

## Modellprojekt

# „Kollaborative Datenauswertung und Virtuelle Arbeitsumgebung“ – VirtAug

## Zwischenbericht

Tanja Schmidt

Peter Bartelheimer

Juni 2010

GEFÖRDERT VOM



Bundesministerium  
für Bildung  
und Forschung

## Inhalt

1.	Einleitung .....	1
2.	Sachstandsbericht.....	2
3.	Daten für die sozioökonomische Berichterstattung.....	4
3.1	Originaldatensätze .....	4
3.2	Syntax.....	11
3.3	Arbeitsdatensätze .....	13
3.4.	Outputs .....	14
4.	Kollaborative Datenanalyse – Anwendungsfälle .....	15
4.1	Datenmanagement-Syntax .....	15
4.1.1	Management von Längsschnittdaten des SOEP.....	15
4.1.2	Datenmanagement bei anderen Datensätzen.....	19
4.2	Generierungs- und Recodierungssyntax .....	21
4.2.1	Generierungs- und Recodierungssyntax in Lebensverlaufsanalysen mit dem SOEP .....	21
4.2.2	Generierungs- und Recodierungssyntax für andere Arbeitspakete .....	23
4.3	Analysesyntax .....	24
4.3.1	Sequenzanalysen-, Optimal Matching und Clusteranalysen mit dem SOEP.....	24
4.3.2	Analysesyntax für andere Arbeitspakete.....	26
4.4	Ergebnissicherung und -austausch .....	27
4.5	Dokumentation und Ergebnistransfer .....	28
4.6	Probleme und ihre Bewältigung im Arbeitsprozess .....	30
5.	Datenschnittstellen einer virtuellen Forschungsumgebung .....	32
6.	Anforderungen an eine virtuelle Arbeitsumgebung für die sozioökonomische Berichterstattung .....	35
6.1	Anforderungen an Langzeitarchivierung und Forschungsdatenarchiv .....	35
6.2	Anforderungen an Metadatenextraktion .....	36
6.3	Anforderungen an Datenkonvertierung .....	37
6.4	Benutzeroberfläche .....	37
7.	Ausblick auf die zweite Projektphase.....	38

ANHANG .....	41
Protokoll des Workshops „Fachwissenschaftliche Anforderungen an eine virtuelle Arbeitsumgebung für SOEB“ – Göttingen, 9. Februar 2010.....	41
Teilnehmende.....	41
Statements zum Input von Tanja Schmidt.....	41
Zusammenfassung des Diskussionsstands.....	43

## **Tabellen und Abbildungsverzeichnis**

Abbildung 1: Beispiel eines Mikrodatsatzes: SOEP Originaldatenmatrix, wide-format	5
Abbildung 2: Zugriff auf die Integrierten Erwerbsbiografien (IEBS)	8
Abbildung 3: Zugriff auf die gepoolten Rentenzugangsstichproben (RZ)	8
Abbildung 4: Zugriff auf das Sozio-oekonomische Panel (SOEP)	9
Abbildung 5: Zugriff auf den Mikrozensus (MZ)	10
Abbildung 6: Zugriff auf Daten der Volkswirtschaftlichen Gesamtrechnung (VGR)	10
Abbildung 7: Multiple Zugriffe auf Originaldaten von „soeb 2“-Partner-Instituten	11
Abbildung 8: Datenmanagement mit dem SOEP	18
Abbildung 9: Geplante Datenrestrukturierung des SOEP	18
Abbildung 10: Ablauf Ferndatenverarbeitung	20
Abbildung 11: Generierung abhängiger Längsschnittvariablen mit dem SOEP	22
Abbildung 12: Kollaborative Entwicklung von Recodierungssyntax mit dem Mikrozensus	24
Abbildung 13: Entwicklung von Analysesyntax für Lebensverläufe	25
Tabelle 1: Ausgewählte Originaldatensätze in „soeb 2“	6

# 1. Einleitung

Von August 2009 bis November 2010 fördert das Bundesministerium für Bildung und Forschung eine fachöffentliche Konzeptphase für ein drittes Verbundvorhaben „Berichterstattung zur Sozioökonomischen Entwicklung Deutschlands“ (soeb 3). Die Durchführung liegt beim Soziologischen Forschungsinstitut (SOFI) an der Georg-August-Universität Göttingen. In einer Reihe fachöffentlicher „Werkstattgespräche“ sollen Themen und Akteure für einen dritten Bericht identifiziert werden. Die Werkstatteergebnisse werden dokumentiert und zu einem Vorschlag für ein Berichtsprogramm und für die Zusammensetzung eines Projektverbunds zu seiner Umsetzung zusammengefasst.

Teil der Konzeptphase ist ein Modellprojekt „Kollaborative Datenauswertung und Virtuelle Arbeitsumgebung“ (VirtAug). Im Modellprojekt soll untersucht und dokumentiert werden, wie die datenbezogene Kooperation von Sozialwissenschaftler/innen an verschiedenen quantitativ-empirisch orientierten Forschungseinrichtungen und insbesondere eine kollaborative Auswertung der Mikrodaten von Forschungsdatenzentren künftig besser organisiert werden kann und welche IT-Verfahren eine solche Arbeitsweise unterstützen können. Dabei sollen fachwissenschaftliche Anforderungen ermittelt und verfügbare IT-Ressourcen für eine angemessene IT-Infrastruktur evaluiert werden. Den „Anwendungsfall“ für diese Entwicklungsarbeit bildet der Forschungsverbund Sozioökonomische Berichterstattung. Ziel ist aber die Entwicklung einer Arbeitsumgebung, die in den Sozialwissenschaften breiter genutzt werden kann. Das Modellprojekt wird am SOFI von Dr. Peter Bartelheimer und Tanja Schmidt durchgeführt.

Dieser Zwischenbericht fasst – nach einem knappen Sachstandsbericht – Ergebnisse aus der ersten Projektphase des Teilprojekts VirtAug zusammen. Anhand von Anwendungsfällen aus der Arbeit am zweiten Bericht zur sozioökonomischen Entwicklung Deutschlands (soeb 2)<sup>1</sup> wurden aus Sicht der Datennutzer/innen funktionale Anforderungen an die IT-Unterstützung kollaborativer Datenanalyse und an eine virtuelle Arbeitsumgebung entwickelt. Abschließend werden Arbeitsaufgaben für die zweite Phase des Modellprojekts skizziert, in der insbesondere technische Umsetzungsmöglichkeiten mit Hilfe der bestehenden D-Grid-Infrastruktur zu prüfen sind.

---

<sup>1</sup> Mehr Informationen: [www.soeb.de](http://www.soeb.de).

## 2. Sachstandsbericht

In der ersten Phase des Teilprojekts VirtAug führte Tanja Schmidt insgesamt elf Expert/innengespräche mit Datennutzer/innen aus dem Forschungsverbund „Berichterstattung zur sozioökonomischen Entwicklung Deutschlands“ sowie mit Vertreter/innen datenhaltender Institute. Die Gesprächspartner/innen waren:

- Dr. Holger Alda, Forschungsdatenzentrum im Bundesinstitut für Berufsbildung (BIBB),
- Dr. Irene Becker, Empirische Verteilungsforschung,
- Dr. Thomas Drowsdowski, Dr. Marc Ingo Wolter, Gesellschaft für Wirtschaftliche Strukturforchung mbH (GWS),
- Dr. Sabine Fromm, Soziologisches Forschungsinstitut (SOFI),
- Tatjana Fuchs, Ewa Sojka, Falko Trischler, Internationales Institut für Empirische Sozialökonomie (INIFES),
- Dr. Anne Hacket, Institut für Sozialwissenschaftliche Forschung (ISF),
- Tatjana Mika, Forschungsdatenzentrum der Rentenversicherung (FDZ-RV),
- Prof. Dr. Jürgen Schupp, Forschungsdatenzentrum Sozio-oekonomisches Panel (SOEP) am Deutschen Institut für Wirtschaftsforschung (DIW).

In den Gesprächen mit den Datennutzer/inne/n wurden die Arbeitsabläufe bei der Datenanalyse und Erfahrungen mit der Zusammenarbeit im zweiten Verbundvorhaben „Berichterstattung zur Sozioökonomischen Entwicklung Deutschlands“ („soeb 2“) rekonstruiert und unter der Fragestellung ausgewertet, welche Anforderungen Nutzer/innen an eine virtuelle Arbeitsumgebung stellen. Die Gesprächspartner/innen der Datenhaltenden Institute wurden zur vorhandenen Dateninfrastruktur und zu Supportmöglichkeiten für die Datennutzer/innen sowie zukünftig geplanten Veränderungen in der Datenbereitstellung befragt.

Gleichzeitig stellte das SOFI Kontakt zum Projekt WissGrid in der D-Grid-Initiative<sup>2</sup> her, die IT-Plattformen für den gemeinsamen Zugriff auf Rechnerkapazitäten, Daten und Programme als universelle Werkzeuge für wissenschaftliche und wirtschaftliche Nutzungen entwickelt. WissGrid<sup>3</sup> soll konzeptionelle Grundlagen für die nachhaltige Nutzung der Grid-Infrastruktur sowie IT-technische Lösungen für verschiedene wissenschaftliche Disziplinen bereitstellen. Ziel der Kontaktaufnahme war es, unter den Projektbeteiligten Partner für eine technische Expertise zu gewinnen, um in der zweiten Projektphase verfügbare IT-Lösungen für eine virtuelle Arbeitsumgebung evaluieren zu können. Peter Bartelheimer nahm am 16. September 2009 am ersten Treffen der WissGrid-AG Langzeitarchivierung in Göttingen teil, Tanja Schmidt am Metadatenworkshop von WissGrid am 17. September 2009 in Göttingen.

---

<sup>2</sup> URL: <http://www.d-grid.de>.

<sup>3</sup> URL: <http://www.wissgrid.de>

Erste Überlegungen zu einer virtuellen Arbeitsumgebung für die sozioökonomische Berichterstattung wurden im Rahmen des Review-Workshops des WissGrid AP Langzeitarchivierung am 28. Januar 2010 im Astrophysikalen Institut Potsdam präsentiert.<sup>4</sup>

Im Anschluss an diese Expert/inn/engespräche fand am 9. Februar 2010 in Göttingen ein erster Projekt-Workshop "Fachwissenschaftliche Anforderungen an eine virtuelle Arbeitsumgebung für soeb 2" statt. Dort wurden die ersten Ergebnisse aus der Bestandsaufnahme präsentiert<sup>5</sup> und mit Forscher/innen des Forschungsverbundes sowie mit Vertreter/innen der datenhaltenden Institute und Experten aus der WissGrid-Initiative diskutiert.<sup>6</sup> Auf Basis dieser Ergebnisse wurde eine Leistungsbeschreibung zur Evaluierung der technischen Umsetzbarkeit einer virtuellen sozialwissenschaftlichen Arbeitsumgebung erstellt (vgl. unten: 7.). Am 12. April 2010 konnte mit der D-Grid Entwicklungs- und Betriebsgesellschaft mbH ein Forschungs- und Entwicklungsvertrag über die Erstellung einer entsprechenden technischen Expertise geschlossen werden. An der Expertise arbeiten 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). Wesentliche Ergebnisse der Expertise sollen am 19. Juli auf einem weiteren fachöffentlichen Workshop zur Diskussion gestellt werden.

---

<sup>4</sup> 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>.

<sup>5</sup> Tanja Schmidt: Kollaborative Datenanalyse und virtuelle Arbeitsumgebung – Erfahrungen und Anforderungen aus der Verbundarbeit an »soeb 2«, Beitrag zum Workshop "Fachwissenschaftliche Anforderungen an eine virtuelle Arbeitsumgebung für »soeb«", 9. Februar 2010, SUB Göttingen.

<sup>6</sup> Vgl. Anlage zu diesem Bericht.

### 3. Daten für die sozioökonomische Berichterstattung

Die sozioökonomische Berichterstattung basiert auf einer Vielzahl verschiedener sozialwissenschaftlicher Datenquellen. In den folgenden drei Abschnitten wird die Arbeitsweise des Forschungsverbands anhand ausgewählter Datenarten dargestellt und erläutert. Dabei sind die Daten selbst zu unterscheiden nach Originaldatensätzen (3.1), Syntaxdateien (3.2), Arbeitsdatensätzen (3.3) und Outputs (3.4).

#### 3.1 Originaldatensätze

Originaldatensätze sind das „Ausgangsmaterial“ der sozioökonomischen Berichterstattung wie jedes quantitativen Forschungsvorhabens in den Sozialwissenschaften. Solche Datensätze werden von den datenhaltenden Instituten, üblicherweise in verschiedenen Datenformaten für verschiedene Statistikpakete<sup>7</sup>, für die Forschung bereitgestellt. In einem solcher Datensatz bilden Variablen (meist in Spalten) für eine Vielzahl von „Fällen“ (meist in Zeilen) eine Datenmatrix. Untersuchungseinheiten („Fälle“) können Personen und Haushalte sein (Individualdaten oder Mikrodaten), Personengruppen oder Länder (aggregierte- oder Makrodaten) oder z.B. Unternehmen. Dabei können Mikrodaten durch Befragungen oder im Rahmen von Verwaltungsprozessen entstehen. Makrodatensätze werden von den datenhaltenden Instituten aus Mikrodaten generiert, so etwa die Daten der Volkswirtschaftlichen Gesamtrechnung durch das Statistische Bundesamt.

Abbildung 1 zeigt den Aufbau eines typischen fallorientierten Datensatzes. Jede Zeile repräsentiert einen Fall, hier eine Person mit allen Informationen in der entsprechenden Zeile, diese Form wird auch „wide-Format“ genannt. Im sog. „long-Format“ stellt jede Zeile die Information einer Person zu einem Zeitpunkt dar, d.h. in Längsschnittdatensätzen können pro Person mehrere Zeilen im Datensatz enthalten sein. Gleiches gilt für das sog. „spell-Format“: hier stellt jede Zeile ein Ereignis mit einem festgelegtem Zeitrahmen dar, wobei für jede Person mehrere Zeilen vorhanden sein können.

Der wissenschaftliche Zugang zu statistischen Einzeldaten (Mikrodaten) unterliegt gesetzlichen Bestimmungen, die sicherstellen sollen, dass Datenschutz und Statistikgeheimnis, also die Anonymität der Auskunftgebenden (Personen, Haushalte oder Unternehmen) gewahrt bleiben. Bundes- und Landesstatistikgesetze regeln die Datenerhebung durch statistische Ämter, Bundes- und Landesdatenschutzgesetze den Umgang mit personenbezogenen Daten in amtlichen und sozialwissenschaftlichen Umfragedaten. und Amtliche Mikrodaten dürfen nach § 16 BStatG wissenschaftlichen Einrichtungen auch ohne Einwilligung der Auskunftgebenden „faktisch anonymisiert“ zur Verfügung gestellt werden, d.h. so, dass

<sup>7</sup> Die häufigsten Statistikformate, die bedient werden, sind SPSS, STATA, SAS, ASCII und zunehmend auch R. Die entsprechenden Dateiformate sind: \*.sav oder \*.por für SPSS, \*.dta für Stata; \*.sas für SAS, \*.Rdata für R. Makrodaten, die über das Internet bezogen werden können (siehe unten), sind häufig auch im Excel-Format \*.xls oder im Austauschformat \*.csv verfügbar.

die Merkmalsträger nur mit unverhältnismäßig hohem Aufwand identifiziert werden könnten. Werden personenbezogene Daten „prozessproduziert“ erhoben, d.h. für Verwaltungsregister (etwa beim Meldeverfahren in der Sozialversicherung), so ist die Weitergabe dieser „Sozialdaten“ für Forschungszwecke nach § 75 SGB X im Einzelfall genehmigungspflichtig.

**Abbildung 1: Beispiel eines Mikrodatensatzes: SOEP Originaldatenmatrix, wide-format**

	hhnr	persnr	vnr	mnr	vgebj	mgebj	vtodj	mtodj	vaort91
1	19	101	-1	-1	-1	-1	-1	-1	trifft nicht
2	19	102	-1	-1	-1	-1	-1	-1	trifft nicht
3	19	103	101	102	1932	1942	-1	-1	trifft nicht
4	27	201	-1	-1	1893	1892	1960	1984	Verstärkt
5	27	202	-1	201	1919	1926	-1	-1	trifft nicht
6	27	203	-1	201	1919	1924	-1	-1	keine Angabe
7	35	301	-1	-1	1934	1946	-1	-1	In einem anderen
8	35	302	-1	-1	1930	1928	1976	-1	Verstärkt
9	43	401	-1	-1	1883	1893	1940	1979	Verstärkt
10	51	501	-1	-1	1884	1889	1962	1955	Verstärkt
11	60	601	-1	-1	1924	1929	1983	-1	Verstärkt
12	60	602	-1	-1	1919	1921	-1	-1	weiter entfernt
13	60	604	603	602	1946	1958	-2	-2	nicht verfügbar
14	60	609102	-1	-1	1926	1926	1973	1987	Verstärkt
15	60	609103	-1	-1	-1	1950	-1	-1	trifft nicht
16	60	609104	-1	609102	1954	1950	2003	-1	trifft nicht
17	60	609105	-1	-1	1939	1940	-1	-1	trifft nicht
18	60	1088902	-1	-1	1931	1930	-1	-1	trifft nicht
19	60	1203002	-1	-1	1951	1950	-1	-1	trifft nicht
20	78	701	-1	-1	1864	1869	1930	1928	Verstärkt
21	86	801	-1	-1	-1	-1	-1	-1	trifft nicht
22	94	901	-1	-1	1920	1924	-1	-1	In einem anderen

!!!

Vars: 65    Obs: 44.787

Für die Sozialwissenschaften gibt es drei Wege des Zugriffs auf Originaldatensätze: Onsite-Nutzung, kontrollierte Datenfernverarbeitung und Scientific-Use-Files. Alle drei wurden im Rahmen der sozioökonomischen Berichterstattung bereits genutzt und müssten durch eine virtuelle Arbeitsumgebung unterstützt werden.

Bei der sog. *Onsite-Nutzung* wird die Analyse der Originaldaten im jeweiligen Forschungsdatenzentrum (FDZ) durchgeführt. Dazu werden dort gering anonymisierte Daten in vollem Stichprobenumfang bereitgestellt. Der Forscher oder die Forscherin kann in einem vereinbarten Zeitraum im FDZ an einem Arbeitsplatz für Gastwissenschaftler/innen unter Aufsicht an den Daten arbeiten. Jedoch dürfen dabei keinerlei Daten physisch aus dem FDZ mitgenommen werden. Nicht nur die Originaldaten, sondern auch die bei der Arbeit generierten Arbeitsdatensätze und die Auswertungsergebnisse (Outputs) verbleiben in einem geschützten Bereich auf dem Rechner des Gastarbeitsplatzes. Output-Tabellen werden seitens des FDZ nach jeder Arbeitssitzung kontrolliert und bei kleinen Zellenbesetzungen gesperrt, um eine Reidentifikation der ausgewiesenen Personen auszuschließen. Erst nach dieser Kontrolle, werden sie dem Nutzer oder der Nutzerin – in der Regel einige Tage nach dem jeweiligen Rechentermin – zugesandt.

Tabelle 1: Ausgewählte Originaldatensätze in „soeb 2“

Name, Beschreibung (Abkürzung)	Datenhaltende Institution	Datenart	Entstehungskontext	Produktion	Möglicher Datenzugang	Datenzugang in „soeb 2“	Nutzung in „soeb 2“
Stichprobe der Integrierten Erwerbsbiographien des IAB (IEBS)	Institut für Arbeitsmarkt- und Berufsforschung (IAB)	Mikrodaten	amtlich und wissenschaftlich	prozessproduziert	Onsite, Fernverarbeitung, SUF	Fernverarbeitung	Einzelnutzung
Zusammengeführte (gepoolte) Rentenzugangsstichproben (RZ)	Rentenversicherung (RV)	Mikrodaten	amtlich	prozessproduziert	Onsite, Fernverarbeitung	Onsite	Eine Partner einrichtung
Sozio-oekonomisches Panel (SOEP)	DIW-SOEP Gruppe	Mikrodaten	wissenschaftlich	Befragung	SUF	SUF	Individuell, kollaborativ
Mikrozensus (MZ)	DESTATIS	Mikrodaten	amtlich	Befragung	Onsite, Fernverarbeitung, SUF	SUF & Onsite	Individuell, kollaborativ
DGB-Index »Gute Arbeit«	DGB Index Gesellschaft	Mikrodaten	wissenschaftlich	Befragung	Sondervereinbarung, SUF	SUF	individuell
Volkswirtschaftliche Gesamtrechnung (VGR) des Statistischen Bundesamtes (DESTATIS)	DESTATIS	Aggregierte Daten	amtlich	prozessproduziert	Internet	Internet, Datenlieferung	Individuell, kollaborativ

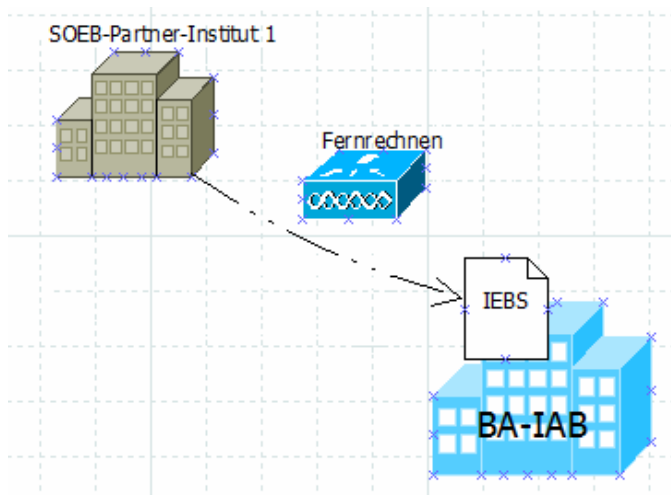
Auch bei der *kontrollierten Datenfernverarbeitung* können die nur formal anonymisierten Daten in vollem Stichprobenumfang analysiert werden. Dazu erhält der Forscher oder die Forscherin einen sog. Strukturdatensatz (Dummy-Datei) mit wenigen (fiktiven) Fällen, der nur im Aufbau und in den Merkmalsausprägungen dem Originalmaterial gleicht. Anhand dieser Dummy-Datei können Auswertungsprogramme (Syntax-Skripte, siehe dazu unten) in den Analyseprogrammen SPSS oder STATA erstellt werden, mit denen das datenhaltende Forschungsdatenzentrum anschließend die Originaldaten auswertet. Wie bei der Onsite-Nutzung erhalten die Datennutzer/innen nach einer manuellen Prüfung auf Geheimhaltung schließlich die Ergebnisse dieser Auswertung in Form einer Liste, in der Tabellen mit zu niedrigen Zellenbesetzungen gesperrt sind.

Beide Zugangswege verlangsamen eine iterative oder explorative Arbeitsweise oder einen „workflow“, in dem mehrere Auswertungsschritte aufeinander aufbauen. Denn Zwischenergebnisse können erst zeitverzögert geprüft werden, und für jeden neuen Arbeitsschritt ist eine neue Arbeitssitzung im FDZ oder eine neue Datenverarbeitung zu verabreden (vgl. unten: 4.3.2).

Dagegen ist mit sog. *Scientific- Use-Files* (SUF) die offsite-Analyse in den Forschungseinrichtungen der Nutzer/innen möglich. Bei den SUF handelt es sich um stärker, jedoch für wissenschaftliche Analysen ausreichend (faktisch) anonymisierte Originaldaten, die auf Antrag und nach Prüfung aller Nutzungsvoraussetzungen (z.B. unabhängige wissenschaftliche Forschung) als CD oder DVD an Datennutzer/innen ausgeliefert werden. Die stärkere Anonymisierung ist mit einem eingeschränkten Informationsgehalt verbunden – Merkmalsausprägungen werden zum Teil gegenüber dem Onsite-File vergrößert, oder es wird (etwa beim Mikrozensus) eine Unterstichprobe als SUF ausgeliefert.

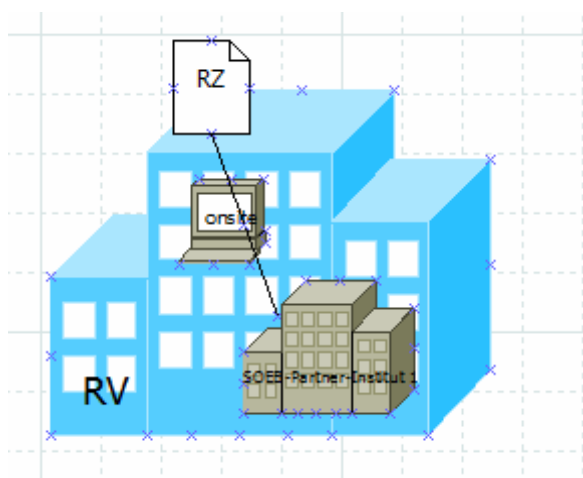
Tabelle 1 zeigt ausgewählte Originaldatensätze, die von einer Forschungseinrichtung in Einzelnutzung oder von mehreren beteiligten Einrichtungen in kollaborativer, gemeinsamer Nutzung für die Arbeit am zweiten Bericht zur sozioökonomischen Entwicklung Deutschlands („soeb 2“) verwendet wurden. Die Abbildungen 2 bis 7 veranschaulichen exemplarisch die unterschiedlichen Wege, in denen Partnerinstitute auf diese Daten zugriffen.

**Abbildung 2: Zugriff auf die Integrierten Erwerbsbiografien (IEBS)**



Die IEBS-Daten wurden von einem Projektpartner über individuelles Fernrechnen beim FDZ des IAB genutzt (Abbildung 2). Partnerinstitut 1 erhielt vom datenhaltenden Institut (FDZ BA-IAB) einen strukturtreuen Datensatz<sup>8</sup> mit nur wenigen Fällen, entwickelte anhand dieses Datensatzes Syntax (vgl. dazu unten: 3.2) und wertete die IEBS per Fernrechnen aus<sup>9</sup>.

**Abbildung 3: Zugriff auf die gepoolten Rentenzugangsstichproben (RZ)**



Da die Arbeit mit den Daten der Rentenversicherung ein hohes Maß an rentenrechtlicher Expertise erfordert und gepoolte Zugangsdaten verwendet werden sollten, die nicht zu den Standardprodukten des FDZ gehören, wurden die Analysen mit der Rentenzugangsstichprobe durch einen Projektpartner und eine Projektpartnerin des FDZ

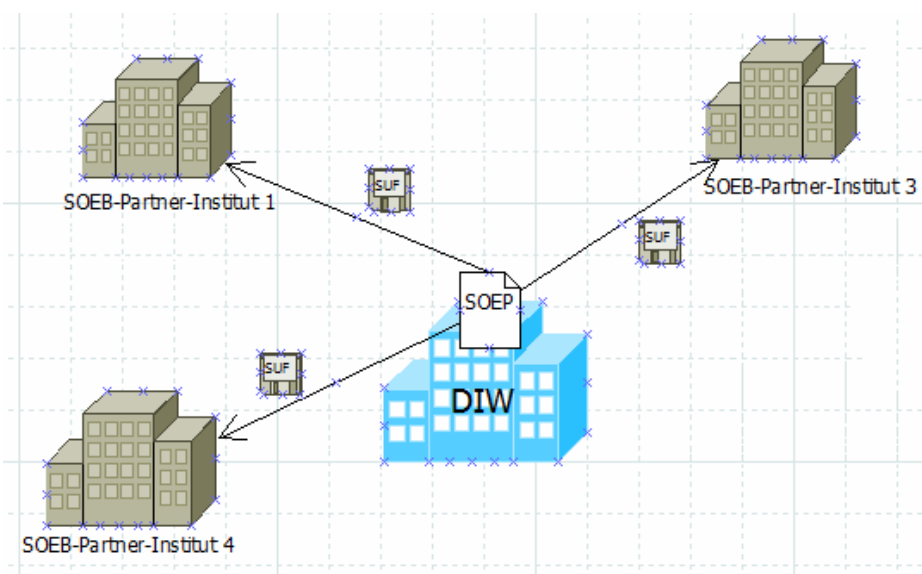
<sup>8</sup> Solche Datensätze entsprechen den Originaldaten in der Variablenstruktur, bestehen jedoch aus einer sehr begrenzten Anzahl von Fällen (max. 1000) mit verfälschten Daten.

<sup>9</sup> Der genaue Ablauf wird ebenfalls weiter unten detailliert beschrieben.

der Rentenversicherung durchgeführt, die als Angehörige des Daten haltenden Instituts on-site Zugriff auf die Daten hatten (Abbildung 3).

Auf das Sozio-oekonomische Panel (SOEP) greifen verschiedene Partnerinstitute jeweils per Einzelnutzungslizenz für den Scientific-Use-Datensatz zu (Abbildung 4). Das DIW sendet den SUF auf einer passwortgeschützten DVD per Post an die Partnerinstitute, wo die Daten auf lokalen Rechnern analysiert werden.

**Abbildung 4: Zugriff auf das Sozio-oekonomische Panel (SOEP)**



Auch der Mikrozensus wurde als Scientific-Use-File (SUF) auf Datenträgern für mehrere Partnerinstitute mit Einzelnutzungslizenzen verfügbar gemacht. Die Institute speichern die SUF-Daten jeweils auf ihre Institutsrechner und werten sie auch dort aus. Zusätzlich analysierte ein Partnerinstitut den Mikrozensus onsite im FDZ eines statistischen Landesamtes (Abbildung 5).

Aggregierte Daten der Volkswirtschaftlichen Gesamtrechnung bezogen verschiedene Partnerinstitute ohne weitere Nutzungsvereinbarungen über das Internet-Portal des Statistischen Bundesamts (DESTATIS, vgl. Abbildung 6)<sup>10</sup>. Für den internationalen Vergleich nutzten einzelne Partnerinstitute Daten anderer statistischer Institutionen (z.B. OECD, Eurostat) teils über deren Internetportale, teils über Datenträger, die sie bestellten, oder über statistische Reihen. Da es sich nicht um Einzeldaten handelt, sind diese allgemein zugänglich; ihre weitere Nutzung unterscheidet sich jedoch nicht wesentlich von der in Abbildung 6 dargestellten Arbeitsweise: In allen Fällen werden die Daten – hauptsächlich als Excel-Dateien (\*.xls, oder \*.csv) – auf die internen Rechner der beteiligten Institute geladen und auch dort verarbeitet.

<sup>10</sup> [www.destatis.de](http://www.destatis.de).

Abbildung 5: Zugriff auf den Mikrozensus (MZ)

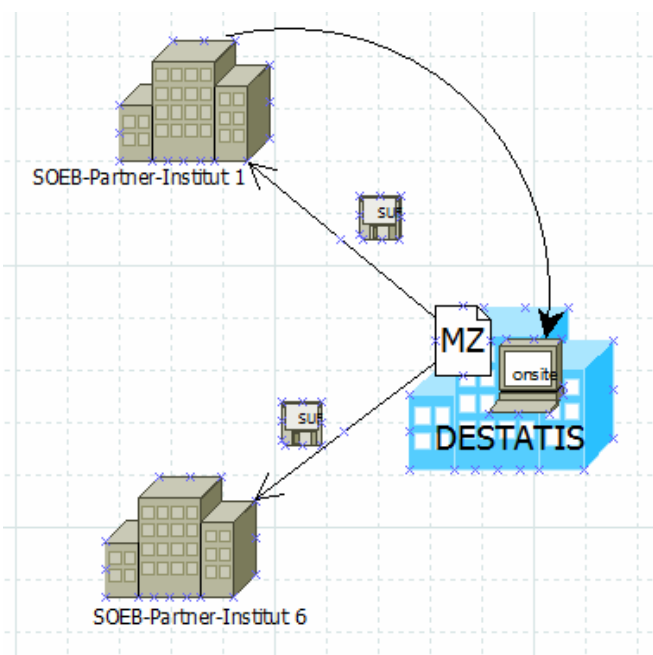
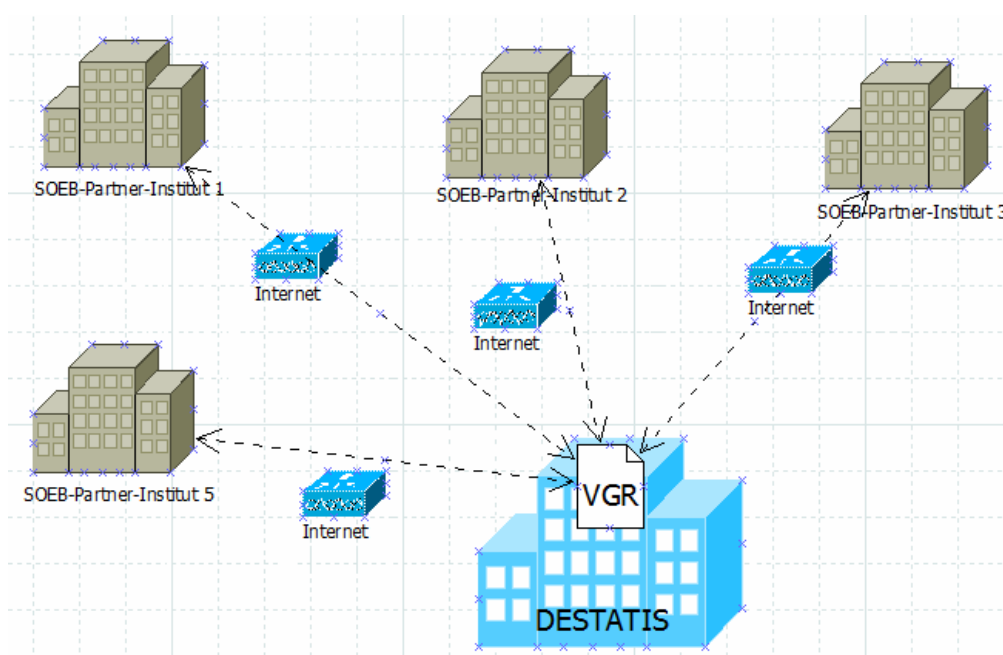


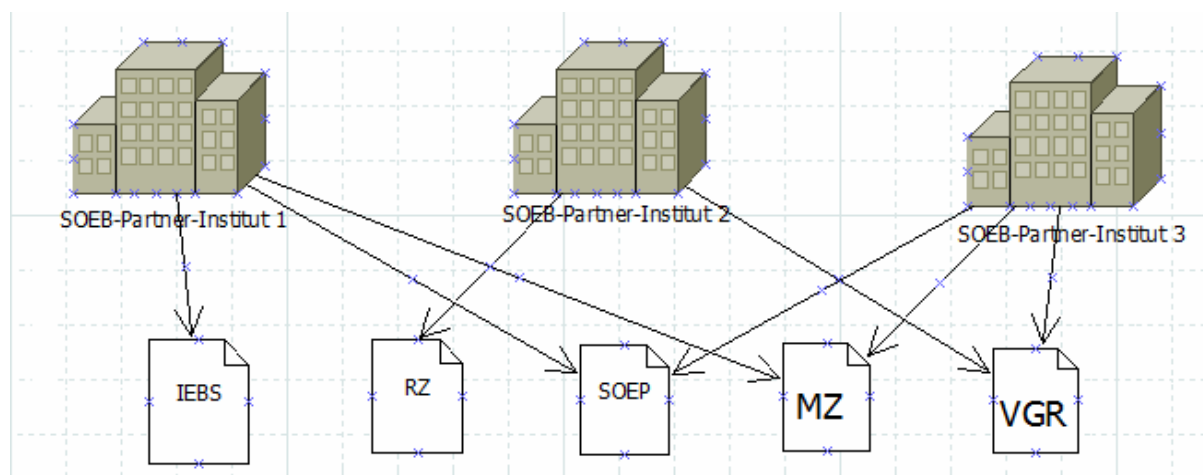
Abbildung 6: Zugriff auf Daten der Volkswirtschaftlichen Gesamtrechnung (VGR)



Mehrere Partnerinstitute nutzen über die hier beschriebenen Zugriffswege gleichzeitig verschiedene Datensätze. Abbildung 7 veranschaulicht, wie sich Datenzugriffe der verschiedenen Partnerinstitute überschneiden: Beispielsweise verwendet Partnerinstitut 1 das SOEP und den Mikrozensus per SUF und analysiert gleichzeitig die IEBS per Fernrechnen. Dazu hat dieses Institut für jede ausführende Person bzw. für das Projekt einen Einzelnutzungsvertrag mit den Daten haltenden Instituten geschlossen,

Nutzergruppen werden bisher nicht berücksichtigt. Gleiches gilt für Institut 3, das für die sozioökonomische Berichterstattung gleichzeitig die SUF's des Mikrozensus und des SOEP mit entsprechenden Einzelnutzungsverträgen nutzt. Institut 2 hingegen greift beispielsweise auf die Rentenversicherungsdaten onsite und auf die VGR-Daten via Internet zu.

**Abbildung 7: Multiple Zugriffe auf Originaldaten von „soeb 2“-Partner-Instituten**



### 3.2 Syntax

Eine Syntaxdatei ist eine ASCII-Datei, die aus mehr oder weniger komplexen Programmweisungen für statistische Anwendungsprogramme besteht.

Inhaltlich lassen sich in sozialwissenschaftlichen Verbundvorhaben wie der sozioökonomischen Berichterstattung drei Gruppen von Syntaxdateien nach ihrer Funktion im Forschungsprozess unterscheiden: Datenmanagementsyntax, Generierungs- und Recodierungssyntax sowie Analysesyntax. Die zahlreichen Syntaxdateien, die in der Arbeit mit quantitativen Datensätzen entstehen, enthalten das Auswertungswissen der empirisch arbeitenden Forschungsteams.

Die Datenmanagementsyntax besteht aus Programmweisungen, die erforderliche Daten in den Originaldatensätzen ansprechen, die zu bearbeitenden und zu analysierenden Variablen und Fälle extrahieren und je nach Bedarf zusammenführen (Matching, record linking). In diesen Operationen (file handling) entstehen Arbeitsdatensätze (vgl. 3.3).

Im nächsten Arbeitsschritt werden mittels Generierungs- und Recodierungssyntax die zuvor ausgewählten und zusammengeführten Originalvariablen je nach Bedarf aufbereitet: durch Abgrenzung der berücksichtigten Fälle, durch Recodierung von Werten und durch Beschriftung von Variablen und Werten. Zudem werden neue Vari-

ablen generiert, beispielsweise durch Verknüpfung mit anderen Variablen oder konstanten Werten.

Über die Analysesyntax werden die auf diese Weise aufbereiteten Daten ausgewertet. Sie enthält auch die Programmanweisungen, welche die zur Ergebnispräsentation benötigten Tabellen und Grafiken erzeugt.

Jedes der verwendeten statistischen Programme (z.B. SPSS, STATA oder R) verwendet seine eigene Befehlsstruktur und erzeugt ein eigenes Dateiformat für Syntax. Beispielsweise lautet die Anweisung, den Wert 1 der Variable X in den Wert 4 einer neu generierten Variable Y zu recodieren, in SPSS:

```
RECODE X (1=4) (ELSE=SYSMIS) INTO Y.
```

In STATA lautet die Anweisung mit dem gleichen Ergebnis:

```
Recode X (1=4) (nonmissing=.), gen(y).
```

Und in R heißt der gleiche Befehl:

```
Y <- ifelse (X==1,4,NA)
```

In diesem Beispiel wird nur ein Befehl dargestellt. Jedoch kann eine Syntaxdatei mehrere tausend Syntaxzeilen mit – meist aufeinander aufbauenden – unterschiedlichen Anweisungen enthalten. Über den Aufbau von Syntaxdateien bestehen in den Sozialwissenschaften kaum gültige Konventionen. Nicht immer werden Anweisungen zu Datenmanagement, Recodierung und Auswertung in getrennten Dateien gehalten. Im Idealfall sind die aufeinanderfolgenden Anweisungen innerhalb der Syntax für andere Nutzer/innen ausführlich kommentiert. Die Kommentierung wird direkt in die Syntaxdatei eingetragen. In SPSS und STATA wird solcher Text, der keine Rechenanweisung enthält, mit einem vorgestellten \* kenntlich gemacht, in R mit #.

Syntaxdateien jeder Art werden dann durch das entsprechende Statistikprogramm ausgeführt (interpretiert), dabei haben SPSS-Syntaxdateien den Dateinamen-Suffix \*.sps, STATA-Syntaxdateien den Suffix \*.do und \*.ado und R-Syntaxdateien \*.r.<sup>11</sup> Die Dateigrößen sind abhängig von der Länge der Syntax, jedoch alles in allem sehr klein, da es sich um Textdateien handelt.

Partnerinstitute in Verbundvorgaben wie der sozioökonomischen Berichterstattung kooperieren vor allem über den Austausch von Syntax. Dazu sind wenigstens elementare Abstimmungen über das verwendete Statistikprogramm und über die Kommentierung erforderlich. In der Arbeit an soeb 2 geschah der Austausch meist per Mail, selten über den internen Bereich der Projekt-Website.

---

<sup>11</sup> Das weniger häufig verwendete Programm „NEWSPELL“ verlangt Commando-Dateien mit dem Suffix \*.cmd.

### 3.3 Arbeitsdatensätze

Vor allem in der Phase des Datenmanagements und der Recodierung entstehen im Forschungsprozess eine Reihe sogenannter Arbeitsdatensätze. Diese werden mit Hilfe der jeweiligen Syntaxprogramme generiert und als externe Arbeitsdatensätze im Format des jeweils verwendeten Statistikprogramms gespeichert<sup>12</sup>. Sie sind grundsätzlich wie Originaldatensätze aufgebaut, enthalten jedoch oft nur einen Teil der ursprünglichen Ausgangsvariablen, dafür aber zusätzliche durch Abgrenzungen und Recodierungen modifizierte Variablen. Ebenso können sie sich durch die Zahl der Fälle von den Originaldaten unterscheiden: Bestimmte Fälle können für die Analyse ausgewählt werden, oder es werden durch das Zusammenführen verschiedener Arbeitsdatensätze zusätzliche Fälle einbezogen. Temporäre Arbeitsdateien werden im Rahmen eines Zwischenschrittes generiert, zur weiteren Verarbeitung nur kurz gespeichert und durch eine Syntax-Anweisung wieder gelöscht. Häufig werden im Forschungsprozess parallel viele Arbeitsdatensätze aus den Originaldatensätzen generiert, abgelegt und in späteren Arbeitsschritten wieder benutzt.

Arbeitsdateien aus den SUF-Dateien legen die Datennutzer/innen auf die lokalen Speicherorte ab. Beim Fernrechnen werden lokale Arbeitsdateien nur auf Basis der kleinen Strukturdatensätze („Spieldatensätze“) erzeugt und gespeichert. Die eigentlichen Arbeitsdateien, die anschließend aus der eingeschickten Syntax erzeugt werden, verbleiben in der Daten haltenden Institution, und die Datennutzer/innen erhalten lediglich die Outputs. Beim Onsite-Rechnen können die Datennutzer/innen in begrenztem Umfang temporäre Arbeitsdateien auf den Rechnern des FDZ zwischenspeichern. Diese sind jedoch ebenso wie die Originaldaten nur onsite zugänglich und werden nach Ende einer Nutzung wieder gelöscht.

Am Ende des Forschungsprozess werden die endgültigen Daten zu den analysierten Fällen einschließlich aller recodierter und neuer Variablen in sog. Ergebnisdatensätzen – wieder im Format des jeweilig verwendeten Statistikprogramms – abgespeichert. Auch diese Datensätze werden durch Syntax generiert.

Für die Zusammenarbeit mit anderen Projektpartner/innen in der sozioökonomischen Berichterstattung können je nach Bedarf ausgewählte Fälle oder Variablen aus den vorliegenden temporären oder Ergebnisdatensätzen in Austauschdatensätzen abgelegt werden. In der Arbeit an soeb 2 wurden solche Datensätze teils per Mail und

---

<sup>12</sup> Dabei entstehen für SPSS-Datensätze entweder: \*.sav oder \*.por Dateien. In STATA werden hauptsächlich \*.dta – Dateien erstellt und in R: \*.Rdata. Es kann und wird häufig auch aus jedem Statistik-Programm nach Excel (\*.xls) exportiert. Zudem werden in Newspell oder Chesea ASCII-Datei-Formate erstellt, wie beispielsweise \*.dat oder \*.raw.

teils über Datenträger, seltener auch über den internen Bereich der Projekt-Website übermittelt.

Je nach Originaldatensatz und Fallauswahl können Arbeitsdatensätze unterschiedlich groß sein. Beispielsweise können Arbeitsdatensätze aus dem SOEP nur wenige KB umfassen; dagegen kann eine Teilstichprobe der IEBS bis zu 10 GB groß sein.

Metadaten, die über die rein datentechnischen Informationen (Größe, Typ und Änderungsdaten) hinausgehen, gibt es für die Datensätze nur, wenn Standards z.B. für Labels zwischen den Kooperationspartnern explizit festgelegt werden. In soeb 2 blieb die Dokumentation von Syntax-Files noch den einzelnen Partner/inne/n überlassen.

### **3.4. Outputs**

Analyseergebnisse werden in der Regel in Tabellenform dargestellt. Diese Tabellen werden mittels Analysyntax erzeugt und entweder im jeweiligen Statistikprogramm in besonderen Ausgabedateien für die weitere Verarbeitung in Text- oder pdf-Dateien formatiert. Outputs können in verschiedenen Formaten vorliegen: in ASCII, html oder pdf oder im output-Format von SPSS (\*.spo oder neuere Versionen). Häufig werden sie zum Austausch zwischen den Partnern eines Forschungsverbunds oder zur weiteren Bearbeitung für Veröffentlichungen und zum Ergebnistransfer in ein Standardtabelleformat wie Excel oder ins Austauschform \*.csv exportiert. Auch hierfür sind wenigstens elementare Verabredungen über einzuhaltende Formatierungen und über den Tabellenaufbau (Spalten- und Zeilenköpfe) erforderlich.

## 4. Kollaborative Datenanalyse – Anwendungsfälle

Verabredungen im Forschungsverbund über die gemeinsame Arbeit am zweiten Bericht zur sozioökonomischen Entwicklung Deutschlands betrafen zunächst den Zchnitt der einzelnen Arbeitspakete, Auswertungskonzepte und den Austausch von Ergebnissen (Austauschdatensätze, Output-Dateien und Textdateien). Wo in verschiedenen Arbeitspaketen mit gleichen Datensätzen oder Analysemethoden gearbeitet wurde, kam es darüber hinaus zur Zusammenarbeit bei der Datenanalyse. In der ersten Phase des Teilprojekts VirtAug wurden diese Arbeitsprozesse und die darin aufgetretenen Probleme bilanziert, um daraus funktionale Anforderungen an eine verbesserte kollaborative Datenanalyse und an eine virtuelle Arbeitsumgebung für die gemeinsame Projektbearbeitung zu entwickeln. Als exemplarischer Anwendungsfall wird in diesem Abschnitt die kollaborative Arbeit mit dem SOEP in den lebensverlaufsorientierten Arbeitspaketen des Berichts als „work flow“ dargestellt; ergänzend wird auf Analysen mit anderen Datensätzen in weiteren Arbeitspaketen eingegangen.

### 4.1 Datenmanagement-Syntax

Bevor die tatsächliche Datenmanagement-Syntax erstellt werden kann, werden im kollaborativen Forschungsprozess Abgrenzungen der gemeinsam oder arbeitsteilig zu bearbeitenden Teildatensätze (Samples) vereinbart. Dies geschah in den stärker kollaborativ angelegten Arbeitspaketen von soeb 2 teils auf Verbundtreffen, teils in kleineren Arbeitsgruppen und telefonischen oder E-Mail-Kontakten. In anderen Arbeitspaketen blieb es bei ad-hoc-Abstimmungen mit der Projektleitung.

#### 4.1.1 Management von Längsschnittdaten des SOEP

In Abteilung 3 des zweiten Berichts zur sozioökonomischen Entwicklung Deutschlands („Lebensverläufe im Umbruch“) kooperierten vier Partnereinrichtungen, von denen drei mit dem SOEP arbeiteten. Besonders zwischen den drei Verbundpartnern, die sich auf SOEP-Daten stützten, fanden Diskussionen und Absprachen zu Sampleabgrenzungen für die jeweiligen Unterkapitel in einem iterativen Prozess mit wiederholten Treffen (mit und ohne Projektleitung) statt. Die weitere Darstellung konzentriert sich auf die Auswertungen mit dem SOEP. Abstimmungen mit dem vierten Projektpartner, der hauptsächlich mit den Rentendaten arbeitete, betrafen vor allem methodische Fragen der Sequenzanalyse.

Die kollaborative Datenauswertung sollte für die drei Lebensabschnitte Junge Erwachsene (17 bis 30 Jahre), Haupterwerbsphase (30 bis 35 Jahre) und höheres Er-

werbsalter (55 bis 65 Jahre) eine Typisierung von Lebensverlaufsmustern ermöglichen. Dazu mussten Spellinformationen<sup>13</sup> aus den Monats- und Jahreskalendarien des SOEP organisiert und ausgewertet werden. Sequenzmuster in diesen Datensätzen wurden mittels eines Optimal-Matching-Verfahrens verglichen und durch Clusteranalyse zusammengefasst. Entsprechend der spezifischen Fragestellungen waren für jeden dieser Lebensabschnitte Zustände und Übergänge im Erwerbsstatus und in der persönlichen Lebensführung (Haushaltszusammenhang, Partnerschaft, Elternschaft) in unterschiedlichen Kombinationen zu berücksichtigen. Diese Aufgaben wurden von den Partnerinstituten eigenverantwortlich gelöst. Die Abstimmung über die Abgrenzung der Längsschnittstichprobe und über die Datenorganisation sollte es jedoch ermöglichen, Generierungs- und Analysesyntax gemeinsam zu nutzen, das gleiche sequenzanalytische Verfahren auf die verschiedenen Datensätze anzuwenden und auf diese Weise auch möglichst vergleichbare Ergebnisse zu erzielen, etwa für die verschiedenen Lebensabschnitte die gleichen Kennziffern zu berechnen.

Syntax für das Datenmanagement erstellten die Verbundpartner überwiegend individuell entsprechend der gemeinsamen Absprachen. Allerdings fand bei Forscher/innen innerhalb der gleichen Verbundinstitute zusätzlich ein interner Informations- und Syntaxaustausch statt. Zu einem späteren Zeitpunkt im Forschungsprozess wurden jedoch auch einzelne kleine Syntaxbausteine, also Programmzeilen einer Syntax, z.B. die korrigierte NETTO- und POP<sup>14</sup>-Abgrenzung, zwischen Partner/inne/n ausgetauscht.

Zunächst kopierten alle VerbundpartnerInnen von der Original-DVD alle Dateien des SOEP<sup>15</sup> auf lokale Festplatten oder institutsinterne Server. Im nächsten Schritt wurden per Syntax aus ausgewählten Originaldateien Untersuchungseinheiten gezogen und damit eine oder mehrere, auch temporäre, Arbeitsdateien erstellt.

Die Auswahl der zu integrierenden Originalvariablen wurde in der Syntax umgesetzt und kommentiert; dabei wurden je nach Forschungsschwerpunkt sehr verschiedene Variablen (die aus unterschiedlichen Originaldateien des SOEP gematcht werden) in die Arbeitsdateien übernommen. Zur Auswahl der Variablen wird häufig das internetbasierte Informationssystem SOEP-Info<sup>16</sup> genutzt. Die ausgewählten Variablen

---

<sup>13</sup> Als „Spells“ werden zusammenhängende Zeiten (hier Monate) bezeichnet, die Individuen im gleichen definierten Zustand (z.B. Erwerbstätigkeit) verbringen.

<sup>14</sup> In der SOEP-Variable NETTO sind Informationen über die Art der Teilnahme an der Befragung hinterlegt, während in der POP-Variable Informationen über die Zugehörigkeit zu einer Population enthalten sind.

<sup>15</sup> Das waren im Jahr 2009 für alle Panel-Jahre inklusive aller generierten Datensätze insgesamt 312 Dateien mit einem Volumen von etwa 1,3 GigaByte. Für jedes weitere Jahr kommen etwa 20 Dateien mit insgesamt etwa 200 MB Volumen dazu.

<sup>16</sup> <http://panel.gsoep.de/soepinfo2008/>

wurden von den Forscher/inne/n von Hand oder mit copy & paste in die Syntax eingefügt. Teilweise wurden von den Forschenden auch ganze Syntax-Abschnitte mit Hilfe des SOEP-Syntax-Generators hergestellt und in die bereits vorliegende eigene Syntax integriert. Die Syntax wurde mit dem jeweiligen Statistikprogramm ausgeführt, und so entstehen (temporäre) Arbeitsdateien.

Im Arbeitsprozess geschehen auch Variablenauswahl und Sampleabgrenzung meist iterativ: Zunächst werden bestimmte Variablen ausgesucht, und nicht selten werden später weitere Variablen ausgewählt und in die Syntax integriert. Nach solchen Korrekturen müssen diese – und alle darauf aufbauenden – Syntaxdateien erneut gestartet werden. Die (temporären) Arbeitsdateien werden dann im weiteren Prozess teilweise miteinander kombiniert und im Format verändert<sup>17</sup>, so dass ausgehend von den Originaldaten am Ende eine Vielzahl von (temporären) Arbeitsdateien entstehen kann. Auf Grund des iterativen Entwicklungsprozesses bei der Datenmanagement-Syntax und der nötigen sequentiellen Abarbeitung der Syntax sind ein modularer Aufbau von Syntax sowie eine ausführliche Dokumentation unerlässlich.

Abbildung 8 stellt die Arbeitsschritte des Datenmanagements mit den SOEP-Daten dar. Arbeitsdateien entstehen aus verschiedenen Abgrenzungs- und Matchinganweisungen auf die Original-SOEP-Dateien und auf in Zwischenschritten erstellte (temporäre) Arbeitsdateien<sup>18</sup>. Alle Abgrenzungen und Matchinganweisungen sind als Programmmanweisungen in Syntaxdateien formuliert. Der modulare Syntaxaufbau ist hier dargestellt durch Syntax 1 bis Syntax 4, wobei das Statistikprogramm Syntax 4 erst nach den Arbeitsschritten aus Syntax 1 bis 3 bearbeiten kann.

Für das notwendige Datenmatching wurde teilweise ein weiteres SOEP-spezifisches Programm genutzt: Mittels „NEWSSPELL“<sup>19</sup> können insbesondere Daten aus Monats- und Jahreskalendarien aufbereitet und organisiert werden. Die hier zu entwickelnde Spezialsyntax im Dateiformat \*.cmd generiert Arbeitsdateien im ASCII-Format, die anschließend wieder über Syntax des jeweiligen Statistikprogramms an die Arbeitsdateien (\*.sav oder \*.dta) gematcht werden.

---

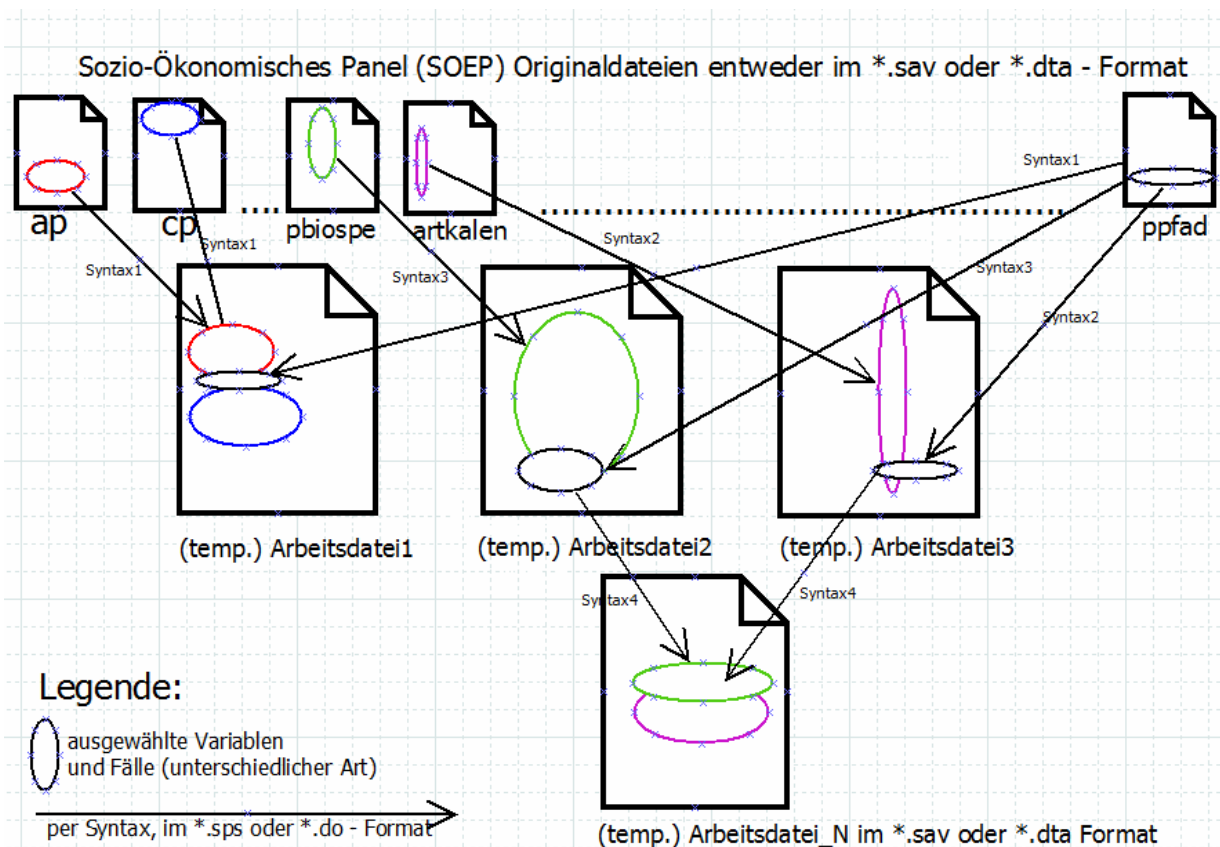
<sup>17</sup> Das heißt, es wird beispielsweise die Datenmatrix transponiert, so dass Zeilen zu Spalten und Spalten zu Zeilen in der Matrix werden.

<sup>18</sup> Diese Dateien können sehr unterschiedlich groß sein, zwischen wenigen KB und bis zu 10 GB.

<sup>19</sup> Siehe auch:

[http://www.diw.de/de/diw\\_02.c.241989.de/sonstiges/\\_soep\\_statistikprogramme.html](http://www.diw.de/de/diw_02.c.241989.de/sonstiges/_soep_statistikprogramme.html)

Abbildung 8: Datenmanagement mit dem SOEP



Das SOEP wird zukünftig in einem „Long-Format“ ausgeliefert werden, das dann aus wesentlich weniger, jedoch deutlich größeren Datensätzen bestehen wird. Dadurch werden andere Datenmanagementarbeiten anfallen. Wie Abbildung 9 zeigt, muss dann zu Beginn einer Analyse auf Personenebene mit etwa 400.000 Fällen gearbeitet werden, was je nach Variablenanzahl zu einer enormen Dateigröße führen kann.

Abbildung 9: Geplante Datenrestrukturierung des SOEP

## Prototypes of \_P and \_H-Files [2008]

	Current data structure [2008]			New data structure [2008] (Long-Format)		
	Files	V(ariables)	N	Files	Items	N
_P, _POST _PAUSL	39	12,018	10,- 23,000	P(l)	2-3	2,292 about 400,000
_H _HOST	26	4,316	6,- 10,000	H(l)	2	683 208,646

Quelle: Krause, Peter (2008): Restructuring the SOEP Database – Perspectives and outcomes, Poster, <http://www.diw.de/sixcms/detail.php/237569>, letzter Zugriff, 22.04.10.

Bei diesem Arbeitsschritt traten vor allem Entscheidungs- und Abstimmungsprobleme in der Zusammenarbeit auf. Es erwies sich zunächst als schwierig, vergleichbare De-

signs zu entwickeln. Durch Testläufe stellte sich heraus, dass zuvor getroffene Festlegungen in der Praxis nicht immer umgesetzt werden konnten. Zuvor kollaborativ festgelegte Arbeitsschritte wurden doch individuell angepasst, also in Syntaxanweisungen umgesetzt. Auch die zeitliche Abstimmung von Arbeitsschritten zwischen den Projektpartner/inne/n erwies sich dabei als Problem.

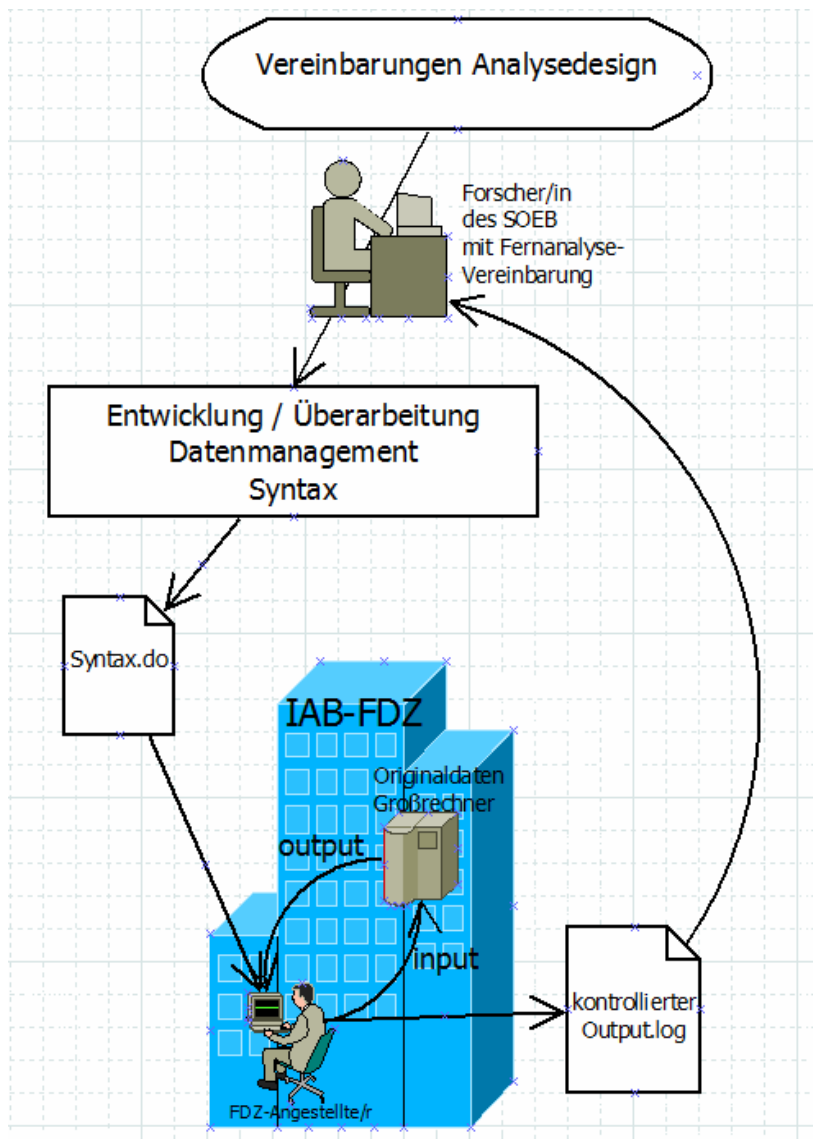
Wie auch auf dem Workshop vom 9. Februar diskutiert, sind abgestimmte Datenanalysen verschiedener Projektpartner/innern nicht nur organisatorisch schwierig. Die wissenschaftliche Freiheit der Beteiligten und ihr Interesse an selbständiger kreativer Forschung sind produktiv und daher zu erhalten. „Ich 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 Infratest. 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.“ (Befragte/r 4, S. 6)

#### **4.1.2 Datenmanagement bei anderen Datensätzen**

Das Datenmanagement für den SUF des Mikrozensus (eine 70%-Stichprobe des vollständigen Mikrozensus-Datensatzes) wurde für einige Analyseschritte zentral von einem Projektpartner in einer Syntax programmiert, die dann von anderen Forschenden für ihre Arbeitspakete übernommen wurden. Gleichzeitig wurden für andere MZ-Auswertungen in den Partnerinstituten auch individuelle Syntaxlösungen verwendet.

Das Datenmanagement für die übrigen verwendeten SUFs, also etwa für die Zeitbudgetstudie (ZB), die Einkommens- und Verbrauchsstichprobe (EVS), den DGB-Index „Gute Arbeit“ und den EU-„Survey on Income and Living Conditions“ (EU-SILC) verantwortete je ein Projektpartner, teilweise in institutsinterner Kollaboration, für das jeweilige Arbeitspaket. Die Syntax hierfür wurde individuell entwickelt und ausschließlich an den lokalen Computern umgesetzt. Notwendige Abstimmungen wurden in der Regel bilateral mit der Projektleitung getroffen.

Abbildung 10: Ablauf Ferndatenverarbeitung



Auch das hochkomplexe Datenmanagement für die IAB-Datensätzen (IEBS, LIAB, IAB-Betriebspanel) übernahm ein Projektpartner im Rahmen eines Sondernutzungsvertrages in Ferndatenverarbeitung. Dazu wurden nach Absprachen und Vereinbarungen im Forschungsverbund anhand eines Strukturdatensatzes mit nur wenigen Fällen die nötigen Datenmanagementprogrammierungen in Syntax umgesetzt, die per Mail an das IAB-Forschungsdatenzentrum übermittelt wurde (vgl. Abbildung 10). Dort wurde die Syntax geprüft und auf die Originaldaten angewendet. Das Ergebnis wurde wieder im FDZ geprüft und als kontrollierter Output an den Projektpartner zurückgeleitet. In der Regel ergibt sich bei diesem Verfahren aus den ersten Outputs zunächst die Notwendigkeit, die Syntax zu überarbeiten und erneut an das FDZ einzuschicken. Zwischen den einzelnen Arbeitsschritten können einige Tage liegen.

## 4.2 Generierungs- und Recodierungssyntax

Nachdem die Arbeitsdatensätze erstellt sind, werden die abhängigen und unabhängigen Variablen operationalisiert und generiert. Abhängige Variablen bezeichnen die zu beschreibenden oder zu erklärenden Größen oder Konstrukte. Ein Konstrukt kann ein komplexes Gefüge verschiedener miteinander verknüpfter Variablen sein. Unabhängige Variablen sind die Informationen, die abhängige Variablen beschreiben oder erklären sollen.

### 4.2.1 Generierungs- und Recodierungssyntax in Lebensverlaufsanalysen mit dem SOEP

In den lebensverlaufsorientierten Arbeitspaketen des Forschungsverbunds bilden sog. „Sequenztypen“, also Folgen zusammengesetzter Zustandsbeschreibungen (z.B. „erwerbstätig“ und „mit Kind“ im Alter von 23) zu jedem Beobachtungszeitpunkt über einem festgelegten Lebensalterszeitraum hinweg die abhängigen Variablenkonstrukte. Beispielsweise ergibt sich im soeb-Arbeitspaket „Junge Erwachsene“ am Ende des Forschungsprozesses ein Verlaufstyp „Frühe Elternschaft mit Erwerbstätigkeit“. Dieser fasst Sequenzmuster junger Menschen zwischen 17 und 30 Jahren zusammen, die sehr früh Eltern werden und in diesem Zeitraum auch erwerbstätig sind.

Welche Zustände und Merkmalskombinationen für die Sequenzanalyse wie zu definieren waren (z.B. Erwerbstätigkeit, Arbeitslosigkeit, Familienstand, Elternschaft, Kind usw.), wurde für die verschiedenen Arbeitspakete der Abteilung „Lebensverläufe im Umbruch“ in Treffen des Forschungsverbundes verabredet. In Generierungs- und Recodierungssyntax umgesetzt wurden diese Vereinbarungen jedoch individuell von den Forschenden. Dabei kam es teilweise zu inhaltlichen Überschneidungen bei den Kapiteln „Junge Erwachsene“ und „Haupterwerbsphase“. Und für diese Kapitel wurde zusätzlich auf das Programm „NEWSPELL“ zurückgegriffen (vgl. oben: 4.1.1); für das Kapitel Altersübergänge wurde hingegen ausschließlich STATA genutzt.

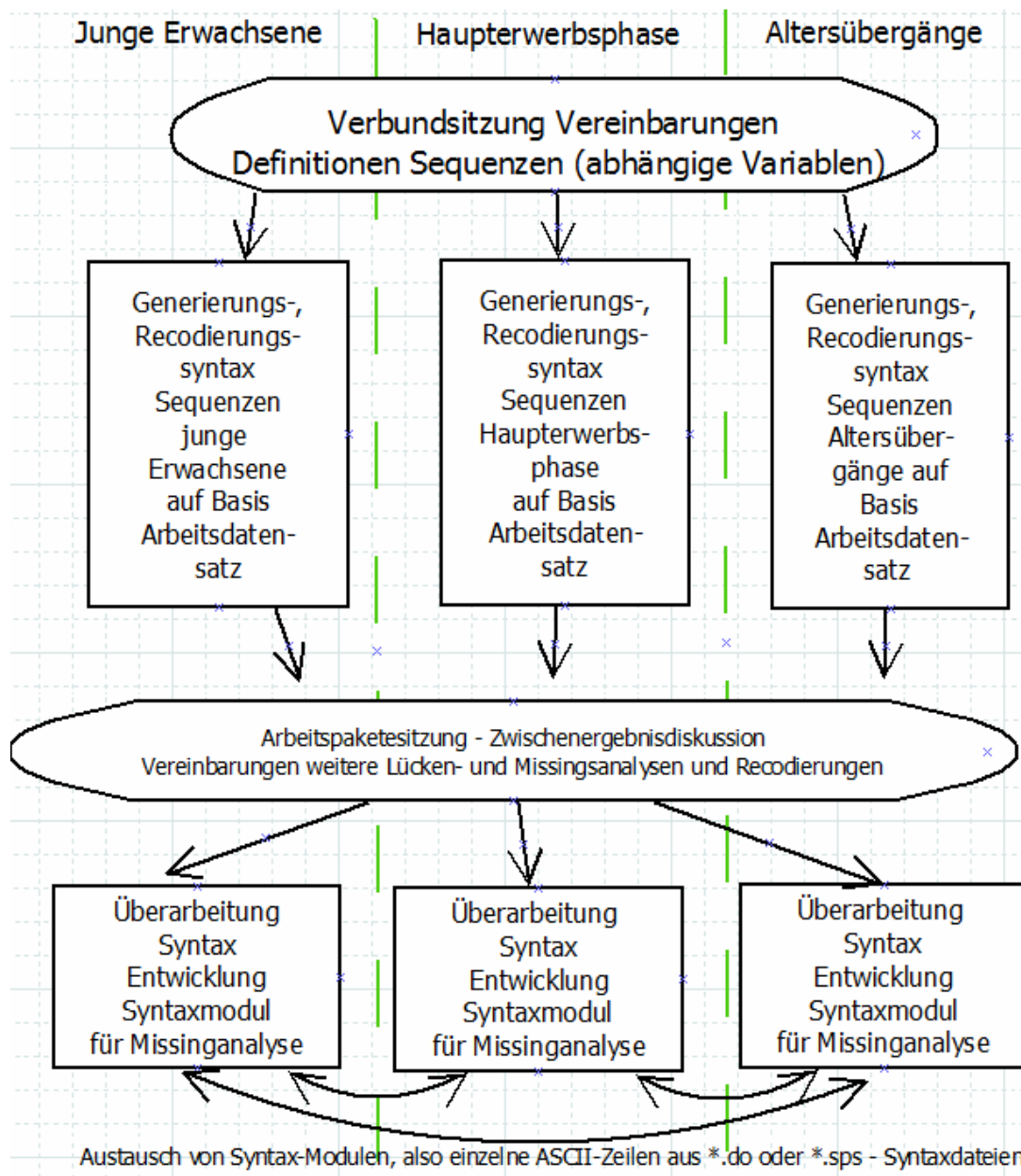
Alle drei Arbeitspakete verwendeten monatliche Kalenderdateien<sup>20</sup> des SOEP, die umfangreich recodiert werden mußten. Die Arbeitspakete „Junge Erwachsene“ und „Altersübergänge“ nutzten zudem auch PBIOSPE, einen retrospektiv erhobenen biografischen Spelldatensatz des SOEP auf Jahresbasis. In jedem Arbeitspaket wurden Arbeitsdateien für verschiedene Zeiträume und mit spezifischen Statuskombinationen erstellt, wozu jeweils eigene Syntaxdateien dienten.

---

<sup>20</sup> In Monatskalendarien sind für jede befragte Person für jeden Monat seit 1984 bestimmte Statusmerkmale, etwa ein Erwerbsstatus, zeilenweise im sog. Spelldatenformat abgespeichert.

Im weiteren Verlauf sprachen die an diesen drei Arbeitspaketen beteiligten Wissenschaftler/innen in einem Arbeitstreffen den Umgang mit Lücken und fehlenden Werten (Missings) in den Sequenzen ab. Alle Beteiligten setzten diese Vereinbarungen anschließend in einer individuellen Recodierungssyntax um.

Abbildung 11: Generierung abhängiger Längsschnittvariablen mit dem SOEP



Syntaxbausteine, also Programmzeilen aus den Recodierungssyntaxen für die abhängigen Variablen, wurden also erst gegen Ende dieses Arbeitsschrittes und damit hauptsächlich zur Validierung und Qualitätssicherung als ASCII-Dateien per Email zwischen den Beteiligten ausgetauscht. Die Generierung und Recodierung der unabhängigen Variablen fand hingegen in jedem Arbeitspaket relativ eigenständig statt;

einzelne Recodierungs- und Abgrenzungsentscheidungen wurden bilateral mit der Projektleitung abgesprochen. Zudem fungierte die Projektleitung in einigen Fällen als „Verteiler“ von Syntax oder Arbeitsdatensätzen für bereits generierte unabhängige Variablen, die in allen Arbeitspaketen verwendet werden sollten.

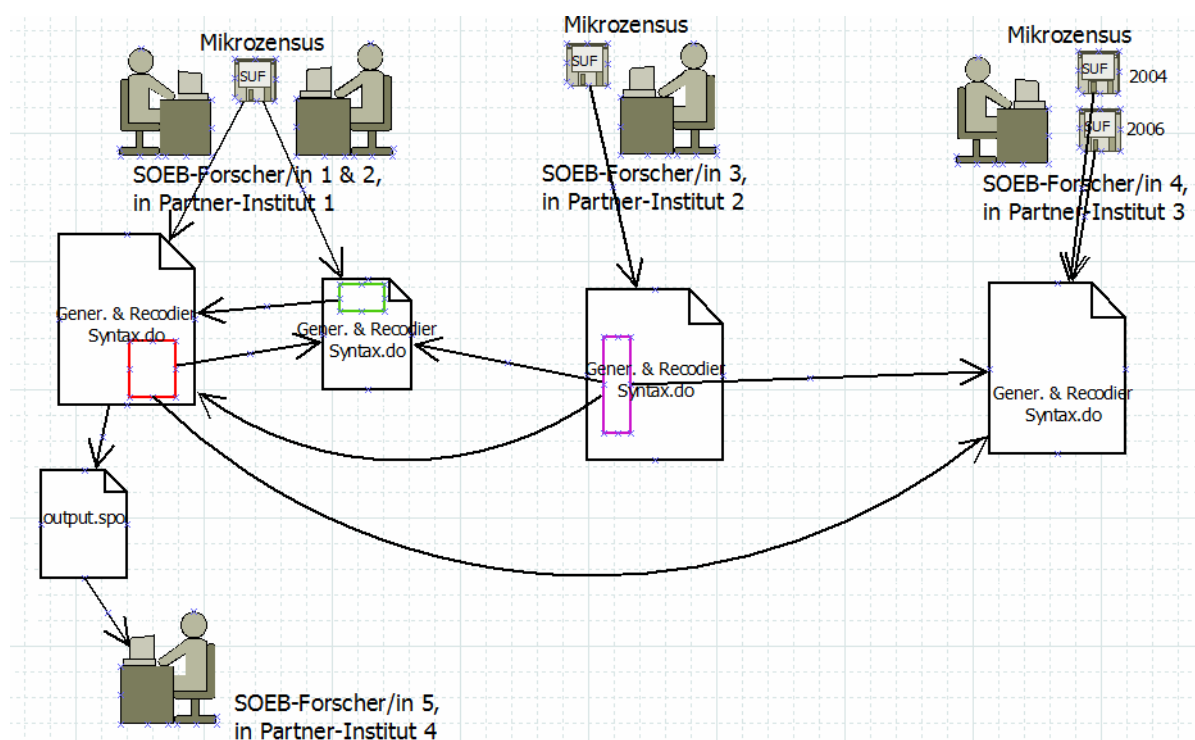
In den Epert/inn/engesprächen problematisierten die beteiligten Wissenschaftler/innen die zeitliche Koordination und die Abstimmungsprozesse in der datenbezogenen Arbeit an den lebensverlaufsorientierten Arbeitspaketen: „Also zuerst mal würde ich mir wünschen, dass gerade am Anfang diese ganzen konzeptionellen Datenfragen alle parallel und gleichzeitig geklärt werden. Nicht: dass am Anfang vielleicht ein paar Haushaltstypologien stehen, dann nach einem halben Jahr das Ganze mit dem Einkommen noch dazukommt, ein Jahr später dann eine Definition für die Datenebene für den Migrationshintergrund, oder so.“ (Interview B6, S. 23)

#### **4.2.2 Generierungs- und Recodierungssyntax für andere Arbeitspakete**

Für den meisten Arbeitspakete beschränkten sich Absprachen zwischen den Partnerinstituten auf die wesentlichen Punkte des jeweiligen Arbeitsprogramms. Die Umsetzung in Syntax und Arbeitsdateien wurde den jeweils verantwortlichen Instituten überlassen. Zu einer stärker kollaborativen Arbeitsweise kam es bei Variablen, die im Rahmen eines Arbeitspakets zu erstellen, aber zugleich für mehrere andere Arbeitspakete zu verwenden waren. Etwa wurden für Haushaltstypologien neben dem SOEP verschiedene Jahrgänge des Mikrozensus-SUF von Projektpartnern teilweise kollaborativ genutzt.

Abbildung 12 zeigt die Verflechtungen eines solchen Arbeitsprozesses. Forscher/in 3 generierte nach Absprache im Verbund eine einzelne Variable (lila Rechteck) und mailte die zugehörige Syntax an Forscher/in 1, 2 und 4. Forscher/in 1 und 2 arbeiteten im gleichen Institut an Variablen, die nur sie für ihr betreffendes Arbeitspaket benötigten. Sie tauschten diese Syntaxmodule nicht nur untereinander aus, sondern mailten sie auch an Forscher/in 4. Zusätzlich mailte Forscher/in 1 eine Ausgabedatei mit Ergebnissen im pdf- oder xls-Format an Forscher/in 5, der/die nur die Ergebnisse weiter verarbeitete und nicht die Syntax selbst.

**Abbildung 12: Kollaborative Entwicklung von Recodierungssyntax mit dem Mikrozensus**



### 4.3 Analysesyntax

In der Analysesyntax werden Anweisungen für die Statistikprogramme festgelegt, die generierte und recodierte Variablen mit uni-, bi- und multivariaten<sup>21</sup> Verfahren auswerten. Je nach Arbeitspaket kamen unterschiedliche Auswertungsverfahren zum Einsatz.

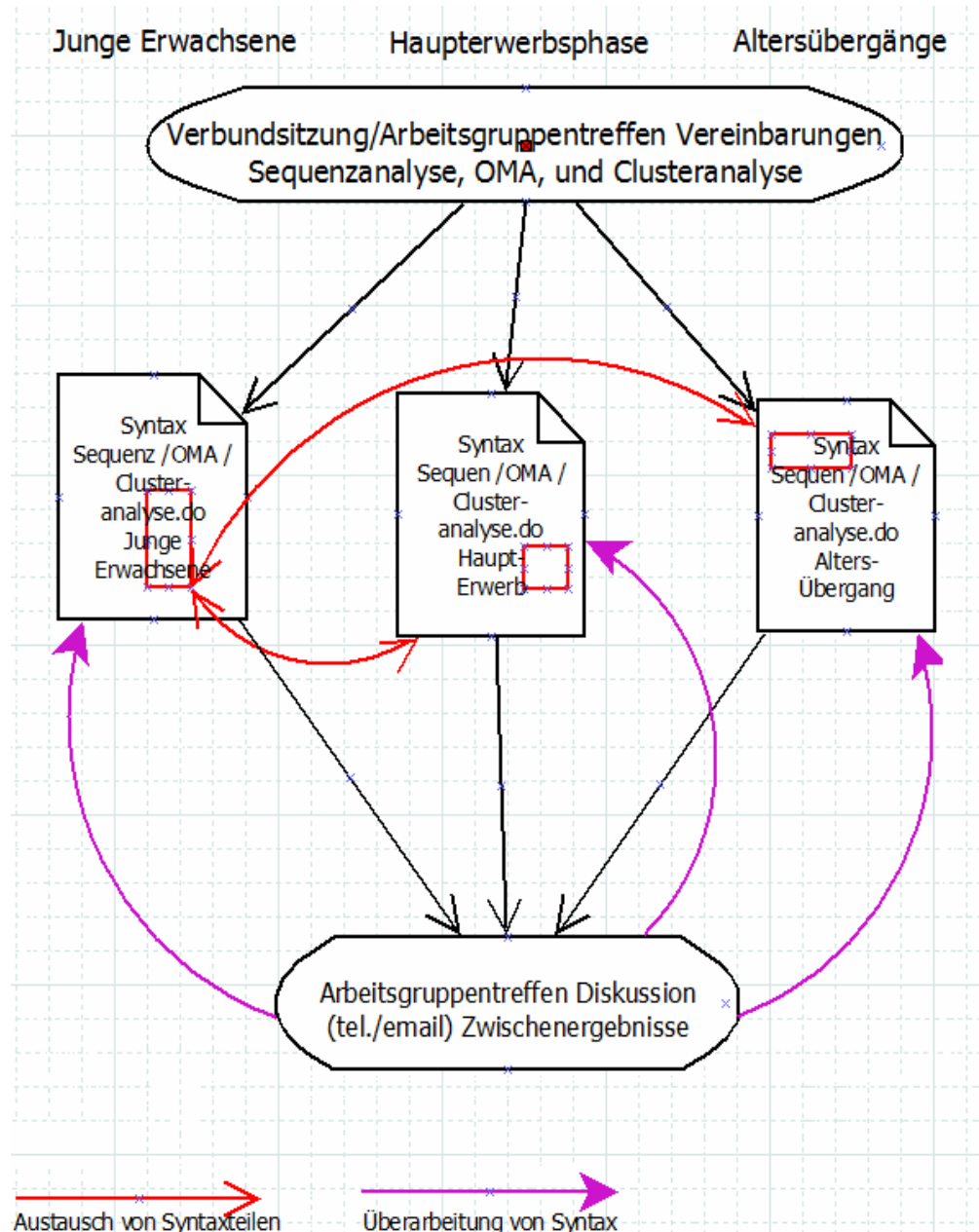
#### 4.3.1 Sequenzanalysen-, Optimal Matching und Clusteranalysen mit dem SOEP

Die drei Arbeitspakete zu „Lebensverläufen im Umbruch“ konnten identische Analyseverfahren umsetzen. Für alle drei Phasen des Lebensverlaufs wurden die über viele Jahre zusammengeführten Informationen der ausgewählten Personen zunächst mittels Sequenzanalyse deskriptiv beschrieben. Daran schloss sich eine Optimal Matching-Analyse (OMA) an, die Sequenzen der Personen miteinander vergleicht und eine sehr große Distanzmatrix für alle Sequenzen aller Personen des Datensatzes generiert. Hierfür konnte ein Zusatzprogramm genutzt werden, das gesondert in STATA implementiert werden muss. Schließlich wurden mittels Clusteranalysen, also durch

<sup>21</sup> Bei einer univariaten Analyse wird nur eine Variable ausgewertet, z.B. in einer Häufigkeitsauszählung. In einem bivariaten Verfahren gelangen zwei Variablen in die Analyse, z.B. in einer Kreuztabellierung, und in einem multivariaten Verfahren werden gleichzeitig mehrere Variablen analysiert, z.B. in einer linearen Regression.

statistische Gruppierung der Distanzen zwischen den individuellen Lebensverläufen mit STATA, typische Verlaufs- und Übergangsmuster für jeden der drei Lebensabschnitte identifiziert. Darüber hinaus war die Berechnung vergleichbarer „Turbulenzkennziffern“ vereinbart; dies geschah für alle drei Arbeitspakete mit dem kostenlosen Software-Tool „CHESEA“. Abbildung 13 zeigt die Kooperationsbeziehungen bei diesen Auswertungsschritten.

**Abbildung 13: Entwicklung von Analysesyntax für Lebensverläufe**



Die in einer Verbandsitzung und in einer Arbeitsgruppe vereinbarten Analysedesigns wurden zunächst individuell für jedes Arbeitspaket umgesetzt. Jedoch wurden bereits während der Entwicklungsphase Teile von Analysesyntax per Email ausgetauscht, von

den beteiligten Wissenschaftler/innen an den jeweiligen Analysebedarf angepasst und in die eigene Syntax implementiert. Für die Clusteranalyse wurde ein einheitliches Verfahren detailliert besprochen und im Arbeitspaket „Junge Erwachsene“ in Syntax umgesetzt, per Email an die Arbeitspakete „Haupterwerbsphase“ und „Altersübergänge“ übermittelt und dort in die jeweilige Syntax integriert. Die abschließende Berechnung der Turbulenzkennziffern wurde in einem Arbeitsgruppentreffen besprochen und beschlossen und mit dem Analysetools „CHESEA“ umgesetzt, wobei auch hierzu Syntax ausgetauscht wurde.

#### **4.3.2 Analysesyntax für andere Arbeitspakete**

Mit der Generierung und Recodierung von Variablen (vgl. oben: 4.2.2) blieb auch die Datenanalyse in den meisten Arbeitspaketen in der Verantwortung jeweils eines zuständigen Partnerinstituts; Syntaxentwicklung und Auswertung fanden individuell auf lokalen Rechnern statt, und es wurden vor allem Outputs ausgetauscht. Wie das SOEP wurde der Mikrozensus sowohl individuell auch kollaborativ ausgewertet. Wie dieser zweite Anwendungsfall zeigt, ist eine solche Arbeitsweise bei Onsite-Nutzung oder Datenfernverarbeitung voraussetzungsvoller und zeitaufwändiger als die Arbeit mit Scientific-Use-Files (SUF).

Beabsichtigt war, Analysesyntax kollaborativ anhand der Mikrozensus-SUF mehrerer Partnerinstitute zu entwickeln, die endgültigen Analysen dann aber mit Onsite-Files durchzuführen. Hierfür gab es vor allem zwei Gründe:

- Da es sich bei den SUF um Unterstichproben mit 70% der Mikrozensus-Fälle handelt, ergeben sich bei der Hochrechnung etwas andere Randverteilungen als in den offiziellen Mikrozensus-Veröffentlichungen des Statistischen Bundesamtes ausgewiesen. Dies sollte wenigstens bei einem Teil der Auswertungen vermieden werden. Nur im Onsite-Verfahren oder in der kontrollierten Datenfernverarbeitung kann aber mit allen Mikrozensus-Fällen gerechnet werden.
- Der Mikrozensus 2007 war in der Redaktionsphase des zweiten Berichts bereits im Onsite-Verfahren nutzbar, als SUF jedoch noch nicht verfügbar. Der Onsite-File sollte möglichst aktuelle Ergebnisse liefern.

An SUF kann gemeinsam entwickelte Syntax immer wieder getestet werden. Jedes Partnerinstitut kann die Auswertungen des anderen reproduzieren. So ist eine gemeinsame Validierung und Qualitätskontrolle bis zu den letzten Auswertungsschritten möglich. Dagegen kann sich bei Onsite-Nutzung oder Datenfernverarbeitung, da sich sowohl die Variablenstruktur als auch die Fallzahlen der Onsite- und der SUF-Files des Mikrozensus unterscheiden, erst im Nachhinein zeigen, ob die Analysen die richtigen Randverteilungen und plausible Resultate ergeben. Fehlersuche und Korrektu-

ren sind nur zu den verabredeten Rechenzeiten in den FDZ möglich. Nach jeder Neuberechnung vergehen wenigstens einige Tage, bis die kontrollierten Outputs aus dem FDZ verfügbar sind. Die Arbeitsdateien und Zwischenergebnisse, die diesen Outputs zugrunde liegen, können zwischen den beteiligten Partnerinstituten nicht ausgetauscht werden. Nach jeder Korrektur sind neue Bearbeitungstermine zu vereinbaren. (Zum Vergleich der möglichen Datenzugangswege siehe auch oben: 3.1.)

Onsite-Files des Mikrozensus wurden an einem FDZ individuell genutzt. An einem zweiten FDZ hatte ein Partnerinstitut Onsite-Nutzung beantragt, um dort auch die gemeinsam erstellte Analysesyntax zu verwenden. Dabei zeigte sich, dass für die erforderlichen Qualitätskontrollen, Korrekturen und Anpassungen beim Zusammenführen verschiedener Syntaxmodule zu Onsite-Auswertungsjobs deutlich mehr Zeit einzuplanen ist. Aus pragmatischen Gründen wurden die kollaborativen Mikrozensus-Auswertungen für den zweiten Bericht schließlich allein mit den verfügbaren SUF erstellt.

#### **4.4 Ergebnissicherung und -austausch**

Die Sozioökonomische Berichterstattung führt quantitative und qualitative Daten aus verschiedenen Quellen zusammen und deutet sie in einem gemeinsamem theoretischen Rahmen (hier: eines Umbruchs des deutschen Produktions- und Sozialmodells). Daher sind während der Forschungsphase die Verarbeitung und der Austausch von Zwischenergebnissen auch für die Verständigung über die Berichtskonzeption, für die gemeinsame Deutung der Ergebnisse und für die Integration verschiedener Befunde bedeutsam. Dieser kontinuierliche Diskussionsprozess, dem für den Erfolg des gemeinsamen Forschungsvorhabens hohe Bedeutung zukommt, fand vor allem in Verbundsitzungen und Arbeitsgruppentreffen, statt in denen aufbereitete Konzepte und Zwischenergebnisse aus Arbeitspaketen präsentiert und diskutiert wurden.

Zur Selbstinformation der Verbundpartner/innen wurde zu Beginn des zweiten Verbundvorhabens auf der projekteigenen Website ([www.soeb.de](http://www.soeb.de)) ein interner Bereich mit passwortgeschütztem Zugang und vorgegebener Verzeichnisstruktur eingerichtet, in dem Arbeitsunterlagen in allen gängigen Dateiformaten abgelegt werden konnten. Die Partnerinstitute nutzten diese Möglichkeit ungleich intensiv und insgesamt eher selten. Die befragten Verbundpartner/innen begründeten dies mit dem höheren Anspruch an allgemein zugängliche Dateien: Dort eingestellte Arbeitsunterlagen hätten auch für Verbundkolleg/inn/en verständlich sein müssen, die nicht direkt an einer Datenauswertung kooperierten, und für solche Aufbereitungen fehlte meist die Zeit. Der interne Bereich wurde daher vor allem von der Projektleitung genutzt, um z.B. vor Verbundsitzungen Entwurfss Fassungen und Zwischenprodukte für alle

Partnerinstitute bereitzustellen. Für kollaborative Datenauswertungen und für punktuelle Kooperationen nutzten die Partnerinstitute fast immer den direkten Kontakt mit einzelnen Adressat/inn/en über Email und Telefon.

Wie in Abbildung 13 gezeigt, wurden zwischen den drei lebensverlaufsorientierten Arbeitspaketen immer wieder Zwischenergebnisse ausgetauscht: Entweder wurden Tabellen- oder Grafik-Outputs aus den Statistikprogrammen (z.B. Dateien im \*.spo oder \*.spv-Format) oder aufbereitete Ergebnissen in Word- oder Excel Dateien (\*.doc- oder \*.xls-Format) übermittelt. Auf Basis dieser Unterlagen wurden dann Syntaxanpassungen oder weitere Verfahren entweder bilateral per Email oder Telefon oder in Arbeitsgruppentreffen der unmittelbar Beteiligten diskutiert.

#### 4.5 Dokumentation und Ergebnistransfer

Der gesamte Forschungsprozesses ist zur Qualitätssicherung zu dokumentieren. Zum einen sind nur Ergebnisse, deren Entstehung wechselseitig überprüfbar nachvollzogen werden kann, für andere Verbundpartner/innen nachnutzbar. Zum anderen verlangen auch die Regeln guter wissenschaftlicher Praxis nicht nur die Sicherung und Aufbewahrung von Primärdaten, sondern auch die Dokumentation von Resultaten und die Reproduzierbarkeit des Wegs zum Ergebnis. „Der primäre Test eines wissenschaftlichen Ergebnisses ist seine Reproduzierbarkeit. Je überraschender, aber auch je erwünschter (im Sinne der Bestätigung einer lieb gewordenen 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.“<sup>22</sup>

In fast allen Arbeitspaketen dokumentierten die Partnerinstitute die empirische Arbeit individuell auf unterschiedliche Weise. Das wichtigste Dokumentationsformat bildeten Syntaxdateien mit Kommentaren zu Arbeitsschritten im ASCII-Format. Für die Kommentierung gab es jedoch keinen einheitlichen Standard. Wie detailliert kommentiert wurde, hing vom individuellen Arbeitsstil ab. Beispielsweise entstanden in einem Falle im Verlauf des Arbeitsprozesses zur Lösung eines Problems fünf verschiedene Syntax-Do-File-Module, wobei jedes Modul eine andere Variante zur Lösung des Problems enthält. Die „beste“ Lösung wurde dann in die endgültige Hauptsyntax übernommen, wobei die übrigen vier Varianten und deren Ergebnisse sowie die darauf basierenden Entscheidungen in Kommentaren beschrieben wurden.

---

<sup>22</sup> Deutsche Forschungsgemeinschaft (DFG): Gute wissenschaftliche Praxis, Denkschrift, Weinheim 1998: 8. Download unter: [http://www.dfg.de/download/pdf/dfg\\_im\\_profil/reden\\_stellungnahmen/download/empfehlung\\_wiss\\_praxis\\_0198.pdf](http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_0198.pdf), Letzter Zugriff 07.06.2010

Einige Verbundpartner/innen verfassten 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 Projektpartner/innen hinterlegt, überwiegend wurden sie bilateral auf Nachfrage per Email ausgetauscht.

Für eine zukünftige kollaborative Arbeitsweise ergibt sich als Anforderung an den Forschungsverbund, Standards für eine detaillierte Dokumentation aller Arbeitsschritte im empirischen Forschungsprozess, von der Operationalisierung der zu analysierenden Konstrukte über Analysedesignentscheidungen bis zur Umsetzung in Syntax zu verabreden und diese Dokumentationen verbundintern zu archivieren.

Präsentation und Transferaktivitäten haben in einem Vorhaben der Sozialberichterstattung noch größere Bedeutung als in Forschungsprojekten, die zunächst auf eine innerwissenschaftliche und fachöffentliche Verwertung von Ergebnissen zielen. Für den Ergebnistransfer nutzt der Verbund zwei Formate: die Projektwebsite und die Buchveröffentlichung des Berichts. Beiträge von allgemeinem Interesse sollten auf der Projektwebsite fachöffentlich zugänglich eingestellt werden. Für den fachwissenschaftlichen Austausch hat sich das Format der ein- oder zweitägigen „Werkstattgespräche“ mit bis zu 50 Teilnehmer/innen bewährt.

Als problematisch bei der Ergebnisdarstellung erwiesen sich die unterschiedlichen Ausgabeformate der verschiedenen Statistikprogramme. Beispielsweise produzieren STATA und SPSS sehr unterschiedliche Tabellen. In der Kooperation zwischen den lebensverlaufsorientierten Arbeitspaketen wurde daher ein Zusatzmodul<sup>23</sup> genutzt, welches einheitliche Tabellen im \*.xls Format erzeugte. In anderen Arbeitspaketen musste dieses Problem nachträglich durch Übertragung der Informationen in Excel gelöst werden.

Die Aufbereitung von Ergebnissen für die Projektwebsite und für die Buchfassung des Berichts lag in soeb 2 beim SOFI als koordinierendem Institut bzw. beim ISF München, das für die redaktionelle Betreuung der Website zuständig war. Redaktionelle und arbeitsorganisatorische Probleme wurden in den beiden Instituten intern gelöst. zu lösen, die sich unabhängig von der mehr oder weniger kollaborativen Datennutzung für alle Arbeitspakete in gleicher Weise stellen. Wie Redaktionsarbeit, Lektorat, Literaturverwaltung sowie die Arbeit mit Formatvorlagen für Texte und Tabellen in künftigen Verbundvorhaben durch eine virtuelle Arbeitsumgebung technisch besser unterstützt werden können und welche Absprachen dabei zwischen den Verbundpartnern getroffen werden müssen, ist in der zweiten Projektphase weiter zu untersuchen.

---

<sup>23</sup> Es handelt sich um ein zusätzliches ado-Programm in Stata.

## 4.6 Probleme und ihre Bewältigung im Arbeitsprozess

Wie die Bilanzierung der empirischen, datenbezogenen Arbeitsprozesse in der Arbeit des Forschungsverbunds am zweiten Bericht zur sozioökonomischen Entwicklung Deutschlands zeigte, stellten sich bei der Zusammenarbeit der Verbundpartner/innen sowohl datentechnische als auch arbeitsorganisatorische Probleme. In der zweiten Projektphase ist daher klar zu unterscheiden zwischen Anforderungen an die Schnittstellen zu Datenhaltern (siehe 5.), an IT-technische Lösungen, die eine kollaborative Arbeitsweise unterstützten können (siehe 6.) und Anforderungen an die Steuerung und Koordination eines großen Verbundprojekts (vgl. dazu 7.)

In der Phase des Zugriffs auf Originaldaten mussten alle Verbundpartner/innen individuelle Nutzungsverträge mit den datenhaltenden Instituten schließen, was zusätzlichen Verwaltungs- und Zeitaufwand darstellte, jedoch nicht anders zu lösen war, da Gruppennutzungsverträge nicht möglich waren.

Im weiteren Verlauf des Forschungsprozesses zeigte sich, dass der Abstimmungsbedarf für den detaillierte Abgleich der anzuwendenden Datenmanagement-, Generierungs-, Recodierungs- und Analyseverfahren im Projektdesign unterschätzt worden war. Uneinheitliche Zeitplanungen und Kommunikationsverfahren, zu knappe zeitliche Ressourcen und fehlende Instrumente für Projektsteuerung erschwerten und verzögerten viele Abstimmungen. Diese Probleme wurden häufig durch bilaterale ad-hoc-Absprachen gelöst.

Auf der Programmier- und Datenebene in der Umsetzung der Analysedesigns zeigten sich im Laufe des Prozesses Übersetzungsprobleme auf Grund der Nutzung verschiedener statistischer Software, wie SPSS, STATA, TDA und Newspell. Für die jeweiligen Programme sind, wie erläutert, unterschiedliche Formen von Syntaxen und verschiedene Arbeitsdatensatzformate nötig. Probleme bei der Migration in andere Formate wurden verbundintern durch bilateralen Austausch, mündliche Syntaxerläuterungen sowie manuelle Konvertierung gelöst.

Einige sehr große Arbeitsdateien und rechenaufwändige Analyseverfahren stellten umfangreiche Anforderungen an die Hardwarekapazitäten und –Leistung. Probleme mit Hardwarekapazitäten ergaben sich insbesondere bei den IAB-Datensätzen, die ausschließlich durch Fernrechnen ausgewertet werden konnten. Durch die großen Fallzahlen des Mikrozensus ergaben sich bei den lokal verfügbaren Rechnerkapazitäten lange Rechenzeiten, die im jeweiligen Institut häufig durch paralleles Arbeiten auf einem zweiten Arbeitsplatzrechner überbrückt wurden.

Hardwareleistungsprobleme ergaben sich insbesondere bei den Sequenzanalysen im Arbeitspaket Lebensverlauf. Diese Prozeduren erfordern enorm viel Arbeitsspeicher und schnelle Prozessoren, die die Einzelplatzrechner der Forscher/innen zum Teil

nicht hatten. Um diese Auswertungen zu ermöglichen, mussten in einigen beteiligten Instituten während des Forschungsprozesses Kapazitäten durch upgrades sowie durch Neuanschaffungen erweitert werden. Falls dies nicht möglich war, wurde auch auf mehreren Arbeitsplatzrechnern parallel gearbeitet.

Vor allem bei der kollaborativen Syntaxentwicklung traten zudem Probleme durch die verschiedenen Dokumentationsstandards auf. Beim Austausch von Syntax mussten nachträglich bilateral zusätzliche Informationen geliefert werden. Zudem waren bilaterale Diskussionen von Syntaxmodulen auch zur Prüfung und Sicherung von der Qualität der Syntax nötig.

Bei der Verbreitung und Verarbeitung von Zwischenergebnissen wurde der interne, passwortgeschützte Bereich der Projektseite uneinheitlich und insgesamt zu wenig genutzt. Hierfür gibt es verschiedene Erklärungen:

- Vorbehalte, unfertige Arbeitsunterlagen verbundintern zugänglich zu machen,
- vergleichsweise schwierige Upload-Verfahren im verwendeten Content-Management-System,
- fehlende Verbindlichkeit dieses Kommunikationswegs.

Alles in allem wurden Zwischenergebnisse und Dokumentation mehrheitlich über bilaterale Kontakte von Verbundpartner/innen und Arbeitsgruppentreffen ausgetauscht und diskutiert. Dabei konnten Probleme bezüglich geistigem Eigentum von Syntax und Qualitätssicherung in direktem Kontakt ausgeräumt werden.

## 5. Datenschnittstellen einer virtuellen Forschungs- umgebung

Die Daten haltenden Institute zeigten in den Expertengesprächen und in den Diskussionen im Rahmen des Workshops Interesse an Lösungen, welche die bestehende Dateninfrastruktur nicht nur für die sozioökonomische Berichterstattung, sondern für alle Nutzer/innen verbessern könnten. Grundsätzlich signalisierten alle befragten Vertreter/innen der Daten haltenden Institute<sup>24</sup>, dass Gruppennutzungslizenzen für einen Forschungsverbund möglich seien. Dabei müssten jedoch die geltenden Datenschutzregelungen gewahrt werden.

Über diese relativ einfach umzusetzenden Änderungen in den Zugangsregeln hinaus ist für das Modellprojekt von Interesse, welche Überlegungen Forschungszentren und Datenservicezentren zur künftigen Datenbereitstellung anstellen. Für die Umsetzung einer virtuellen Arbeitsumgebung ist davon auszugehen, dass sich die Schnittstelle zur Forschungsdaten-Infrastruktur gleichzeitig auch von der Anbieterseite her verändern wird.

Für die Längsschnittstudie Sozio-oekonomisches Panel setzt die SOEP-Gruppe des DIW bereits die Restrukturierung der SOEP-Daten von derzeit über 200 Dateien in eine sehr kleine Anzahl von Dateien (etwa fünf) um. Ebenfalls in naher Zukunft sollen die Daten des SOEP ausschließlich online weitergegeben werden. Ausserdem evaluiert das DIW in Zusammenarbeit mit dem Datenservicezentrum in Bielefeld (Stefan Liebig) derzeit die nachträgliche Erhebung von Betriebskontextdaten, die mit den personenbezogenen SOEP Mikrodaten verknüpft werden. Auch die Ergebnisse weiterer angelagerter Organisationsbefragungen (beispielsweise Kita-Informationen) sollen SOEP-Personendaten zugeordnet werden. In Kooperation mit der DLR testet die SOEP-Gruppe derzeit, wie verschiedenste Raumb Beobachtungsdaten auf Basis von GPS-Codes an die SOEP-Geocodes angefügt und wie den Datennutzer/innen weitere kontextuelle Informationen zur Verfügung gestellt werden können.

Betreffen diese Entwicklungen das Datenangebot selbst, so soll das neu entwickelte Tool „Panelwhiz“<sup>25</sup> für das SOEP und andere Paneldatensätze<sup>26</sup> die Stichprobenzusammenstellung und -Abgrenzung mit gleichzeitiger Syntaxgenerierung unterstützen. Es soll zukünftig auch die Option bieten, selbst generierte Syntax online ein-

<sup>24</sup> Diese Angaben beruhen auf Aussagen des DIW, des BIBB, des FDZ der Rentenversicherung sowie des FDZ des statistischen Landesamtes Niedersachsen (LSKN) auf dem Workshop am 9. Februar 2010.

<sup>25</sup> <http://www.panelwhiz.eu/>

<sup>26</sup> Panelwhiz unterstützt auch die Erstellung von Paneldatensätzen des FDZ des IAB. Voraussetzung für die Nutzung ist, dass ein Datennutzungsvertrag geschlossen wurde und die Daten selbst auf der lokalen Festplatte vorliegen.

zustellen, mit der Zusicherung, dass die Urheber/innen bei Verwendung durch Dritte zitiert werden.

Im Bereich der Datenserviceleistungen werden nicht nur im DIW, sondern auch in anderen Daten haltenden Instituten (GESIS-ZA, DESTATIS), in neu gegründeten Datenservicezentren (MicrodataLab und Missy in der GESIS, International Data Service Center im IZA) und in anderen Einrichtungen Projekte zur Implementation verbesserter Metadateninformationen sowie zum Ausbau des virtuellen Datenzugriffs und der Datenanalyse durchgeführt. So wurde im Projekt "Questionnaire Development Document Support" (QDDS 3) am Institut für Soziologie der Universität Duisburg-Essen (Prof. Dr. R. Schnell) wurde eine Software entwickelt, mit welcher der gesamte Entstehungsprozess der Fragen in Fragebögen nachvollzogen werden kann. Auch hier prüft die SOEP-Gruppe, inwieweit dieses Verfahren in das SOEP-Informationssystem integriert werden kann. Auch die Metadatenanwendung DDI 3<sup>27</sup> wird für verschiedene Anwendungsfälle bereits getestet und umgesetzt<sup>28</sup>. Beispielsweise entspricht das Mikrozensus-Informationssystem „MISSY“<sup>29</sup> dem Dokumentationsstandard von DDI zur Beschreibung von Sozialdaten. GESIS ist zudem Mitglied bei CESSDA (Council of European Social Science Data Archives), einem zentralen Datensatzinformationssystem für europäische sozialwissenschaftliche Daten. Dadurch lassen sich auch deutsche Daten und insbesondere Metadaten, die für europäische Datensätze wie beispielsweise das Eurobarometer erhoben wurden, online zentral durchsuchen. Auch die SOEP-Gruppe prüft derzeit, inwieweit DDI zukünftig für das derzeitige SOEP-Info-System genutzt werden kann.

Für soeb 3 wäre auch zu prüfen, welche Teile oder Entwicklungsbereiche von DDI 3 in einer virtuelle Arbeitsumgebung für die Partnerinstitute eines Verbundvorhabens von Nutzen sind und ob und wie sie implementiert werden können.

Nach Implementation geeigneter Datenzugangs- und -Verwaltungssoftware sowie netzwerkfähiger Open-Source-Statistiksoftware (z.B. R) wären auch datenschutzkonforme neue Lösungen beim Fernrechnen denkbar (wie beispielsweise im National Opinion Research Center (NORC) in Chicago).

Für die Entwicklung einer virtuellen Arbeitsumgebung für die sozioökonomische Berichterstattung ist die SOEP-Gruppe des DIW als Datenhalter und aufgrund der vielfältigen Aktivitäten und Erfahrungen bei der Verbesserung der Datenbereitstellung

---

<sup>27</sup> Data Documentation Initiative (DDI): URL: <http://www.ddialliance.org/>; letzter Zugriff 17.04.10.

<sup>28</sup> Auf der ersten europäischen DDI-Konferenz im Dezember 2009 in Bonn wurden einige dieser Projekte vorgestellt. Vgl. [http://www.iza.org/conference\\_files/eddi09/2009\\_12\\_11\\_online\\_program.pdf](http://www.iza.org/conference_files/eddi09/2009_12_11_online_program.pdf) letzter Zugriff 17.04.10.

<sup>29</sup> <http://www.gesis.org/missy-test/>

ein wichtiger Partner. 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. Zudem ist die SOEP-Gruppe Mitglied der Working Group on Future Data Access des Rates für Sozial- und Wirtschaftsdaten (RatsWD) und kooperiert nicht zuletzt über das CNEF (Cross national equivalent file) international mit Instituten, die Längsschnittdaten halten. In der zweiten Projektphase soll daher die SOEP-Gruppe des DIW zur Mitwirkung an der Entwicklung von Konzepten für die Schnittstellen einer virtuellen Arbeitsumgebung zur Dateninfrastruktur eingeladen werden. Sie kann darüber hinaus Interessen der Daten haltenden Einrichtungen an Dokumentationsstandards für Forschungsdaten in das Modellprojekt einbringen.

## 6. Anforderungen an eine virtuelle Arbeitsumgebung für die sozioökonomische Berichterstattung

Auf Basis der beschriebenen Arbeitsabläufe in der sozioökonomischen Berichterstattung bei der kollaborativen und individuellen Projektarbeit mit Mikrodaten ergeben sich funktionale Anforderungen an eine virtuelle Arbeitsumgebung für die sozioökonomische Berichterstattung, die in den nächsten Abschnitten entsprechen der technischen Terminologie von WissGrid und D-Grid dargestellt werden.

### 6.1 Anforderungen an Langzeitarchivierung und Forschungsdatenarchiv

Die bisherige Bestandsaufnahme der zu unterstützenden Arbeitsprozesse hat gezeigt, dass kollaborative Arbeit an sozialwissenschaftlichen Mikrodaten wesentlich iterativ stattfindet. Die gleichzeitige Arbeit an einem Datensatz oder komplexere „work flows“ mit verteilten Rechenaufgaben dürften eher die Ausnahme sein. Die Regel ist der Austausch über verschiedene Zwischenergebnisse und Arbeitsstände zwischen Verbundpartnern und die Nachnutzung von Syntax und generierten Variablen. Daher bilden der Aufbau eines Forschungsdatenarchivs und Verfahren der Selbstinformation aktuell den Kern der zu entwickelnden Arbeitsumgebung.

Langzeitarchivierung und Nachnutzung sind für alle projektbezogenen Daten und Dokumente zu unterstützen. Der größte Bedarf besteht bei der nachnutzbaren Archivierung von Arbeitsdatensätzen und Syntaxdateien. Die Architektur einer virtuellen Arbeitsumgebung muss offen sein für neue, zu Projektbeginn möglicherweise noch nicht bekannte Originaldatensätze. Diese und die dazu erstellten Arbeitsdatensätze und Syntaxdateien sollten problemlos über eine Benutzeroberfläche integriert werden können. Sofern die datenschutzrechtlichen Bestimmungen dies erlauben, sollte der lesende Zugriff auf Syntaxdateien und Arbeitsdatensätze auch nach Projektende möglich sein, so dass die abgelegten Daten (Syntax und Arbeitsdatensätze) für Re-Analysen oder als Ausgangsdaten für weitere Analysen verwendbar wären. Um das Auffinden der abgelegten Daten und den Datenzugriff zu erleichtern, sollte sich der Datenbestand über eine einfache Eingabemaske nach technischen und kontextbezogenen Metadaten durchsuchen lassen. Vor allem für die Syntaxdateien wäre die Möglichkeit zur Volltextsuche sinnvoll.

Für einen Teil der Originaldatensätze und der daraus erstellten Arbeitsdatensätze gelten Löschfristen der Daten haltenden Einrichtungen. Daher sind automatisierte Löschroutinen nach Projektende technisch wünschenswert. Allerdings stehen diese

Auflagen in einem Spannungsverhältnis zum Interesse an nachnutzbarer Langzeitar-  
chivierung – hier besteht inhaltlicher Diskussionsbedarf.

Die Zugangsrechte aller Projektpartner/innen müssen sich detailliert personenbe-  
zogen definieren lassen, damit das Arbeiten in einer virtuellen Arbeitsumgebung den  
Datenschutzbestimmungen und den Nutzungsbedingungen für die verschiedenen  
Originaldatensätze genügt. Im weiteren Projektverlauf müssen daher alle technischen  
Möglichkeiten der Verwaltung von Zugangsrechten geprüft und ausführlich darge-  
stellt werden. Datensicherheit umfasst Zutritts-, Zugangs-, Zugriffs-, Weitergabe-, Ein-  
gabe-, Auftrags- und Verfügbarkeitskontrolle und den Schutz gegen Datenverlust.

Ablage und Zugriffs auf Syntax und Arbeitsdatensätze und deren Sicherung sind in  
einer virtuellen Arbeitsumgebung fortlaufend zu dokumentieren. Zu prüfen ist daher  
die technische Umsetzbarkeit von Ablage- und Zugriffssystemen mit Sub-Version-  
Control-Systemen, d.h. mit Softwarelösungen, die den Entstehungsprozess verschie-  
dener Versionen von Datensätzen steuern und dokumentieren. Dabei wäre es wün-  
schenswert, technische und kontextbezogene Metadaten als Versionsinformationen  
hinterlegen zu können.

Im Interesse einer besseren Unterstützung von Fernrechenanalysen ist zu prüfen  
und darzustellen, wie sich open-Source-Statistikprogramme (z.B. GridR) integrieren  
lassen und wie GRID-Workflow-Rechnerkapazitäten genutzt werden können<sup>30</sup>.

## 6.2 Anforderungen an Metadatenextraktion

Nicht nur für die Dokumentation des Forschungsprozesses, sondern auch für den Aus-  
tausch von Syntax-Dateien zwischen den Projektpartnern und für deren Nachnutzung  
in der laufenden Arbeit ist eine umfangreiche technische und kontextbezogene Me-  
tadatenextraktion aus den Arbeitsdateien und den Syntaxdateien nötig. Daher soll die  
Leistungsfähigkeit von Ablage- und Zugriffssystemen, Sub-Version-Control-Systemen  
und Metadatenmanagement für die Ablage und automatisierte Metadatenextraktion  
(Formatcharakterisierung) von Syntax für Statistikprogramme geprüft und dargestellt  
werden: Gibt es Software-Lösungen, die aus den Syntaxdateien automatisiert Syntax-  
beschreibungen –generiert und verwaltet? Syntaxbeschreibungen und andere Meta-  
daten (z.B. über Arbeitsdatensatzversionen) sollten durchsuchbar sein und auf die  
entsprechende Quelldatei (also Syntax oder Arbeitsdatensatz) verlinkt sein.

In der bisherigen Verbundpraxis waren Syntax- und Datendokumentation nicht  
standardisiert. Die Dokumentation geschah in unterschiedlicher Form meist innerhalb  
der individuellen Syntax. Für eine automatisierte Metadatenverwaltung sollte ein Do-

---

<sup>30</sup> Siehe hierzu <http://cran.r-project.org/web/packages/GridR/index.html>

kumentationsstandard entwickelt werden, der nach Muss-, Kann- und optionalen Metadatenanforderungen unterscheidet.

### **6.3 Anforderungen an Datenkonvertierung**

Aus Sicht der fachwissenschaftlichen Anwender/innen ist es wünschenswert, Arbeitsdatensätze (idealerweise auch Syntax) aus einem Statistikformat in andere Formate und in verschiedene Programmversionen zu konvertieren. Hierzu ist zu klären, ob bereits entsprechende Konvertierungstools bestehen und welche Dienste zur Formatkonvertierung, -validierung und charakterisierung nötig und möglich sind.

Dabei soll die Interoperabilität der Dienste untereinander und mit dem Forschungsdatenarchiv gewährleistet sein. Es ist zu prüfen, ob die Dienste als Grid-Jobs oder als Web-Services integriert werden sollen, welche Schnittstellen benötigt werden, und welche Prozessschritte im Provenienzdienst protokolliert werden sollen. Zudem ist zu prüfen und darzustellen, welche Informationen extrahiert werden (können) und in welchem Format die erzeugten Status- und Prozessmeldungen vorliegen.

### **6.4 Benutzeroberfläche**

Um eine kollaborative Arbeitsweise zu unterstützen, soll für ein neues Verbundvorhaben eine gemeinsame interne, interaktive Projekt-Benutzeroberfläche eingerichtet werden, die nur über persönliche Authentifizierung erreichbar ist und Zugriff auf alle Ressourcen und alle beschriebenen Dienste einschließlich der Suche in Syntaxdateien und Arbeitsdatensätze erlaubt.

Da Gruppen-Lizenzen für die Statistikprogramme SPSS und STATA zu teuer wären, sollte das open-source Statistikprogramm R innerhalb der virtuellen Arbeitsumgebung optional verfügbar sein und ausgehend von der Projekt-Benutzeroberfläche genutzt werden können. Es sollte die Möglichkeit bestehen, alle statistischen Jobs, insbesondere kapazitätsintensive Jobs, wie Sequenzanalysen und Optimal Matching im Grid bearbeiten zu lassen.

Die Benutzeroberfläche die gemeinsame Dateibearbeitung- und verwaltung unterstützen. Sie sollte ein Benachrichtigungssystem zu Veränderungen an Dateien und interne Kommunikationsmöglichkeiten (ähnlich wie bei SKYPE) einschließen. Über eine Schnittstelle bzw. ein integriertes WIKI sollten ausgewählte Inhalte kommuniziert und unaufwändig auf eine externe öffentliche Website (soeb.de) eingestellt werden können.

## 7. Ausblick auf die zweite Projektphase

Die Bestandaufnahme der Erfahrungen, die der Forschungsverbund Sozioökonomische Berichterstattung bei der Nutzung der arbeitsteiligen und kollaborativen Arbeit mit sozialwissenschaftlichen Mikrodaten Dateninfrastruktur für sein Berichtsvorhaben gesammelt hat, zeigt ein erhebliches Potenzial für eine bessere IT-technische Unterstützung arbeitsteiliger und kollaborativer Analysen von Mikrodaten in großen sozialwissenschaftlichen Forschungsverbänden. Die Entwicklung einer solchen virtuellen Arbeitsumgebung ist jedoch kein bloß technisches Vorhaben. An der Schnittstelle zwischen sozialwissenschaftlichen Nutzer/inne/n und Forschungsdatenzentren verändert das IT-System zum einen die wissenschaftliche Datennachfrage und reagiert zum anderen auf Veränderungen im Datenangebot und in der Datenbereitstellung durch Forschungsdatenzentren und Datenservicezentren. In der Projektorganisation verlangt die gemeinsame Nutzung einer technischen Infrastruktur verbindlichere inhaltliche Absprachen und eine intensivere und komplexere prozessbegleitende Koordination und Steuerung von Arbeitsprozessen.

Von April bis Juli 2010 evaluiert die D-Grid Entwicklungs- und Betriebsgesellschaft mbH im Rahmen eines Forschungs- und Entwicklungsvertrags mit dem SOFI, welche Systemanforderungen sich aus den funktionalen Anforderungen an eine virtuelle Arbeitsumgebung ergeben und wie weit Ressourcen und Technologien, die im Rahmen der D-Grid-Initiative und in den an WissGrid beteiligten wissenschaftlichen Communities Anwendung finden, für die Umsetzung eines solchen IT-Systems genutzt werden können. Diese „technische Expertise“ soll auch Aufschluss darüber geben, welche Funktionen mit vertretbarem technischem Entwicklungsaufwand rechtzeitig für ein drittes Verbundvorhaben sozioökonomischer Berichterstattung bereitgestellt werden können und welche einer späteren Ausbaustufe vorbehalten bleiben müssen.

Gleichzeitig sollen in der zweiten Projektphase die Funktionen einer virtuellen Arbeitsumgebung mit Einrichtungen der Dateninfrastruktur unter der Fragestellung besprochen werden, wie die Datenschnittstelle zu gestalten ist und wie das IT-System datenschutzkonform umgesetzt werden kann. Mit der SOEP-Gruppe im DIW sollen insbesondere zu den folgenden Fragen Inputs aus Sicht eines Forschungsdatenzentrums eingeholt werden:

- zum Diskussionsstand Daten haltender Institute bezüglich technischer Umsetzungsmöglichkeiten für Remote Data Access,
- zu rechtlichen Aspekten dieser Form des Datenzugriffs,
- zum Diskussionsstand zu Metadatenstandards, etwa zu den Einsatzmöglichkeiten von DDI für soeb 3,

– zur beabsichtigten Weiterentwicklung der Datenbereitstellung im FDZ des DIW. Neben der SOEP-Gruppe sollen andere amtliche und wissenschaftliche Forschungszentren weiter in die Diskussion einbezogen werden.

Die Grundzüge einer Architektur für eine virtuelle sozialwissenschaftliche Arbeitsumgebung und Gestaltungsmöglichkeiten für die Schnittstelle zur Dateninfrastruktur sind auf einem zweiten Projekt-Workshop am 19. Juli 2010 in Göttingen zu diskutieren. Dieser Zeitpunkt ist so gewählt, dass die Endfassung der technischen D-Grid-Expertise die Ergebnisse des Workshops noch berücksichtigen kann.

Im Verlauf der ersten Projektphase wurde deutlich, dass die Entwicklung und Umsetzung einer virtuellen Arbeitsumgebung nicht nur IT-technisch angegangen werden kann, sondern eine Intensivierung und Verstärkung der Verbundkoordination erfordert. Nicht zuletzt um eine „Einbettung“ der technischen Architektur in einen Vorschlag für die künftige Arbeitsweise des Verbunds und eine Ausarbeitung der nicht-technischen Anforderungen an kollaborative Arbeitsprozesse zu ermöglichen, wurde daher die Konzeptphase für ein drittes Verbundvorhaben „Berichterstattung zur Sozioökonomischen Entwicklung Deutschlands“ (soeb 3) kostenneutral bis Ende November 2010 verlängert.

In der bisherigen Verbundarbeitsweise fand Steuerung im Wesentlichen auf zwei Wegen statt:

- durch eine formale Aufgabenteilung zwischen den Verbundpartnern und durch vertikale Kommunikation der Projektleitung mit den Bearbeiter/inne/n der einzelnen „Arbeitspakete“,
- durch Arbeitstreffen und persönliche Kontakte.

Dokumentation, gemeinsame Nutzung von Daten, Syntax und Methodenkompetenz blieben weitgehend den einzelnen Verbundpartnern überlassen.

Während die technische Expertise betont, dass eine virtuelle Arbeitsumgebung bestehende Vorgehensweisen und Arbeitsprozesse unterstützen soll, ohne eine bestimmte Arbeitsweise technisch vorzugeben, bildet die Nutzung einer gemeinsamen IT-Plattform unter arbeitsorganisatorischen Gesichtspunkten einen eigenen, neuen Steuerungsmodus. Daher muss die Konzeptphase auch den nicht-technischen Entwicklungsbedarf berücksichtigen. Sowohl die Expert/inn/eninterviews mit bisherigen Verbundpartnern als auch der ständige Austausch mit den Bearbeitern der technischen Expertise haben Probleme und Grenzen der bisherigen Verbundkoordination offen gelegt. Die wirksame Nutzung einer virtuellen Arbeitsumgebung setzt die Einhaltung verbindlicher Dokumentationsstandards und Konventionen zur Datenablage voraus und erzeugt ein größeres Maß an Transparenz zwischen den parallel bearbeiteten Arbeitspaketen. Sie wird von den Projektpartner/inne/n nur akzeptiert werden,

wenn die praktischen Vorteile eine Nutzung attraktiv machen und wenn die Systemarchitektur die berechtigten Ansprüche an individuellen Gestaltungsspielraum bei der wissenschaftlichen Arbeit (etwa die Wahl des Statistikprogramms, die spontane und problembezogene horizontale Vernetzung sowie die Kontrolle über die Weitergabe von Zwischenergebnissen) respektiert. In der zweiten Projektphase müssen daher auch die Anforderungen an die professionelle Steuerung und Koordination eines großen Verbundprojekts und die Konsequenzen für Ressourcenbedarf und Zeitplanung erörtert werden. Benötigt wird nicht nur eine technische Systemarchitektur, sondern zugleich ein dem Forschungsprozess angemessenes Koordinations- und Steuerungsmodell. Gelingt eine „Einbettung“ in ein Modell der Projektsteuerung nicht, kann die Nutzung einer virtuellen Arbeitsumgebung in der Forschungspraxis individuell leicht umgangen werden, und sie wird ihren praktischen Nutzen nicht entfalten können.

## ANHANG

### Protokoll des Workshops „Fachwissenschaftliche Anforderungen an eine virtuelle Arbeitsumgebung für SOEB“ – Göttingen, 9. Februar 2010

#### Teilnehmende

Ilya Agapov, Peter Bartelheimer, Irene Becker, Sarah Cronjäger, Thomas Drosdowski, Rita Hoffmeister, Jens Ludwig, Frank Schlünzen, Tanja Schmidt, Jürgen Schupp, Ewa Sojka, Michael Stegmann, Falko Trischler, Klaus-Peter Wittemann.

#### Statements zum Input von Tanja Schmidt

##### *Jürgen Schupp (DIW)*

- Auch die Halter wissenschaftsträger Daten haben ein Interesse an der Entwicklung von Standards für die Dokumentation, die eine problembezogene Recherche ermöglichen. Dabei sind Verbesserungen der bestehenden Serviceleistungen wie SOEP-Info in den Blick zu nehmen, ebenso die „Paradaten“ der beauftragten Erhebungsinstitute, die Fachinformationssysteme und Datenarchive.
- Andere Initiativen sind zu berücksichtigen: z.B. CESSDA 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 Informations-

- infrastruktur“ der Leibniz-Gemeinschaft (WGL) im Auftrag der Gemeinsamen Wissenschaftskonferenz (GWK) teil.
- DFG-Projekt von Rainer Schnell Questionnaire 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.

*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 SOEP-Info 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.