Data Warehouses and Business Intelligence for Social Sciences

Aims and Possibilities of a Data Warehouse within the National Educational Panel Study (NEPS) in Germany

David Schiller
National Educational Panel Study (NEPS)
University of Bamberg, Germany

IASSIST, 03.06.2010, Cornell University, Ithaca (NY)
Outline

1. The National Educational Panel Study (NEPS)
   - Aims
   - Design
   - Challenges for Data Access

2. Data Warehouse Solutions for the NEPS
   - Possibilities
   - The shopping basket

3. The next steps
Aims of the NEPS

- Educations is a key factor in modern societies.
- Most surveys on education are collecting only cross sectional data.
- Huge demand for data to describe developments.
Aims of the NEPS

- Educations is a key factor in modern societies.
- Most surveys on education are collecting only cross sectional data.
- Huge demand for data to describe developments.
- The NEPS is supported by the German Federal Ministry of Education and Research and
  will provide longitudinal data to explain
  - the development of competencies,
  - the impact of learning environments,
  - educational decisions,
  - the impact of migration background,
  - and the returns of education
- over the life course.
Organizational Background

- To collect high-quality data the NEPS-Team is divided into special groups with special functions.
- Five *pillars* are responsible for the different theoretical topics,
- eight *stages* care about different intervals within the life course (the pillars and the stages have to develop the questionnaires and the tests) and
- a *methods department* which is accountable for the survey management (how to get to the people), the data warehouse (how to store the data) and the user service (how to bring the data to the researchers).
Multicohort Sequenzen Design

- Six different starting cohorts (each with corresponding pilot studies every year) and many additional samples (149 studies at the moment).
- Not only target persons also teachers, school administration and parents will be consulted.
- There will be questionnaires and tests for competencies.
- In total more than 60,000 people will be followed through their life course.
Need for easy and comfortable Data Access I

But we have a:

- Complex data structure with different cohorts, waves, target persons and additional samples.
- Researchers from different disciplines.
- High diversity in research interests (maybe related to the topics of the pillars mentioned earlier).
- Need for disclosure control.
- Need to provide a high quality description of the data.
Need for easy and comfortable Data Access II

Therefor the NEPS will build:

- A Data-Enclave,
- a User-Rights-Management-System and
- a Metadata-Editor (based on DDI).
Need for easy and comfortable Data Access III

How to structure the access to the data.

- With every wave (of the six panels) things will get more complicated.
- A big number of people will be followed individually (because they dropped out of their original panel or have left the school that is in the NEPS-sample).
- Common file servers and even relational databases will belittle the user-friendliness and the research potential of the data.
- With Data-Warehouse-Solutions and Methods of Business Intelligence a better research environment can be offered for the users.
Data Warehouse and Business Intelligence for the NEPS I

- Via integration services data in different formats (e.g. different studies) can be imported and controlled (e.g. for implausible values).
- During the import process metadata information can be connected to the data.
- Data from different sources can be matched.
- With data stored in the DWH tools of business intelligence can be used for analysis (like data mining or Reporting).
Data Warehouse and Business Intelligence for the NEPS II

- Or one can export the data into various formats for different statistical packages (such as Stata, SPSS, R)
- Results and new computed variables can be stored for later use by the scientific community.
- New solutions for statistical disclosure control (based on the DWH) can be integrated. For example by creating synthetic data or during the output control process.
The Shopping-Basket-Concept

- Instead of working with big datasets (and the dropping of variables)
- the user can select single variables or sets of variables and
- put them into his shopping basket.
Selecting Data

1. Log into the Enclave.
2. Pick single variables or sets of variables into your shopping basket.
3. (Sets of variables consist of variables belonging to special topics, like migration background or learning environments.)
4. Use the statistical tools within the DWH to analyze the data or
5. export the dataset into your private workspace within the enclave in the desired format (SPSS, Stata, R).
Support

To help the user by selection variables the following options can be used:

- Metadata for each variable will be shown (e.g. Information about distributions or about the theoretical background of the variable).
- Via data mining new trends within the data can be found.
- Statistical analysis within the DWH can be used (e.g. Clusteranalysis or Correlations).
- Results can be saved for other users.
- The user can make comments on variables for other users.
Security

- Because of the danger of disclosure each variable or set of variables will have a defined security level. Reaching from unproblematic to high disclosure risk.
- Only defined combinations of variables will be allowed.
- If you want an forbidden combination, you have to contact the user-service.
- What you do in the enclave will be logged.
- Output will be controlled.
Further topics

- Selections can be saved and they can be found via unique identifiers.
- So other researcher can check the results and do further calculations.
- New computed variables can be integrated in the DWH so others can use them.
Next steps

- Make sure, that data will be available by Summer 2011 (adults) and Summer 2012 (the rest).
- Test the shopping basket concept with real data. That means:
  - Find the correct security levels and possible combinations.
  - Connect the metadata to the variables.
  - Enable analysis within the DWH.
  - Find solutions for In- and Export.
- Provide standard shopping baskets for the scientific community.
Future Research

- Automatic Output Control.
- Developing techniques to create Synthetic Data.
- For full use of the DWH - merge data from other sources (like other studies, administrative data, paradata and so on).
- Extend the Data-Enclave (Community, Reports on the NEPS-data etc.).
The National Educational Panel Study...
The National Educational Panel Study...

...to be continued.
The National Educational Panel Study...

...to be continued.

Thank you for listening.