



Anwendungsfall Sozialwissenschaften: Kollaborative Datenauswertung in virtueller Arbeitsumgebung

Begutachtung des WissGrid AP 3
28. Januar 2010, AIP Potsdam

Peter Bartelheimer
Tanja Schmidt
SOFI, Göttingen



Sozioökonomische Berichterstattung (soeb)

- Berichterstattung zur sozioökonomischen Entwicklung Deutschlands
 - Umbruch in Wirtschafts- und Lebensweise
 - Auswirkungen auf Wohlfahrt und Ungleichheit
 - Wohlfahrtsmessung – BIP ist kein Wohlfahrtsmaß
- Forschungsverbund sozialwissenschaftlicher Institute mit Mitteln der Forschungsförderung (BMBF)
 - Neue Brücken zwischen Sozialforschung und Sozialberichterstattung
 - Übertragung von Forschungsergebnissen in Konzepte und Indikatoren für Dauerbeobachtung
- Mehr Information: www.soeb.de



Methoden und Datengrundlage

- Nutzung neuer Infrastruktur (Sozial- und Wirtschaftsdaten)
 - Mikrodatenzugänge über Forschungsdatenzentren
 - prozessproduzierte Längsschnittdaten, Panelstudien
- Sekundärnutzung amtlicher oder prozessproduzierter Mikrodaten, z.B.
 - Sozio-oekonomisches Panel (GSOEP)
 - Mikrozensus
 - Integrierte Erwerbsbiografien, Linked Employer-Employee-Daten
- Integration von Mikrodatenanalysen in makroökonomische Modelle
 - Nutzung von Szenariotechniken und Projektionsmodellen



- Erster Bericht zur sozioökonomischen Entwicklung Deutschlands (soeb 1)
 - Arbeit und Lebensweisen (SOFI u.a.) 2005)
- Zweiter Bericht sozioökonomischen Entwicklung Deutschlands (soeb 2)
 - Teilhabe im Umbruch (im Erscheinen, 2010)
- Dritter Bericht sozioökonomischen Entwicklung Deutschlands (soeb 3)
 - Fachöffentliche Konzeptphase – Werkstattreihe 1. Hj. 2010



Modellprojekt VirtAug für soeb 3

- Modellprojekt „Kollaborative Datenauswertung und Virtuelle Arbeitsumgebung“ – VirtAug
 - Laufzeit August 2009 bis Juli 2010 im Rahmen der Konzeptphase für soeb 3
 - Förderung durch BMBF
 - Projektdurchführung: SOFI (Göttingen)
 - Dr. Peter Bartelheimer (Projektleitung), Tanja Schmidt
 - Kooperation mit Forschungsdatenzentrum Sozio-oekonomisches Panel (DIW)
 - Unterauftrag zur technischen Evaluation der Grid-Entwicklung



Ziele des Modellprojekts VirtAug

- Anforderungsprofil für virtuelle Vernetzung von Verbundpartnern (Forschungseinrichtungen) an verschiedenen Standorten zur
 - Sekundäranalyse von Forschungsprimärdaten
 - Archivierung und Dokumentation von Arbeitsergebnissen
- Prüfung bestehender IT-Entwicklung in der Grid-Community auf ihre Eignung
- Erarbeitung eines ersten Umsetzungskonzeptes für die Arbeit an soeb 3
- Ggf. Vorarbeiten für einen "Anwendungsfall Sozialwissenschaften" in der D-Grid-Initiative



Langzeitarchivierung ...

- Langzeitarchivierung von Primärdaten durch Datenhalter (FDZ)
- Anforderungen an Langzeitarchivierung von Arbeitsdateien
 - Regeln guter wissenschaftlicher Praxis (DFG 1998)
 - Originaldaten (Primärdaten), die Grundlage von Veröffentlichungen sind, sollen zehn Jahre aufbewahrt werden
 - Zugangsmöglichkeiten für berechnigte Dritte
 - Überprüfung von Arbeitsergebnissen, Reproduzierbarkeit
 - Arbeit nach Wissensstand („lege artis“)
 - Regeln für Sicherung und Aufbewahrung von Primärdaten:)
 - Antragsanforderungen Datenmanagementkonzepte als Antragsbestandteil in Forschungsförderung (z.B. DFG)



... und mehr: kollaborative Datenauswertung

- Arbeitsteilige Arbeit an Forschungsdatensätzen
 - Aufbereitung von Arbeitsdateien, Data Matching (Beispiele SOEP, LIAB)
 - Gemeinsame Erstellung modularer Auswertungssyntax
 - Transfer spezieller Datenexpertise
 - Nutzung von Daten, Syntax in verschiedenen Arbeitspaketen
 - Konvertierung in andere Statistikprogramme
 - Archivierung von Outputs, Übertragung in Excel
- Nachnutzbarkeit
 - für spätere Arbeitsphasen: Aktualisierung, lange Reihen
 - für berechnete Dritte



Sekundärnutzung sozialwissenschaftlicher Mikrodaten

- Forschungsdateninfrastruktur für Einzelnutzer
 - Grundfiles auf viele Forschungsdatenzentren / Datenhalter verteilt
 - Nutzung für Forschung in drei Verfahren
 - Fernrechnen
 - Onsite-Rechnen
 - faktisch anonymisierte Scientific Use Files (SUF)
 - Zugangsbeschränkung in Einzelnutzungsverträge von Verbundpartner/inne/n



Probleme im arbeitsteiligen Arbeitsprozess

- Kein gemeinsamer Zugriff auf Grunddaten
 - Onsite-Files, SUF nicht identisch
 - Probleme bei explorativer Datenanalyse, Fehlersuche, Übernahme von Syntaxmodulen
- Gemeinsame Nutzung generierter Arbeitsdateien kann Datenschutzbelangen und Nutzungsregeln der Datenhalter widersprechen
- Beschränkte PC-Speicher- und Rechenkapazität
 - bei großen Datensätzen (z.B. IEBS; LIAB)
 - bei komplizierten Längsschnittverfahren (z.B. Sequenzanalysen, OMA)

Anwendungsfall Mikrozensus



- Partnerinstitute arbeiten mit verschiedenen Scientific Use Files (Unterstichproben)
 - Gemeinsame Syntax ergibt keine identischen Ergebnisse
- Nutzungsvertrag eines Partners mit FDZ für Onsite-File
 - Onsite-Rechnen einer Person zu festgesetzten Zeiten / Terminen mit An- und Abreise
 - Für aktuellsten Mikrozensus (2007) nur Onsite-File
 - Abweichende Datenstruktur von Onsite-File und SUF
 - Zellenbesetzungen können nicht vorab überprüft werden
 - „Spieldatensätze“ lassen keine inhaltliche Kontrolle von Auswertungssyntax über Randverteilungen zu
 - Kein gemeinsamer, gleichzeitiger Zugriff auf Onsite-Outputs



Die (geplante) Grid-Forschungsumgebung (1)

- Forschungsdatenarchiv
 - Gemeinsamer Zugriff auf Grunddaten (?) –
Regelungsbedarf mit Forschungsdatenzentren
 - Archivierung und Nachnutzung von Arbeitsdateien
 - Datenschutz und beschränkte Zugriffsrechte
 - entsprechend Nutzungsverträgen mit Datenhaltern
 - entsprechend Interessen der Partnerinstitute
 - Archivierung und Dokumentation von Syntax und Outputs (Gemeinsame Dokumentationsregeln)
 - Archivierung und Dokumentation von Arbeitsunterlagen, Arbeitspapieren



Die (geplante) Grid-Forschungsumgebung (2)

- Datenkonvertierung
 - Forschungsdaten (Datenformate, z.B. SPSS, Stata)
 - Auswertungssyntax für verschiedene Statistikprogramme
- Provenienzdienst, Datenextraktion, Valisierung
 - Welche Datensätze sind vorhanden? (Dokumentationsstandards)
 - Wer hat Datensätze, Syntax erstellt?
 - Wie wurde Syntax geprüft?
- Projekt-WIKI zur internen net-basierten Kommunikation zwischen Kooperationspartnern



Die (geplante) Grid-Forschungsumgebung (3)

- Unterstützung von Textproduktion
- Schnittstelle zur Projekt-Website:
Ergebnisdokumentation
 - Zugriff auf Outputs durch Dritte
- Offene Fragen
 - Gemeinsame Lizenz für Statistikprogramme auf Grid-Rechner?
 - Gemeinsame Nutzung von Rechnerkapazität?
 - „Geistiges Eigentum“ an Syntax, Arbeitsdateien



Nächste Schritte in VirtAug

- Workshop am 9.2.2010: Forschungsverbund, Forschungsdatenzentren, WissGRID (LZA, Fachberater)
- Spezifikation von Anforderungen an Grid-Forschungsumgebung (Februar 2010)
- Vergabe einer technischen Expertise zu Nutzungsmöglichkeiten bestehender Grid-Entwicklungen, Abschätzung von Entwicklungsbedarf (ca. März 2010)
- Projektförmiges Umsetzungskonzept für Verbundprojekt Dritter Bericht zur sozioökonomischen Entwicklung Deutschlands (soeb 3) (Juli 2010)

Exkurs – Ein möglicher anderer Anwendungsfall



- Qualitative Empirie in Verbundinstituten (z.B. Fallstudien, vergleichbar MediGRID)
- Langzeitarchivierung, Nachnutzung von
 - Korrespondenz
 - Fallbezogenen Dokumenten (für Dokumentenanalysen)
 - Erhebungsinstrumenten (z.B. Leitfäden, ausgefüllte Erhebungsbögen, Beobachtungsprotokolle)
 - Gesprächsmaterial (Audiodateien, Rohtranskripte, anonymisierte Transkripte)
 - Memos und Protokollen
 - Identifizierungs- und Pseudonymisierungslisten
- Nachnutzung, kollaborative Nutzung derzeit nicht geregelt