgreifen. Der Grund ist, dass die Syntaxdateien Referenzen auf die Ein- und Ausgangsdaten einer Berechnung beinhalten. Diese werden ebenso innerhalb einer Jobbeschreibung benötigt, um dem Grid die Möglichkeit zum so genannten Staging, also dem Zur-Verfügung-Stellen der benötigten Dateien auf dem Zielsystem, zu geben. Dabei müssen die Pfade zu den Dateien gegebenenfalls auf Grid-interne Pfade angepasst werden. All dies soll, wenn möglich, transparent für den Benutzer passieren. Im Idealfall ist der Unterschied zwischen lokaler und Grid-basierter Berechnung nur in wenigen Punkten, z.B. in nur einem Flag beim Starten der Berechnung, sichtbar.“ (Technische Evaluation: 79).

Output-Kontrolle – auf Rechnern des FDZ, wo auch die Daten verbleiben. Der Metadaten-Editor der VFU greift auf diese Daten nur virtuell zu.

4.7 Kollaboration und Kommunikation

Um die kollaborative Arbeit und die Kommunikation zwischen Wissenschaftler/innen zu unterstützen, sollen Technologien wie Instant-Messaging, Voice over IP oder Videotelefonie eingesetzt werden können. Bereits existierende Systeme wie beispielsweise Adobe® Acrobat® Connect™ Pro, welche für Web-Konferenzen konzipiert sind, können dazu in die VFU integriert werden. Zusätzlich kann über die Datenverwaltungswerkzeuge gemeinsam auf in der Arbeitsumgebung verwaltete und dokumentierte Materialien zugegriffen werden.

> „Zur Dokumentation der Projektarbeit werden üblicherweise zentrale Systeme benötigt, in denen von jedem Teammitglied Informationen abgelegt und abgerufen werden können. Was mit einem Dokumentenschrank mit vielen Ordnern in der realen Welt möglich ist, kann in der digitalen Welt durch Wikis oder zentrale Dateiverwaltungssysteme ersetzt werden. Letzteres wird durch die Datenverwaltungswerkzeuge der virtuellen Arbeitsumgebung bereits abgedeckt. Wikis sollten zusätzlich integriert werden. Wichtig ist jedoch, sowohl bei einem Dokumentenschrank als auch bei Wikis oder anderen Dateiverwaltungssystemen, dass es teamweite Vorgaben für deren Ordnung geben muss, und dass diese Vorgaben eingehalten und deren Einhaltung überprüft werden.“ (Technische Evaluation: 80).

Zusätzlich soll das gleichzeitige Arbeiten mehrerer Projektbeteiligter an gemeinsamen Syntaxdateien durch einen kollaborativen Syntaxeditor in der VFU unterstützt werden. Damit lassen sich Eingaben und Änderungen in einer Syntaxdatei gleichzeitig von allen Teilnehmer/innen beobachten und verfolgen.

> “Unter paralleler Nutzung eines Kommunikationssystems wie z.B. Skype entsteht somit eine Zusammenarbeit, die einer lokalen, gemeinschaftlichen Arbeit sehr ähnlich ist.“ (Ebd.)

4.8 Konfiguration und Verwaltung

In der Benutzeroberfläche der VFU müssen Werkzeuge zur individuellen Konfiguration und Verwaltung zu finden sein, mittels derer Nutzer/innen eigene Einstellungen wie beispielsweise das Erscheinungsbild der grafischen Oberfläche oder Verzeichnispfade festlegen. Für verschiedene Nutzer/innen oder Gruppen können die Nutzungsrechte für Zugriffe auf bestimmte Dateien oder Verzeichnisse und ausgewählte Funktionen, sowie konkrete Vorgaben für die allgemeine Kategorienhierarchie von Dateien und grundlegende Metadatenstrukturen festgelegt werden.

4.9 Publikation

Künftige Verbundprojekte sollten sich nach Möglichkeit am Leitbild von Open Access[50] orientieren, mit öffentlichen Mitteln finanzierte Forschungsergebnisse über die Projekt-Website (hier www.soeb.de) und andere geeignete institutionelle und fachliche Repositorien unentgeltlich und barrierefrei im Internet zugänglich zu machen. Die VFU soll daher vor allem die Veröffentlichung von Ergebnissen in digitaler, frei zugänglicher Form unterstützen. Veröffentlichungen in Printmedien (Buch- und Journal-Veröffentlichungen) bleiben ein wesentlicher Teil einer Publikationsstrategie, zielen jedoch vor allem auf eine stärker verdichtete oder wissenschaftlich spezialisierte Ergebnisdarstellung oder behandeln konzeptionelle und methodische Fragen.

Um die Veröffentlichung von Ergebnissen auf der projekteigenen Website zu unterstützen, müssen die „Publikationswerkzeuge" der VFU „Zugriffe auf jene Systeme bieten, mit denen die Projektwebsite aufgebaut und gepflegt wird" (Technische Evaluation: 84). „Handelt es sich hierbei um Content-Management-Systeme, so muss Zugriff auf deren Konfigurationsoberfläche bestehen. Sind es aber eher Wikis, so reicht ein Zugang zur entsprechenden Einstiegsseite. [...] Die Zugriffe müssen dabei einen einfachen Transfer von bereits erstellten Teilen der Veröffentlichung in die Systeme zur Websiteverwaltung erlauben. Demnach sollte es einfach möglich sein, z.B. eine Grafik oder einen Text über die Datenverwaltungswerkzeuge aus der virtuellen Arbeitsumgebung herauszuholen und über die Verwaltung der Projektwebsite zu veröffentlichen. Außerdem sollten Reviewfunktionalitäten der eingesetzten Websitesysteme, sofern sie von diesen angeboten werden, aus der virtuellen Arbeitsumgebung heraus nutzbar sein. Hierdurch kann sichergestellt werden, dass mehrere Mitglieder eines Forschungsverbunds eine geplante Publikation vor deren tatsächlicher Veröffentlichung verifizieren." (Ebd.)

[50] http://www.open-access.net/

5. Schlussfolgerungen und Empfehlungen

5.1 Forschungsnahe und projektbegleitende Entwicklung

Virtuelle Forschungsumgebungen (VFU) unterscheiden sich von anderen Komponenten der Informationsinfrastruktur dadurch, dass sie an der tatsächlichen Nutzung von Daten und Diensten in der wissenschaftlichen Forschungspraxis ansetzen an diese durch technische Unterstützung auch verändern. Daher können VFU nicht „generisch" entwickelt werden, sondern nur forschungsnah und projektbegleitend. Die Entwicklung einer (VFU), die eine kollaborative Nutzung der Infrastruktur an Wirtschafts- und Sozialdaten unterstützt, muss an praktischen Problemen im Forschungsprozess ansetzen. Die schrittweise Entwicklung und Umsetzung einer Systemarchitektur („Top down") muss immer wieder mit Nutzer/innen rückgekoppelt werden („bottom up"). Die Möglichkeiten der VFU sollen im Rahmen von Projektarbeit erkundet werden, und ihre Komponenten müssen ihren Nutzen in der Forschung erweisen.

Hierzu müssen Infrastrukturentwicklung und sozialwissenschaftliche Forschung unmittelbar verknüpft werden. Das Infrastrukturprojekt, das den Prototyp einer VFU für Forschungsdaten fachlich zu konzipieren und technisch umzusetzen hat, soll sich auf einen sozialwissenschaftlichen Anwendungsfall beziehen, der für die zu unterstützende Forschungspraxis exemplarisch ist. Das Verbundvorhaben Dritter Bericht zur sozioökonomischen Entwicklung Deutschlands kann ein solcher Anwendungsfall sein, da es eine intensive und arbeitsteilige Nutzung von sozial- und wirtschaftswissenschaftlichen Mikrodaten beabsichtigt. In die Entwicklung sind typische Quer- und Längsschnittfiles ebenso einzubeziehen wie Datensätze, die Personen- und Organisationsdaten verknüpfen.

Das Infrastrukturprojekt benötigt gegenüber dem Forschungsprojekt einen zeitlichen Vorlauf, damit der Verbund, für den es den Prototyp der VFU konzipiert, zu Beginn seiner Arbeit bereits Komponenten, Datenzugangswege und Konzepte vorfindet. Andererseits kann dieser Vorlauf nicht zu groß sein, da das Forschungsprojekt mit seinen inhaltlichen Fragestellungen für das Infrastrukturprojekt die einzubeziehenden Forschungsdaten definiert und die Arbeit an ihnen bestimmt.

Das Projekt und der Projektfortschritt sollten sowohl den quantitativ-empirischen Sozialwissenschaften als auch den Einrichtungen der Dateninfrastruktur zu geeigneten Zeitpunkten durch Transferaktivitäten vermittelt werden. Von besonderer Bedeutung für das Vorhaben sind die Arbeitsgruppe „Future Data Access" beim RatSWD und andere Vorhaben des „Remote Access" zu Mikrodaten.

5.1 Schwerpunkt und Kernelemente einer VFU

Die Schnittstelle zu den Forschungsdaten bildet nicht nur eine notwendige Umweltbedingung für die zu entwickelnde VFU. Hier stellen sich Probleme der Forschungspraxis, deren Lösung die VFU unterstützen soll. Indem sie den Datenfernzugriff auf Offsite- und Onsite-Forschungsdaten erleichtert und in eine gemeinsame Arbeitsumgebung mit Metadaten mit „Syntax Sharing“ integriert, soll sie Formen der datenbezogenen Kooperation anstoßen, die bislang offenbar eher auf die Arbeit mit Offsite-Scientific-Use-Files beschränkt blieben.

Die VFU soll zwei Schnittstellen zu Forschungsdaten unterstützen: eine für Offsite-Daten, auf die berechtigte Nutzer/innen in einer gemeinsamen Arbeitsumgebung zugreifen können, und eine zweite, bei der Rechenoperationen – mit oder ohne Output-Kontrolle – in den Dateneinrichtungen stattfinden und lediglich die Syntax und die Metadaten zu diesen Onsite-Daten im Grid verwaltet werden. Ein zweiter Schwerpunkt soll auf der Archivierung, Dokumentation und Nachnutzung/ kollaborativen Nutzung von Auswertungssyntax und auf der Entwicklung eines dafür geeigneten Metadatenstandards liegen.

Aus dieser Schwerpunktsetzung ergibt sich, welche weiteren Kernkomponenten vorrangig in die Entwicklung einzubeziehen sind:

- eine Sicherheitsarchitektur mit personalisierten Zugangsrechten und definierten virtuellen Organisationen, die Nutzungsbeschränkungen für einbezogene Forschungsdaten technisch umsetzt und den Nutzer/innen Kontrolle über ihre Arbeitsergebnisse belässt,
- die notwendigen Funktionen der Datenverwaltung, der Datenbearbeitung (Syntax- und Metadaten-Editoren) und des Datenvergleichs, die gemeinsame Arbeit an Syntax unterstützen,
- die für eine grafische Benutzeroberfläche erforderlichen Konfigurationswerkzeuge.

5.3 Kooperation mit Einrichtungen der Dateninfrastruktur

Aus dem Projektziel, eine VFU für die vernetzte Arbeit mit Forschungsdaten zu entwickeln, ergibt sich die Notwendigkeit, ausgewählte Dateneinrichtungen von vornherein als Partner/innen in das Infrastrukturprojekt einzubeziehen. Wünschenswert ist eine Konstellation, in der die Dateneinrichtungen zugleich mit eigenen Forschungsbeiträgen am Verbundvorhaben beteiligt sind und die Entwicklung der VFU sowohl aus der Perspektive der Datenbereitstellung als auch aus der Nutzungsperspektive bewerten.

Die Kooperation mit den einbezogenen Dateneinrichtungen sollte sich also nicht nur auf die Entwicklung der Datenschnittstelle für „Remote Data Access“ und „Remote Processing“ und auf Anforderungen an die Sicherheitsarchitektur beschränken, sondern auch auf die Integration dieser Datenzugangswege in die gemeinsame Arbeitsumgebung, auf gemeinsame Metadatenstandards von Datenhaltern und Nutzer/innen und auf fachliche Lösungen für das Editieren von Metadaten und Syntax und für die nachnutzbare Archivierung.

Das Infrastrukturprojekt soll mit einer begrenzten Zahl von FDZ begonnen werden, deren Daten voraussichtlich im Verbundvorhaben benötigt werden, soll aber für die Einbeziehung weiterer Dateneinrichtungen offen sein.

5.4 Datenzugang und Datenschutz

Für alle Forschungseinrichtungen, die in einer VFU kooperieren, müssen Einzelnutzungsverträge mit den Dateneinrichtungen bestehen. Die datenbezogene Kooperation wird in der VFU auf virtuelle Organisationen beschränkt, deren Mitglieder über eine solche Nutzungsberechtigung für den jeweiligen Datensatz verfügen. Die Sicherheitsarchitektur und die Datenverwaltung der VFU muss für beide Nutzungsszenarien, die des Zugriffs auf Daten innerhalb der VFU und die des Datenfernzugriffs, technisch sicherstellen, dass die personelle Nutzungsbeschränkung sowie die Befristung, Zweckbindung und Genehmigungspflicht für die Nutzung der Forschungsdaten entsprechend den Bedingungen des jeweiligen FDZ gewahrt sind. Daher bleibt die Archivierung von Forschungsdaten (Ausgangs- und Arbeitsdaten) in der VFU auf solche Datensätze beschränkt, deren Nutzungsbedingungen die Verwaltung in einer gemeinsamen Arbeitsumgebung zulassen. Für die Onsite-Nutzung von Forschungsdaten wird eine Langzeitarchivierung in den FDZ selbst angestrebt.

5.5 Arbeitsumgebung und Arbeitsprozess

In seiner Koppelung mit einem Forschungsverbund als „Anwendungsfall“ führt das Infrastrukturprojekt unterschiedliche Arbeitskulturen zusammen: die von IT-Entwickler/inne/n und Datendienstleister/inne/n mit der von empirisch orientierten Sozialwissenschaftler/inne/n. Es handelt sich also nicht nur um eine technische Entwicklung, sondern zugleich um eine sozialwissenschaftliche Aufgabe und um ein soziales Experiment.

Von den Projektbeteiligten des Forschungsverbunds wird die Bereitschaft erwartet, sich an Metadatenerfassung und daten- und syntaxbezogener Zusammenarbeit in der gemeinsamen Arbeitsumgebung zu beteiligen. Die Nutzung der Komponenten einer VFU kann jedoch nicht erzwungen, sondern nur durch überzeugende praktische Lösungen erreicht werden.

Die VFU soll individuell und disziplinär verschiedene Arbeitsweisen der beteiligten Wissenschaftler/innen unterstützen, also die Vielfalt der Arbeitsweisen im Verbund als produktive Ressource erhalten und Standards setzen, die nicht standardisierend wirken. Die Projektbeteiligten sollen über den Zusammenschluss zu „virtuellen Organisationen“ in einem abgestuften System projektinterner Öffentlichkeit die Kontrolle über ihre Arbeitsergebnisse behalten und festlegen können, für wen diese wann einsehbar und editierbar sind.

In der VFU werden bestimmte Verhaltensanforderungen an die Nutzer/innen technisch „hinterlegt“. Diese müssen sich durch den praktischen Nutzen für die Forschungsarbeit legitimieren lassen. Daher muss das Infrastrukturprojekt Nutzer/inn/en bei der Arbeit in der

VFU und bei der Einhaltung der Standards wirksam unterstützen. Für die fachliche Begleitung, die Moderation und den technischen Support der VFU und ihrer interaktiven Komponenten sind ausreichende Personalressourcen vorzusehen.

Die VFU bildet für das Management des Forschungsverbunds kein Steuerungsinstrument, sondern eine zusätzliche Steuerungsaufgabe. Der Einsatz der VFU kann die organisatorischen Probleme der Koordinierung und Leitung eines Forschungsverbundes nicht lösen. Die bessere technische Unterstützung formalisiert aber viele Aufgaben der Verbundkoordination und macht sie damit sichtbarer. Und sie schafft - wenigstens in der Entwicklungsphase – zusätzlichen Organisations- und Steuerungsaufwand: So stellt die Entwicklung des Kategoriensystems für Metadaten auch an die Verbundkoordination fachliche Anforderungen, und ein Teil der Moderationsaufgaben für die VFU ist im Rahmen des Verbunds zu leisten, der in der VFU arbeitet.

6. Literaturverzeichnis

Alexander von Humboldt-Stiftung/Deutscher Akademischer Austauschdienst/Deutsche Forschungsgemeinschaft/Fraunhofer-Gesellschaft/Helmholtz-Gemeinschaft Deutscher Forschungszentren/Hochschulrektorenkonferenz/Leibniz-Gemeinschaft/Max-Planck-Gesellschaft/Wissenschaftsrat (2008): Schwerpunktinitiative „Digitale Information“ der Allianz-Partnerorganisationen, Berlin, 11. Juni 2008. http://www.allianzinitiative.de/fileadmin/user_upload/keyvisuals/atmos/pm_allianz_digitale_information_details_080612.pdf

Brandt, Maurice/Zwick, Markus (2011): Improvement of data access – The long way to remote data access in Germany. Statistische Ämter des Bundes und der Länder. Wiesbaden: FDZ-Arbeitspapier Nr. 39. http://www.forschungsdatenzentrum.de/publikationen/veroeffentlichungen/39.asp

Deutsche Forschungsgemeinschaft (2010): Informationsverarbeitung an Hochschulen – Organisation, Dienste und Systeme. Empfehlungen der Kommission für IT-Infrastruktur 2011-2015. Bonn. http://www.dfg.de/download/pdf/foerderung/programme/wgi/empfehlungen_kfr_2011_2015.pdf

Deutsche Forschungsgemeinschaft (DFG): Gute wissenschaftliche Praxis, Denkschrift, Weinheim 1998. http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_0198.pdf.

Dickmann, Frank/Enke, Harry/Harms, Patrick (2010): Technische Evaluation der Grid-Technologie für das Modellprojekt Kollaborative Datenauswertung und virtuelle Arbeitsumgebung – VirtAug. SOEB Arbeitspapier 2010-1. http://www.soeb.de/fileadmin/redaktion/downloads/expertise_virtaug.pdf

Forschungsverbund Sozioökonomische Berichterstattung (Hg.) (2012): Berichterstattung zur sozioökonomischen Entwicklung in Deutschland. Teilhabe im Umbruch. Zweiter Bericht. Redaktion: Peter Bartelheimer, Sabine Fromm, Jürgen Kädtler. Wiesbaden: VS Verlag.

Gemeinnützigen D-Grid Entwicklungs- und Betriebsgesellschaft mbH (Hrsg) (2010): D-Grid. Die Deutsche Grid-Initiative. Vorstellung der neuen Projekte. http://www.d-grid-ggmbh.de/downloads/BroschuereAHM2010.pdf (01.08.2011)

Gentzsch, Wolfgang (2007): Grid-Computing und die deutsche D-Grid-Initiative. In: Neuroth, Heike/ Gentzsch, Wolfgang: 9-11.

German Data Forum (RatSWD) (Hrsg.) (2010): Building on Progress. Expanding the Research Infrastructure for the Social, Economic, and Behavioral Sciences (Vol. I, II). Opladen/Farmington Hills: Budrich UniPress Ltd.

Gregory, Arofan/ Heus, Pascal/ Ryssevik, Jostein (2010): Metadata. In: German Data Forum (RatSWD): 487-508.

High Level Expert Group on Scientific Data (2010): Riding the wave. How Europe can gain from the rising tide of scientific data. Final report of the High Level Expert Group on Scientific Data. A submission to the European Commission. October 2010. http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf

Huschka, Denis/Oellers, Claudia/Ott, Notburga/Wagner, Gert G. (2011): Datenmanagement und Data Sharing: Erfahrungen in den Sozial- und Wirtschaftswissenschaften. Working Paper Series des Rates für Sozial- und Wirtschaftsdaten 184, http://www.ratswd.de/download/RatSWD_WP_2011/RatSWD_WP_184.pdf.

Kommission Zukunft der Informationsinfrastruktur (zitiert als KZII) (2011): Gesamtkonzept für die Informationsinfrastruktur in Deutschland. Empfehlungen der Kommission Zukunft der Informationsinfrastruktur im Auftrag der Gemeinsamen Wissenschaftskonferenz des Bundes und der Länder. April 2011. http://www.allianzinitiative.de/fileadmin/user_upload/KII_Gesamtkonzept.pdf.

Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik (Hg.) (2001): Wege zu einer besseren informationellen Infrastruktur; Gutachten der vom Bundesministerium für Bildung und Forschung eingesetzten Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik. Baden-Baden: Nomos.

Kreuter, Frauke/ Casas-Cordero, Carolina (2010): Paradata. In: German Data Forum (RatSWD): 509--529.

Lajos Herpay, Lajos/, Sonja Neweling, Sonja/Uwe Schwiegelshohn, Uwe (Hg.) (2009): D-Grid. Die Deutsche Grid-Initiative. Vorstellung der Projekte. Dortmund. http://www.d-grid-ggmbh.de/downloads/BroschuereAHM2009.pdf

Mochmann, Ekkehard (2010): e-Infrastructure for the Social Sciences. In: German Data Forum (RatSWD): 265-285.

Neuroth, Heike (2010): WissGrid -Grid für die Wissenschaft. In: Gemeinnützigen D-Grid Entwicklungs- und Betriebsgesellschaft mbH (Hrsg) (2010): D-Grid. Die Deutsche Grid-Initiative. Vorstellung der neuen Projekte. http://www.d-grid-ggmbh.de/downloads/BroschuereAHM2010.pdf (01.08.2011)

Neuroth, Heike/ Gentzsch, Wolfgang (Hrsg.) (2007): Die D-Grid-Initaitive. Göttingen: Universitätsverlag Göttingen.

Neuroth, Heike/Aschenbrenner, Andreas/Lohmeier, Felix (2007): e-Humanities - eine virtuelle Forschungsumgebung für die Geistes-, Kultur- und Sozialwissenschaften. In: Bibliothek. Forschung und Praxis, 3 (2007), S. 272-279.

PARSE.insight (2009): Insight into digital preservation of research output in Europe. Survey Report. FP7-2007-223758 PARSE.Insight; Deliverable: D3.4

Publications Office of the European Union (2011): ESFRI Strategy report on research infrastructures – Roadmap 2010. Luxembourg, ISBN: 978-92-79-16828-4, doi:10.2777/23127

Rat für Sozial- und Wirtschaftsdaten (Hrsg.) (2011): Auf Erfolgen aufbauend. Zur Weiterentwicklung der Forschungsinfrastruktur für die Sozial-, Verhaltens- und Wirtschaftswissenschaften. Empfehlungen des Rates für Sozial- und Wirtschaftsdaten (RatSWD). Opladen/Farmington Hills: Barbara Budrich.
http://www.ratswd.de/publ/KVI/Empfehlungen_Auf_Erfolgen_aufbauend.pdf

Sennett, Richard (2011: Alles furchtbar einfach. In: Blätter für deutsche und internationale Politik, 56. Jg., Heft 7: 99-109.

Soziologisches Forschungsinstitut (SOFI / Institut für Arbeitsmarkt- und Berufsforschung (IAB) / Institut für sozialwissenschaftliche Forschung (ISF) / Internationales Institut für empirische Sozialökonomie (INIFES) (Hg.) (2005): Berichterstattung zur sozioökonomischen Entwicklung in Deutschland. Arbeit und Lebensweisen. Erster Bericht. Wiesbaden: VS Verlag.

Winkler-Nees, Stefan (2011): Anforderungen an wissenschaftliche Informationsinfrastrukturen. Working Paper Series des Rates für Sozial- und Wirtschaftsdaten 180,
http://www.ratswd.de/download/RatSWD_WP_2011/RatSWD_WP_180.pdf

Wissenschaftsrat (2011): Übergreifende Empfehlungen zu Informationsinfrastrukturen. Berlin, Wissenschaftsrat, Drucksache 10466-11, 2011.01.28;
http://www.wissenschaftsrat.de/download/archiv/10466-11.pdf (04.09.2011).

Anhang 1 Erster Workshop „Fachwissenschaftliche Anforderungen an eine virtuelle Arbeitsumgebung für soeb“

Göttingen, 9. Februar 2010

Teilnehmer/innen

Dr. Ilya Agapov (DESY-IT, Hamburg), Dr. Peter Bartelheimer (SOFI, Göttingen), Dr. Irene Becker (Johann Wolfgang Goethe-Universität, Frankfurt), Sarah Cronjäger (SOFI, Göttingen), Dr. Thomas Drosdowski (GWS, Osnabrück), Rita Hoffmeister (FDZ im Landesbetrieb für Statistik und Kommunikationstechnologie Niedersachsen (LSKN), Hannover), Jens Ludwig (SUB, Göttingen), Dr. Frank Schlünzen (DESY-IT, Hamburg), Tanja Schmidt (Schmidt-Sozialforschung, Berlin), Prof. Dr. Jürgen Schupp (DIW, Berlin), Ewa Sojka (INIFES, Augsburg), Dr. Michael Stegmann (FDZ, Deutsche Rentenversicherung, Würzburg), Falko Trischler (INIFES, Augsburg), Dr. Klaus-Peter Wittemann (SOFI, Göttingen).

Programm

Bei dem 1. Workshop handelt es sich um ein Arbeitstreffen aus Mitarbeiter/innen des Forschungsverbunds Sozioökonomische Berichterstattung, ausgewählter Einrichtungen der Dateninfrastruktur und Vertretern des WissGrid-Projekts. Das Programm beinhaltet: 1) Eine kurze Einführung (Peter Bartelheimer, Tanja Schmidt); 2) eine kurze Präsentation von TextGrid (Jens Ludwig); 3) eine Präsentation der Bedürfnisse und Ideen aus der SOEB II zur kollaborativen Arbeit an Datensätzen und virtuellen Zusammenarbeit (Tanja Schmidt); 4) Kommentare der SOEP-Gruppe (Jürgen Schupp), des FDZs im LSKN (Rita Hoffmeister), des FDZs der Deutschen Rentenversicherung (Michael Stegmann) und eines Fachberaters von WissGrid (DESY) (Frank Schlünzen); 5) Diskussion.
Die Präsentation von Tanja Schmidt ist abrufbar unter:
http://www.soeb.de/fileadmin/redaktion/downloads/VirtAug/Präsentation_Schmidt_Workshop_1_VirtAug.pdf

Statements zum Input von Tanja Schmidt

Prof. Dr. Jürgen Schupp (DIW)

– Auch die Halter wissenschaftsgetragener Daten haben ein Interesse an der Entwicklung von Standards für die Dokumentation, die eine problembezogene Recherche ermöglichen. Dabei sind Verbesserungen der bestehenden Serviceleistungen wie SoepInfo in den Blick zu nehmen, ebenso die „Paradaten“ der beauftragten Erhebungsinstitute, die Fachinformationssysteme und Datenarchive.

- Andere Initiativen sind zu berücksichtigen: z.B. CesTa im Rahmen von ESRI, Questionnaire Development Document Support (QDDS III), oder PanelWhiz.
- Auch Primärdatenerheber und Datenarchive/Bibliotheken haben Interesse an Standards für Langzeitarchivierung.
- Auch für die Abteilung SOEP am DIW fragt sich: Ist das ausreichend, was auf der Daten-CD ist? Für Nutzer/innen sind auch Referenzwerte, Fragenkontext und Referenzstudien wichtig.
- Je mehr generierte Variablen auf Web-Portalen hinterlegt werden, desto wichtiger wird für die wissenschaftlichen Nutzer/innen, dass die in ihnen hinterlegten Vorannahmen dokumentiert sind und nachvollzogen werden können. Auch Imputationen sollten gut dokumentiert sein. Zukünftig werden Peer Reviewer wissenschaftlicher Beiträge ebenfalls Zugang zu Daten und Syntaxen haben wollen.
- Nutzergruppenlizenzen sind für SOEP möglich. Gruppenlizenzen für Statistikpakete seien dagegen ein „bottleneck".
- Für das Thema wichtige Entwicklungen und Initiativen:
- Die Abteilung Sozio-oekonomisches Panel (SOEP) nimmt an den Arbeitsgruppen „Forschungsdaten", „Hosting / Langzeitarchivierung" und „Informationskompetenz/Ausbildung" innerhalb der „AG Informationsinfrastruktur" der Leibniz-Gemeinschaft (WGL) im Auftrag der Gemeinsamen Wissenschaftskonferenz (GWK) teil.
- DFG-Projekt von Rainer Schnell Questionaire Development Document Support (QDDS3.)
- SOEP-Info – künftig in einer Perl-XML-Umgebung.
- In PanelWhiz kann künftig generierte Syntax eingestellt werden; wer hochlädt, erhält die Versicherung, dass er / sie bei Verwendung der Syntax zitiert wird („Zitieren als Währung").
- Jürgen Schupp gibt zu bedenken, dass einige der andiskutierten Lösungen für eine virtuelle Arbeitsumgebung ein „Weg zurück zur Zentralisierung" von Datenanalysen sein könnten. Die Vielfalt der Analysen sei produktiv und daher zu erhalten.

Rita Hoffmeister (LSKN)

Rita Hoffmeister verweist auf die gesetzlichen Grundlagen für die Arbeit der Forschungsdatenzentren (FDZ). Entscheidend ist der Datenschutz auf Basis des Bundesstatistikgesetzes.

Fernanalyse und Onsite-Analyse greifen nicht auf die gleichen Datensätze zu. Für On-Site-Daten wird es sicherlich keine Zugriffslösungen in einer Grid-Umgebung geben, für Scientific-Use-Files (SUF) erscheint dies eher möglich. Die Zugriffsberechtigung für (SUF) muss aber sicherstellen, dass das FDZ alle Nutzer/innen kennt: Wo liegen die Daten, wer hat Zugriff, wie gut ist das System nach außen geschützt (Firewall)?

Nutzergruppen sind im Prinzip möglich, jedoch müssen alle Partner „unabhängige wissenschaftliche Einrichtungen" sein: Universitäten, angegliederte Einrichtungen. Geprüft

wird diese Eigenschaft nach § 16 Abs. 6 StatG z.B. anhand von Satzungen. Forschung muss unabhängig von Weisung Dritter sein.

Das Metadateninformationssystem der FDZ der Länder sei noch verbesserungsfähig.

Dr. Michael Stegmann (FDZ-RV)

Die FDZ-Zugangsregeln zu den Rentenversicherungsdaten ähneln denen der amtlichen Statistik; zusätzlich sind hier Matching und record linking ausgeschlossen.

Zur technischen Seite weist Michael Stegmann darauf hin, dass der Versuch, gekaufte Software im Verbund anzuwenden, sich möglicherweise nicht lohnt. Weder SPSS noch Stata seien multiprozessorfähig. Stata und R arbeiten nur mit dem Arbeitsspeicher, würden also nicht den Grid-Prozessor nutzen.

Zur rechtlichen Seite spricht Michael Stegmann den Sozialdatenschutz für prozessproduzierte Daten an. Er möchte daher „eine Lanze fürs Fernrechnen brechen". Man kann auf den kompletten Datenbestand zugreifen und selbst Arbeitsdateien ablegen. Nutzer/innen können an Standard-SUF oder Themenfiles „üben" und dann ihre Auswertungen auf diesem Wege durchführen.

Schließlich verweist Michael Stegmann auf den Widerspruch zwischen Löschfristen und Langzeitarchivierung.

Frank Schlünzen (DESY-IT)

Im Grid werden verteilte Ressourcen genutzt, nicht unbedingt ein großer Rechner oder Prozessor.

Kollaboratives Arbeiten setzt eine eindeutige ID voraus – entweder über DFN-ID, oder ein persönliches Zertifikat (über DFN, oder Registrierungsautoritäten z.B. Registrierungsstelle mit „Medium Assurance", also Lichtbildausweis). Dies sei sicherer als eine eindeutige IP, die nicht ausreichen würde. Zertifikate sind 1 Jahr gültig.

Außerdem werden durch den Attributserver Rechte verwaltet – dies ist allerdings betreuungsintensiv.

Sensitive Daten können in „Trusted-Zones" abgelegt werden. Daten zentral zu speichern, sei dem Verbund offenbar nicht so wichtig die die Dokumentation und der Austausch von Syntax.

Verteiltes Datenmanagement z.B. mit Fedora oder Irods erlaubt es, auf sichere Art und Weise auf die Daten zuzugreifen.

Im Grid gibt es entweder „Fat-Clients", also Programme, die mit dem GRID selbst sprechen oder „Thin Clients", also einfache Web-Portale. WIKI wäre dabei völlig problemlos integrierbar.

Auch Subversion-Server-Dienste mit Benachrichtigungsfunktion sind einfach zu implementieren.

Weitere Diskussionsbeiträge

- In der Arbeit des Verbunds am zweiten Bericht fehlte oft die Zeit für eine gute Dokumentation – hierfür gibt es keine technische Lösung.
- Nicht alle Probleme lassen sich vermeiden – die Standardisierung von Datenauswertungen hat Grenzen. Einzelne Forscher/innen brauchen Freiheit für ihre Lösungen (z.B. für den Umgang mit imputierten Daten.
- Verteilungstreue „Spieldatensätze“ würden das Fernrechnen erleichtern.

Zusammenfassung des Diskussionsstands

Aufgabenstellung des Projekts
Ziel des Abschlussberichts ist die Beschreibung eines Prototyps für eine auf Arbeitsprozesse des Forschungsverbund zugeschnittene virtuelle Arbeitsumgebung, keine „generische“ Lösung für die quantitative Sozialforschung. Erwartet wird aber, dass der Bericht Probleme von allgemeiner Bedeutung behandelt und für andere Verbundvorhaben nutzbare Lösungen bietet.

Akteure
Einrichtungen, die Primärdaten halten, sind nicht nur unter dem Gesichtspunkt des Datenzugangs zu beteiligen, sondern sollen auch ihre Interessen an besseren Datendienstleistungen einbringen. Die virtuelle Arbeitsumgebung soll auch ihre Datendokumentation verbessern helfen (z.B. SOEP: Erweiterung von SOEPinfo um Infos zu Fragenkontext und Generierung, PanelWhiz zur Archivierung und Nachnutzung von Arbeitsdateien; LSKN: Verbesserung des Metadaten-Infosystems der FDZ der Länder).
Offene Frage: Erhebungsinstitute, Vernetzungsstrukturen der Datenhalter, Datenarchive einbeziehen?

Forschungsdatenarchiv
Der Schwerpunkt der Entwicklung soll auf Archivierung, Dokumentation und Nachnutzung/ kollaborativer Nutzung von Auswertungssyntax liegen. Syntax soll zu den Daten verlinkt sein, auf die sie angewendet wurden. Archivierung und Dokumentation soll für Verbundpartner während der Projektlaufzeit verpflichtend sein. Qualitätssicherung durch Herstellung von (Fach-)öffentlichkeit, nicht als eigenständige Moderationsleistung.

Archivierung von Arbeitsdateien und Output soll technisch möglich sein. SUF sollen für Nutzergruppen in die Arbeitsumgebung (Archivstruktur) integriert sein. Ziel ist aber nicht eine zentralisierte und standardisierte Arbeitsweise der Verbundpartner/innen.

- Kann Remote-Onsite-Rechnen (gespiegelte Datenverarbeitung) in FDZ-Datenbeständen ermöglicht werden? Wenn ja, wäre dieser Nutzungsweg in die Arbeitsumgebung zu integrieren.
- Kann explorative Syntaxentwicklung durch verteilungstreue Strukturdatensätze unterstützt werden?
- Mögliche Anreize für freiwillige Archivierung von Arbeitsdateien, Outputs?
- Regelungen für Archivierung zum Zugang von Syntax nach Ende der Projektlaufzeit?
- Löschfristen für Originaldaten und Arbeitsdateien und Archivierungsvorschriften (gute wissenschaftliche Praxis) sind widersprüchlich.

Arbeitsumgebung

Die Arbeitsumgebung soll vor allem ein SubVersion Control System und ein Metadatenmanagement (automatische Metadatenextraktion) bieten, ebenso einen Dienst zur Verwaltung von Zugangsrechten (u.a.: Nutzungsverträge für Scientific-Use-Daten; Accounting für Zugriffe), Konvertiermöglichkeiten von Arbeitsdatensätzen und Syntax zur Nachnutzung und ein Wiki.

Kommerzielle (lizenz- und kostenpflichtige) Software (z.B. SPSS, Stata) soll nicht Teil der Arbeitsumgebung sein, Open-Source-Software (z.B. R) soll integriert sein.

Offene Fragen:

- Welchen Zusatznutzen versprechen Rechnerverbünde? („Rechnen, wo die Daten liegen.“)
- Welche Dienste zur Unterstützung von Textproduktion?

Nutzungs- und Zugangsrechte

Nutzung setzt eine eindeutige ID voraus, befristete und verlängerbare persönliche Zertifikate (IP-Adresse reicht nicht).

Für Scientific-Use-Daten sind Nutzergruppen zu bilden: Zugang zu Arbeitsdateien nur für Verbundpartner mit Einzelnutzungsrechten. FDZ müssen wissen, wo Daten liegen.

Daten auf Grid-Rechnern müssen nach außen geschützt sein.

Wird ins Forschungsdatenarchiv eingestellte Syntax genutzt, müssen die Urheber zitiert werden.

Verabredungen

- Protokoll und Zwischenbericht an alle Beteiligten.
- Anhand des Zwischenberichts Gespräch mit RatSWD suchen.
- WissGrid bietet erstes Architekturkonzept an. Angebot für technische Evaluation wird eingeholt.

Anhang 2 Zweiter Workshop „Virtuelle Arbeitsumgebung für sozioökonomische Forschung und Berichterstattung“

Göttingen, 19. Juli 2010

Teilnehmer/innen

Dr. Peter Bartelheimer (SOFI, Göttingen), Stefan Bender (FDZ, Forschungsdatenzentrum der Bundesagentur für Arbeit im IAB, Nürnberg), Sarah Cronjäger (SOFI, Göttingen), Prof. Dr. Frank Dickmann (Universitätsmedizin Göttingen), Dr. Thomas Drosdowski (GWS, Osnabrück), Dr. Harry Enke (WissGRID, Astrophysikalisches Institut Potsdam), Anke Gerhardt (FDZ der Statistischen Ämter der Länder), Peter Gietz (Gap-SLC-Projekt), Christian Grimme (Institut für Roboterforschung, TU Dortmund), Heidi Hanekop (SOFI, Göttingen), Patrick Harms (Abteilung Forschung & Entwicklung der Niedersächsischen Staats- und Universitätsbibliothek Göttingen), Rita Hoffmeister (FDZ im Landesbetrieb für Statistik und Kommunikationstechnologie Niedersachsen (LSKN), Hannover), Alexia Meyermannn (Datenservicezentrum für Betriebs- und Organisationsdaten, Universität Bielefeld), Hans Nerlich (Projektträger im Deutschen Zentrum für Luft- und Raumfahrt e.V. Umwelt, Kultur, Nachhaltigkeit, Bonn), Tanja Schmidt (Schmidt-Sozialforschung, Berlin), Prof. Dr. Uwe Schwiegelshohn (Institut für Roboterforschung, TU Dortmund), Dr. Michael Stegmann (FDZ, Deutsche Rentenversicherung, Würzburg).

Programm

Das Programm beinhaltet: 1) Eine kurze Einführung (Peter Bartelheimer, Tanja Schmidt); 2) Architekturskizze einer virtuellen Arbeitsumgebung: Struktur und Elemente der virtuellen Arbeitsumgebung (Patrick Harms); 3) Diskussion; 4) Schnittstelle zur Dateninfrastruktur: Datenbereitstellung und –sicherheit in der virtuellen Arbeitsumgebung (Harry Enke); 5) Diskussion und Feedback von Forschungsdateneinrichtungen; 6) Kollaborative Arbeitsprozesse: Nichttechnische Voraussetzungen für die Nutzung einer virtuellen Arbeitsumgebung (Peter Bartelheimer); 7) Diskussion; 8) Kommentar zum Diskussionsstand (Hans Nerlich); 9) Abschlussdiskussion.

Folgende Präsentationen sind unter der Projektwebsite abrufbar:

- Peter Bartelheimer/ Tanja Schmidt „Einführungspräsentation“:
- http://www.soeb.de/fileadmin/redaktion/downloads/workshop_2_einfuehrungspraesentation_bartelheimer_schmidt.pdf
- Peter Bartelheimer/ Tanja Schmidt „Nichttechnische Voraussetzungen kollaborativer Arbeitsprozesse“:
  http://www.soeb.de/fileadmin/redaktion/downloads/workshop_2_praesentation_bartelheimer_schmidt.pdf

- Harry Enke:
- http://www.soeb.de/fileadmin/redaktion/downloads/workshop_2_praesentation_harry_enke.pdf
- Patrick Harms:
- http://www.soeb.de/fileadmin/redaktion/downloads/workshop_2_praesentation_patrick_harms.pdf

Panel 1 - Architekturskizze einer virtuellen Arbeitsumgebung

Original- und Forschungsdaten

- Vorgenommene Änderungen in der VAU sind dokumentiert und nachvollziehbar. Damit lassen sich u.a. auch die wissenschaftlichen Arbeitsprozesse dokumentieren.
- Der Datenzugriff in der VAU lässt sich pro Person festlegen.
- Das IAB bietet bereits einen Dokumentationsservice an, so dass bestimmte Datensätze von Wissenschaftlern an das IAB zur Sicherung und zur Replikation gegeben werden können. Damit wird die gute wissenschaftliche Praxis gesichert. In der VAU können dazu Metadaten verbleiben, die auf den Datensatz bei dem IAB verweisen. Die Vertreterin des FDZ der Statistischen Ämter der Länder hält dies für eine sinnvolle Erweiterung des Leistungsspektrums der FDZ.
- Eine Standardisierung der Metadaten wird als sinnvoll erachtet (DDI und andere Community-Standards). FDZ IAB führt DDI-Standards schrittweise ein; zunächst schreiben die Nutzer/innen in eine Metadatenbank.
- Bei „Panelwhiz“ garantiert der beauftragte Träger keine Nachhaltigkeit. Auf die muss aber Verlass sein.
- Der Projektträger sieht durch den verbesserten Zugang zu den Forschungsdaten die Möglichkeit, bessere Forschungsergebnisse zu erzielen. Die bessere Forschung soll als Ziel im Vordergrund stehen. Damit können auch nicht technisch versierte Nutzer von den Vorteilen der VAU überzeugt werden. Besserer Zugang zu Daten hat bereits auf die Datenproduzenten zurück (Bedingung: Kritik zulassen).
- soeb: Fokus soll auf Unterstützung von Arbeitsprozessen liegen
- Das Hauptgeschäft der sozioökonomischen Forschung findet nach der Datenschnittstelle zu den FDZ statt. Daher bestehen durch den Einsatz der VAU Vorteile für die Wissenschaftler auch ohne Änderungen bei den FDZ:

Nachweisbarkeit

Optimierungspotenzial im Forschungsprozess → z.T. Rechenleistung aber vor allem interne Zusammenarbeit z.B. an gemeinsamer Syntax

Nutzungsrechte können transparent verwaltet werden und ermöglichen die bessere Administration in Arbeitsgruppen

- Das Nationale Bildungspanel (NEPS) diskutiert über ähnliche Lösungen wie das Modellprojekt – es wird empfohlen, dies zu berücksichtigen.
- Dabei ist zu berücksichtigen, dass die VAU nicht die organisatorischen Probleme der Koordinierung und Leitung eines Forschungsverbundes lösen kann (sh. Panel 3). Hier soll gemeinsam mit den FDZ an möglichen Lösungen gearbeitet werden. Ziel ist der Aufbau einer Sicherheitsachitektur, die Datenprovider akzeptieren.
- Das FDZ der Statistischen Ämter der Länder bittet um Einbindung.
- Es ist auch zu klären, welche Sanktionen bei Verstößen gegen die Nutzungsbedingungen innerhalb der VAU zur Anwendung kommen. Andere Statistik-Lösungen außer R sollen erst zu einem späteren Zeitpunkt ebenfalls in die VAU integriert werden.
- soeb/FDZ: Klären, was bei Nutzungsverstößen passiert.
- Die Erfahrung in der soeb zeigt, dass teilweise für Berechnungen leistungsfähige IT-Komponenten benötigt werden. „Processing“ steht aber bei der Entwicklung einer VAU nicht im Zentrum.
- Einzelrechnerkosten haben sich verändert: heute sind die Betriebs-, Wartungs- und Administrationskosten um ein Vielfaches höher, so dass eine effiziente Auslastung der Rechenkapazitäten für eine effektive Verwendung von Fördermitteln notwendig ist.

Panel 2 - Schnittstelle zur Dateninfrastruktur

Dateninfrastruktur

- Erste Schritte in der Zusammenarbeit von verschiedenen FDZ mit anderen FDZ wurden soeben getan, weiteres ist in Planung.
- soeb/FDZ: Zusammenarbeit zur Erarbeitung von Anbindungs- und Integrationslösungen von Originaldaten in und an die VAU.
- Verschiedene FDZ (IAB und FDZ der stat. Landesämter) zeigten Interesse weiterhin in den Prozess der Entwicklung der VAU eingebunden zu werden und an der weiteren Entwicklung von Lösungen mitzuarbeiten
- Nicht nur die nationale Perspektive ist von Bedeutung, auch OECD, EU und Standards der Institutsorganisationen sind relevant.
- Erste Schritte sollen mit der VAU bereits im Testbetrieb gegangen werden. Die FDZ zeigten Interesse an der Mitarbeit an den Policies. Dabei sollte die Einbindung in kleinen Schritten vorgenommen werden.
- soeb: ISO/BSI-Standards berücksichtigen
- Das NEPS (Nationales Bildungspanel) beschäftigt sich auch mit diesen Problemen, sollte ebenfalls in den Prozess eingebunden werden

- Weitere datenhaltende Institute, NutzerInnen und ExpertInnen sollen in die Planung, Entwicklung und Vorbereitung zur Nutzung einbezogen werden.

Sicherheit

- Digitale Zertifikate ermöglichen die eindeutige Identifikation einer Person in IT-Umgebungen. Die Problematik, dass sichergestellt ist, dass die Person, die das Zertifikat nutzt, auch die berechtigte/richtige Person ist, kann nur durch physische Prüfung am Standort gelöst werden. Entsprechende Verifikationsmechanismen (z.B. FDZ BA: „FDZ in FDZ“) werden daher pro Standort benötigt.
- soeb: Differenzierung für allgemeine Lösung notwendig.
- Es bestehen auch in den Geisteswissenschaften und der Medizin sehr hohe Anforderungen an den Datenschutz und die Datensicherheit.
- Datensicherheit nach ISO-Norm oder Vorgaben des BSI stellt Datenschützer der FDZ i.d.R. zufrieden.
- Die PKI in D-Grid ist durch EUGrid/PMA akzeptiert und hat damit EU-weite Akzeptanz erfahren.
- Die Sicherheitslösungen der VAU basierend auf der Grid Security Infrastructure (GSI) sollten Datenschützer zufriedenstellen. Remote Zugriff kann mit GSI realisiert werden, sofern Remote Zugriff nicht von vornherein ausgeschlossen ist.
- FDZ BA: Eine Zertifizierung ist für unsere Datenschützer nicht ausreichend für remote access. Auch die gemeinsame Nutzung von Rechnerkapazität wäre beim IAB rechtlich nicht zu machen. Auslieferung von SUF über das Web ist dagegen kein Problem (sicherer als Post).
- Eine „single sign-on“-Lösung ist nicht absehbar. Eher nutzerbasierter als rollenbasierter Zugang.
- Andere Sicherheitstechniken wie in GAP-SLC-Projekt wären evtl. auch interessant, sind zu prüfen
- Eine Einschätzung / Expertise von juristischen Personen zum Datenschutz- und zur Datensicherheit sollten bei der weiteren Entwicklung berücksichtigt werden
- Exp.: Evtl. Prüfen von SLC-Sicherheitstechnik

VAU Allgemein

- Die VAU ist die Lösung für die Zukunft. Bei der Implementierung kann von der soeb als Anwendungsfall ausgegangen werden und für das Projekt die benötigte VAU aufgebaut werden. Standards sind für die soeb ebenfalls von großer Bedeutung.
- Eine Differenzierung zwischen der sozioökonomischen Berichterstattung und allgemeiner empirischer sozialwissenschaftlicher Forschung ist nötig um, ausgehend von soeb

später auf die gesamte empirische sozialwissenschaftliche Forschung übertragen zu können.

- Eine prototypische Entwicklung wird als sinnvoll angesehen, um die Entwicklung gemeinsam mit den Nutzenden schrittweise voranzutreiben.
- Der bottom-up-Ansatz soll nach Möglichkeit verfolgt werden und mit top-down verknüpft werden: Schrittweises Vorgehen mit Rückkopplung der NutzerInnen im Entwicklungsprozess.
- Weitere Datenlieferanten, Nutzer und Experten sollen in die Planung, Entwicklung und Vorbereitung zur Nutzung einbezogen werden. „Wir gehen gern mit, wollen eingebunden sein, und wenn das Ressourcen kostet, wollen wir die haben." FDZ Länder: Man müsse im Detail rechtlich prüfen, ob die Mikrozensus-Vertragsbedingungen eine Arbeit in einer VAU zulassen. Angeregt wird ein weiterer Workshop zu den rechtlichen Fragen mit FDZ.
- Zu Beginn von soeb3 sollte eine Basis-VAU zur Verfügung stehen und Überarbeitungsmöglichkeiten während des Forschungsprozesses gegeben sein.
- soeb: Prototypische Entwicklung anwenden.

Panel 3 - Kollaborative Arbeitsprozesse

Projektkoordination

- Es wird durch den Einsatz der VAU ein höherer Organisations- und Steuerungsaufwand in der soeb erwartet. Die Vorteile durch bessere Forschungsleistung werden jedoch auch sehr hoch eingeschätzt.
- Der Einsatz der VAU erfordert Veränderungen in der Arbeitsweise der soeb. Die Community muss diesen Entwicklungsprozess durchlaufen, um die Vorteile der VAU nutzen zu können.

Fach-Community

- Die Arbeitsteilung in der Community bzw. den Projekten wird sich ändern. Kollaboration setzt durch die VAU früher an.
- Nutzer sind nicht bereit, Beschränkungen zu akzeptieren. Sie wollen „geheime Projekte, bekannte Software".
- Bisherige Erfahrungen mit Syntax-Sharing sind nicht ermutigend. Zitiert zu werden, ist kein starkes Incentive (time lag für Veröffentlichungen bis zu drei Jahren). Das Sharing der Syntax bietet jedoch den Vorteil, dass Doppelentwicklungen vermieden werden können.
- Syntax-Sharing als Bedingung für die Teilnahme am Forschungsverbund zu realisieren wäre eine Möglichkeit, übt jedoch viel Zwang aus. Besser wäre Akzeptanz.

Panel 4 - Abschlussdiskussion

- Für ein Verbundprojekt soeb 3 wird eine Verbesserung im Datenzugang und den Datennutzungsmöglichkeiten angestrebt, ebenso verbesserte kollaborative Arbeitsmöglichkeiten.
- Die Möglichkeiten der VAU sollen im Rahmen der Anwendung/Projektarbeit erkundet werden. Dies sollte parallel passieren. Auch wenn dies mehr Aufwand bedeutet könnte zunächst auf einem niedrigerem Level beginnen und sukzessive ausgebaut werden. Einige der jetzt zu beteiligenden FDZ sollten in einen neuen Verbund einbezogen sein. Praktische und rechtliche Probleme des Datenzugangs und der Vernetzung sollten für einen neuen Verbund geklärt werden; remote access steht noch nicht zur Debatte. Eine VAU sollte projektbegleitend mit entwickelt werden; „auf niedrigem Niveau einfach mal anfangen“.
- Dabei wurden einige Ziele bereits erreicht: Vorbehalte abgebaut, Datenproduzenten & Forschungsverbund an einem Tisch und Bereitschaft zur Zusammenarbeit.
- Ein Workshop zu rechtlichen Fragen noch während der Konzeptphase für soeb 3 könnte sinnvoll sein.

Anhang 3 Dritter Workshop „Rechtliche Aspekte der Nutzung von Forschungsdaten“

Göttingen, 05. November 2010

Teilnehmer/innen

Dr. Peter Bartelheimer (SOFI, Göttingen), Stefan Bender (FDZ BA, IAB, Nürnberg), René Büttner (SOFI, Göttingen), Sarah Cronjäger (SOFI, Göttingen), Dr. Harry Enke (WissGRID, Astrophysikalisches Institut Potsdam), Tatjana Mika (FDZ, Deutsche Rentenversicherung, Berlin), Hans Nerlich (Projektträger im Deutschen Zentrum für Luft- und Raumfahrt e.V. Umwelt, Kultur, Nachhaltigkeit, Bonn), Michael Schmaus (FDZ der Statistischen Landesämter, Düsseldorf), Tanja Schmidt (Schmidt-Sozialforschung, Berlin), Prof. Dr. Jürgen Schupp (DIW, Berlin), Dr. Michael Stegmann (FDZ, Deutsche Rentenversicherung, Würzburg), Dr. Heike Wirth (GESIS, Mannheim).

Programm

Das Programm beinhaltet: 1) Eine kurze Einführung (Peter Bartelheimer, Hans Nerlich); 2) Veränderte Arbeitsstrukturen in einer virtuellen Arbeitsumgebung (Tanja Schmidt); 3) Datenbereitstellung und Sicherheit in einer virtuellen Arbeitsumgebung (Harry Enke); 4) Erster Diskussionsblock zu Fragen der Vertragsgestaltung mit Empfänger/in der Daten und Datenschutzkonzepte beteiligter Forschungseinrichtungen; 5) Zweiter Diskussionsblock zu Fragen der personellen Nutzungsbeschränkung, Befristung der Nutzung, Zweckbindung der Datennutzung, Anforderungen an Auswertungsprogramme und Metadaten; 6) Ergebnissicherung, Zwischenfazits des Diskussionsstands.

Folgende Präsentationen sind unter der Projektwebsite abrufbar:

- Peter Bartelheimer/ Tanja Schmidt „Veränderte Arbeitsstrukturen in einer virtuellen Arbeitsumgebung“: http://www.soeb.de/fileadmin/redaktion/downloads/pr_sentation_bartelheimer_schmidt_workshop_3_05112010.pdf
- Harry Enke:
- http://www.soeb.de/fileadmin/redaktion/downloads/enke_2010_11_05_workshop_sofi_datenbereitstellung.pdf

Zur Expertise und zu den Präsentationen

Der Begriff „Originaldaten“ sollte nicht für SUF verwendet werden, da es sich um anonymisierte Daten handelt. Einheitlich sollte in den Darstellungen von Ausgangsdaten gesprochen werden. Zudem sollte unterschieden werden zwischen „personenbezogen“ versus „Personen beziehend“.

Welchen Sicherheitsstandards Provider von Speicherplatz bieten, ist nicht generell durch das GRID festgelegt. Provider einer virtuellen Arbeitsumgebung sollten ein account management für virtuelle Organisationen bieten.

Die Vertreter/innen der an der Diskussion beteiligten FDZ zeigten generelle Bereitschaft, an der Entwicklung mitzuwirken, verwiesen aber auf aus ihrer Sicht zentrale Datenschutzprobleme. Sie stellten übereinstimmend fest, dass für alle Forschungseinrichtungen, die in einer virtuellen Arbeitsumgebung (VAU) kooperieren, Einzelnutzungsverträge mit den (berechtigten) Dateneinrichtungen bestehen müssen. Die Vertragsgestaltung setzt verantwortliche (aufsichtspflichtige) Antragsteller voraus, die in einer hierarchischen Arbeitsorganisation unterbinden können, dass Mitarbeiter/innen Datenschutzanforderungen verletzen, und die eine Nutzung ausschließlich für den vertraglich bestimmten Forschungszweck an lokalen Standorten sicherstellen. Die Verträge der amtlichen FDZ die Zahl und die Identität der Nutzer/innen fest; auch andere FDZ verlangen die Benennung der Stellen und Personen, die personenbezogene Daten bearbeiten. Die virtuelle Organisation kommt als Vertragspartner/in nicht in Betracht. Gruppen- oder Verbundverträge können auch für die beteiligten Nutzer/innen ein erhöhtes Haftungsrisiko bedeuten: Bei einem individuellen Verstoß gegen den Datenschutz (z.B. einer unzulässigen Datenverknüpfung) würden alle Verbundeinrichtungen von der Datennutzung ausgeschlossen.

Ein FDZ wägt ab, dass eine VAU einerseits die vertraglich unterstellte Hierarchie schwächt, andererseits aber einen weniger privaten Raum für die Datennutzung schafft (Syntax liegt an einem öffentlichen Ort). Ein anderes FDZ kann sich vorstellen, dass SUF in einer VAU gehalten werden und laufen, wenn diese gegen nicht nutzungsberechtigte Dritte abgegrenzt ist (nicht jedoch z.B. regionale Kontextdaten, die das Risiko der De-Anonymierung erhöhen würden. Oder der Provider wäre selbst eine Forschungseinrichtung − in der Diskussion blieb offen, ob das z.B. auf die GWDG zuträfe − und als solche Verbundpartner.

Konsens besteht darüber, dass der „Idealzustand“ für eine VAU Datenfernverarbeitung (remote access processing) über eine remote-access-Plattform wäre. Die Daten würden das FDZ gar nicht verlassen, damit ist auch die Gefahr durch unsichere Leitungen gebannt. Zukunftsträchtig wäre eine gemeinsame Remote-Access-Architektur (nächster Schritt „FDZ in FDZ, Zugangskontrolle ohne Verletzung von Persönlichkeitsrechten). In dieser Richtung zeichnen sich Parallelentwicklungen ab (IAB-Ausschreibung Metadatenbanksystem, EU-Projekt Data Without Boundaries, DDI-Datendokumentation).

Ein weiteres rechtliches Problem liegt darin, ob bestehende Datenschutzkonzepte beteiligter Forschungseinrichtungen auf eine VAU übertragbar sind. Ist dies nicht der Fall, müsste das Sicherheitskonzept für die Datenaustauschplattform von den FDZ, die Daten bereitstellen sollen, zeitaufwändig und mit unsicherem Ergebnis geprüft werden.

Zweiter Diskussionsblock: Nutzungsbeschränkungen, Auswertungsprogramme, Metadaten
Aus Sicht der FDZ sind für die Lösung von Fragen der personellen Nutzungsbeschränkung, Befristung und Zweckbindung zwei Szenarien zu unterscheiden: Liegen die Forschungsdaten nur bei einem zentralen Provider oder auch auf individuellen Workstations. Was ist mit Arbeitsdateien? Wann dürfen Mikrodaten auf die Workstation gespeichert werden? Lässt sich abgleichen, was auf dem Repository, was auf Workstations vorhanden ist? Kann man einen gemeinsamen Ort des Datenaustauschs in individuelle Datennutzungsverträge mit Partnereinrichtungen einer VAU aufnehmen? Was für die Nutzer/innen am einfachsten wäre, ist datenschutzrechtlich am kompliziertesten.

Partnereinrichtungen einer VAU müssten als berechtigte Nutzer/in anerkannt sein. Mit dem Provider einer VAU, der nicht zur scientific community gehört, hätten die FDZ kein Vertragsverhältnis. Forscht der Provider nicht selbst und lägen die Daten bei ihm, betriebe er eine rechtlich nicht zulässige Datenvorratshaltung. Eine Lösung könnte darin bestehen, dass Provider Vertragspartner werden – etwa ein Rechenzentrum als An-Institut, so dass ggf. Vertrag mit der Universität geschlossen werden könnte („vertragstechnisch am schönsten"). Alternativ wären die Provider (z.B. ein Rechenzentrum) zu zertifizieren. Die Zertifizierung müsste sicherstellen, dass Mitarbeiter/innen des Rechenzentrums die Daten nicht einsehen können und ISO-Normen für Datensicherheit eingehalten werden; zu den Anforderungen könnten auch sichere Leitungen gehören.

Lösungen für schwach anonymisierte Daten wären noch komplizierter.

Für das Verbot der Datenzusammenführung würde die Datenhaltung auf einem Rechner kein neues Problem darstellen.

Personelle Nutzungsbeschränkung und Genehmigungsvorbehalt wären in der skizzierten Architektur gewährleistet: Man kann gruppenbasierte Personenrechte nachvollziehen. Die Protokollierung durch eine dritte Instanz erlaubt sogar eine bessere Nutzer/innen/-kontrolle – eine Qualität, die man gegenüber den FDZ herausstellen sollte.

Für die FDZ ist die Befristung der Nutzung entscheidend. In einem größeren Projekt sollte die Nutzung für vier oder fünf Jahre beantragt werden. Überlässt die bisherige Vertragsgestaltung die tatsächliche Entscheidung, wann welche Daten gelöscht werden, den Nutzer/inne/n überlässt, wäre dies in einer VA dies technisch gesichert und nachvollziehbar.

Einige FDZ bieten den Nutzer/inn/en nach Ablauf der Nutzungsfrist für Arbeitsdateien Langzeitarchivierung (zehn bis 20 Jahre) und Wiederaufrufbarkeit. Dabei sichert die von den Nutzer/innen entwickelte Syntax die Nachnutzbarkeit; Programme müssen auf den Ursprungsdatensätzen lauffähig sein. Die FDZ verweisen auf ihre Kompetenz bei Datenmanagement, Versionierung und Versionskontrolle (einzigartig zitierbar über DOI, Persistent Identifyer).

Ein offener Diskussionspunkt blieb, ob eine Archivierung beim Bundesdatenarchiv eine weitere Alternative zur individuellen Archivierung darstellt.

Verwiesen wurde darauf, dass in der amtlichen Statistik die Nutzungsgebühren erhöht werden, da die FDZ sich teilweise über Gebühren finanzieren müssen, was auch Veränderungen für die möglichen Nutzungsfristen nach sich zieht.

Metadaten und Auswertungssyntax unterliegen keinen Vertragsbeschränkungen. Bei SUF haben Nutzer/innen völlige Freiheit bei Programmierung und Dokumentation; für Fernrechnen und Onsite-Nutzung bestehen Anforderungen an die Dokumentation. Die FDZ beklagen, dass an Informationen über generierte Variablen, Wissen über Datenmängel und Datenqualität bislang kaum etwas an die Daten haltenden Einrichtungen zurückkommt. Die FDZ bekommen auch die vertraglich zustehenden Publikationen nicht. Metadatensysteme mit Wiki-Elementen setzen generell eine andere Bereitschaft der Nutzer voraus, sich an „kollektiven Gütern“ zu beteiligen.

Anhang 4 Vierter Workshop „Auswirkungen einer Virtuellen Arbeitsumgebung auf Arbeitsabläufe in der sozioökonomischen Forschung und Berichterstattung“

Göttingen, 15.Dezember .2010

Teilnehmer/innen

Dr. Peter Bartelheimer (SOFI, Göttingen), Mara Boehle (GESIS, Mannheim), René Büttner (SOFI, Göttingen), Sarah Cronjäger (SOFI, Göttingen), Dr. Thomas Drosdowski (GWS, Osnabrück), Dr. Harry Enke (WissGRID, Astrophysikalisches Institut, Potsdam), Christian Gerhards (Universität Bielefeld), Patrick Harms (WissGRID, SUB Göttingen), Prof. Dr. Jürgen Kädtler (SOFI, Göttingen), Dr. Florian Köhler (FDZ LSKN Niedersachsen, Hannover), Jens Ludwig (WissGRID, SUB Göttingen), Hans Nerlich (PT DLR, Bonn), Tanja Schmidt (Schmidt-Sozialforschung, Berlin), Ewa Sojka (INIFES, Stadtbergen), Prof. Dr. Christoph Weischer (Universität Münster).

Programm

Das Programm beinhaltet: 1) Eine kurze Einführung (Peter Bartelheimer); 2) Arbeiten im Verbund – Praktische Probleme in Arbeitsabläufen, technische und nichttechnische Lösungen (Peter Bartelheimer); 3) Ein Tag in der virtuellen Arbeitsumgebung – Architekturskizze und Anwendungsfall (Tanja Schmidt); 4) Diskussionsrunde zu den Funktionalitäten: Konfigurations- und Verwaltungswerkzeuge (u.a. Rechteverwaltung), Datenbezogene Arbeitswerkzeuge, Werkzeuge für die Ausbaustufe (Kollaborations-, Datenanbieter- und Publikationswerkzeuge); 5) Ergebniszusammenführung und –sicherung, Abschlussdiskussion.
Folgende Präsentationen sind unter der Projektwebsite abrufbar:

- Peter Bartelheimer „Arbeiten im Verbund – nichttechnische und technische Anforderungen“: http://www.soeb.de/fileadmin/redaktion/downloads/pr_sentation_bartelheimer_workshop_4_15122010.pdf
- Tanja Schmidt „Ein Tag in der virtuellen Arbeitsumgebung – Architekturskizze und Anwendungsfall“: http://www.soeb.de/fileadmin/redaktion/downloads/pr_sentation_schmidt_workshop_4_15122010.pdf

Diskussionsergebnisse

Auf Basis der Präsentationen ergaben sich durch Kartenabfrage verschiedene Diskussions-Cluster.

Cluster 1: Virtuelle Arbeitsumgebung (VAU)
Es handelt sich dabei um keine fertige Anwendung, sondern um ein Komponentensystem mit Modulen, der Zugang geschieht via Browser über ein Portal, existierende Komponenten liegen gebündelt beim Provider und es werden möglichst kleine Pakete gepackt. Die Module können über eine Benutzeroberfläche an- und abgewählt werden.

Das System ist absturzsicher: Es wird „backend" durch professionellen Provider und überinstitutionelle Ressourcen sichergestellt. Dementsprechend ist zwar möglich, dass Einzelkomponenten abstürzen, aber die Daten wären nicht weg. Die Implementation der VAU hat noch nicht begonnen.

Welche Instrumente sind sinnvoll für die Zusammenarbeit? Das Instrument der Versionsverwaltung, kann als kollaborative Komponente gesehen werden, denn jede/r Projektpartner kennt und kann auf den letzten Stand der anderen Projektpartner/innen zugreifen. Nutzer/innen loggen sich auf ihrem Desktop in das Eingangsportal ein um die Komponenten nutzen zu können. Dabei können parallel auch lokale Programmkomponenten genutzt werden (z.B. auch in TextGrid).

Der zentrale Server steht bei einem überregionalen Provider wie der GDWG; bei diesen sind dann professionelle Leistungen einzukaufen.

Cluster 2: Datenschnittstelle
Es gibt nicht nur Probleme, sondern auch Lösungen – interessant für die gesamte Daten-Landschaft.
Die FDZ haben Interesse an der Zertifikats-Infrastruktur für remote access und auch an aktiver Forschungsbeteiligung.

Cluster 3, andere, aggregierte Daten
Wo enden Beschränkungen aufgrund des Datenschutzes? Das heißt, wann findet der Übergang zu öffentlichen Daten in der VAU statt? Diese Frage stellt sich besonders bei Betriebsdaten – wer hat die Rechte an diesen Daten? Insbesondere bei Organisationsdaten von Betriebsfallstudien wurde für den Betriebszugang Anonymität vorausgesetzt. Daraus entstehen Identifikationsprobleme. Die VAU setzt damit ein Umfeld voraus, das durch die Datenhalter definiert wird. Neue „Umgebungsprobleme" entstehen dann durch andere, weitere Datenhalter, z.B. Betriebsdaten.

Urheber- und Nutzungsrechte: öffentlich finanzierte Daten sind für die Wissenschaft frei.

Aggregatdaten bedürfen einer detaillierten Beschreibung, was einen großen Aufwand bedeutet. Und insbesondere Organisationsdaten haben datenschutztechnisch spezielle Anforderungen und andere „Umgebungsprobleme“.

Metadaten sollen nicht nur technische Entstehungsinformationen enthalten, sondern müssen die Daten auch fachwissenschaftlich charakterisieren.

VAU muss einerseits den Datenzugang für geschlossene Arbeitsgruppen unterstützen, andererseits aber auch die Arbeit in gemeinsam benutzten Bereichen. Es muss ein Wissen bestehen, welche Daten es im Projekt sonst noch gibt. Die Möglichkeit, sich zu informieren, kann nur technisch unterstützt werden (Datei- und Literaturverwaltung). Eventuell können die Nutzer/innen auch durch Meldungen aus dem System hingewiesen oder erinnert werden. Jedoch bleibt die Frage offen, wer das moderiert. Die VAU kann auf jeden Fall die Transparenz in arbeitsteiligen Zusammenhängen erhöhen, jedoch ist dies nicht rein technisch lösbar.

Denn die Technik führt auch zu neuen Problemen. Dabei ist der Koordinationsaufwand mit IT-Problemen zusammen zu sehen. Jedoch werden dadurch, dass technische Umsetzungen notwendig sind, Organisationsanforderungen bewusster, und der Koordinationsaufwand wird sichtbarer.

Insgesamt betrachtet befördern die technischen Abläufe kulturellen Wandel im Arbeiten und Publizieren, es gibt kaum noch Einzelautoren.

Welche bisher bereits formalisierte Teilarbeitsprozesse können durch die neue technische Basis gewinnen? VAU ist besonders da effektiv, wo vorher schon viel Standardisierung vorlag.

Cluster 4, Metadaten

Der DDI−Standard sollte bei Erstellung von Metadaten nicht in ganzem Umfang angewendet werden. DDI-Datensätze von externen Daten können auch importiert und nachgenutzt werden. Andererseits können selbst erstellte Metadaten im DDI-Format für andere Anwendungen exportiert werden. Problematisch ist nur, festzulegen, welche Informationen in die Metadaten eingehen sollen. Wenn man zu viel von den Nutzenden verlangt, bekommt man nichts. Evtl. Schlagworte vorgeben? Man wird sich dauerhaft damit beschäftigen müssen. Möglich wäre, auf die bereits aufbereiteten Metadaten von definierten Datensätzen der FDZ, z.B. Mikrozensus (auch über Microdata Lab) via links zuzugreifen und diese gezielt durchzuarbeiten. Außerdem lässt sich Prozesswissen nicht über Daten formalisieren.

Metadatensysteme sind nie fertig; Felder werden „ missbraucht“, mehrere Informationen werden in ein Feld eingetragen.

Welche Metadaten werden standardmäßig abgefragt? Welche technischen und fachlichen Daten werden extrahiert? „Kommentarfelder auszählen“.

Wer bekommt die Lizenz zum Löschen?

Cluster 5 „Syntax Sharing“

Es ist zu unterscheiden, ob iterativ oder parallel an Syntax gearbeitet wird. Bei beiden Anwendungsfällen treten Koordinationsprobleme auf, die nicht rein technisch zu lösen sind.

In der VAU sollten so viele generierte Variablen wie möglich verfügbar sein – eventuell nicht nur als Ergebnis aus dem Arbeitsprozess, sondern auch als Datenservice-Funktion. Hier ist die Servicequalität der VAU gefragt, z. B. wenn sich Variablennamen im SOEP ändern?

Schlussdiskussion

Virtuelle Forschungsumgebungen stoßen auf Akzeptanzprobleme in Sozial- und Geisteswissenschaften. Es sollte aufgezeigt werden, dass sie ein nützliches Instrument sind. Gegenstand einer neuen Bekanntmachung des BMBF (eHumanities, bis 30.09.2011) sind sowohl Infrastrukturprojekte als auch Forschungsprojekte.

Ideal wären Anwendungsfälle, wie in den Textwissenschaften, in denen durch VAU Fragen beantwortet werden können, die vorher nicht zu beantworten waren, und noch besser, die vorher gar nicht gestellt wurden.

Eine VAU ist sinnvoll als eine Möglichkeit für bestimmte komplexe Projekte, ihre Nutzung sollte jedoch keine Verpflichtung, sondern freiwillig sein.

Der verbesserte Datenzugang wäre ein wesentlicher Vorteil, den eine VAU allen sozialwissenschaftlichen Projekten bietet. Sonst ist eine VAU nur sinnvoll für Verbundprojekte. Darüber hinaus könnte sie Serviceleistungen für einzelne Datensätze (Modell: MISSY) oder Datenmanagement im Zeitverlauf unterstützen.

Durch die VAU wird die Schnittstelle zu den Daten von der „Verbraucherseite“ her gestärkt. Durch paralleles überörtliches Arbeiten lokal verteilter Wissenschaftler/innen nähern sich zwei unterschiedliche Akteursgruppen an, nämlich die Datenbesitzer/innen und die Datennutzer/innen.

Zusammenfassend ist immer zu bedenken, dass es nicht allein um technische Möglichkeiten geht, sondern um eine gemeinsame sozialwissenschaftliche „Story“.