

- **Kollaborative Datenanalyse und virtuelle Arbeitsumgebung – Erfahrungen und Anforderungen aus der Verbundarbeit an »soeb 2«**

Workshop

“Fachwissenschaftliche Anforderungen an eine virtuelle Arbeitsumgebung für »soeb«,

09.02.10, SUB Göttingen

SOFI

Soziologisches Forschungsinstitut
an der Universität Göttingen

Tanja Schmidt

 **soeb.de**

■ Inhalt

■ Erfahrungen aus der Arbeit an »soeb 2«

■ Anforderungen an kollaborative Analysen und virtuelle Arbeitsumgebung

■ Offene Fragen

■ Erfahrungen aus »soeb 2« - kollaborativ genutzte Originaldatensätze (1)

■ Mikrodatensätze

- SOEP (DIW) – Einzelnutzungsverträge
- Mikrozensus, verschiedene Jahrgänge (DESTATIS)
 - Einzelnutzungsverträge für faktisch anonymisierte SUF
 - Onsite-Nutzungsverträge

■ Aggregierte Daten (ohne Nutzungsverträge)

- AMECO (OECD)
- OECD, Eurostat

■ Erfahrungen aus »soeb 2« - individuell genutzte Originaldatensätze (2)

■ Mikrodatensätze

- Integrierte Erwerbsbiographiestichprobe (IAB); Linked Employer-Employee (IAB); Betriebspanel (IAB); Daten zu Arbeitsbedingungen (BIBB-IAB / IAB BAUA) ; gepoolte Renten-zugangsstichproben (RV); Versicherungskontenstichprobe Längsschnitt (RV); Einkommens- und Verbrauchsstichprobe (StBA); Zeitbudgetstudie (StBA); DGB-Index »Gute Arbeit« (DGB Index Gesellschaft); EU-SILC »Survey on Income and Living Conditions« (EU);
- Dafür individuelle Einzelnutzungsverträge für SUF, Onsite-Nutzung oder Fernrechnen

■ Aggregierte Daten

- Volkswirtschaftliche Gesamtrechnung (DESTATIS)
- EU-Strukturindikatoren (Eurostat)

■ Erfahrungen aus »soeb 2« – Originaldatensätze (3)

■ Vielzahl von verschiedenen Originaldatensätzen

■ Verschiedene Nutzungsrechte und -formen:

- Scientific Use Files (SUF)
- Onsite
- Fernrechnen

■ Ausschließlich individuelle Nutzungsverträge

- Nutzergruppen werden bisher nicht berücksichtigt

■ Erfahrungen aus »soeb 2« – Datenmanagement

■ Kooperationen beim Datenmanagement

- Absprachen zu Datensatzabgrenzungen bei SOEP und Mikrozensus (Quer- und Längsschnitt)
- Detaillierte Festlegung aller Abgrenzungsmerkmale

■ Individuelle Datenmanagement-Entscheidungen von Verbundpartner/inne/n

- Verschiedene Formen von Querschnittdatensätzen
- Unterschiedliche Längsschnitt-Ausschnitte

■ Erfahrungen aus »soeb 2« – Analyse (1)

■ Kooperationen beim Analysedesign

- Harmonisierte Analysedesigns für Quer- und Längsschnittanalysen vor allem mit dem SOEP und dem Mikrozensus
- Detaillierte Absprachen

■ Methodenbezogene Kooperationen

- Faktoren- und Clusteranalysen, Sequenzanalysen, Optimal Matching in Kooperationen durchgeführt
- Detaillierter Abgleich der Verfahren nötig

■ Individuell festgelegte Analysedesigns und Methoden

- Verschiedene Quer- und Längsschnittanalysen
- Vielfältige uni-, bi- und multivariate Auswertungsverfahren

■ Erfahrungen aus »soeb 2« – Analyse (2)

■ Softwarebezogene Kooperationen

- Gemeinsame Nutzung spezieller Software-Tools (z.B. ausgewählte Ados oder CHESA)
- Detaillierte Absprachen

■ Individuell genutzte Software

- Diverse Statistikpakete (SPSS, STATA, TDA, NEWSPELL)
- Übersetzungsprobleme im Verbund

■ Erfahrungen aus »soeb 2« – Analyse (3)

■ Hardwarebezogene Kooperationen

- Meist institutsinterne Kooperationen mittels Netzwerkplatten - große Datenspeicherkapazität
- Keine geteilte Rechnerleistung

■ Hardwarebezogene individuelle Lösungen

- Individuelle Rechnerleistungen
- Vielfach Probleme bei größeren Fallzahlen oder aufwändigen Verfahren

■ Erfahrungen aus »soeb 2« – kooperative datenbezogene Arbeitsprozesse (1)

■ Auf Arbeitsdatensätze bezogene Kooperationen

- Übermittlung von Teilarbeitsdatensätzen mit Matchingvariablen
- Übermittlung einzelner Variablen
- Persönliche Arbeitstreffen in kleineren Gruppen
- Absprachen meist auf bilateraler Ebene

■ Probleme:

- Verschiedene Arbeitsdatensatzabgrenzungen
- Verschiedene Statistiksoftware
- Verschiedene Datenformate

■ Erfahrungen aus »soeb 2« – kooperative datenbezogene Arbeitsprozesse (2)

■ Auf Syntaxentwicklung bezogene Kooperationen

- Bilaterale Absprachen
- Bilateralen Austausch von Syntaxen
- Wechselseitige sowie gemeinsame Entwicklung von Syntax

■ Probleme:

- Verschiedene Statistiksoftware
- Verschiedene Syntaxversionen und deren Dokumentation
- Verschiedene Datenformate, Matching
- Prüfung und Sicherung von Qualität
- Verschiedene Dokumentationsstandards
- Urheberrechtliche Fragen

■ Erfahrungen aus »soeb 2« – kooperative datenbezogene Arbeitsprozesse (3)

■ Kooperative Nutzung von Zwischenergebnissen

- Bilateraler Austausch auf Nachfrage
- Austausch von Outputs bzw. Excel-Dateien mit aufbereiteten Zwischenergebnissen
- Kurze Berichte auf Projektwebsite

■ Probleme:

- Aufwand für Aufbereiten für die Projektwebsite
- »Work in Progress«
- Qualitätssicherung

■ Erfahrungen aus »soeb 2« – kooperative datenbezogene Arbeitsprozesse (4)

■ Kooperation bei Ergebnisdarstellung

- Gemeinsame Nutzung von Output-Syntax
- Bilateraler Austausch fertiger Texte, Excel-Tabellen und Grafiken
- Anpassen der Ergebnisdarstellung an gemeinsames Format

■ Probleme:

- unterschiedliche Ausgabeformate
- Uneinheitliches Vorgehen

■ Erfahrungen aus »soeb 2« – kooperative datenbezogene Arbeitsprozesse (5)

■ Kooperationen bei Dokumentation des Forschungsprozesses

- Dokumentation findet häufig in der Syntax statt
- Austausch der Syntax auf Nachfrage
- Austausch von Arbeitspapieren

■ Probleme

- Uneinheitliche Syntaxdokumentation

■ Anforderungen (1)

■ Forschungsdatenarchiv

- Gemeinsamer Zugriff auf Arbeitsdatensätze
 - Regelungsbedarf mit Forschungsdatenzentren
- Archivierung und Nachnutzung von Arbeitsdateien aller Art
- 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

■ Anforderungen (2)

■ Datenkonvertierung

- Forschungsdaten (Datenformate, z.B. SPSS, STATA)
- Auswertungssyntax für verschiedene Statistikprogramme

■ Provenienzdienst, Datenextraktion, Validierung

- Welche Datensätze sind vorhanden? (Dokumentationsstandards)
- Wer hat Datensätze, Syntax erstellt, überarbeitet?
- Wie wurde Syntax geprüft?

■ Arbeitsumgebung

- Projekt-WIKI zur internen net-basierten Kommunikation
- Unterstützung von Textproduktion
- Schnittstelle zur Projekt-Website: Ergebnisdokumentation
 - Zugriff auf Outputs durch Dritte

■ Offene Fragen

- Gemeinsamer »virtueller« Zugang zu Originaldaten
 - Eine neue Nutzungsform für Nutzergruppen (registrierte Nutzer)?
- »Geistiges Eigentum« an Syntax und Arbeitsdateien
- Gemeinsame Nutzung von Rechnerkapazität?
- Gemeinsame Lizenzen für Statistikprogramme auf Grid-Rechner?

■ Mehr ...

■ <http://www.sofi-goettingen.de>

- Soziologisches Forschungsinstitut (SOFI)
an der Georg-August Universität Göttingen

■ <http://www.soeb.de>

- Berichterstattung zur sozioökonomischen Entwicklung in
Deutschland

■ <http://www.wissgrid.de/index.html>

- WissGrid – Grid für die Wissenschaft