
Peter Bartelheimer

Daten in Virtuellen Forschungsumgebungen

Symposium Forschungsdateninfrastrukturen
Helmholtz-Zentrum Potsdam, GFZ
22. Januar 2013

■ Datenbereitstellung, Virtuelle Forschungsumgebung – zwei Komponenten der IT-Forschungsinfrastruktur

■ Forschungsdateninfrastruktur

- Ziel: Mikrodaten aus Bevölkerungsumfragen und Verwaltungsregistern berechtigten wissenschaftlichen Nutzer/innen verfügbar zu machen
- Leitbild: Zugriff, (Nach-) Nutzung, Kontrolle netzbasiert unmittelbar (»seamless«) zugänglich machen
- Problem Datengeheimnis: Daten sind unterschiedlich stark anonymisiert – Kontrolle von Zutritt, Zugang, Zugriff, Weitergabe und Output

■ Virtuelle Forschungsumgebung (VFU)

- Ziel: Wissenschaftliche Arbeitsprozesse vernetzen
- Leitbild: Kooperative Forschung unabhängig vom Ort zu gleicher Zeit ohne Ressourcen- und Zugangsprobleme
- Problem Workflow: Individualisierte Arbeitsprozesse analysieren und typisieren, um sie zu unterstützen

■ Entwicklungsbedarf auf beiden Seiten der Datenschnittstelle

-
- **Das Projekt: Virtuelle Forschungsumgebung für die sozioökonomische Berichterstattung (VFU soeb 3)**

 - **Projektpartner für die Entwicklung der VFU**
 - SOFI mit IT-Partnern GESIS, GWDG, SUB
 - Forschungsdatenzentren (FDZ-IAB, FDZ-RV, FDZ-SOEP)

 - **Funktionen der VFU in typischen Workflows der Datennutzung**
 - IT-Portal mit Funktionen zur Unterstützung des gesamten Workflows bei der Nutzung sozial- u. wirtschaftswissenschaftlicher Mikrodaten durch
 - Kollaborationstools
 - Datenmanagement
 - Benutzer- und Rollenmanagement
 - Datenzugang
 - Metadaten- und Syntaxeditoren

 - **Derzeit: Einführungs- und Entwicklungsphase**
 - Juni 2012 bis Februar 2014

-
- **Anwendungsfall: Dritter Bericht zur sozio-ökonomischen Entwicklung in Deutschland (soeb 3)**
 - **Verbundvorhaben 2013 - 2015 in Vorbereitung (Antragslage)**
 - Bis zu 24 beteiligte wissenschaftliche Einrichtungen mit bis zu 29 Arbeitspaketen (AP)
 - Nutzung von 66 verschiedenen Forschungsdatensätzen, Datenbanken
 - Neun der beim RatSWD akkreditierten FDZ als Datengeber
 - An VFU-Entwicklung und Verbund beteiligt: FDZ-IAB, FDZ-RV, FDZ-SOEP
 - **Schwerpunkte der Datennutzung**
 - Neue sozial- und wirtschaftswissenschaftliche Panel-Datensätze
 - 13 Datensätze kollaborativ genutzt, davon sieben intensiv (> zwei AP)
 - **Projektbegleitende Weiterentwicklung der VFU**
 - Operative Nutzung im Verbund
 - Teilnahme von Verbundpartnern in Nutzungsstudien
 - **Ziel: Bereitstellung einer »nachhaltigen« VFU-Plattform**
-

■ Welche Daten («Objektklassen») und Metadaten?

■ Unveränderliche Daten

- Forschungsdaten (Ausgangsdaten)
 - Datensätze und Dokumentation («Studien«)
 - Metadaten auf Studien-, Datensatz und Variablenebene
- Statistikprogramme (z.B. R), statistische Packages (z.B. ado-Files)

■ Im Workflow erzeugte Daten)

- Syntaxdateien, Syntax-Memos, Metadaten zu Syntax
- Output-Dateien, Metadaten zu Outputs, z.B. Log-Dateien
 - Tabellen, Grafiken, Textdateien (diverse Datenformate)

■ Speicherorte, Orte der Nutzung außerhalb der VFU

- Forschungsdatenzentren, Daten(service)einrichtungen
- Internet, Intranet
- Lokale Workstation, Intranet

-
- **An der Datenschnittstelle: Wie kommen Forschungsdaten in die VFU?**
 - **Bereitstellung für lokale Nutzung**
 - Public / Scientific Use Files –
 - **Bereitstellung für Rechnen in Dateneinrichtungen**
 - Gastarbeitsplätze: interaktiver lesender Zugriff auf Originaldaten, Outputkontrolle
 - Kontrolliertes Fernrechnen, zwei Varianten
 - Versenden von Syntax («job submission«, »Remote Execution«), Übermittlung kontrollierter Outputs
 - »Remote Data Access«: interaktiver lesender Zugriff über gesicherte Verbindung («safe centers«), Outputkontrolle
 - **Jedes FDZ gestaltet Fernrechnen anders**
 - **Anforderungen an Authentifizierung, Rechte- u. Rollenverwaltung**
 - Zugangskontrolle, Zugriffskontrolle, Weitergabekontrolle

■ VFU muss alle Datenzugangswege unterstützen

■ Rechnen in der VFU als Variante externer Bereitstellung

- Public Use Files (PUF) / Scientific Use Files – nach Vereinbarung
- Nutzung der Metadaten und Syntax in der VFU beim Rechnen in der VFU – mit R oder lizenzierten Statistikprogrammen

■ »Job Submission«

- Einheitliche Nutzungserfahrung durch Integration der verschiedenen Schnittstellen in VFU-Portlet?
- Nutzung der Metadaten und Syntax in der VFU für Job-Erstellung
- Historisierung der Jobs anhand von Metadaten

■ »Remote Data Access«

- Anmeldung über Aufruf der VFU an gesicherten Standorten oder zertifizierten Forschungseinrichtungen
- Nutzung der Metadaten und Syntax in der VFU beim interaktiven Fernrechnen

-
- **Hinter der Datenschnittstelle: Workflow und VFU**

 - **Aufarbeiten des Forschungsstands**
 - Arbeit mit Literatur: Recherche, Exzerpte, Bibliografie
 - **Arbeit mit Forschungsdatensätzen**
 - Planung, Organisation, Dokumentation und Ausführung
 - Datensatzgenerierung
 - Recodierung und Analysen
 - Ergebnispräsentation
 - Datensicherung
 - **Upload und Download**
 - von Dokumenten, Syntax, Outputs
 - **Editieren**
 - von Syntax und Metadaten
 - **Ergebnisdokumentation, Publikation (Open Access)**
-

■ Hinter der Datenschnittstelle: Wie kommen Syntaxdateien in die VFU?

■ Syntaxeditor als technische Entwicklungsaufgabe

- Funktionen des Erstellen und Nutzung von Syntax in der VFU
- Upload / Download zwischen lokaler Workstation und VFU
- Versionierung und Dateivergleich
- Ggf. Volltextsuche

■ Soziale Aspekte – Arbeitsweise in der VFU

- Wer lädt welche Syntax hoch und wann?
 - Lauffähige (vollständige) Syntax oder Syntaxbausteine
- Kontrolle über eigenes Arbeitsprodukt (geistiges Eigentum)
 - Rechte- und Rollenverwaltung: dateibezogene Rechte
- Verhaltensanforderungen
 - z.B. Namenskonventionen, »kontrollierte Vokabulare«
- Integrität zwischen lokaler Workstation und VFU?

■ Hinter der Datenschnittstelle: Wie kommen Metadaten in die VFU?

■ Metadatenschema als konzeptionelle Entwicklungsaufgabe

- Welcher Metadatenstandard (z.B. DDI 3.1, DDI 2.5)
- Für Metadaten zu Syntax kein etablierter Standard
- Dokumentationstiefe: Einschluss von Variablenebene?

■ Metadateneditor als technische Entwicklungsaufgabe

- Integration in VFU
 - Metadatenverwaltung in SQL-Datenbank
 - Verknüpfung mit »Objekten« in Dateiverwaltung
- Suchfunktion: strukturierte Suche und Volltextsuche
- Metadaten: Eingabe und Extraktion: z.B. aus Syntax, aus Datensätzen

■ Soziale Aspekte – Arbeitsweise in der VFU

- Verhaltensanforderungen und individuelle Arbeitsweise
 - Z.B. Erfassungsaufwand, Dokumentationstiefe, kontrollierte Vokabulare

■ Zum Schluss

■ Zwei Perspektiven auf VFU

- Kritisch für Akzeptanz datenhaltender Einrichtungen: Kontrollanforderungen
- Kritisch für Akzeptanz durch Nutzer/innen: intuitive individuelle Anwendung

■ Nicht alle Daten stehen in der VFU

- Leistungsversprechen der VFU: einheitliche Arbeitsumgebung
- Schrittweise mehr Daten integrieren

■ Entwicklungsbedarf auf beiden Seiten der Datenschnittstelle

- »Forschende FDZ« in VFU-Entwicklung und Nutzungsstudien integrieren

■ »Koevolution« technischer und sozialer Entwicklung

- Individualität der Forschungspraxis unterstützen
- Entwicklung als transdisziplinäre Herausforderung
- Selbstreflexion sozialwissenschaftlicher Datennutzung:
 - Analyse und Typisierung des Workflow ≠ Standardisierung

■ **Mehr ...**

■ **<http://www.soeb.de>**