

Integrierte Metadatenverwaltung als ein Schlüsselkonzept für datenbankbasierte statistische Analysen

Abstract:

Für eine wirtschaftliche und sichere Datenhaltung mit der Option eines schnellen Datenzugriffs ist der Einsatz von Datenbanksystemen heutzutage Standard. Zunehmend betrifft diese Art der Datenverwaltung und Zugriffsorganisation auch den Bereich der Sozialwissenschaften, wo in wachsendem Maße größere Datenvolumina zur wissenschaftlichen Auswertung anstehen.

Der Zugriff auf Datenbankdaten wird von den gängigen Statistiksystemen unterstützt. Jedoch scheinen hier bei weitem noch nicht alle Möglichkeiten ausgeschöpft. Auch wenn z.B. mit Wizard-Einsatz (bei SPSS) oder entsprechenden Kommando-Optionen DB-Zugriffe realisierbar sind, ist der Einsatz von Datenbanksystemen bislang mit der Einschränkung verbunden, dass nur Analysedaten, nicht aber die besonders in der sozialwissenschaftlichen Forschung wichtigen Zusatzdaten wie Variablen- und Valuelabels verwaltet werden. Darüber hinaus kommt bislang der Vorteil relationaler Datenbanksysteme, über die Möglichkeit der Verknüpfung von Datenbanktabellen mehrere Datenbestände gleichzeitig im Zugriff zu haben, bei gängigen Statistikpaketen kaum zum Tragen.

Im Vortrag soll aufgezeigt werden, dass durch einen (zumindest im Kontext der Statistiksysteme) neuartigen Ansatz zur integrierten Metadatenverwaltung ein Leistungsgewinn möglich ist.

Durch das Vorhalten geeigneter Informationen zu den einzelnen Datentabellen – hierzu zählen DB-Zugriffsinformationen, Spezifikationen der Tabellendaten (Typ, Feldname,...) sowie auch Variablen- und Valuelabels (alle unter dem Begriff Metadaten zusammengefasst) - kann dem Anwender zu jedem Zeitpunkt eine Informationssicht zur Verfügung gestellt werden, auf deren Basis er in der Lage ist, Daten aus den verschiedenen Quellen auszuwählen und deren Filterung und Zusammenspiel in komfortabler Weise zu veranlassen. Filter und Arbeitsskripte sind in die integrierte Arbeitsumgebung ebenso einbindbar wie diverse Dokumentationsobjekte.

Als Ergebnis einer Datenauswahl liefert das System ein Syntaxskript, welches – von SPSS gestartet – den gewünschten Arbeitsfile erzeugt. Alternativ kann eine direkte DB-Abfrage initiiert werden, deren Ergebnisdaten in unterschiedliche Ausgabekanäle leitbar sind.

Hinsichtlich der Datenbereitstellung für das Statistiksystem SPSS sei hervor gehoben, dass mittels eines einfachen Auswahldialogs bei Bedarf mehrstufige Matchings realisierbar sind, wobei die zu koppelnden Daten jeweils aus komplexen

SQL-DB-Abfragen resultieren können. Die Ergänzung bereits existierender sav - Files durch DB-Daten stellt auch kein Problem dar.

Der verwendete Ansatz wird mit dem vorzustellenden Servicewerkzeug (PMT) für die DB-basierte Arbeit mittels SPSS umgesetzt. Daten für Längsschnittauswertungen oder komplexere Querschnittsanalysen lassen sich hier mit hoher Effektivität bereitstellen. Prinzipiell kann durch die schnelle Reproduzierbarkeit der Datenzugriffe auf umfangreiche (sav-) File-Archive verzichtet werden. Dies entspricht gegenwärtig nicht der allgemein üblichen Praxis, die PMT-Metadatenverwaltung ist auf eine solche Arbeitsweise aber bereits eingerichtet.

Neben der Unterstützung für SPSS ist zukünftig auch eine Anbindung an andere Systeme möglich. Aktuell kann PMT z.B. für die Auswahl von DB-Daten zur Auswertung mit Excel herangezogen werden und dabei neben dem gebotenen Komfort bei der Datenauswahl auch ergänzende Metadaten als zusätzlichen Service liefern. Bei ordnungsgemäßem Einsatz des Systems hat der Anwender alle verfügbaren Daten sofort im Überblick und, eine funktionierende DB-Anbindung vorausgesetzt, auch im Zugriff.

PMT liefert ein ganzes Bündel an Serviceleistungen:

- Unterstützung einer sachbezogenen, strukturierten Gliederung und entspr. Selektionsmöglichkeiten für Variablen (über Tabellen-/Datenbestandsgrenzen hinausgehende Sicht auf Sachinhalte),
- Verwaltung von Filtern und Verarbeitungsskripten mit automatischer Bereitstellung dafür benötigter Variablen bei Aufruf,
- Verwaltung nutzerdefinierter Variablen, die - nach üblichem Verfahren mit SPSS-Mitteln berechnet - automatisch in der DB abgelegt werden können (z.B. für die Herstellung einer einheitlichen Arbeitsbasis für eine Arbeitsgruppe),
- Verwaltung und automatische Bereitstellung benötigter Variablen- und Valuelabels,
- Export-/Importroutinen für den Metadatenaustausch und Teamkommunikation,
- automatische Realisierung eines Ordnungsrahmens (Systematisierung, Archivierung ohne Zusatzaufwand) wird auf eine optimale Unterstützung des Anwenders orientiert.